Thèse n° 9083

The Role of Adaptivity in Source Identification with Time Queries

Présentée le 14 juillet 2022

Faculté informatique et communications
Laboratoire de la dynamique de l'information et des réseaux 2
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

Gergely ODOR

Acceptée sur proposition du jury

Prof. M. Kapralov, président du jury
Prof. P. Thiran, directeur de thèse
Prof. R. Lambiotte, rapporteur
Prof. D. Mitsche, rapporteur
Prof. L. Zdeborová, rapporteuse

École
polytechnique
fédérale
de Lausanne

2022

To the Water Bearer at the Mountain Pass

Acknowledgements

Fortunately, I have many people to thank for having a positive impact on me during my PhD. First, I would like to express gratitude towards my advisor, Prof. Patrick Thiran, for the opportunity to write this thesis, and for his constant support during these past few years. Patrick, thank you for always being there for me, and for accommodating my requests, even when they were very unusual. The second person I would like to thank is Prof. Júlia Komjáthy, whom I had the fortune to collaborate with on multiple papers. Juli, thank for going out of your way to help advance my scientific career, and for caring for me as a person overall. Third, I would like to thank Prof. Márton Karsai and Prof. László Lovász, who have invited me to contribute in their COVID-19 modelling team at the beginning of the pandemic, which has been an incredibly exciting and fruitful experience.

I express my gratitude towards Prof. Bob West, Prof. Balázs Ráth, Prof. Gábor Pete, Prof. Negar Kiyavash, Prof. Daniel Cullina, Prof. Matthias Grossglauser, Prof. Zoltán Király and Prof. Tamás Móri for the many interesting discussions on various research projects, and for teaching me the ways of becoming a successful scientist. I also thank Prof. Michael Kapralov, Prof. Renaud Lambiotte, Prof. Dieter Mitsche and Prof. Lenka Zdeborova for accepting to be on my thesis committee, and for reviewing my thesis. I especially want to thank Prof. Dieter Mitsche for helping to develop some of the unpublished results in the thesis, and for carefully checking the writing of the thesis on multiple occasions.

At EPFL, I was fortunate to supervise several masters or bachelor student projects and summer internships. Some of these projects have turned into published papers, and constitute core chapters of this dissertation. Thank you Jana, Miguel, Stanislas, Victor, Satvik, Nicolas, Isabela and Farzad for being amazing students, and for the opportunity for getting to know you. In the COVID-19 team, thank you Domonkos, Csegő and Dani for collaborating with me, and sharing the load of being a junior scientist in a large research group.

At the INDY Lab, I had the pleasure of meeting many interesting fellow PhD students. Thank you Victor, William, Mladden, Masha, Aswin, Daniyar, Lars, Farnood and Lucas for the relaxing tea breaks and your camaraderie. Arnout, it has been an honor to share an office with you, and have extremely fast paced discussions about pretty much anything involving academia or else. I also want to thank Brunella, who graduated after the first six month of my time at INDY, but had a lasting impact on my PhD. Without your initial help, and your very related thesis, which I have spent a lot time reading, I probably never would have started working on source identification in the first place.

I was lucky to meet many new friends in Switzerland. Bence and Jakab, it has been a blast to be your roommates for our first year in Lausanne! I am not sure I would have made it this far without you. And your couches, which I had the privilege to sleep on numerous times. Dirk, thank you for being an amazing neighbor, and for being my running buddy! Thank you Nóri, Peti, Danó, Ilcsi, Michelle and Elene for the fun hikes, bike trips and dinners. This list could go on for a long time, and there is no way I can include everyone in just a few pages.

I am fortunate that I had the opportunity to spend one year of my PhD in Hungary at the Rényi Institute and the Budapest University of Technology and Economics on the SNSF DocMobility Fellowship. I am grateful for the friendships that started during this time. Aranka, Octave, Emma, Marco, Gábor, Diego, Roberto, Vilas, Caio, Konrad, Olle, Sarah and Stefan, thank you for the relaxing lunches, and sharing the difficulties of reading disturbing news about the world almost every day. I also thank my friends and teachers from high school, especially Laci, Sanyi, Bunthy and Gergő, who kept me motivated in the first few years of the PhD, and even helped me think about my research. The year in Hungary involved several administrative and logistical challenges, I thank Angela, Patricia and Eileen from EPFL, Gábor from BME, and Dezső, Miklós and Anna from the Rényi Institute for helping me.

I am grateful that as a PhD student, I still had the opportunity to pursue my passion for theatre. I thank the English Theater Company at EPFL & UNIL, and the Metro Works Theater School in Budapest for taking me in, and providing the opportunity for me to learn and perform.

I thank my family, especially my mother Viktória, my father Géza, and my stepfather Laci for always supporting my carrier as a researcher.

Finally, I thank my lovely wife, Dóri. It is not too much of a stretch to say that half of the motivation to finish this thesis came from the will that we wanted to be at the same place, and the other half from the gratitude for succeeding to be together.

Lausanne, June 29, 2022 G. Ó.

Abstract

Understanding epidemic propagation in large networks is an important but challenging task, especially since we usually lack information, and the information that we have is often counter-intuitive. An illustrative example is the dependence of the final size of the epidemic on the location of the initially infected agents (sources): common sense dictates that the most dangerous location for the sources is the largest city, but the second chapter of the thesis shows that this holds true only if the epidemic is just above the infection threshold.

Identifying the initially infected agents can help us better understand the epidemic. The focus of this thesis is on identifying the very first infected agent, also called the source or patient zero. According to the standard assumptions, a few agents reveal their symptom onset time, and then it is our goal to identify the source based on this information, together with full knowledge of the underlying network. Unfortunately, even if we can choose the set of agents that are queried about their symptom onset time, the number of queries required for reliable source identification is too large for practical applications. In this thesis, we carefully assess if this issue can be mitigated by introducing adaptivity to the standard assumptions. Our main goal is to study the reduction in the query complexity if the queries can be chosen adaptively to previous answers, but we also investigate whether adaptively querying the edges can relax the full knowledge assumption on the network.

Providing rigorous proofs about source identification with time queries is difficult. A notable exception is when the infection is passed with a known, deterministic delay from each agent to all of its neighbors, in which case the number of required non-adaptive and adaptive queries are equivalent to well-known notions in combinatorics; the metric dimension (MD) and the sequential metric dimension (SMD), respectively. We extend previous results in the field by computing the MD of a large class of random trees, where adaptivity can significantly reduce the query complexity, and the SMD of Erdős-Rényi random networks, where the reduction is found to be small, at most a constant factor. We address the case of non-deterministic diffusion processes for the first time in the mathematical literature: on the path graph, we observe a striking, double logarithmic decrease in adaptive query complexity compared to the non-adaptive case.

Our analysis on the robustness of the MD to adding a single edge to specially constructed and d-dimensional grid networks suggests that even small changes in the network could easily derail source identification algorithms. This is concerning since it is difficult to obtain a perfect dataset about the entire contact network in practice. Inspired by recent implementations of contact tracing, we propose new source identification assumptions, where not only the symptom onset times, but also the edges of the network are queried by the algorithm, resulting in less, but potentially higher quality information. We propose two local search algorithms that outperform state of the art identification algorithms tailored to the new assumptions, and we analytically approximate their success probabilities on realistic random graph models. The adaptive assumptions enable us to evaluate our algorithms on a COVID-19 epidemic simulator: the first time that source identification algorithms are tested on such a complex dataset.

Résumé

Comprendre la propagation épidémique dans les grands réseaux est une tâche importante mais difficile, en particulier parce que nous manquons généralement de données, et que les données dont nous disposons sont souvent contre-intuitives. Un exemple éloquent est la dépendance de la taille finale de l'épidémie à la localisation des agents infectés initiaux (sources) : le bon sens veut que la localisation la plus dangereuse pour les sources soit la plus grande ville, mais le deuxième chapitre de la thèse montre que cela n'est vrai que si l'épidémie est juste au-dessus du seuil d'infection.

L'identification des agents infectieux initiaux peut nous aider à mieux comprendre l'épidémie. L'objectif de cette thèse est d'identifier le tout premier agent infecté, également appelé source ou patient zéro. Selon les hypothèses standard, quelques agents révèlent l'heure d'apparition de leurs symptômes. Notre objectif est alors d'identifier la source sur la base de ces données, ainsi que d'une connaissance complète du réseau sous-jacent. Malheureusement, même si nous pouvons choisir l'ensemble des agents qui sont interrogés sur leur heure d'apparition des symptômes, le nombre de requêtes d'information nécessaires pour une identification fiable de la source est trop important pour les applications pratiques. Dans cette thèse, nous évaluons soigneusement si ce problème peut être atténué en introduisant une adaptabilité aux hypothèses standard. Notre objectif principal est d'étudier la réduction de la complexité des requêtes si les requêtes peuvent être choisies de manière adaptative aux réponses précédentes, mais nous examinons également si un processus adaptatif du choix des arêtes interrogés peut relaxer l'hypothèse de connaissance complète sur le réseau.

Il est difficile de donner des preuves rigoureuses sur l'identification de la source avec des requêtes temporelles. Une exception notable est lorsque l'infection est transmise avec un délai déterministe, connu de chaque agent à tous ses voisins, auquel cas le nombre de requêtes non adaptatives et adaptatives requises est équivalent aux notions bien connues en combinatoire ; la dimension métrique (Metric Dimension, MD) et la dimension métrique séquentielle (Sequential Metric Dimension, SMD), respectivement. Nous étendons les résultats précédents dans le domaine en calculant la MD d'une grande classe d'arbres aléatoires, où l'adaptabilité peut réduire considérablement la complexité des requêtes, et la SMD des réseaux aléatoires d'Erdős-Rényi, où la réduction est faible, au plus un facteur constant. Nous abordons le cas des processus de diffusion non déterministes pour la première fois dans la littérature mathématique : sur le graphe de chemins, nous observons une diminution frappante (doublement logarithmique) de la complexité des requêtes adaptatives par rapport au cas non-adaptatif. Notre analyse de la robustesse de la MD à l'ajout d'une seule arête à des réseaux spécialement construits, et de réseaux en grille de dimension d suggère que même de petits changements dans le réseau pourraient facilement faire dérailler les algorithmes d'identification de source. Ceci est préoccupant, car il est difficile d'obtenir un ensemble de données parfait et complet sur le réseau de contacts dans la pratique. Inspirés par les progrès récents de la recherche des contacts, nous proposons de nouvelles hypothèses pour l'identification des sources, où non seulement les heures d'apparition des symptômes, mais aussi les arêtes du réseau sont interrogées par l'algorithme, ce qui donne moins d'informations, mais potentiellement de meilleure qualité. Nous proposons deux algorithmes de recherche locale qui surpassent les algorithmes d'identification de l'état de l'art, adaptés aux nouvelles hypothèses, et nous approchons analytiquement leurs probabilités de succès sur des modèles de graphes aléatoires réalistes. Les hypothèses adaptatives nous permettent d'évaluer nos algorithmes sur un simulateur d'épidémie de COVID-19, testant pour la première fois des algorithmes d'identification de source sur un ensemble de données aussi complexes.

Contents

III Stochastic Spreading 109

Introduction Part I

1 Motivation, Previous Work and Contributions

Many interesting natural phenomena can be modelled by *spreading processes* on *networks*. An important example is disease spreading from infectious individuals to susceptible ones, where the nodes of the network represent individual humans, and the edges between them represent physical interactions. If we model the spread of a new disease, usually caused by a spontaneous mutation, the process starts from a single individual (also called *patient zero*) represented by the *source* node of the network. Since human contact networks can be very large, the process may spread to many nodes through the edges of the network before an outside observer can even detect that the spreading has started. By that time, the information about the identity of the source node is often lost, even though this information could be very useful in some applications. In source identification, our goal is precisely to recover this lost information, whenever possible.

The identification of the source of an epidemic can be useful while planning our response as a society, since any information on the disease is crucial in uncertain times [122]. For example, source identification can aid contact tracing efforts [52, 206], and understanding how the mutation occurred can give information on how dangerous the outbreak is [134, 152]. Besides epidemics, related applications of source identification includes rumor spreading between individual humans [215], a virus spreading in a computer network [243], train delay propagation [168], and food-borne disease outbreaks [169].

Historically, the algorithmic framework for source identification in epidemics was proposed by Pinto, Thiran and Vetterli [198], inspired by the pioneering work of Shah and Zaman [215] on rumor spreading. In the framework of [198], a few *sensor* nodes reveal the time of their symptom onset after the epidemic has spread to the entire network. With these symptom onset times, and also full knowledge about the network and the epidemic model, the algorithmic task is to recover the identity of the source with as high probability as possible.

Surprisingly, given the importance of the problem and the relatively large literature on the topic, we are not aware of any instance where source identification algorithms have been applied in real epidemics (beyond proof of principle studies), including during the COVID-19 epidemic. An important criticism of the original framework of Pinto, Thiran and Vetterli [198] is that the contact network G is fully known, which is unrealistic (let alone because of privacy concerns). Moreover, and even if G is fully known, it is very difficult to find the source exactly unless a large fraction (20-50%) of the population act as sensors, which is unrealistic in the case of epidemics, when the source is searched in a large population [221]. In this dissertation, we show that these two issues can be mitigated if we introduce an extra assumption, that the algorithm can place sensors, and can gather information about the network in an *adaptive* way. Our analysis shall include both rigorous theorems and data-driven simulations to make our points precise and applicable at the same time!

1.1 General Related Work in Epidemics

Epidemics are typically modelled by assigning the individuals of the population into compartments with labels such as “Susceptible”, “Infected” or “Removed/Recovered”, and then the individuals progress through these compartments according to the same set of rules [182]. The most elementary example is the Susceptible-Infected (SI) model, where each infected individual infects susceptible contacts with probability β at each timestep. The Susceptible-Infected-Removed (SIR) model is a simple extension of the SI model where infected individuals are removed (due to recovery or death) with probability μ at each timestep. The process can start from a single infected node, which we then refer to as the *source*, or multiple infected nodes, which we then refer to as *seed nodes*. Originally, these standard models operated under the homogeneous mixing (mean-field) assumption, where the underlying network is a complete graph, and the dynamics of the spread can be analyzed by differential equations [182]. In mean-field models, the basic reproduction number $R_0 = \beta/\mu$ is sufficient to describe the dynamics of the spread: for $R_0 < 1$ the infection dies out quickly, while for $R_0 > 1$ the infection spreads exponentially. With the appearance of the Internet and modern data collection methods, our understanding of the underlying contact network became more accurate. Since it is unlikely that we can have access to the true contact network of the population during a real epidemic because of privacy concerns, the common approach is to infer some of the more important properties of the true contact network, and build random network models that also possess these properties. The first, and most studied property of social networks is degree heterogeneity; usually it is assumed that a few people have many more contacts than most people [23]. Epidemics in random networks with a given degree distribution can still be understood via differential equations [195]. However, if the spatial properties of the contact networks (due to the underlying geographical effects) are also taken into account, analysing the epidemic spread becomes very challenging, and is usually done via numerical simulations. If the size of the population is about the size of a city or smaller, then the numerical models tend to be agent-based: each individual is assigned some demographic properties such as age, home location, work location, etc., and all of their individual actions are simulated together with the disease spreading [180]. For larger populations (such as regions, countries or the world), more aggregated methods are used.

A notable example of an aggregated method that incorporates spatial properties of the contact network is the *metapopulation model*: the population is divided into subpopulations, inside which the disease spreads according to the usual homogenous mixing assumption, and between which the individuals may travel as if they were performing a random walk in the network [60]. The average number of individuals travelling between subpopulations and the average number of contacts in each subpopulation are inferred from mobility datasets and questionnaires. In case of a world-wide model, subpopulations are divided based on the closest major airport [60]. The added value of the metapopulation approach on the world-wide level is that a disease may spread very differently in different countries due to cultural and seasonality effects, and that the mobility between countries is fairly limited. In a country-wide or (resp., a city-wide) model, the subpopulations are cities (resp., districts). The added value of the metapopulation approach in smaller countries and cities is less obvious since the seasonality or cultural differences are weaker and the mobility inside a country is relatively strong. Nevertheless, the metapopulation approach allows for a study of different mobility patterns [64] and geographically targeted interventions [157], and has been shown to produce a closer fit to historical datasets, compared to the mean-field approach with a single population [105]. In this dissertation, we make another argument for why it is important to incorporate the spatial properties of the contact network even in a country-wide simulation.

Contribution 1. *[published in paper [187]] In Chapter 2, we compare two scenarios based on the geometric distribution of the initial infection seeds of the epidemic motivated by historical data. In one scenario, the seeds are concentrated on the central nodes of a network, while in the second one, they are spread uniformly in the population. Comparing the final size of the epidemic started from these two initial conditions in metapopulation networks, we find evidence for a switchover phenomenon: When the basic reproduction number R_0 is only slightly above 1, more individuals become infected in the first seeding scenario, but for larger values of R_0, the second scenario is more dangerous [187]. This finding challenges the intuition that the epidemic from the best-connected nodes of a network would intuitively lead to the largest outbreak [143], and highlights the importance of tracking the geographic location of the initial seeds.*

The primary goal of the models and results mentioned so far is to make predictions, and identify potential scenarios about the future. The focus of this dissertation is on the reverse problem; on uncovering past events about the history of the epidemic, which is also very important for our scientific understanding. If adequate resources are available, the most accurate tool to uncover the history of an epidemic is through *phylogenetic reconstruction*, which requires DNA samples of the disease [106, 239]. In source identification our goal is more extreme, we aim to find only the very first patient that developed the disease, and we would like to do it without collecting DNA samples, based on information that is easier to obtain at a large scale. Since our goal is to find the very first patient, and since metapopulation networks do not have network information at the individual level, in the rest of this dissertation we restrict ourselves to the individual based SI or SIR epidemic models, and their variants. We mention, that a related version of the source identification problem, where the goal is to find the country (or meta-node) where the epidemic originated, has been studied via metapopulation networks in a different line of work [44].

1.2 Source Identification Frameworks

We start by reviewing the related work in source identification. We visualize the citation network of the most important papers in Figure 1.2. Additional reviews with a different focus can be found in [131], and in the PhD theses [221, 129].

Originally, source identification was introduced in the context of rumor spreading by Shah and Zaman in their pioneering papers [214, 215]. Translating to the language of epidemics for clarity, in the framework of [215], an epidemic spreads over a network G that is completely known to us. As illustrated in Figure 1.1 (a), we observe a *snapshot* of the epidemic, which means that every individual reveals if they are infected or not at some given time (not too early, because then the problem is trivial, nor too late, because then the problem is impossible). Shah and Zaman find the ML estimator of the source on regular trees, which they call the *rumor center.* Subsequent works propose different centrality-based estimators, such as a local version of the rumor center [74], the *Jordan infection-center* [258], and an estimator based on the *minimum-description-length principle* [199]. In search of more applicable results, centrality-based estimators have been extended to the case of *partial observation*, when only a subset of nodes reveal their infection state. This subset is selected either randomly [166, 257] or strategically to maximize the performance of the subsequent estimation [212]. Another way to generalize the snapshot framework of Shah and Zaman is to allow multiple diffusion sources, which may start the infection at different times [128]. More recently, Lokhov et. al. and Altarelli et. al. developed general estimation algorithms based on Dynamic Message Passing (DMP) [161] and Belief Propagation (BP) [14, 15]. Both of these algorithms can be easily adapted to different source identification frameworks by simple changes to their initial equations. However, these algorithms tend to be slower than centrality-based estimators and they assume that the underlying network is (at least locally) tree-like.

So far, we only talked about variants of the snapshot-based framework, where nodes only provide binary information about their infection state. In the application of epidemics, it is reasonable to assume that if someone reveals their infection state, they will provide additional information too, for instance their symptom onset time. Pinto, Thiran and Vetterli formalized this idea in the so called *sensor-based* or *source identification with time queries* framework [198], which is the focus of this dissertation. The framework was proposed for a variant of the SI epidemic model with general transmission delay distributions: an infected node v_1 can infect a susceptible node v_2 to which it is connected by edge $\mathrm{e} = (v_1, v_2)$ after some (possibly random) time $w(\mathrm{e})$ drawn from the *transmission delay distribution* $\mathcal{W}$. After the epidemic has spread to the entire network, a small subset of nodes $S \subset V$, which we call hereafter *queries, query nodes* or *sensors*, are assumed to reveal their symptom onset time (see Figure 1.1 (b)). It is also

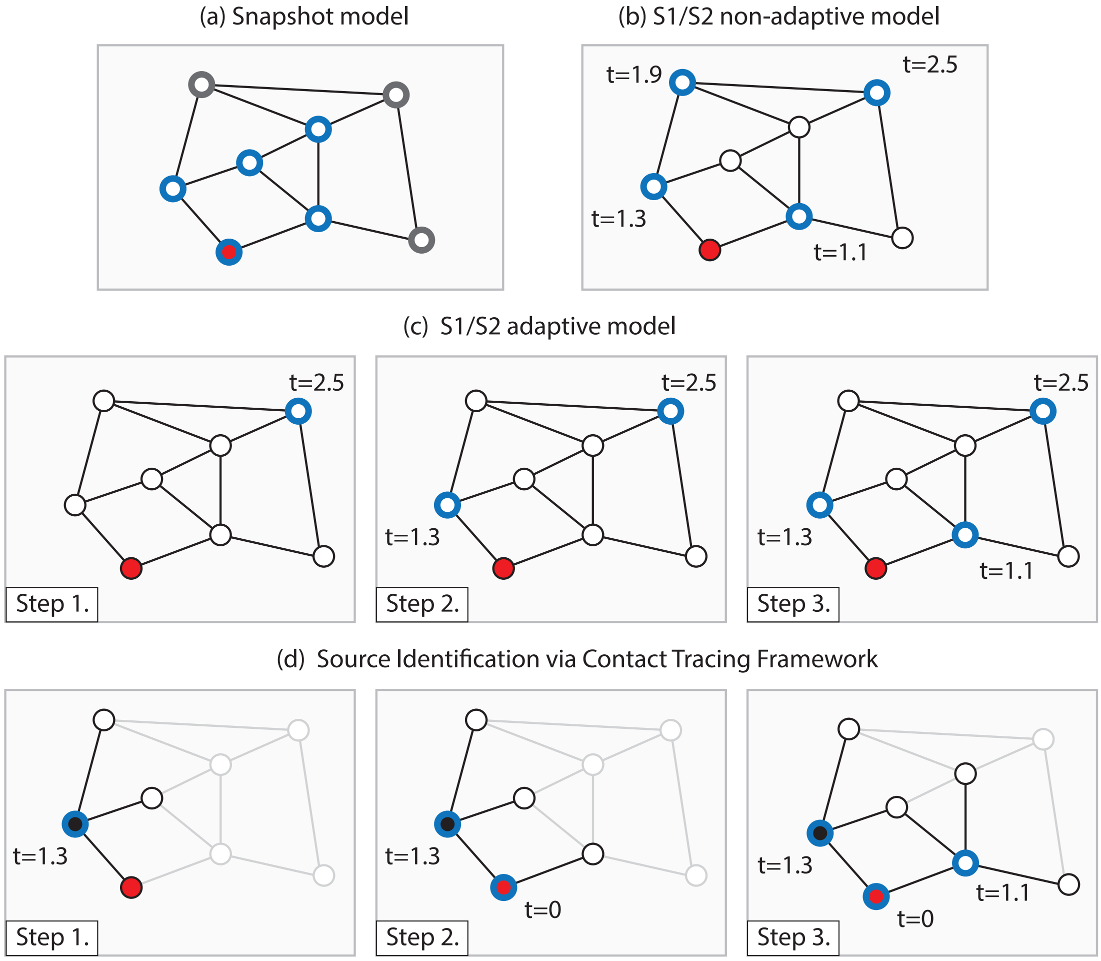


Figure 1.1: Illustration of the different source identification frameworks reviewed in Section 1.2. In each network, the source node (unknown to the algorithm) is marked in red. (a) In the snapshot framework, each node gives binary information: whether it is infected at the moment when the snapshot is taken (blue stroke) or not (grey stroke). (b) In the S1/S2 non-adaptive framework, each node is assumed to have already been infected, and the sensor/query nodes (blue stroke) report their infection time (relative to each other in the S1 framework, or relative the infection of the source node in the S2 framework). (c) The adaptive version of the S1/S2 framework differs form the non-adaptive version in that the sensor/query selection proceeds in multiple steps, and in each step, the algorithm can use the information provided by the observations from the previous steps. (d) In the Source Identification via Contact Tracking Framework (SICTF), the source identification task starts when the first node is hospitalized (marked in solid black), and initially most of the network is unknown to the algorithm (marked in grey). Then, in each step, the algorithm proceeds to explore the network and query the infection state of the nodes in an adaptive way. Queried nodes reveal their symptom onset time (if they are symptomatic), and the infection time of the source node is not known (similarly to the S1 framework).

assumed that the network G and the transmission delay distribution $\mathcal{W}$ is known. Then, the algorithmic task is to recover the identity of the source with as high probability as possible. There is one more assumption in [198] that we have not mentioned so far: in [198], query nodes also reveal the neighbor from whom they received the infection. This information is difficult to obtain directly in the context of epidemics, and most follow-up works do not assume that this information is available. The resulting modified version of [198], which we call S1, is the most popular framework in source identification with time queries, and has been the subject of a long list of papers. An important caveat of framework S1 is that the time of the first infection is not known. This is a realistic assumption in most applications (especially in epidemics), however, in some cases it is instructive to study a model where the time of the first infection is known, which we call framework S2.

We also mention that there exist frameworks where a different kind of information is obtained about queried nodes compared to the S1 and S2 frameworks. Fanti et. al. consider the opposite problem of source obfuscation, when the goal is to spread some information in the network in a decentralized way so that an outside observer cannot identify where the information originated from [88]. These source obfuscation models are used to anonymize transactions on the Bitcoin network [35]. In this case, the query nodes reveal all information they receive during the decentralized spreading; e.g., these "control packets" may contain instructions about which nodes should get the information next. On the contrary, Kumar et. al. considers a source identification framework, which is very similar to S1, except only the relative infection times of the query nodes are revealed [151].

In this dissertation, we focus on the S1 and S2 frameworks. The research on models S1 and S2 is driven by the following three questions of increasing complexity [250]:

(i) given the answers to a fixed set of queries, how can we estimate the source?

(ii) given the maximum number of queries that we can ask, which queries should we choose so that we can solve the estimation problem as accurately as possible?

(iii) if we want to correctly identify the source, at least how many queries do we need?

We note that in theoretical research, (ii) and (iii) are difficult to separate, however, in applied research (ii) is often solved before (iii).

Most of papers in the field address the estimation problem (i): Pinto et al. found that in their framework, if the sensors are already selected, the maximum likelihood estimator of the source has a closed form solution when the underlying network is a tree, and the transmission delay distribution is Gaussian. For general graphs, it is difficult to find an algorithm with any theoretical guarantees, although we note that many heuristics have been developed [119, 158, 192, 193, 217, 230, 246, 256].

Several previous works address problem (ii) (see [191, 225, 252, 253], and the references therein) by algorithms of heuristic nature. These types of sensor selection problems are typically NP-hard, and only approximation algorithms, such as submodular approximation algorithms [51], can be provided with theoretical guarantees. For problem (iii), most papers find by simulations that a linear fraction about (20%-50%) of the sensors are needed for identifying the source with high accuracy [191, 221].

Providing rigorous proofs in source identification with time queries is particularly challenging. A notable exception is when the transmission delay distribution $\mathcal{W}$ is deterministic, i.e., when the infection is passed with a known, deterministic delay from each individual to all of its neighbors. In this case, if the first infection time is also known (framework S2), problem (i) is trivial, problem (ii) is equivalent to finding a *resolving set* in a graph (a set of nodes such that the distance to those nodes is enough to uniquely determine the identity of an unknown node) [248], and problem (iii) is equivalent to the *metric dimension* (MD) problem [218]. We review previous works on the MD in detail in Chapter 3. For the purposes of this introduction, it is enough to mention that in the context of random graphs, the studies on the MD have been limited to Erdős-Rényi graphs [36], uniform random trees [177], and random geometric graphs [159]. This dissertation adds a new class to this list of random graph distributions, whose the MD has been studied.

Contribution 2 (published in paper [147]). *In Chapter 4, we provide rigorous results for the MD of general classes of random trees: critical Galton-Watson trees conditioned to have size n, and growing general linear preferential attachment trees.*

1.3 Limitations of the Frameworks S1 and S2

1.3.1 The Query Complexity

One of the main criticisms of source identification algorithms is that the number of queries required to find the source is large. Although this has not been shown theoretically for stochastically spreading epidemics for any network before this dissertation, it is widely accepted that source identification with time queries is possible only if no less than a constant fraction of the population are queried, which makes it questionable that the developed algorithms are applicable in real-world scenarios. To remedy the situation, a recent research direction suggests to abandon the task of exact identification of the source, and to replace it by the computation of confidence sets around it, which can be done with fewer queries (see [45, 140] for the snapshot-based, and [65] for the time query settings). However, if our goal is to find the source exactly, without querying a prohibitively large fraction of the population, the only viable options is to modify the assumptions of source identification framework.

In the case of rumor spreading, it is often assumed that multiple cascades are observed [90, 201, 240], which can significantly decrease the number of required queries. But it is not a realistic assumption if we are modelling epidemics on the individual level.

In epidemics, a promising approach is to allow some of the queries to be selected adaptively to previous answers. Louni and Subbalakshmi [164] studied a source identification framework where the queries are selected in two steps: the first set of sensors help determine the community (densely connected subgraph) where the source is located, and the second set of sensors (placed strategically inside the candidate community) help to find the exact identity of the source, thus reducing the number of required queries. Of course, allowing three or four steps for the query selection can reduce the number of required queries even further. The extreme version of the problem is when each step consists of the selection of exactly one query, and we receive the observation for the most recent query after each step (see Figure 1.1 (c)). This framework, which we call the *adaptive setting* or *adaptive framework*, was introduced by [249, 250]. Heuristic adaptive strategies have been studied by Spinelli, Celis and Thiran [222, 224] by simulations; they show a large reduction in the number of required queries compared to the non-adaptive case, especially in real networks.

For theoretical proofs in the adaptive S1/S2 framework, it is useful to start with the case of deterministically spreading epidemics. Zejnilovic et. al. [249] addressed the adaptive, deterministic case from an algorithmic point of view. In the literature on combinatorics, the corresponding adaptive version of the metric dimension (MD) is called the sequential metric dimension (SMD). We provide the definition, and review the literature related to the SMD in Section 3.3. Thereafter, we compare the SMD with the MD in several classes of random graphs in Section 3.4, which is a key section in this dissertation; Section 3.4 summarizes many of the quantitative results in Chapters 4 and 5, and motivates Chapters 6 and 7.

In our literature review in Section 3.3, we found that before this dissertation, the SMD has not been studied in any family of random graphs.

Contribution 3. *[published in paper [188]] In Chapter 5, we tighten the previously known upper bound by [36] on the MD, and provide matching upper and lower bounds on the SMD of Erdős-Rényi random graphs. We find that the ratio of the SMD and the MD is between* 1 *and* 1/2*, which implies that the role of adaptivity in Erdős-Rényi graphs is fairly small.*

Comparing the MD and SMD is an important goal, because there is previous research on these notions in combinatorics, and because they model the source identification problem in epidemics with deterministic edge delay distribution. Eventually, the goal is to generalize these results to a wide class of edge delay distributions $\mathcal{W}$, which would be more relevant in real-world scenarios. Before this dissertation, rigorous results on the number of required queries with stochastic $\mathcal{W}$ have been lacking both in the non-adaptive and the adaptive versions of the source identification problem.

Contribution 4 (published in paper [156])**.** *In Chapter 6, we prove upper and lower bounds on the adaptive and non-adaptive query complexities of source identification on the path network for Gaussian edge delay distributions $\mathcal{W}$. We observe a striking, double logarithmic decrease in adaptive query complexity compared to the non-adaptive case, hinting that adaptivity plays a very important role in stochastic epidemics.*

We discuss extensions of Contribution 4 to more general edge delay distributions and network models in Chapter 6. The most challenging part of the proofs is the lower bound in the adaptive setting; we discuss the information theoretic connections at the beginning of Chapter 6.

1.3.2 The Knowledge About the Network

We believe that the most concerning, and incidentally less studied, assumption in source identification papers is the full knowledge of the contact network. Previous works reconstruct or approximate the contact network from digital mobility traces [207], from piecing together on small subnetworks [179, 247], or from previously observed events within population that use the same network [104]. Despite these efforts, reconstructing the contact network with absolute certainty remains very hard, and even potentially undesirable because of privacy concerns [79]. Because of this lack of data-availability, algorithms in the source identification literature are typically tested on synthetic datasets instead of realistic epidemic datasets.

We only know about one paper that studies the effect of imperfections in the network data on the source identification task. Zejnilovic et al. [251] defines a generalization of the MD, the *extended metric dimension*, which models the difficulties of source identification in a deterministic epidemic, when $k-1$ edges are unknown, without which the network splits to k connected components. A strength of [251] is that their extended resolving sets can identify the source even without knowing where the missing edges are, however, their approach does not work in the case when the missing edges do not split the graph into disconnected components. For this dissertation, we were interested in understanding a more extreme scenario: how much difference can the uncertainty about a single edge can make, if that edge can be anywhere in the graph?

Contribution 5 (published in paper [173])**.** *In Chapter 7, we study how much the metric dimension (MD) can change if one edge is added to the graph. We draw conclusions about the implications of our results to source identification: there are graphs which are very sensitive to even a single edge change (the MD increases exponentially), however, in most graphs we expect smaller changes. We prove that in d-dimensional grid graphs, the MD at most doubles, and for 2-dimensional grid graphs we almost completely understand the effect of a single edge addition.*

Chapter 7 suggests that the uncertainly about the network can have a drastic effect on the source identification problem, which makes us reconsider the original assumption that the entire network is known to the algorithm. We adopt a more realistic assumption that only those edges (contacts) are known, which can be confirmed by a government monitoring agency with very high certainly, without violating the privacy of the entire population. We formalize this intuition in the final technical chapter of this dissertation.

> **Contribution 6** (published in paper [189])**.** *In Chapter 8, inspired the recent implementations of contact tracing algorithms, we propose a new framework for source identification, where the (possibly time dependent) edges of the network are adaptively queried, in addition to the adaptive querying of the infection state of the nodes (see Figure 1.1 (d)). We call this the Source Identification via Contact Tracing Framework (SICTF). We develop new local search algorithms for the SICTF, which outperform state of the art algorithms in simulations, and we provide analytical bounds on their success probabilites. We benchmark our algorithms on a COVID-19 simulator [162], which is the first time source identification algorithms are tested on such a complex and realistic dataset.*

1.4 Overview

The core of this dissertation is constituted by the six contributions introduced above, each of which is devoted its own chapter. The chapters are ordered based on how much novelty they bring to the field of source identification, in relaxing the limitations discussed in Sections 1.3.1 and 1.3.2. Chapter 2 on the switchover phenomenon of different seedings mainly serves as motivation. In Chapter 3, we review the relevant results on the role of adaptivity in deterministic epidemic models, and we include the specific contributions in Chapters 4 and 5. These results can be considered to be more standard, because the analysis is performed on well-researched combinatorial notions (the metric dimension and its adaptive variant, the sequential metric dimension). However, the proofs in these sections are challenging, and they bring several new theoretical tools to the field of source identification. In Chapter 6, we analyze source identification in stochastic epidemics rigorously for the first time in the literature, both in the adaptive and the non-adaptive version of the problem. The tools developed in this chapter are novel and pave the way for further analysis in more complex network structures. In Chapter 7, the robustness of metric dimension is studied to understand how uncertainty in the network structure affects source identification. Finally, Chapter 8 introduces a list of novel contributions, including adaptive queries to uncover the network structure, asymptomatic patients who do not report their infection time even if they are queried, and local algorithms that outperform state of the art source identification algorithms.

One of the key features of this dissertation is its multidisciplinary nature. Indeed, Chapters 4, 5, 6 and 7 are based on theoretical papers, whereas Chapters 2 and 8 have a strong applied component besides the theoretical analyses. More precisely, in the order of "appliedness", Chapter 2 belongs to the field of network science; Chapters 8 and 6 belong to applied and theoretical computer science, respectively; Chapters 4 and 5 both belong to the area of ran-

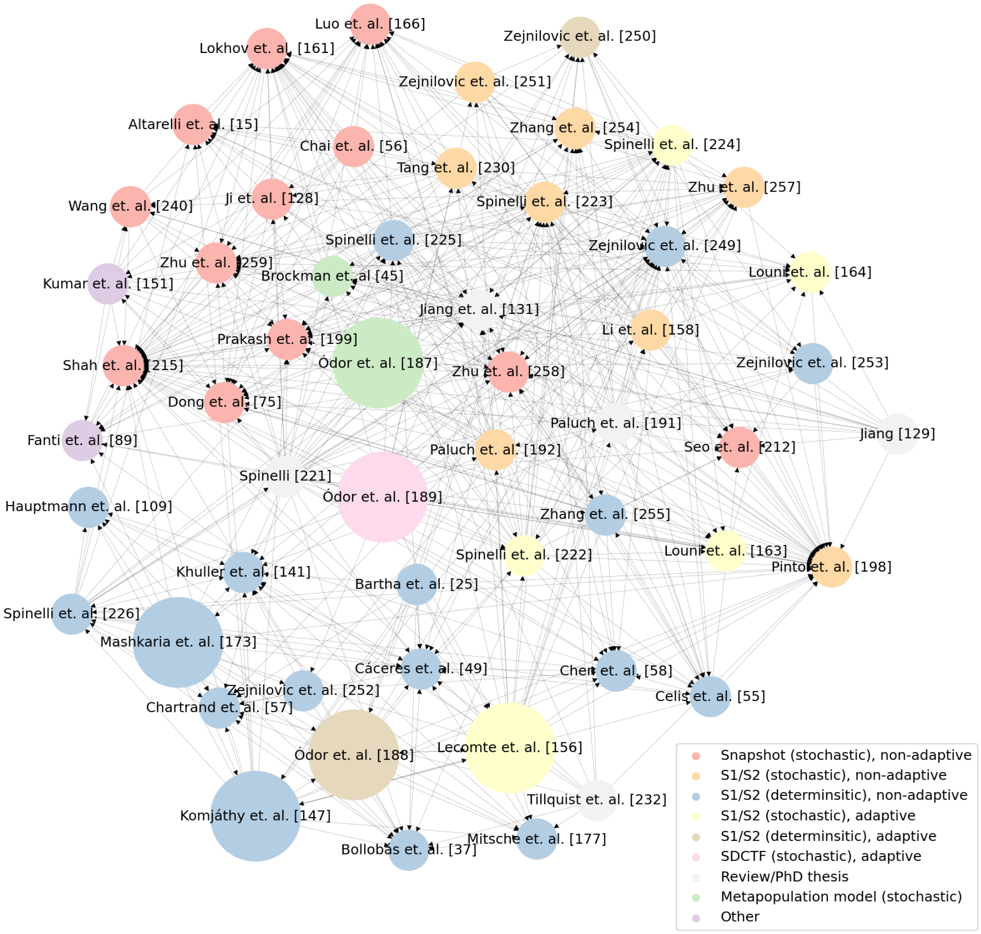


Figure 1.2: The citation network of the 55 most important papers related to this dissertation. The included papers were manually labeled based on source identification framework (or the epidemic model) adopted in them, and the colors of the nodes are determined based on these labels (see legend). Methodology: The papers included in the figure were selected by an iterative procedure in the spirit of bootstrap percolation: we start with the 6 papers $I_0 = \{[147, 156, 173, 187, 188, 189]\}$ that contain the main contributions of the dissertation (the nodes corresponding to these papers have a larger radius in the figure). Thereafter, if in the beginning of iteration i we had included the set of papers I_i, we constructed the undirected citation network [10] induced by I_i, together with the papers cited by I_i, and the papers that cite I_i, and we included the nodes with degree at least $|I_i|^{0.6}$ to I_{i+1}. We stopped the process after 5 iterations, when we had $|I_5| = 58$. After manual inspection, 4 papers that were not closely related were removed and the paper [44] was added to obtain a final list of 55 papers. Then, the papers were embedded into 20 dimensions via the Doc2Vec Natural Language Processing method [155] based on their abstracts, and into 2 dimensions via a subsequent Principle Component Analysis. Finally a force-directed graph drawing algorithm and minimal manual fine-tuning resulted in the citation network shown in the figure.

dom graphs, but Chapter 4 takes a more probabilistic approach, whereas Chapter 5 is more combinatorial, and finally, Chapter 7 is almost purely combinatorial.

To show the relationship between the different chapters, and how they fit into the related literature in an intuitive way, we collected related papers in an automatized way and embedded their citation network with a data-driven method (see Figure 1.2). The embedding was done based on a natural processing algorithm applied to the abstracts, to which we applied Principle Component Analysis (PCA). Since PCA is a method to find the biggest variability in the data, the best way to read Figure 1.2 is to start with the articles at the extremes. We would like to highlight the three pioneering papers by Shah and Zaman [215] (on the left), by Pinto, Thiran and Vetterli [198] (on the right), and by Bollobás, Mitsche and Prałat [36] (at the bottom). These three directions (left, right, bottom) in the embedding in Figure 1.2 correlate with the source identification framework used in the papers, which is shown by the coloring of the nodes (done by manual labelling). Indeed, red nodes use the model of [215] and are clustered to the left, orange/yellow nodes use the model of [198] and are clustered to the right, and blue nodes use the model of [36] and are clustered at the bottom. A closer inspection shows that the directions actually correspond to the scientific fields, which the papers were written in. For example, there are some papers colored blue at the top part of the figure; these correspond to the applied computer science papers which use the deterministic model. The six papers which constitute the core of the this dissertation are shown in Figure 1.2 with an increased radius. The four papers [147, 188, 156, 173] at the bottom correspond to the theoretical chapters 4, 5, 6 and 7, respectively. Their proximity to the paper of Pinto, Thiran and Vetterli [198] shows how close these chapters are to the original paper that started the field of source identification with time queries. The two papers [189, 187] corresponding to Chapters 2 and 8, respectively, are in the middle, because the former mainly serves as motivation for the rest of the chapters, while the latter proposes a new source identification framework and new research directions, and these two papers are not drawn to any of the extremes identified by the PCA algorithm.

The map in Figure 1.2 is intended to put this dissertation into perspective, and also to help understand the future directions, which will be presented in Chapter 9, the conclusion of the thesis.

2 The Switchover Phenomenon

In this chapter, we study how the location of the seed nodes affect the outcome of an epidemic in a metapopulation model, introduced in Section 1.1. We consider two scenarios for the location of the seeds, which we motivate by historical data about the COVID-19 pandemic in Hungary. We perform this data analysis in Section 2.1, after which we present our simulation and theoretical results in Section 2.2, followed by a discussion in Section 2.3. The methods used in the simulations, and the proofs are included in Appendix A.

This chapter is based on the publication [187] by Ódor, Czifra, Komjáthy, Lovász and Karsai.

2.1 Rationale and Data Analysis

Whether a local epidemic becomes a global pandemic depends on several conditions. Biological [195], environmental [227] and behavioral [97] factors are important but the final outcome of the epidemic is also strongly determined by the size and location of the seed population where it originates from [175, 61, 16, 143, 63, 99]. If the epidemic strikes first at an isolated place with low population density and few local transportation connections, it may become rapidly extinct without causing a major breakout. The dynamics can be entirely different if the epidemic starts from a well connected, more populated place where it can survive and spread to the rest of the population more easily. Although this is the broadly accepted picture, we challenge this intuition and show that seeding an epidemic from the most tightly connected core of a network does not always lead to a larger epidemic in the long run, in terms of the number of final infected people: If the disease transmits easily, seeding the spreading from nodes selected uniformly at random from the network could reach a larger population.

Among many factors, similar processes could act in the background during the early phase of the COVID-19 pandemic: Even though the circulating SARS Cov-2 epidemic variants had similar transmission profiles, the number of infections differed significantly in subsequent waves of the pandemic in several countries [41, 78, 241]. This was especially true for Hungary with an order of magnitude more daily number of detected cases observed at the peak of

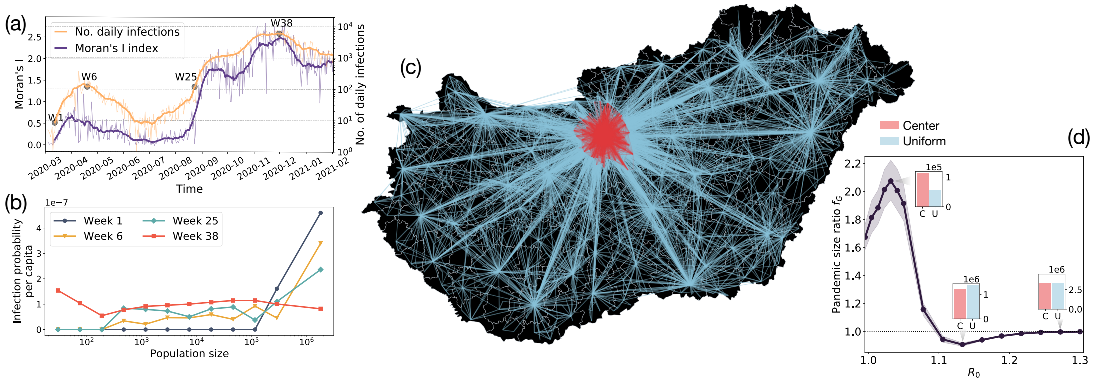


Figure 2.1: Data-driven observations of the switchover phenomenon. (a) Dynamics of the number of daily infections (orange) and the Moran's I index (purple) for Hungary. Indicated time points match the observation weeks in panel b. (b) Distribution of per capita infection probabilities in settlements of different sizes at different observation times (in weeks). (c) Commuting network map of Hungary with settlements larger than 1000 inhabitants and commuting links with more than 25 travelers depicted. Central Hungary (called Center) is highlighted with red. (d) Pandemic size ratios $f_G(R_0, s)$ measured between the endemic sizes of simulated SIR epidemic processes seeded from s towns selected from the center or uniformly at random from the whole metapopulation network. Epidemic seeded from the center may lead to larger outbreaks for small R_0 basic reproduction numbers (left bar plot), while uniform seeding results more infections for larger R_0 (middle bar plot). For very large R_0 values, differences due to different seeding strategies disappear (right bar plot).

the second wave as compared to the first outbreak (see Figure 2.1 (a)). Reasons behind this variation could be the effect of several factors. This includes seasonal effects as people may have spent more time outside during the first wave [176]; Regulations were followed less strictly during the second wave that may have potentially induced a larger number of contacts per person transmitting the disease [203]; The testing capacities also developed considerably since the beginning of the pandemic, allowing for more observations during the second wave; Further, while the first wave of the epidemic was boosted by institutional outbreaks (e.g. in hospitals and care homes) that were easier to identify and contain [46], the second wave circulated freely in the population without effective control [13].

The global and local mobility of people are among the most important driving factors behind the spatial spread of most diseases [149, 21, 229]. How people commute locally or travel between cities and countries can be well represented by mobility networks. Concentrating on Hungary, we consider a spatial mobility network (see Figure 2.1 (c)) describing the average number of daily commuters, who travel to work and school between 1398 settlements with populations larger than 1000 inhabitants according to the 2016 Hungarian micro-census [1]. From epidemic data we can follow the daily number of new COVID-19 infection cases in each of these settlements to explore their spatiotemporal distribution in this geometric network. The analysis of the epidemic on this structure sheds light on a so-far neglected effect associated to the different initial seeding conditions of the virus, which may contributed to the emerging large differences between the first and the second waves.

The first wave started in March 2020 in Hungary (W1 in Figure 2.1 (a)). As in many countries, the disease arrived to the country via international air-travel and first landed in larger cities [205, 139, 91, 135] resulting in outbreaks clumped around highly populated areas. This is evident from Figure 2.1 (b), where the per-capita infection probability at the beginning of the first wave (week 1) indicates that infection cases were concentrated in cities with the largest populations. To further demonstrate how much of the infection spreading can be attributed to everyday mobility (as opposed to atypical mobility patterns, such as going on a vacation), we computed the Moran's I index on this network (for definition see Appendix A.1). This is a spatial auto-correlation function, which has been previously used to measure the spatial association of the COVID-19 infections by [135]. Looking at the time dependence of the Moran's I index (on Figure 2.1 (a)), during the beginning of the first wave (W1) the index indicates low spatial correlation, meaning that infected cases were concentrated only in a few places during this initial stage of the epidemic. In contrast, the second wave in Hungary (and Europe) emerged after the summer season, and was potentially induced by people coming back from holidays bringing back the virus to their local community, and thus re-starting the pandemic from a significantly different initial condition. Indeed, at the beginning of the second wave (at the end of August 2020 in Hungary (see Figure 2.1 (a)), new infected cases were distributed more homogeneously all around the country. On the one hand, this is evident from Figure 2.1 (b) where the corresponding probability distribution (week 25) is more stretched towards smaller population, as compared to week 1. On the other hand, the same conclusion can be drawn from Figure 2.1 (a) (W25) where the Moran's I index starts to grow rapidly from a state where infections were even more homogeneously distributed than at the peak of the first wave (W6), although the infection numbers were comparable. This homogenization of infected cases continued during the unfolding of the second wave leading to a fully uniform distribution – corresponding to population densities – at the peak (W38 in Figure 2.1 (a) and b). Surprisingly, the first wave that started from the most tightly connected, central, and largest populations led to significantly smaller number of infections as compared to the second wave, that reached an order of magnitude more people, even though it was initiated from more uniformly distributed populations of the network.

To better understand this phenomenon we build an SIR *metapopulation network* [60], introduced in Section 1.1, using the spatial commuting network of Hungary [1]. We consider n nodes, which represent populations of individuals (which we also call towns or settlements from now on), connected by weighted edges, encoding the number of people traveling between them. In our simulations, the epidemic evolves in two phases in each iteration. During the reaction phase, individuals inside each town mix homogeneously, according to the mean-field SIR dynamics (see Section 1.1). Subsequently, during the diffusion phase, individuals (possibly infected) are selected with probability p_m to move to neighboring nodes in the metapopulation network, this way migrating the epidemic to other towns (for a more formal definition see Appendix A.1). Note that we concentrate on the conventional SIR model to demonstrate a new phenomenon. However, based on preliminary simulations (not shown), our observations hold for more realistic models too, including the SEIR model, which is an SIR model enriched with an addition compartment of exposed (E) state, which describes the reaction scheme of the SARS-Cov-2 disease better.

To capture the observed structural distinction of the central towns in case of the spatial commuting network of Hungary, we identify a *central node set* $\mathcal{C}$, containing the districts of Budapest and its suburbs (red nodes in Figure 2.1 (c)), which represent about 30% of the total population of the country [6]. While this definition of $\mathcal{C}$ relies on the specific urban structure of Hungary, we could find more general definitions for $\mathcal{C}$, that are based solely on the network structure. The simplest formal definition would be to take a prescribed number of nodes with the highest degrees. We could also use the core of the network for this purpose (for definition see Appendix A.1), which is obtained by repeatedly deleting all nodes with the lowest degrees as long as only nodes with prescribed degrees remain. Later we will exploit all these definitions to identify $\mathcal{C}$ when studying different types of model network structures.

Once we selected $\mathcal{C}$, we consider two initial conditions to seed the SIR process in the metapopulation network, starting the spreading from the same number of towns and individuals in both cases. In one case, we choose s $(< |\mathcal{C}| \ll n)$ towns randomly from the central set $\mathcal{C}$, while in the other case we choose s towns uniformly at random from the whole network. To initiate the spreading, we infect a small i_0 fraction of the total population selected uniformly at random from the chosen s towns, irrespective of their size. This way, for both seeding strategies (centralized or uniform), each seeded town is infected on average with the same number of agents (i_0/s fraction of the total population). To observe the relative effects of the two seeding scenarios, we look at the *experimental pandemic size ratio* $f_G(R_0, s)$, that we define as the ratio of average final infection sizes of epidemic processes seeded from central or uniformly randomly selected towns (for a related but more formal definition, see Section 2.2 on the theoretical results below). Interestingly, as shown in Figure 2.1 (d), we find $f_G(R_0, s) > 1$ for small $R_0 \simeq 1$, which means that the epidemics seeded from the central set $\mathcal{C}$ leads to larger outbreaks. However, as we increase R_0, the fraction $f_G(R_0, s)$ falls under 1, thus seeding from uniformly random selected towns over the whole country induces a larger outbreak. This *switchover phenomenon* appears in the slightly super-critical regime, where R_0 is not too large, and where the epidemic never reaches the total population, but stays clustered around the seeded towns until it dies out. The differences in the infected cluster sizes induced by the two seeding scenarios lead to the observed switchover phenomenon in this regime. Finally, if R_0 grows larger, the difference between these seeding scenarios vanishes as the epidemic reaches essentially the whole population in each case.

The switchover phenomenon challenges the commonly accepted intuition that the size of the epidemic is always the largest if seeded from the best connected sub-graph, or from the largest degree nodes of a network. In the remaining sections of this chapter, we show that the switchover phenomenon is ubiquitous in various network models, and argue that the geometric nature of the underlying network plays an important role in amplifying the size of the phenomenon. We perform data-driven and synthetic simulations of spreading processes on real, geometric, and random metapopulation networks and provide a rigorous proof of the phenomenon after mapping it to a bond percolation problem. Our results open a new research direction towards understanding why real epidemics started from seemingly similar conditions may have significantly different outcomes.

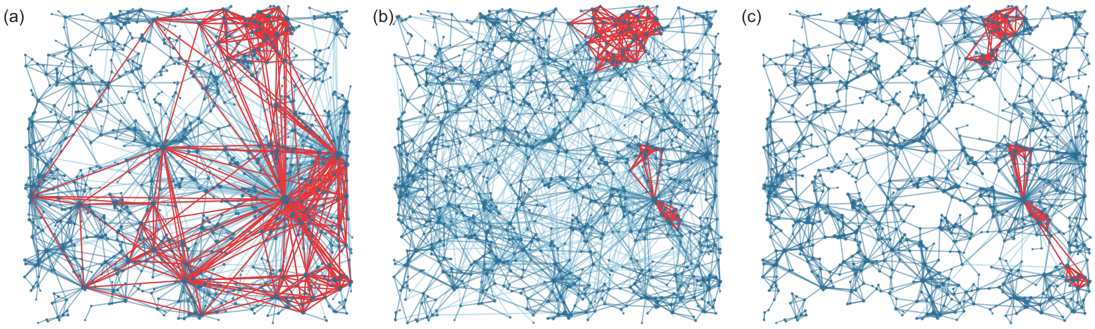


Figure 2.2: Geometric Inhomogeneous Random Graph (GIRG) models. (a) Model networks with connection parameter $\tau = 2.5$ and geometry parameter $\alpha = 2.3$. In case $\tau < 3$ the network appears with high degree variability and dominant hubs connected via (light blue) long-range edges. (b) By increasing $\tau > 3$ (here $\tau = 3.5$) and decreasing $\alpha = 1.3$, hubs' sizes reduce and long-range interactions become more random. (c) With parameter $\alpha = 2.3$ (and $\tau = 3.5$) the network is strongly geometric with (dark blue) short-range interactions and no long-range links. Networks were generated over the same $N = 1000$ nodes randomly distributed in a unit square and the highest k-cores of each graph are colored in red.

2.2 Results

2.2.1 Simulation Results

In the previous section, we demonstrated the existence of the switchover phenomenon on a metapopulation model parametrized by the Hungarian commuting network, which is a spatially embedded geometric network (see Figure 2.1 (c) for Hungary) featuring various structural heterogeneities (for a detailed data description see Appendix A.1). Geometric constraints inducing commuting connections at various distances, link weights coding the daily commuting frequencies between towns, the number of commuting connections of each settlement (also called the node degree in the network), or the size of the different towns are all network characteristics taking values ranging over orders of magnitudes. These properties may all contribute to the emergence of the observed switchover phenomenon of simulated spreading processes (an SIR model in our case), with central vs random seeding in the meta-network.

To identify which underlying network characteristics are the most important to play a role, we use random reference network models [182]. We homogenize the network in different ways to remove certain structural heterogeneities, and compare the outcome of the experimental pandemic size ratio of simulated spreading processes on the randomized structures to our observations on the empirical network (see blue dotted curve in Fig 2.3 (a)). First, to reduce the effects of weight heterogeneities, we reset edge weights to the mean weight of all outgoing edges of each node (see green diamond curve in Fig 2.3 (a)). Although this way of homogenization changes somewhat the pandemic size ratio function, it does not have dramatic effects on the observed phenomena. Second, to remove the effects of heterogeneous town sizes and the varying number of commuting individuals from different settlements, we set each town's population to the system average ($\overline{N} = 6581$) and choose the fraction of commuters to be the same (i.e. to $p_m = 0.001$) for each town. Interestingly, this way of homogenization makes the switchover phenomenon even stronger (see red squared curve in Fig 2.3 (a)). Finally, we re-shuffle the ends of network links using the configuration network model [182]. This removes any structural correlations from the network beyond degree heterogeneity, including geometric effects such as long distance connections, the central-periphery structure, structural hierarchy, and locally dense sub-graphs. Due to this shuffling process the switchover phenomenon disappears, or becomes too small to be observed (see yellow triangle curve in Fig 2.3 (a)), indicating that geometric correlations play a central role behind its emergence.

Geometric Inhomogeneous Random Graphs. The specific effect of an underlying geometry can be studied by using geometric network models, opening directions for an analytical description of the phenomenon. *Geometric Inhomogeneous Random Graph* (GIRG) models [43] provide a good framework to generate structurally heterogeneous synthetic metapopulation structures embedded in geometric space (for detailed definition see Appendix A.1). GIRGs have two robust parameters that control the qualitative features of the emerging network. The parameter τ determines the variability of the number of neighbors of individual nodes (smaller values of τ correspond to more variability, while keeping the average degree the same). This is apparent when comparing Figure 2.2 (a) to 2.2 (c) where all parameters of the simulated network structures are identical, only τ is increased gradually, leading to the disappearance of hubs, i.e., nodes with large number of neighbors. The other robust parameter of GIRG, α controls the number of long-range connections in the network coding the possible travels between far-apart nodes. If $\alpha \simeq 1$, many long-range edges appear resembling an ageometric (or mean-field) structure (see Figure 2.2 (b)), but when α is increased, the number of long-range contacts are reduced, and the network exhibits a more apparent underlying geometry as shown in Figure 2.2 (c). The values of (τ, α) determine different *universality classes* of GIRGs with respect to average distance in the network (see more detailed definition and explanation in Appendix A.1).

To distinguish the central set $\mathcal{C}$ from the rest of the network, we adopt here the concept of *core decomposition* (for definition see Appendix A.1). Formally, this procedure provides the highest core as a sub-graph with nodes having at least k neighbors inside the core, for the largest possible k (for definition see Appendix A.1). Similarly to the data-driven simulations,

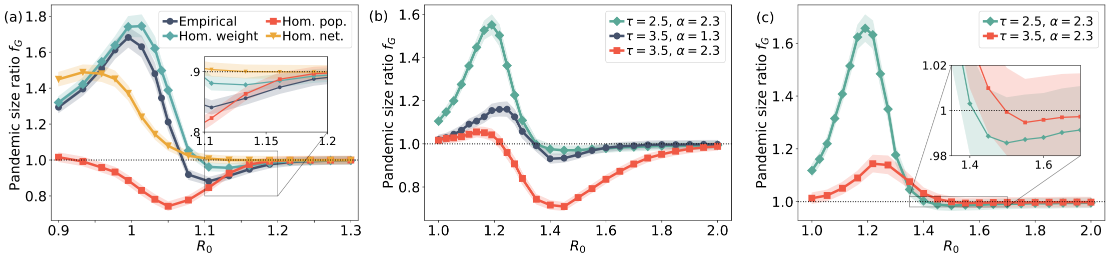


Figure 2.3: The pandemic size ratio f_G as a function of R_0. a) Simulation results on the real commuting network of Hungary and its three homogenized versions as explained in the main text. Each data point is an average computed from 150 independent simulations, shown with 81% confidence interval. For each of them, initially $s = 97$ settlements are selected according to one of the seeding scenarios (central or uniform). Then, we infect $i_0 = 0.0005$ fraction of the 10^7 agents in the total population, and we distribute these agents in the s settlements uniformly at random, irrespective of the size of the settlements. b) Pandemic size ratio computed on three geometric inhomogeneous random graph models corresponding to the three main universality classes with respect to graph distance. When $(\tau, \alpha) = (2.5, 2.3)$, the core is de-localized, the switchover is weak. When $(\tau, \alpha) = (3.5, 1.3)$ or $(3.5, 2.3)$, the underlying geometry is more apparent, the core is localized, so the parameter range for R_0 where random seeding is more dangerous ($f_G < 1$) is more spread-out. On c), we see f_G on configuration model networks, where the switchover phenomenon appears weaker. For b),c) the size of the metapopulation networks are $n = 1000$ and each town is set with N=2000 individuals. Each pandemic size ratio data point is computed on 25 networks, with 35 simulations on each network, with $i_0 = 0.0005$ and $s = 30$. In all simulations we set $p_m = 0.001$, which means that 0.1% of the population moves in each iteration.

we start the spreading process from two seeding conditions: by initially selecting s towns within the highest core of the metapopulation network (corresponding to the central set $\mathcal{C}$), or by selecting the same number of towns uniformly at random from the whole structure, and infecting on average the same number of agents (i_0/s fraction of the total population) in each selected town in both scenarios.

We observe a similar but stronger switchover phenomenon of the pandemic size ratio f_G in GIRGs as compared to the data-driven simulations. As seen in Figure 2.3 (b), the "shape" of $f_G(R_0, s)$ as the function of R_0 strongly depends on the network properties controlled by the parameters of the model. With network parameter $\tau \geq 3$, the modeled epidemic processes, which were initiated from uniform random seeds, reach larger populations: this is reflected by f_G (blue curve in Figure 2.3 (b)) falling well below 1 for a broad range of R_0. This is because the hubs in the network have relatively smaller degrees compared to the case when $\tau < 3$. They are too far away from each other to form direct connections, therefore the highest cores are localized around some of them, as demonstrated in Figure 2.2 (b) and (c). Although high degree seed nodes in these cores should have an advantage to effectively induce a larger outbreak, this effect is not strong enough to compensate for the disadvantage of starting the infection from a localized setup. Beyond localized cores, long range interactions also have important effects on the network structure. Rare long range connections (induced by higher α values) reduce the number of edges leaving the localized cores, which leads to networks with dominant local geometric structures (as shown in Figure 2.2 (c)). This makes it even harder for the infection to spread from a localized setup. Thus, for $\tau \geq 3$, increasing α enhances the danger of the random seeding scenario, as evident from Figure 2.3 (b) where the (red) curve with $\tau = 3.5$ and $\alpha = 2.3$ dives below 1 more than a similar curve with $\alpha = 1.3$. Finally, when $\tau \in (2,3)$ (see green curve in Figure 2.3 (b)), the pandemic size ratio f_G goes well above 1 for $R_0 \simeq 1$ values, and goes barely below 1 for larger R_0. In this case the highest degree nodes are so dominant that they connect to each other even when they are spatially remote, this way they induce a de-localized core (see Figure 2.2 (a)). Simulations on these networks with de-localized cores resemble the phenomenon that is closest to our data-driven simulations (see Fig 2.3 (a)) where the effects of the geometry are somewhat reduced due to the inter-connectedness of larger cities all over a country. For $\tau \in (2,3)$, the parameter α does not have a significant effect on the network structure.

For comparison, we also study the phenomenon on meta-networks sampled from the *configuration model*, a uniform distribution over networks with a given power-law degree sequence. The configuration model has no underlying geometry and features heterogeneity only in its degree distribution, parametrized by the exponent of the power-law distribution τ (see details in Appendix A.1). For a fair comparison with the results obtained on GIRGs, we take $\tau = 2.5$ to obtain a configuration model with plenty of hubs, and $\tau = 3.5$ for a model with reduced degree heterogeneity. These cases correspond to two different universality classes (in both GIRGs and in configuration models) with respect to average distance (see Appendix A.1). To keep the average degree and the number of nodes the same for the configuration model networks as in GIRGs, we obtain them by swapping randomly the links of the GIRG structures, while keeping the total number of connections for each node the same. Interestingly, when larger hubs are present in the structure (the case of $\tau = 2.5$ in Figure 2.3 (c)) the switchover phenomenon is recovered, even though the structure is fully uncorrelated. However, the switchover appears weaker, similar to the case on GIRGs where the effect of the geometry is suppressed due to the high inter-connectedness of the network. We provide a heuristic explanation of this observation during the derivation of our theoretical results below.

In summary, our simulation results demonstrate that while the emergence of the switchover phenomenon requires only degree heterogeneity in the network, it is certainly amplified by geometric correlations of the underlying structure.

2.2.2 Theoretical Results

To explain rigorously the switchover phenomena we developed a mathematical framework relying on percolation theory.

Epidemics and percolation on metapopulation networks

The pandemic size (i.e., the final number of recovered individuals) of a SIR model with deterministic, unit recovery time (e.g. a day) on a (non-meta) network G has a useful connection with the commonly used simple mathematical framework of *bond percolation*. In such a SIR model, every edge of the network G transmits the disease at most once, when one endpoint is infected but the other is still susceptible. Equivalently, one may decide about every edge *in advance*, independently with probability p, whether it will do so. This is called *retaining the edge*, and p is then the retention probability of the model. The retained edges form the percolated random subgraph G^p of G. If a set S of nodes is selected as infected seeds in the network, then the epidemic will spread exactly over the connected components (also called clusters) of G^p that contain at least one node of S.

Metapopulation models are more difficult to treat mathematically, but a fundamental result by [25, 62] connects the behavior of SIR on metapopulation models to bond percolation. Following their arguments, once a large outbreak occurs in a town A, the proportion of infected people within the town *concentrates* around some $r_\infty \in (0,1)$ (called local outbreak ratio). Infected people during the local pandemic carry the infection to a neighboring town B and cause a large outbreak there with a certain – computable – probability:

$$p_{AB} = 1 - \exp\left(-\frac{N p_m w_{AB} r_\infty \left(1 - \frac{1}{R_0}\right)}{\mu}\right), \tag{2.1}$$

where N is the (initial) number of individuals in each town A and B (assumed to be identical), p_m is the fraction of commuters, w_{AB} is the edge weight scaling the number commuters between towns A and B, R_0 is the basic reproduction number, and μ is the recovery rate. Note that the dependence on A and B can be neglected if we assume an unweighted network. Since herd immunity is reached in each town after the first large local epidemic outbreak of size $r_\infty N$, later infections to a town are no longer able to cause macroscopically visible outbreaks. Therefore, after time-rescaling, the towns themselves go through an $S \to I \to R$ progression with unit recovery times and infection probability p given by equation (2.1). Consequently, the metapopulation model can be approximated by a simple SIR model on the network of towns, and in turn with a bond percolation process with retention probability p.

The connection between metapopulation models and bond percolation allows us to understand the switchover phenomenon of the pandemic size ratio using a theoretical analysis of percolation cluster sizes, which have been extensively studied both in the mathematics and physics literature for various network models, because they show a remarkable *phase transition* in the edge retention probability p. At a *critical value* p_c two phases are separated, where for $p < p_c$ all clusters are small, while for $p > p_c$ a single *giant cluster* emerges that contains a positive proportion of all nodes, while all other clusters are small. The critical parameter p_c depends only on the structure of the network G. For some networks, p_c can only be measured using numerical simulations. However, for the configuration model, the critical p_c can be explicitly computed, given the degree-distribution of the network, as $p_c = \mathbb{E}[\deg(v)]/\mathbb{E}[\deg(v)(\deg(v)-1)]$, which is asymptotically nonzero when $\tau > 3$. Using equation (2.1) the critical parameter p_c translates back to a critical basic reproduction number $R_c^{\text{glob}} > 1$ for the infection process. For $R_0 < 1$ the epidemic is sub-critical already within a single town, while for $1 < R_0 < R_c^{\text{glob}}$ the epidemic is super-critical within towns but sub-critical globally in the meta-network (hence outbreaks containing only a few towns are possible). Finally, for $R_0 > R_c^{\text{glob}}$ the epidemic is super-critical in the entire network.

Beyond percolation cluster sizes, we also need to understand how the different seedings (central or uniform) interact with the clusters to explain the switchover phenomenon of the pandemic size ratio. Slightly deviating from the experimental setup, where central seeding corresponded to the highest core, here we define the central seeding set $\mathcal{CI}_0(s)$ as the *s highest degree nodes*. This can be done as the two definitions are strongly correlated in the network models we focus on in this section [94, 127, 165]. For the uniform seeding, just as earlier, we choose the seed set $\mathcal{UI}_0(s)$ as s nodes sampled uniformly at random in the whole network. In both setups we look at $\mathbb{E}_p[\mathbf{Cl}(\mathcal{CI}_0(s))]$ and $\mathbb{E}_p[\mathbf{Cl}(\mathcal{UI}_0(s))]$, the *average percolation cluster sizes* of the initially infected nodes, when edges are retained with probability p. This corresponds to the average number of populations that experience local large outbreaks in the two seeding scenarios. The *percolation pandemic size ratio function* is then defined as the ratio of these two averages:

$$f_G(p,s) = \mathbb{E}_p[\mathbf{Cl}(\mathcal{CI}_0(s))]/\mathbb{E}_p[\mathbf{Cl}(\mathcal{UI}_0(s))], \tag{2.2}$$

similarly to the earlier defined experimental function.

We define two approaches to the switchover phenomenon on a meta-network of n cities. In the *weak switchover phenomenon*, we require that there exists a seed count $s \le n$ and link-retention probabilities $0 < p_1, p_2 < 1$ with

$$f_G(p_1,s) > 1+c, \quad \text{and} \quad f_G(p_2,s) < 1-c, \tag{2.3}$$

for some constant c that might depend on the network size. Meanwhile, in the *strong switchover phenomenon*, we require that the constant c *does not depend on the network size n* and thus holds across a whole model class (e.g. GIRG or configuration model with fixed degree heterogeneity). When the switchover occurs for a seed count s, we say that the switch happens at retention probability p_{switch} if $f_G(p,s) > 1$ for $p < p_{\text{switch}}$, while $f_G(p,s) < 1$ for $p > p_{\text{switch}}$.

While the switchover phenomenon in GIRGs is hard to study analytically due to the lack of percolation theory developed for this model, we borrow concepts from a simpler conventional network model, called Stochastic Block Model (SBM), to observe the strong switchover phenomenon. The SBM is able to mimic the central and rural areas of a town network, since it contains a 'hidden geometry': We group towns into two sets of central or rural areas. Within areas we assume ageometric random networks, i.e., each pair of nodes is connected with the same probability, while the edge density between the two areas is lower.

Theorem 2.2.1. *In the Stochastic Block Model with appropriately scaled parameters and* $s_n = \Theta(n)$ *the strong switchover phenomenon happens. (For a proof see [187].)*

In case of the ageometric configuration model, we are able to prove the weak switchover, already observed experimentally in Figure 2.3 (c). Further, we are able to give quantitative bounds on c in (2.3) as a function of the size n of the population network G, the parameter τ expressing the prevalence of hubs, and the initial seed number $s = s_n$ that may also depend on the network size n.

Theorem 2.2.2. *In the configuration model with exponent* $\tau \in (2,4)$ *and* $1 \ll s_n \ll n$ *the weak switchover phenomenon appears with* p_{switch} *slightly above the critical percolation parameter* p_c*. (For a proof see the Appendix A.2.)*

While Theorem 2.2.2 is valid for $\tau \in (2,4)$, the two regimes $\tau \in (2,3)$ and $\tau \in (3,4)$ quantitatively differ. In the former case, also called the scale-free regime, p_c tends to zero as the network size n grows and the region of the parameter space where seeding from nodes selected uniformly randomly is more dangerous is described by different linear equations compared to the $\tau \in (3,4)$ case. We believe that the switchover phenomenon disappears when $\tau > 4$ as hubs become too small and separated from each other to produce the desired effect.

Theorem 2.2.2 is proven (partially based on some non-rigorous, but well-accepted results in the network science community [183, 59]) in Appendix A.2. Here we give a *heuristic explanation* of the switchover phenomenon, which also plays a key role in the proofs:

a) Below the percolation threshold, as demonstrated in Figure 2.4 (a), the connected components of the central area seed nodes (indicated as red nodes in Figure 2.4 (a)) will be much larger than the connected components of the uniformly randomly selected seed nodes. However, they do not yet form a giant component. Nodes selected uniformly at random (blue nodes in Figure 2.4 (a)) are not likely to be in these large components, hence the union of the connected components of seeds selected uniformly at random from the network will be smaller than the union of the connected components of the central seeds. For very small values of p, seeding the highest degree nodes is the most dangerous initialization of the pandemic on every graph.

b) Slightly above the percolation threshold, there will be a single "giant component" of nodes experiencing a local pandemic, containing most of the central nodes. Thus when seeding starts from the central nodes, their components will contain this giant component and a small portion of smaller components. In the uniform seeding scenario, it is likely that a few of the random seed nodes will also belong to the giant component, while the other seeds, spread out randomly over the rest of the network, will be contained in lots of additional smaller components. Hence the union of the components of the uniformly chosen nodes will be larger, as shown in Figure 2.4 (b).

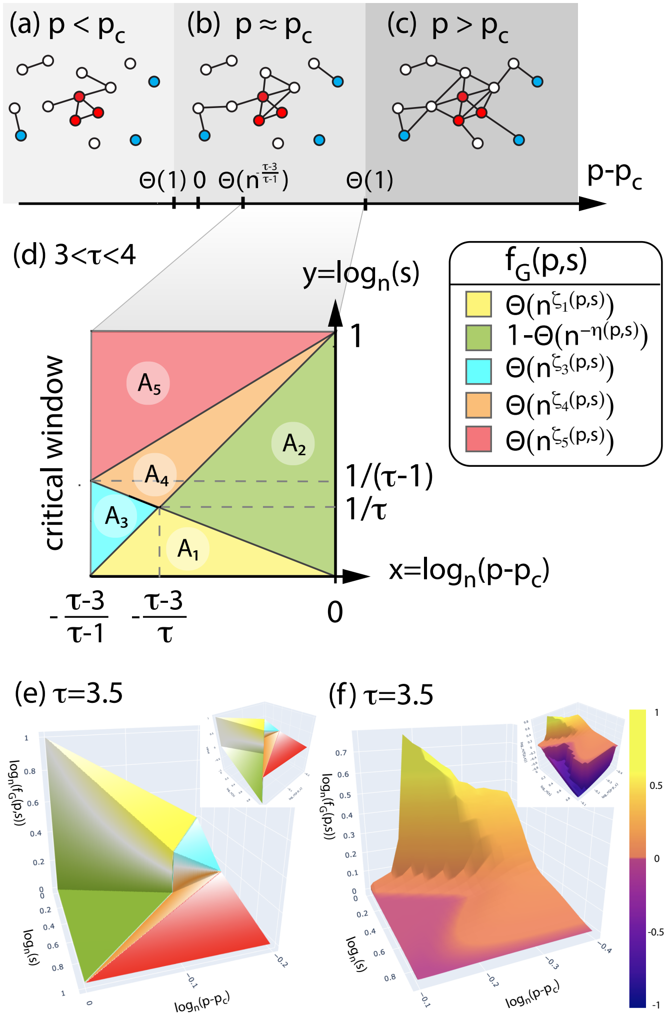


Figure 2.4: Panels (a), (b) and (c) show the heuristic explanation of the switchover of the pandemic size ratio function f_G in the configuration model. Panel (d) shows the phase diagram of the function f_G for $3 < \tau < 4$, for values of p slightly above the percolation threshold and for various values of s. The asymptotic of f_G is different in parameter regions with different colors. The precise values of ζ_i in the legend of panel (d) are included in Appendix A.1, and the phase diagram for $2 < \tau < 3$ is included in Appendix A.2. Panel (e) shows the 3D plot the limit function of $\log_n(f_G)$ for $\tau = 3.5$, as the number of nodes in G tends to infinity, and panel (f) shows the corresponding simulation results on configuration model networks with $n = 10^7$ nodes (each datapoint is an average of 1000 independent percolation instances on 10 independent random networks, after outlier removal is performed as explained in Appendix A.2). The coloring on panel (e) follows the coloring on the phase diagram on panel (d). Since in the configuration model we only have weak switchover, the green part of the surface $\log_n(f_G)$ (which corresponds to $f_G < 1$) converges to 0. For a visualization of the precise deviation of $\log_n(f_G)$ below 0, in the inset of panels (e) and (f), we plot the function $\log_n(f_G)$ for $f_G > 1$, and $-\log_n(1 - f_G) - 1$ for $f_G < 1$.

c) Well above the percolation threshold (see Figure 2.4 (c)), there is essentially only one connected component, thus each node gets infected regardless of the position of the seed nodes in the network.

The phenomenon in b) is stronger when there are relatively few edges leaving the central area, which can be due to lack of long range interactions amplifying local geometric effects (as observed in GIRGs, see Figure 2.3). However, in Theorem 2.2.1 the geometry induced by the two blocks is already enough to cause a strong switchover. On the contrary, there is nothing to limit the number of edges leaving the central area in the configuration model, (the degree-degree correlation coefficient is close to 0 [160, 113, 117]), hence the switchover phenomenon is weak.

Quantitative results for the configuration model

For geometric networks with various node degree distributions, *critical exponents* have been already proposed earlier [183, 59], with some of them proven rigorously for the configuration model (possibly with power-law degree distribution) [71, 72, 70, 236], as well as for rank-1 inhomogeneous random graphs [31, 32, 237] and for Erdős-Rényi graphs [73]. Based on these results, we can prove that, after appropriate scaling, the pandemic size ratio f_G of the configuration model (simulated on Figure 2.3 (c) for a fixed number s of seeds) converges to a two-dimensional limit function, which can be precisely determined. To state our result, let us re-parametrize $f_G(p,s)$ as a function of x, y where $p = p_c + n^x$ for $x \in (-(\tau-3)/(\tau-1), 0)$ and $s = s_n = n^y$ for $y \in (0,1)$, i.e., we consider $\widetilde{f_G}(x,y) := f_G(p_c + n^x, n^y)$. For $\tau \in (3,4)$, we divide the parameter space into five triangular regions A_1-A_5 illustrated on Figure 2.4 (d) (defined precisely in Appendix A.2). For $\tau \in (2,3)$, the picture is similar, except there is an extra triangular region A_6.

Theorem 2.2.3. *On the configuration model with exponent $\tau \in (2,4)$, $\log_n(\widetilde{f_G}(x,y))$ converges to a function $\zeta(x,y)$. On each triangular region A_i, $\zeta(x,y)$ can be expressed as a linear function $\zeta_i(x,y)$ specific to A_i.*

Theorem 2.2.3 implies that the percolation pandemic size ratio $f_G(p_c + n^x, n^y) = \Theta(n^{\zeta_i(x,y)})$ on A_i for each i = 1-5. We give the formula for each ζ_i in Appendix A.1, and the proof of Theorem 2.2.3 in Appendix A.2. A three-dimensional illustration of ζ can be seen in Figure 2.4 (e). The limiting function ζ is discontinuous at the boundary line between A_1 and A_2 and between A_3 and A_4, respectively. These discontinuities correspond to a discontinuous phase transition of the system's behavior at those boundaries. Curves in Figure 2.3 correspond to horizontal cross-sections of the two-dimensional f_G function for fixed s values. We experience this phase transition on the curves of Figure 2.3 dropping steeply from above 1 to below 1 when slightly increasing R_0. Our result implies that the curves will look steeper and steeper as n, the size of the network increases.

We also show that $f_G(p_c + n^x, n^y) = 1 - \Theta(n^{\eta(x,y)})$ in region A_2. Hence, the scaling $\log_n(\widetilde{f_G})$ in

Theorem 2.2.3 is not appropriate in region A_2, which is reflected by the fact that the limiting function ζ_2 is identically 0 in this regime. To be able to compute by how much f_G ((2.2)) goes below 1 in A_2, i.e., how much more dangerous uniform seeding can be compared to central seeding, we extract the limiting exponent $\eta(x, y)$ using a different normalization for f_G, and give the formula in Appendix A.1. This different normalization is used on the area A_2 (green) in the inset of Figure 2.4 (e), which in turn demonstrates that f_G falls below 1 in this regime. Finally, we validate our theoretical results for the configuration model by simulations in Figure 2.4 (f) [8]. Despite the finite size of the simulations ($n = 5 * 10^6$), the resemblance to the theoretical predictions is already apparent.

2.3 Discussion

Different seedings of an epidemic can lead to significantly different outcomes depending on the actual value of the basic reproduction number R_0. While R_0 is defined by the biological parameters of the spreading disease, it is only one factor determining the effective reproduction rate R_t, which characterizes the actual speed of reproduction during an ongoing epidemic. In case of an influenza like disease, R_t depends on many other factors including the actual number of interactions of people, the actual interventions, the self-protection measures (e.g. masks, sanitizing, etc.) or even the seasonal variance of temperature and humidity. Considering the distribution of the initial seeds and the actual R_t values, our theory may suggests counter-intuitive effects during the consecutive waves of a pandemic. This could be the case in Hungary, where the first wave of the COVID-19 pandemic was initiated from large, well connected towns, while social distancing was very effective at the time, causing a smaller actual R_t value during this period. Thus the clumped initial seeds and the somewhat low R_t could set relatively favorable conditions for the epidemic to reach a larger population, as compared to a uniformly seeded situation. Meanwhile, at the beginning of the second wave, seeded towns were more distributed all around the country, while social distancing was not followed rigorously. This induced larger R_t values, which yet again set relatively easier conditions for the epidemic to reach a larger population, now seeded from a uniform initial state.

In this chapter we studied the effects of epidemic seeding on geometric metapopulation networks. We were interested in the long-term behavior of spreading processes and showed that the relative danger of infecting a larger population when starting the process from the core or uniformly at random in a network has a non-monotonous dependence on R_0. We explored an entirely new switchover phenomenon and demonstrated them on real and synthetic networks via numerical simulations. We provided a rigorous proof for the existence of this phenomenon on a large class of random graphs, while we are confident that our theory can be extended for a more general class of graphs, which satisfies certain structural constraints. Importantly, we identified the spatial geometry of the underlying structure as an important amplifying factor of the switchover phenomenon.

We build our theory on some results [183, 59], which are broadly accepted by the network science community, yet it has not been proven rigorously for all network structures (for exceptions see Appendix A.2). This implies certain limitations for our results, although assuming these results to hold, our proposed theory has been derived rigorously. In addition, we took some assumptions for the simplicity of our presentation but their generalization is possible. We demonstrated experimentally the switchover phenomenon on directed networks, while we assume an undirected structure in our theory, which can be extended for directed structures easily. Moreover, we conjecture that the observed phenomenon occurs in most networks where a "central" region can be meaningfully distinguished in the structure. While metapopulation networks are generally used to model spreading phenomena at global spatial scales where towns are well separable, they provide a useful tool to study epidemics at shorter spatial distances too [105]. In our case, we applied metapopulation networks on the country level but we carefully separated the scales of flows inside and between towns (by setting $p_m = 0.001$), this way evidently separating towns from each other. We concentrated on the conventional SIR model for the demonstration of the switchover phenomenon, but this observation holds for more realistic models, including the SEIR model with an addition compartment of exposed (E) state, better capturing the reaction scheme of the SARS-Cov-2 disease. Our goal in this chapter was to identify, verify, and mathematically prove the existence of the observed switchover phenomenon, not to provide predictions. Accordingly, we worked with the simplest models and ignored the effects of possible interventions, seasonal weather conditions, superinfection events, permanent residence changes, etc. It indicates the fundamental nature of our observations, that surprisingly they occurred even under these simplified circumstances.

Beyond scientific merit, our results may contribute to better designs of epidemic forecasts and intervention strategies in a country during an ongoing pandemic. We highlight the importance to follow not only the rate but also the spatial distribution of new infection cases of a spreading disease or its variants during the early phase of an epidemic. This could lead to new testing strategies, which disclose the spatial distribution of the epidemic during its initial phase, as this was the case in some countries (like Denmark [114]) from the beginning of the COVID-19 pandemic. Based on these early-time observations our theory provides understanding about the long-term consequences of an epidemic by considering the commonly overlooked convoluted effects of epidemic seeding and the geometric structure of human populations and mobility.

Deterministic Spreading Part II

3 The Metric Dimension and Related Notions

In this chapter, we review the combinatorial notions related to the metric dimension (MD), with a special focus on the MD itself, and its adaptive variant, the sequential metric dimension (SMD). Sections 3.1 and 3.3 give the formal definitions of these two notions, and introduce notation, which will be useful in the later chapters. Finally, in Section 3.4 we compare the difference between the MD and the SMD in the random graph models where they have already been studied, and we use these findings to give a detailed overview on the role of adaptivity in source identification with time queries, when the epidemic spreads deterministically.

This chapter can be seen as a second introduction for the Chapters 4, 5 and 7.

3.1 Metric Dimension

The metric dimension (MD) was defined by Slater [218] and independently by Harary and Melter [107]. It is a notion of dimension that can be applied to any finite graph (or finite discrete metric space), and it is based on the idea that in a d-dimensional vector space, every vector can be uniquely identified by d coordinates (and the basis vectors). In the MD, the coordinates of the nodes are determined by distances on the graph from a distinguished set of nodes, called the *metric basis*.

Definition 3.1.1 (MD)**.** *Let $G = (V, E)$ be a simple connected graph, and let $d(v, w) \in \mathbb{N}$ be the length of the shortest path between nodes v and w. For $R = \{w_1, \dots, w_{|R|}\} \subseteq V$ let $d(R, v) \in \mathbb{N}^{|R|}$ be the vector whose entries are defined by $d(R, v)_i = d(w_i, v)$. A subset $R \subseteq V$ is a* resolving set *in G if $d(R, v_1) = d(R, v_2)$ holds only when $v_1 = v_2$. A resolving set of minimum cardinality is called a* metric basis*, and the cardinality of these metric bases is the* metric dimension *of G.*

Incidentally, the MD is equivalent to the minimum number of sensors needed in the non-adaptive S2 source identification model introduced in Section 1.2 if the transmission delay distribution $\mathcal{W}$ is deterministic [248]; in this case we are allowed to directly observe the distance between the sensors and the source. Besides source identification, resolving sets have a wide range of applications, including robot navigation [141, 216], computational chemistry

[56], and network discovery [26].

Computing resolving sets or even the metric dimension for general graphs is shown to be NP-hard [141] and it is approximable only up to a factor of $\log(N)$ [26, 109]. The MD of specific deterministic graph families has been extensively studied, we refer to [48, 202] for a list of references. For tree graphs, the MD can be written as the difference of the number of leaves and so-called exterior major vertices of the tree (vertices of at least degree 3 that have a line-graph leading to a leaf), both of which can be computed in linear time [141]. We mention that the MD has deep connections to the automorphism group of the graph G [20, 47, 98], and hence the graph isomorphism problem [19].

From the probabilistic point of view, there has been a recent effort to understand the *asymptotic behavior* of MD of random graph families as their sizes tend to infinity. A pioneering work [36] determines the asymptotics of MD of Erdős-Rényi random graphs. In this Law of Large Numbers (LLN) type of result, the authors showed a surprising non-monotonous zig-zag phenomenon of the metric dimension as the average degree increases from bounded to linear in the graph size. A recent work by Lichev, Mitsche and Prałat [159] analyzed the metric dimension of random geometric graphs of n nodes in the unit square. A central limit theorem (CLT) type result for uniform random trees was determined in [177], and also for sub-critical Erdős-Rényi random graphs. In Chapter 4, we extend previous work in this direction by providing LLN type results for the MD for general classes of trees: critical Galton-Watson trees conditioned to have size n, and growing general linear preferential attachment trees. The former class includes uniform random trees, which was analyzed in [177] using analytic combinatorics. Our probabilistic approach reproduces the LLN result with a very short proof, and in higher generality. The latter class includes Yule-trees (also called random recursive trees), m-ary increasing trees, binary search trees, and positive linear preferential attachment trees (we introduce these random trees in detail in Chapter 4).

3.2 Related Notions

There are several notions related to the MD, we refer to [232, 153] for recent reviews. Here, we survey the related literature from the point of view of source identification.

The combinatorial counterpart of the S1 source identification model introduced in Section 1.2, when the time of the first infection is not known, is the double metric dimension (DMD) [49]. In this case, we require that all vectors in the set $\{d(R, v^\star) + C \mid v^\star \in V, C \in \mathbb{Z}\}$ are different (the unknown constant C models the unknown time of the first infection). Although the MD and the DMD can be very different in certain deterministic graph families, they seem to behave similarly in most random graphs [226]. The algorithmic aspects of the DMD in the source location context were investigated in [54, 57], and the DMD of Erdős-Rényi random graphs was computed by [226]. We also mention that the version of the DMD corresponding to the case of multiple sources has been studied in [254].

Faulty or non-respondent sensors in source identification can be modelled via the *fault-tolerant resolving sets* R[110], for which $R \setminus \{v\}$ is also a resolving set for every $v \in R$, thus tolerating the fallout of any sensor. Fault-tolerant resolving sets were later generalized to tolerate the failure of k sensors in the *k-metric dimension* [86, 219].

Noise in the propagation delays can be addressed through the *truncated metric dimension* (TMD) [233], also called *threshold metric dimension* [24]. In these notions, the resolving sets R have to distinguish every pair of nodes v and u based on distances $d_t(w_i, v)$ and $d_t(w_i, u)$ for $w_i \in R$, where $d_t(w_i, v) = \max(d(w_i, v), t+1)$. For threshold $t = 1$, the sensors can only distinguish between neighbors and non-neighbors, and the TMD becomes (almost) equivalent to *identifying codes*, which have also been analyzed in Erdős-Rényi random graphs [95]. In Chapter 6 we show how the TMD can be connected to source identification in the stochastic epidemic models introduced in Chapter 1 in the path graph. We are not aware of such connections in more complex graph models.

Certain variants of the MD can also be used to address uncertainly in the network structure itself. In [251], we are given k connected graphs and it is assumed that $k-1$ edges are missing between them, which would connect all k components into a single one. The *extended metric dimension* is the number of landmarks we need to distinguish any pair of nodes, no matter where the $k-1$ edges are.

As a first step towards modelling the uncertainty in the network structure, it can be useful to understand the robustness of the MD to changing a few *known* edges. The question of how much the MD of a graph can change on the addition of a *single* edge has been first studied for trees, where Chartrand et. al. [56] found that on the addition of an arbitrary edge, the MD cannot increase by more than one, and cannot decrease by more than two. The result has been proved later in [84]. In [83], the change of the MD on a single edge or vertex addition or deletion is studied in general graphs. The authors find that, similarly to trees, the decrease of the MD on edge additions cannot be more than two, however, the increase is not bounded by any constant in general graphs. The latter statement is supported by an example graph, where the addition of a single edge doubles the MD. In Chapter 7, we give a different example graph, where adding a single edge exponentiates the MD. We believe that such an increase is possible only very special (in a sense very heterogeneous) graphs, and that in most cases the MD at most doubles. We prove this doubling upper bound for d-dimensional grid graphs, and perform an even more refined analysis for the case of $d = 2$ in Chapter 7.

Not to be confused with the TMD, the *threshold dimension* of a graph G is the minimum MD that can be achieved by adding an arbitrary number of edges to G [178]. Obviously, adding too many edges will bring G close to the complete graph, whose MD is the largest possible among graphs on n nodes ($\mathrm{MD}(G) = n-1$), but the authors show that for some graphs G it is possible to add edges in a smart way to significantly reduce the MD. We note that Geneson and Yi have constructed connected graphs H and G such that $H \subset G$ and the ratio of the metric dimensions of H and G is arbitrarily large [101]. The authors of [178] also connect the threshold dimension with the dimension of the Euclidean space in which the graph can be embedded.

Since in this dissertation our goal is to study the role of adaptivity, we focus on the MD, and its adaptive variant: the sequential metric dimension.

3.3 Sequential Metric Dimension

In the literature on combinatorics, the adaptive version of the metric dimension problem (i.e., the deterministic S2 framework) is called the sequential metric dimension (SMD) [210]. To define the SMD, we adopt the vocabulary of binary search. Let $v^\star \in V$ be the *target* node. The target node is unknown to us, but for a set *queries* $R \subseteq V$ the distance $d(R, v^\star)$ is known.

Definition 3.3.1 (candidate targets)**.** *Given a set of queries R, the set of candidate targets for the graph G is*

$$\mathcal{T}_R(G) = \{v \in V \mid d(R, v) = d(R, v^\star)\}.$$

Our goal is to identify $v^\star$, which means that we would like to find a set R with $\mathcal{T}_R(G) = \{v^\star\}$, or equivalently $|\mathcal{T}_R(G)| = 1$ (as $v^\star \in \mathcal{T}_R(G)$ must always hold). Recall, that for a resolving set R we have $|\mathcal{T}_R(G)| = 1$ for every $v^\star \in V$. In contrast, in the adaptive case, a (potentially) different R is constructed for every $v^\star$; in the j^{th} step, we select query w_j based on the distance information revealed by $R_{j-1} = \bigcup_{k=1}^{j-1} w_k$, and we still aim for $|\mathcal{T}_{R_j}(G)| = 1$.

Definition 3.3.2 (SMD)**.** *Let* ALG(G) *be the set of functions*

$$g : \{(G, R, d(R, v^\star)) \mid R \subseteq V, v^\star \in V\} \to V.$$

The sequential metric dimension (SMD) of G is the minimum $r \in \mathbb{N}$ such that there is a query selection algorithm $g \in$ ALG(G), for which if we let $R_0 = \emptyset$ and $R_{j+1} = R_j \cup g(G, R_j, d(R_j, v^\star))$, then $|\mathcal{T}_{R_r}(G)| = 1$ for any $v^\star \in V$.

Such combinatorial search problems are often posed in the context of algorithmic two-player games. A well-known and related game is the Cops and Robber game, where in each round the Cops select a subset of nodes, and try to catch a Robber moving strategically on the edges of the graph, as introduced in [185, 200]. Recently, different versions of the Cops and Robber game has been studied where each time a node is selected, it reports its distance to the robber [40, 76, 108, 208, 209]. The SMD differs form this version of the Cops and Robber game in that the Robber is not allowed to move between different rounds [210], because this assumption would not interpretable in the case of epidemics. In this context, Bensmail et. al. generalized the SMD [29] to the case when k queries need to be selected before the answers to the queries are revealed.

The SMD can also be interpreted as a two-player game, but this interpretation slightly differs from the intuition behind the Cop and Robber games, therefore we call the two players Player

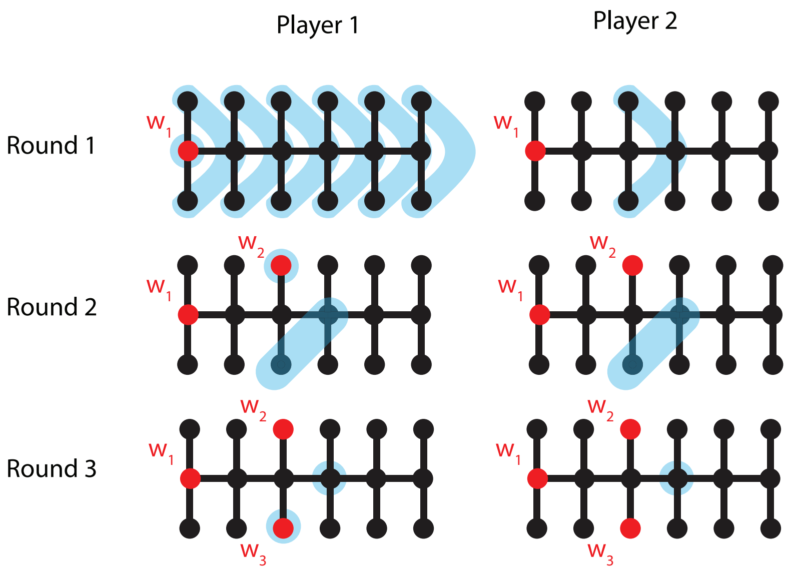


Figure 3.1: An example of how the SMD can be interpreted as a two-player game. In the j^{th} round, Player 1 creates the set R_j by adding a sensor node w_j (marked in red) to R_{j-1}. The sensor w_j partitions the current candidate target set $\mathcal{T}_{R_{j-1}}(G)$ based on distances (marked in blue). In turn, Player 2 must provide a distance from w_j to a feasible but not necessarily predetermined source node, which is equivalent to selecting one of the blue sets. Player 1 tries to reduce and Player 2 tries to increase the total number of rounds until the end of the game, which happens when $\mathcal{T}_{R_j}(G)$ shrinks to a single element. In this example, the game ends in 3 rounds if both players play optimally. Hence, the SMD of this "comb graph" of size 18 is also 3. In fact, the SMD of the "comb graph" of any size $n \geq 9$ is still 3, in sharp contrast with the MD of the same graph, which is $n/3$.

1 and Player 2. In each step, Player 1 selects a query and tries to reduce as fast as possible the candidate set to a single element. Player 2 must then provide an observation that is consistent with at least one of the target nodes. If there are multiple such observations, Player 2 can choose one to try to make the game as long as possible. In this setting Player 2 does not decide on the source $v^\star$ in advance, but must always be consistent with the observations that have been revealed so far (i.e., $\mathcal{T}_{R_j}(G)$ can never be empty). Since every predetermined source $v^\star$ can be found this way by Player 1, and since for every set of answers provided by Player 2 there is a node that could have been the source, the SMD can be seen as the number of steps the game takes if both players play optimally. See Figure 3.1 for an example of how the two-player game corresponding to the SMD is played.

3.4 The Role of Adaptivity in the Metric Dimension

From Definitions 3.3.1 and 3.3.2, it is clear that $1 \leq \text{SMD} \leq \text{MD} \leq N$, as being able to adaptively select the queries only gives Player 1 more power in the two-player game described at the end of Section 3.3. The focus of this section is on quantifying how much more power Player 1 has, depending on the underlying graph.

The comb graph in Figure 3.1 is an example where adaptivity has an extremely important role: the MD is linear in the size of the graph whereas the SMD is constant. Of the large class of random trees analyzed in Chapter 4, the ones with bounded degrees (m-ary recursive trees, Galton-Watson trees with bounded offspring distributions) are also examples where adaptivity has an important role (almost linear in N). Indeed, in Chapter 4 we show that the MD is $\Theta(N)$ in these trees, and there is a previous result by Kim el. al. ([142] Theorem 1.2), which says that the SMD is $O(\Delta \log(N))$ on graphs with maximum degree Δ.

The role of adaptivity seems to be less drastic, but still quite important, in random geometric graphs (RGG). Lichev, Mitsche and Prałat ([159] Theorem 5.2) recently showed that the metric dimension of RGGs of N nodes in the unit square with connectivity range r is

$$\Omega\left(\max\left(\frac{1}{r^2}, \frac{N^{2/3} r^{4/3}}{\log^{1/3}(N)}\right)\right) \tag{3.1}$$

for $1/\sqrt{N} \ll r \leq 1/4$. As noticed by Lecomte, Ódor and Thiran [156], we may apply the upper bound by Kim el. al. [142] on the SMD of RGGs. This upper bound, together with the well-known result that the maximum degree of RGGs is $\Delta = O(Nr^2 \log(N))$ with high probability [196], implies that the SMD of RGGs is $O(Nr^2 \log^2(N))$. Since the lower bound on the MD in equation (3.1) is asymptotically larger than the upper bound $Nr^2 \log^2(N)$ on the SMD for $1/\sqrt{N} \ll r \leq 1/4$, we observe a relatively large (sublinear polynomial in N) role of adaptivity in that range.

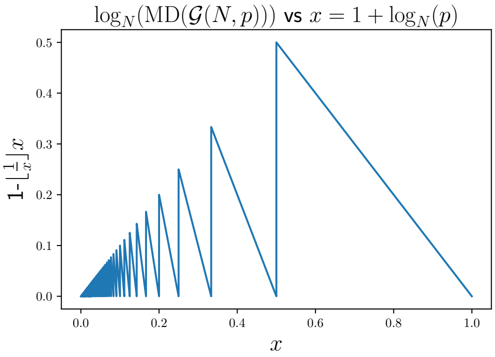


Figure 3.2: The “zig-zag” function identified by [36] for the MD of Erdős-Rényi graphs (with appropriate scaling). In Chapter 5, we prove that the SMD follows asymptotically the same function (after the same scaling).

Finally, we consider the case of Erdős-Rényi random graphs $\mathcal{G}(N,p)$. Bollobás, Mitsche and Prałat [36] found that $\text{MD}(\mathcal{G}(N,N^{x-1})) = N^{1-\lfloor 1/x \rfloor x + o(1)}$, which means that the MD is a ("zig-zag" shaped, see Figure 3.2) power of N, unless $1/x$ is an integer, in which case the MD is a constant times $\log(N)$. In Chapter 5, we prove a similar bounds for the SMD in Erdős-Rényi graphs. To ease notation, we express our main results on the SMD in terms of the MD. The precise formulation and the proof of our main theorem are given in Section 5.5. In its crudest form, our main result says that the ratio of the SMD and the MD is between 1 and 1/2 a.a.s, which implies that the role of adaptivity in Erdős-Rényi graphs is fairly small (constant in N).

We are able to make our results more precise by computing the leading constant of the ratio for all values of $p \gg \frac{\log^5(N)}{N}$. We find that the leading constant is 1 for almost all values of p, except near the values where the MD is logarithmic (which correspond to the zero values in Figure 3.2). Since for such values the dependence of this leading constant on p and N is rather complicated, we simply denote it by $F_\gamma(p,N)$ in Theorem 3.4.1, and we defer the precise definition of $F_\gamma(p,N)$ to Remark 5.5.2. It is shown in Remark 5.5.2 that $F_\gamma(p,N)$ can take any value in the interval $(1/2,1)$, which implies that there are values of p for which the SMD is strictly smaller than the MD.

Theorem 3.4.1. *Let $N \in \mathbb{N}$ and $p \in [0,1]$ such that $\frac{\log^5(N)}{N} \ll p$ and $\frac{1}{\sqrt{N}} \ll 1-p$. Let G be a realization of a $\mathcal{G}(N,p)$ random graph. Then,*

$$1 \geq \frac{\text{SMD}(G)}{\text{MD}(G)} = F_\gamma(p,N) + o(1) \geq \frac{1}{2} + o(1) \quad \textit{if } (Np)^{i+1} = \Theta(N) \textit{ for } i \in \mathbb{N} \tag{3.2}$$

$$\frac{\text{SMD}(G)}{\text{MD}(G)} = 1 - o(1) \qquad \textit{otherwise,} \tag{3.3}$$

hold a.a.s, where F_γ is a function of p,N that is explicitly expressed in Remark 5.5.2.

The proof of this theorem requires a thorough analysis of the SMD of Erdős-Rényi graphs, which has been published in [188], and also an improvement on the upper bound for the MD of Erdős-Rényi graphs, which appears in this thesis for the first time. As a side-result, this improved upper bound closes the gap between the upper and the lower bounds on the MD of Erdős-Rényi graphs in [36], which has been an open problem for the past 10 years.

See Figure 3.3 for simulation results confirming Theorem 3.4.1 on the role of adaptivity.

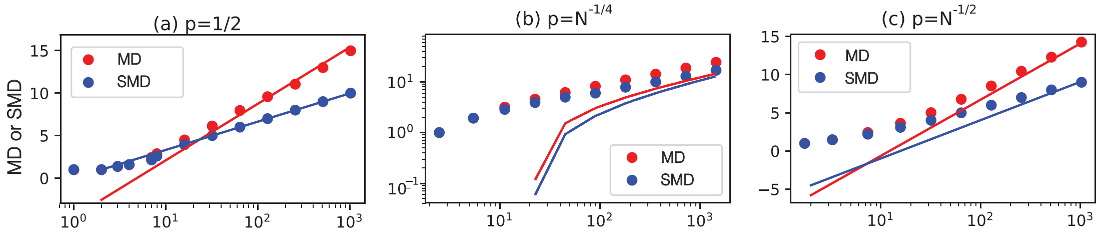


Figure 3.3: The red and blue dots show the approximated value of the MD and the SMD of simulated Erdős-Rényi graphs computed by the toolbox [186] averaged over 100 iterations (confidence intervals are too small to be plotted). The slope of the red and blue lines is computed by Theorem 5.5.1 and the intercept is chosen to fit the last few data points. On (semi-log) plots (a) and (c) we have $(Np)^{i+1} = \Theta(N)$ for $i = 0$ and $i = 1$ respectively. For such parameters the MD and the SMD are both logarithmic and there is a constant factor difference between them. On the contrary, on (log-log) plot (b) we have $(Np)^{i+1} \neq \Theta(N)$ for all $i \in \mathbb{N}$, and for such parameters the MD and the SMD grow as a power of N.

4 Metric Dimension of Random Trees

In this chapter, we provide law of large numbers (LLN) type results about the metric dimension (MD) of two general distributions on trees: general linear preferential attachment trees, and conditioned critical Galton-Watson trees. See Chapter 3 for a general introduction about the MD and its connection with source identification.

We summarize the main results of this chapter in Section 4.1, and we precisely define the notions that we use in Section 4.2. In Section 4.3 we explain the general methodological background about the main theoretical tools used in this chapter: Crump-Mode-Jagers trees, fringe trees and subtree properties. Finally, we prove our results in Section 4.4.

This chapter is based on the publication [147] by Komjáthy and Ódor.

4.1 Summary of Results

We briefly describe the families of random trees that are studied in this chapter. A general linear preferential attachment tree is a model with two parameters, $\rho > 0$ and $\chi \in \mathbb{R}$, which is constructed as follows. We start with a single root vertex. When there are i vertices, we attach the $(i+1)$-st vertex to one of the existing vertices $v \le i$ with probability proportional to

$$\rho + \chi \deg_i(v), \tag{4.1}$$

where $\deg_i(v)$ is degree of vertex v after i vertices have been added. Clearly, due to the normalization, only the ratio ρ/χ matters, and for the rest of the chapter, without loss of generality, we only consider $\chi \in \{-1, 0, 1\}$. When $\chi = -1$, we require ρ to be an integer.

We explain now why this class of trees contain m-ary increasing trees, binary search trees, and uniform recursive trees as well as rich-get-richer trees, that are the 'usual' linear preferential attachment trees. When we take $\rho = m$ and $\chi = -1$, we obtain the *m-ary increasing tree*: In the original definition of an *m-ary increasing tree*, each vertex has the potential to have m labeled offspring. The tree starts with a single vertex (the root) at step 1, and at each step a new vertex arrives. When the tree has i vertices, a new vertex can attach to $mi-(i-1)=(m-1)i+1$ possible places, since out of the mi possible places, $i-1$ are already taken (only the root does not have a parent). An *m-ary increasing tree* with n vertices is constructed by starting with a single root vertex, and placing the $(i+1)$-st vertex uniformly randomly among the $(m-1)i+1$ possible places [116]. The probability that the $(i+1)$-st vertex connects to vertex $v \le i$ is thus proportional to $m-\deg_i(v)$. Hence, we recognize equation (4.1) with $m=\rho$ and $\chi=-1$.

For $\rho=2$ and $\chi=-1$, the binary increasing tree corresponds to another well-known tree: the random *binary search tree*, an object that gained attention in computer science. In (the original definition of) a binary search tree, each vertex can store a single key and can have at most two children. The keys can be thought of as i.i.d. uniform random variables on $[0,1]$ (this is a representation used by Devroye in [69]). Initially, the first key K_1 arrives and is placed at the root. This makes the root a *full* vertex. Upon filling, every vertex creates two *potential* vertices, one on the left and one on the right, that can receive a key each. These potential vertices do not count as part of the tree yet, only once they contain a key and become full vertices. After the tree has i keys, the $(i+1)$-st key K_{i+1} arrives and is compared to the key in the root. If $K_{i+1}<K_1$, it is pushed to the left (otherwise to the right). Then it is compared to the key occupying the vertex that is the left (resp. right) child of the root, and again pushed left (resp. right) if it is less (resp. larger) than the key in that vertex. The procedure continues until the key finds a potential vertex and occupies it. Since only the permutation of the keys matters, it can be shown that when the tree has i full vertices, and therefore $i+1$ potential vertices, the $(i+1)$-st vertex is equally likely to be placed at any of these potential vertices. Therefore, the probability that a full vertex with $v \le i$ with degree[1] $\deg_i(v)$ gets a new child in step $i+1$ has probability $(2-\deg_i(v))/(i+1)$, and we get back a formula proportional to equation (4.1) with $\rho=2$ and $\chi=-1$.

A similar construction exists for $m>2$, called the m-ary search tree, where each vertex can store up to $m-1$ keys. This tree, however, is not equivalent to the m-ary increasing tree [116], and we omit studying them further in this dissertation. Binary search trees are also the tree-representation of the Quicksort algorithm [145]. Many of their properties are well studied, including Law of Large Numbers and Central Limit Theorems, see e.g. [69, 96], such as the proportion of k-protected nodes or subtree sizes.

The *random recursive tree* is constructed analogously to the previous construction, except that there is no dependence on the degree: starting with a single root vertex, the $(i+1)$-st vertex attaches by an edge uniformly to each of the i vertices already present. This case corresponds to $\rho=1, \chi=0$ in equation (4.1). Random recursive trees have a natural correspondence to binary search trees, and so they are often treated together [144]. They are also called Yule-trees, as they can be naturally embedded in a Yule-process, and hence they have connections to phylogenetic trees [34].

The ‘usual’ linear preferential attachment tree, also called *rich-get-richer* tree, is constructed

[1] Potential vertices do not contribute to the degree, only full vertices do.

by taking $\rho > 0$, $\chi = 1$. In this case the $(i+1)$-st vertex attaches to $v \le i$ with probability proportional to $\rho + \deg_i(v)$. The $\rho = \chi = 1$ case corresponds to the positive linear preferential attachment tree, which was informally introduced by Barabási and Albert [23], although they allowed general graphs, not only trees. This is the model that produces power-law degree distributions [37], see also van der Hofstad [111] and the survey [116]. Positive linear preferential attachment trees have already been studied in the context of source identification [132], with the difference that the authors of [132] consider snapshot-based source identification, whereas the MD is connected to source identification with time queries (see Chapter 1).

The survey [116] gives and excellent overview of the literature on various properties of all these growing trees, hence we refer the reader there for further literature.

Our main results can be summarized in the following two meta-theorems.

Theorem 4.1.1 (Meta-theorem about growing trees)**.** *Let $(\mathcal{T}_n^{(\rho,\chi)})_{n\ge1}$ be a sequence of random growing general linear preferential attachment trees with n vertices, with growth parameters $\rho > 0$ and $\chi \in \{-1,0,1\}$, with $\rho \in \mathbb{N}$ when $\chi = -1$. Then*

$$\frac{\mathrm{MD}(\mathcal{T}_n^{(\rho,\chi)})}{n} \xrightarrow{a.s.} c_{(\rho,\chi)}, \tag{4.2}$$

where $c_{(\rho,\chi)} \in (0,\infty)$ is a constant that we determine explicitly.

We mention that our method provides almost sure LLN for a much larger class of random growing trees. This class is the class of trees that can be embedded in a Crump-Mode-Jagers branching process with finite Malthusian parameter; e.g. sub-linear preferential attachment trees, m-ary search trees, fragmentation trees, etc. We refer the reader to various classes of such trees to the survey of Janson and Holmgren [116].

Our second result is motivated by reproducing LLN of the metric dimension of *uniform random trees* [177]. A uniform random tree on n vertices is a tree that is chosen uniformly at random (u.a.r.) from the possible n^{n-2} labeled trees on n vertices. As mentioned before, LLN and even CLT for the MD of uniform random tree was proved in [177] using analytic combinatorics. We are able to reproduce the LLN result with a very short proof, which offers a better generalization to other random tree distributions. A uniform random tree has the same distribution as a Galton-Watson branching process, with Poisson offspring distribution with mean 1, conditioned to have total progeny n, see e.g. [111, Proof of Theorem 3.17]. Therefore it is equivalent to determine the MD of conditioned GW trees.

A *Galton-Watson tree* is a random tree defined by the offspring distribution ξ taking values in $\mathbb{N} = \{0,1,\ldots\}$. Initially a single individual (vertex) is born, which becomes the root of the tree, and the root gives rise to ξ children. Thereafter, each newly born individual samples its own independent copy of ξ and gives rise to that many new children, and the process continues recursively. We consider Galton-Watson trees conditioned to have n vertices, so we must assume that $\mathbb{P}(\xi = 0) \neq 0$, otherwise the process never ends. We will assume that the Galton-Watson trees are critical, i.e., $\mathbb{E}[\xi] = 1$, which is also fairly natural for conditioned Galton-Watson trees (see Remark 3.1 of [126]), since in this case a non-trivial limiting measure on trees exists (called the incipient infinite tree).

Theorem 4.1.2 (Conditioned Galton-Watson trees)**.** *Let $\mathcal{GW}_n$ be a sequence of critical Galton-Watson trees conditioned to have n vertices, with offspring distribution ξ, where* $\mathbb{E}[\xi] = 1$ *and* $\mathbb{E}[\xi^2] < \infty$*. Let $p_k = \mathbb{P}(\xi = k)$ for $k \in \mathbb{N}$. Then*

$$\frac{\mathrm{MD}(\mathcal{GW}_n)}{n} \xrightarrow{p} p_0 - 1 + G_\xi\left(1 - \frac{p_0}{1-p_1}\right) + \frac{p_1 p_0}{1-p_1}, \tag{4.3}$$

where $G_\xi(x) = \sum_{n=0}^{\infty} p_n x^n$ is the probability generating function of ξ evaluated at x.

As a corollary of this theorem, by substituting $\xi = \mathrm{Poi}(1)$ we recover the result of [177] on uniform random trees.

Corollary 4.1.1. *The metric dimension of a uniform random tree $\mathcal{U}_n$ on n vertices satisfies the following Law of Large Numbers:*

$$\frac{\mathrm{MD}(\mathcal{U}_n)}{n} \xrightarrow{p} \mathrm{e}^{-1} - 1 + \mathrm{e}^{-\frac{1}{\mathrm{e}-1}} + \frac{\mathrm{e}^{-1}}{\mathrm{e}-1} \approx 0.14076941.$$

Methodology. The metric dimension of a given fixed tree can be computed explicitly using the number of leaves of the tree and the number of exterior major vertices, i.e., vertices of at least degree 3 that have a line-graph leading to a leaf, see Theorem 4.2.1 below.

The novel insight in our proofs is that both the asymptotic proportion of leaves as well as that of exterior major vertices of random trees $\mathcal{T}$ can be computed using results from the *fringe tree literature* initiated by Aldous in [12]. A fringe tree of a rooted tree, in plain words, is the random subtree obtained by choosing a vertex u.a.r. in the tree and taking its subtree pointing away from the root. The distribution of fringe trees is shown to converge for a large class of trees. So, fringe trees of a *rooted tree* $\mathcal{T}$ help us to compute the asymptotic proportion of vertices v in $\mathcal{T}$ that have a certain property $\mathcal{P}$, with the limitation that $\mathcal{P}$ must be a *subtree-property.* A subtree property is any property that depends only on the subtree of $\mathcal{T}$ rooted at v pointing away from the root. It is easy to see that being a leaf is a subtree-property. While strictly speaking being an exterior major vertex is not a subtree-property, we find a subtree property that serves as a good proxy.

The use of fringe tree-methodology allows us to use probabilistic arguments that are often much shorter than the analytic-combinatorial arguments used in [177]: the proportion of fringe-trees satisfying a given subtree property converges. Moreover, since the fringe distribution of several general random tree families is known [116, 126], our proofs are quite general. Our results hold for critical Galton-Watson trees with a finite variance degree distribution (which include, among others, uniform random trees, Motzkin trees, random binary trees) and all linear preferential attachment trees (which include, among others, binary search trees,

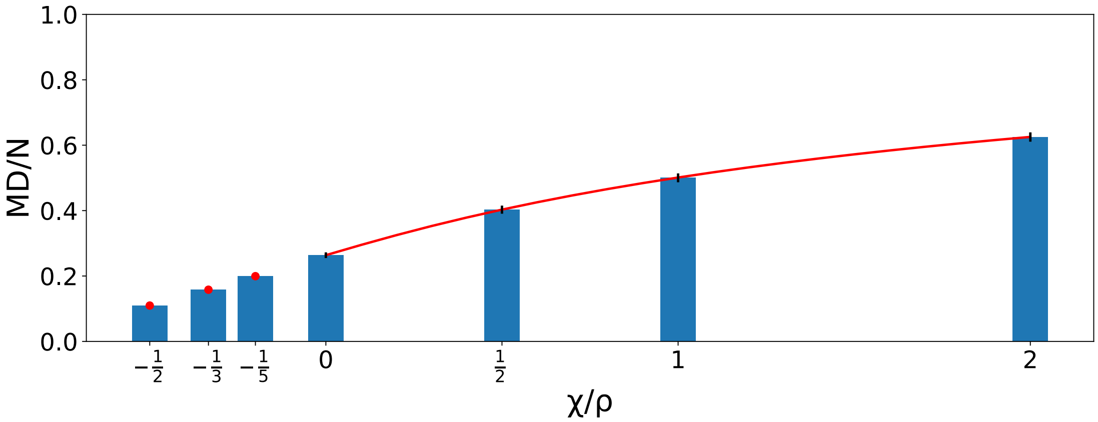


Figure 4.1: The red dots and the red line show $c_{(\rho,\chi)}$ as a function of χ/ρ based on our theoretical results in Theorems 4.2.2, 4.2.3 and 4.2.4. The blue bars show simulation results for $c_{(\rho,\chi)}$. We show the average of the normalized MD of 1000 independently simulated random trees with 1000 nodes. Unless they are too small to be visible on the plot, we also show the 95% confidence intervals for the simulation results on top of the bar plots.

random recursive trees, positive linear preferential attachment trees) [75].

The fringe tree literature has CLT type results, which suggests that many of our results in this chapter can also be extended to a CLT. In particular, the CLT of metric dimension for binary search trees and uniform recursive trees should be a consequence of the CLT proved in [115]. For the other cases, this is not a trivial extension, and we leave it for future work.

4.2 Definitions and Numerical Values for $c_{(\rho,\chi)}$

We start by expressing the MD of fixed trees explicitly. First we need a few definitions.

Definition 4.2.1 (Leaves and exterior major nodes)**.** *Let us denote by* $\deg(v)$ *the degree of a node* $v \in V$*. We say that a node* $v \in V$ *is a leaf if* $\deg(v) = 1$*, and is a major node if* $\deg(v) \geq 3$*. If a major node* $v \in V$ *has a path to a leaf that only contains degree-two vertices besides the beginning and the end of the path (i.e., a line-graph), we say that* v *is an exterior major node. Let us denote the set of leaves of* G *by* $L(G)$ *and the set of exterior major nodes of* G *by* $K(G)$*.*

The following theorem characterises the MD of a fixed tree.

Theorem 4.2.1 (Metric dimension of trees [218])**.** *Consider a fixed tree* T*. If* T *is a path graph, then* $\mathrm{MD}(T) = 1$*. Otherwise,*

$$\mathrm{MD}(T) = |L(T)| - |K(T)|. \tag{4.4}$$

We refer the reader to [218] for a proper proof, but we explain the formula heuristically. It is not hard to see that if two or more leaves are attached to a major node by line-graphs, then the vertices at equal distance from the major node on these lines are indistinguishable by sensors that do not fall into these lines. Hence, all but one of the terminal leaves of such lines have to be sensors.

Now we state our more detailed results about families of trees growing according to general linear preferential attachment schemes, that is, we refine Theorem 4.1.1 and express the limiting constant $c_{(\rho,\chi)}$ of the MD explicitly. Some of the numerical values acquired from the Theorems 4.2.2, 4.2.3 and 4.2.4 below are shown in Figure 4.1 along with numerical approximation given by computer simulations.

Random Binary Search Tree and m-ary Increasing Trees

Recall that an m-ary increasing tree is equivalent to a general linear preferential attachment tree with $\rho = m$ and $\chi = -1$, and that for $m = 2$, an m-ary increasing tree is equivalent to a random binary search tree.

We write

$$\gamma(s,t) = \int_0^t x^{s-1} \mathrm{e}^{-x}\,\mathrm{d}x \tag{4.5}$$

for the lower incomplete gamma function, and

$$\binom{m}{i,j} = \frac{m!}{i!j!(m-i-j)!} \tag{4.6}$$

for the generalized binomial coefficient.

Theorem 4.2.2 (MD of m-ary increasing trees). *Let $(\mathcal{T}_n^{(m,-1)})_{n\geq 1}$ be a growing sequence of random m-ary increasing trees with n vertices. Then*

$$\frac{\mathrm{MD}(\mathcal{T}_n^{(m,-1)})}{n} \xrightarrow{a.s.} \sum_{j=1}^{m} \frac{m-1}{(m-1+j)m^j}\binom{m}{j} + \sum_{\substack{0\leq i+j\leq m\\ i\neq 0}} A_{i,j}\gamma\left(\frac{i+j}{m-1}+1, \frac{im}{m-1}\right), \tag{4.7}$$

where for all $(i,j) \in \mathbb{N}^2$ with $i+j \leq m$ and $(i,j) \neq (1, m-1)$

$$A_{i,j} = \frac{(-1)^i}{m^{i+j}}\binom{m}{i,j}\mathrm{e}^{\frac{im}{m-1}}\left(\frac{m-1}{im}\right)^{\frac{i+j}{m-1}+1}, \tag{4.8}$$

except for $(i,j) = (1, m-1)$, where

$$A_{1,m-1} = \left(1 - \frac{m}{m^m}\right)\mathrm{e}^{\frac{m}{m-1}}\left(\frac{m-1}{m}\right)^{\frac{m}{m-1}+1}. \tag{4.9}$$

In particular, for the binary search tree ($m = 2$), this expression evaluates to

$$\frac{\mathrm{MD}(\mathcal{T}_n^{(2,-1)})}{n} \xrightarrow{a.s.} \frac{3\mathrm{e}^4 - 48\mathrm{e}^2 + 233}{384} \approx 0.1096868681. \tag{4.10}$$

χ/ρ	-1/2	-1/3	-1/4	-1/5	0	1/2	1	2	10
$\mathrm{MD}(\mathcal{T}_n^{(\rho,1)})/n$	0.10969	0.15812	0.18377	0.19953	0.26371	0.40304	0.50120	0.62535	0.87501

Figure 4.2: The table shows numerical values of the MD of general linear preferential attachment trees for some parameters χ/ρ. The parameter values $-1/2, 0$ and 1 correspond to the binary search tree, the random recursive tree, and the positive linear preferential attachment tree respectively.

We provide two proofs of this theorem for $m = 2$ below in Section 4.4: a combinatorial proof and a probabilistic proof. The probabilistic proof is more robust, and we are able to generalize that proof for $m > 2$ and other types of attachments rules.

Random Recursive Tree

As mentioned in Section 4.1, a random recursive tree is constructed by attaching each new node uniformly randomly to one of the existing nodes. It is also a special case of a general linear preferential attachment tree with parameters $\rho = 1, \chi = 0$.

Theorem 4.2.3 (MD of random recursive trees)**.** *Let $\mathcal{T}_n^{(1,0)}$ be a sequence of random recursive trees with n nodes. Then*

$$\frac{\mathrm{MD}(\mathcal{T}_n^{(1,0)})}{n} \xrightarrow{a.s.} \mathrm{e}\left(\int_1^{\mathrm{e}} x^{-1}\mathrm{e}^{-x}\,\mathrm{d}x + \gamma(2,1)\right) - 1 \approx 0.263709059. \tag{4.11}$$

Rich-Get-Richer Trees

Theorems 4.2.2 and 4.2.3 covered general linear preferential attachment trees with $\chi \in \{-1, 0\}$. In the next theorem it suffices to state the result with $\chi = 1$. These trees are often called rich-get-richer trees, as new nodes are more likely to attach to nodes with higher degrees.

Theorem 4.2.4 (MD of rich-get-richer trees)**.** *Let $\mathcal{T}_n^{(\rho,1)}$ be a sequence of linear preferential attachment trees with n nodes and $\chi = 1$, $\rho > 0$. Then*

$$\begin{aligned}\frac{\mathrm{MD}(\mathcal{T}_n^{(\rho,1)})}{n} \xrightarrow{a.s.} -1 + \int_0^{\infty} (\rho+1)\mathrm{e}^{-x(\rho+1)}\left(1 + \frac{\mathrm{e}^{x+\frac{\rho}{\rho+1}(1-\mathrm{e}^{-(\rho+1)x})} - \mathrm{e}^{x}}{\rho}\right)^{-\rho} \mathrm{d}x \\ + \int_0^{\infty} (\rho+1)\mathrm{e}^{-x(\rho+1)}\mathrm{e}^{-\rho x+\frac{\rho}{\rho+1}(1-\mathrm{e}^{-(\rho+1)x})}\,\mathrm{d}x.\end{aligned} \tag{4.12}$$

The $\rho = \chi = 1$ case corresponds to the positive linear preferential attachment tree, introduced by [23]. For positive linear preferential attachment trees we can use Theorem 4.2.4 and a numerical integration software [121] to obtain the following result.

Corollary 4.2.1. *Let $\mathcal{T}_n^{(1,1)}$ be a sequence of positive linear preferential attachment trees with n* *nodes. Then*

$$\frac{\mathrm{MD}(\mathcal{T}_n^{(1,1)})}{n} \xrightarrow{a.s.} c_{(1,1)} \approx -1 + 0.679824 + 0.821372 = 0.501196.$$

4.3 Method and Discussion

In this section we introduce fringe-trees and general results on their convergence, we explain the embedding of trees growing in discrete times into Crump-Mode-Jagers branching processes, and relate the metric dimension to subtree properties.

4.3.1 Fringe Trees

For the rest of the chapter, all trees T are considered to be rooted, which simply means that they have a special vertex denoted by $\mathrm{root}(T)$. In rooted trees, every vertex $v \in T \setminus \{\mathrm{root}(T)\}$ has a parent, which is the first vertex on the path from v to $\mathrm{root}(T)$. For any vertex $v \in T$, let T_v be the subtree of T rooted at v, that is, the connected subtree of T that contains v after removing the parent of v (as a special case $T_{\mathrm{root}(T)} = T$). If we sample v uniformly at random from T, we say that the random tree T_v is a *random fringe tree* of T. When T is a deterministic tree, this definition is quite straightforward. However, we are interested in the case when T itself is random, and in this case defining random fringe trees requires more care.

Definition 4.3.1. *For rooted trees S and T let $n_S(T)$ be the cardinality of $\{v \mid T_v = S\}$ and for a rooted tree property $\mathcal{P}$, let $n_{\mathcal{P}}(T)$ be the cardinality of $\{v \mid T_v \in \mathcal{P}\}$.*

When T is deterministic, $n_S(T)/|T|$ defines the random fringe tree distribution. When $\mathcal{T}$ is random, we can think of the sampling of $\mathcal{T}$ and v as a joint random event, which again gives rise to a distribution over trees. This is called the annealed fringe tree distribution. In this chapter, we are interested in the *quenched fringe tree distribution*. In the quenched version, we think of $n_S(\mathcal{T})/|\mathcal{T}|$ as a distribution that is itself random. Since we are interested in the convergence of fringe tree distributions as the sizes of the trees tend to infinity, we are going to focus on the convergence of the random variables $n_S(\mathcal{T}_n)/|\mathcal{T}_n|$ (almost surely (a.s.) or in probability (p)).

We also defined the seemingly more general notion of $n_{\mathcal{P}}(T)$, however, in our applications whenever we can say something about the convergence of $n_S(\mathcal{T}_n)/|\mathcal{T}_n|$, we have a similar result for $n_{\mathcal{P}}(\mathcal{T}_n)/|\mathcal{T}_n|$. In fact, since working with subtree properties will be very convenient for computing the MD (see Lemma 4.3.2), we only state the results from the fringe tree literature on $n_{\mathcal{P}}(T)$.

Theorem 4.3.1 ([12], Theorem 1.2 of [126])**.** *Let $\mathcal{GW}_n$ be a sequence of Galton-Watson trees conditioned to have n vertices, with offspring distribution ξ, where $\mathbb{E}[\xi] = 1$ and $\mathbb{E}[\xi^2] < \infty$. Let $\mathcal{F}$ be the* unconditioned *Galton-Watson tree with the same offspring distribution. Then, for* *every subtree property $\mathcal{P}$,*

$$\frac{n_{\mathcal{P}}(\mathcal{GW}_n)}{n} \xrightarrow{p} \mathbb{P}(\mathcal{F} \in \mathcal{P}). \tag{4.13}$$

The previous theorem applied to any Galton-Watson tree with $\mathbb{E}[\xi] = 1$ and $\mathbb{E}[\xi^2] < \infty$. The next theorem only applies to a single family of growing trees, the binary search tree. We will use it to give a combinatorial proof of the LLN of the MD of binary search trees (second part of Theorem 4.2.2).

Theorem 4.3.2 ([12])**.** *Let $(\mathcal{T}_n^{(2,-1)})_{n\geq 1}$ be a growing sequence of binary search trees of size n. Then, for every subtree property $\mathcal{P}$,*

$$\frac{n_{\mathcal{P}}(\mathcal{T}_n^{(2,-1)})}{n} \xrightarrow{p} \sum_{k=1}^{\infty} \frac{2}{(k+1)(k+2)} \mathbb{P}(\mathcal{T}_k^{(2,-1)} \in \mathcal{P}). \tag{4.14}$$

In words, this theorem says that the fringe-tree distribution of a random binary search tree is again a random binary search tree with a *random size*: the probability that the size of the fringe-tree is k is $2/((k+1)(k+2))$. A similar statement can be made for random recursive trees, however, we do not include this statement as it will not be used in our proofs. Instead we introduce a more powerful theorem which will help to strengthen the convergence to almost sure, treat m-ary increasing trees for general $m \geq 2$, random recursive trees, and linear preferential attachment trees.

4.3.2 Crump-Mode-Jagers Trees and Fringe Trees

A *Crump-Mode-Jagers* (CMJ) branching process generalizes, among many other random tree models, m-ary increasing trees and random recursive trees. Heuristically speaking, CMJ branching processes provide a method of embedding trees growing in discrete steps into a corresponding continuous time process. The CMJ process is defined by a point process $\Xi = (\xi_1, \xi_2, \ldots)$, called the reproduction process. At time zero, a single vertex is born, which becomes the root of the tree, and the children of the root are born at time $\xi_1, \xi_2, \cdots$. Similarly, each vertex v born at time t_v has an independent copy of Ξ denoted as $\Xi_v = \xi_{v,1}, \xi_{v,2}, \cdots$, and the offspring of v are born at time $t_v + \xi_{v,1}, t_v + \xi_{v,2}, \cdots$. So far we defined a branching process that grows over time. We obtain a random tree from this branching process by stopping the process at time τ and taking only the vertices (individuals) that have already been born. The stopping time τ can depend on the tree (very often τ is the time the n^{th} individual is born), or it can be an independent random variable.

Definition 4.3.2 (Linear preferential attachment reproduction process)**.** *Let the reproduction process $\Sigma_{\rho,\chi} = (\xi_1, \xi_2, \ldots,)$ with parameters $\rho > 0$ and $\chi \in \{-1, 0, 1\}$ be a linear preferential attachment reproduction process if*

$$\xi_j - \xi_{j-1} \sim \operatorname{Exp}(\rho + \chi(j-1)) \tag{4.15}$$

are independent exponential random variables, with the convention that $\xi_0 := 0$ *(but it does not count as a birth event). If* $\chi = -1$*, let us also assume* $\rho \in \mathbb{N}$ *and let us truncate the process to* ρ *terms (i.e.* $\Sigma_{\rho,\chi} = (\xi_1, \dots, \xi_{\rho-1}, \xi_\rho)$*), which ensures that the exponential random variables in* (4.15) *are well-defined.*

Lemma 4.3.1. *A CMJ tree with a linear preferential attachment reproduction process* $\Sigma_{\rho,\chi}$ *stopped when it reaches n vertices has the same distribution as a linear preferential attachment tree with n vertices and parameters* ρ *and* χ*.*

This lemma is due to the memoryless property of exponential random variables; the proofs can be found in [116, Sections 6.3, 6.4].

The interesting property of CMJ trees is that the fringe tree distribution of the random CMJ tree stopped at n vertices is again a random CMJ tree, with the *same reproduction process*, stopped at a random time that is independent of the number of vertices. This independence of the stopping time will be heavily exploited in our proofs. In this chapter, we only use the results on the fringe trees of linear preferential attachment trees. We refer to [116] for the general statement on CMJ trees.

Theorem 4.3.3 ([123, 181],Theorem 5.14 of [116])**.** *Let* $(\mathcal{T}_n^{(\rho,\chi)})_{n\geq 1}$ *be a growing sequence of linear preferential attachment trees with n vertices and parameters* $\rho > 0$ *and* $\chi \in \{-1, 0, 1\}$*. Let* $\mathcal{F}$ *be the corresponding CMJ tree stopped at random time* $\mathrm{Exp}(\rho+\chi)$*. Then, for every subtree property* $\mathcal{P}$*,*

$$\frac{n_{\mathcal{P}}(\mathcal{T}_n^{(\rho,\chi)})}{n} \xrightarrow{a.s.} \mathbb{P}(\mathcal{F} \in \mathcal{P}). \tag{4.16}$$

4.3.3 Expressing the Metric Dimension with Subtree Properties

In this section we reduce the metric dimension of trees to counting subtrees with certain properties. Recall Theorem 4.2.1 that expresses the MD of a tree as the difference between the number of leaves and the number of exterior major vertices.

Definition 4.3.3. *Let* $\mathcal{P}_L$ *be the subtree property that the subtree is a single vertex, that is, a leaf. Let* $\mathcal{P}_K$ *be the subtree property that the root has degree at least* two *and at least one of its subtrees is a line-graph to a leaf (a single vertex is considered to be a line).*

Lemma 4.3.2. *For any sequence of trees* T_n*, with* $|T_n| \to \infty$ *and* $\mathcal{P}_L, \mathcal{P}_K$ *given by Definition 4.3.3,*

$$\frac{\mathrm{MD}(T_n)}{|T_n|} = \frac{|L(T_n)| - |K(T_n)|}{|T_n|} = \frac{n_{\mathcal{P}_L}(T_n)}{|T_n|} - \frac{n_{\mathcal{P}_K}(T_n)}{|T_n|} + \frac{\varepsilon}{|T_n|}, \tag{4.17}$$

where $\varepsilon \in \{-1, 0, 1\}$*.*

Proof. We are going to show the equivalent statement that for any deterministic rooted tree T, we must have $n_{\mathcal{P}_L}(T) = |L(T)|$ and $|n_{\mathcal{P}_K}(T) - |K(T)|| \leq 1$. The equality $n_{\mathcal{P}_L}(T) = |L(T)|$ follows from the definition. Next we show that $|n_{\mathcal{P}_K}(T) - |K(T)|| \leq 1$ (see also Figure 4.3).

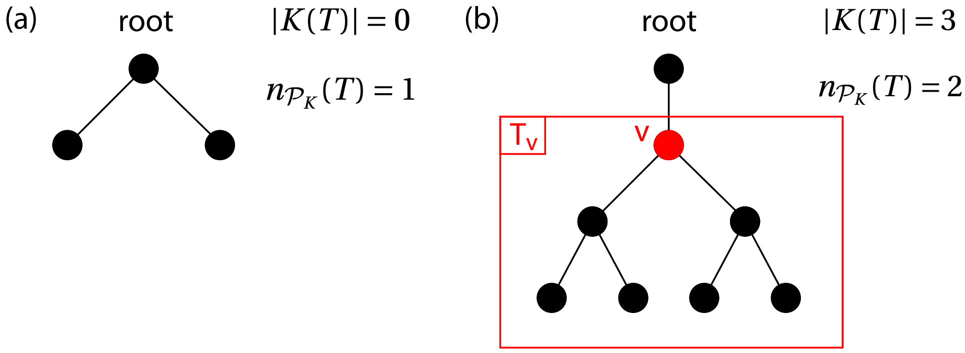


Figure 4.3: Illustration for the proof of Lemma 4.3.2. The subfigures (a) and (b) show the smallest trees where $n_{\mathcal{P}_K}(T) - K(T) = \pm 1$, respectively. The inequality $n_{\mathcal{P}_K}(T) - |K(T)| > 0$ holds only for trees in which the root has degree 2, and the root has a line-graph to a leaf. In this case, the root has property $\mathcal{P}_K$, but it does not count into $K(T)$ since it has degree 2. The inequality $n_{\mathcal{P}_K}(T) - |K(T)| < 0$ holds only for trees in which the root that has degree 1, and the first descendant of the root with degree 3 (node v) has no other line-graph to a leaf. In this case v counts into $K(T)$, but it does not have property $\mathcal{P}_K$.

If $v \in V$ is not the root of T, then $T_v \in \mathcal{P}_K$ implies $v \in K(T)$. This is because v must have at least two children by the property $\mathcal{P}_K$ and a parent vertex since v is not the root, which means that v has degree at least three. By the definition of $\mathcal{P}_K$, T_v contains a line-graph to a leaf. Hence $n_{\mathcal{P}_K}(T) - 1 \leq |K(T)|$.

For the other direction, we argue that $v \in K(T)$ implies $T_v \in \mathcal{P}_K$, except for at most one vertex $v \in V$. This is because v has degree at least three by the exterior major vertex property, two of which must be the children of v in T_v. Moreover, the path of degree two vertices to a leaf ensured by the exterior major vertex property must be a subtree that is a path in T_v, unless the path of degree two vertices to a leaf is through the parent of v. This can only happen if all ancestors of v have degree two, $\operatorname{root}(T)$ has degree one or two, and if $\operatorname{root}(T)$ has another subtree that does not contain v, this must be a line-graph. In other words, $\operatorname{root}(T)$ can have only one subtree with a major vertex, and v must be the first major vertex on this subtree, if such a v exists. Hence $|K(T)| - 1 \leq n_{\mathcal{P}_K}(T)$. □

In all of our proofs we will combine Lemma 4.3.2 with either Theorem 4.3.1, 4.3.2 or 4.3.3. Since $\mathbb{P}(\mathcal{F} \in \mathcal{P}_L)$ is an easy computation in all cases, most of the difficulty will come from computing $\mathbb{P}(\mathcal{F} \in \mathcal{P}_K)$, where $\mathcal{F}$ is a random tree having the limiting fringe tree distribution (see formulas (4.13), (4.14) and (4.16)). To compute $\mathbb{P}(\mathcal{F} \in \mathcal{P}_K)$, it will often be useful to condition on the degree of the root of $\mathcal{F}$, and another event $\mathcal{E}$, that will be the ringing time of the doomsday clock $\operatorname{Exp}(\rho + \chi)$ in Theorem 4.3.3. Recall that for any non-negative discrete random variable Y we denote by

$$G_Y(x) = \sum_{n=0}^{\infty} \mathbb{P}(Y = n) x^n$$

the probability generating function of Y evaluated at x.

Lemma 4.3.3. *Let κ be the degree of* $\mathrm{root}(\mathcal{F})$. *If v is an offspring of* $\mathrm{root}(\mathcal{F})$, *let B_v be the event that $\mathcal{F}_v$ is a line-graph. Suppose that for some event $\mathcal{E}$ the indicators of B_v, conditioned on κ and $\mathcal{E}$, are independent and identically distributed Bernoulli random variables with parameter q. Then*

$$\mathbb{P}(\mathcal{F} \in \mathcal{P}_K \mid \mathcal{E}) = 1 - G_{\kappa|\mathcal{E}}(1-q) - q\mathbb{P}(\kappa = 1 \mid \mathcal{E}). \tag{4.18}$$

Proof. Let A be the event that the root has at least two offspring, and B_i be the event that the root of the i^{th} subtree is born, and the subtree is a line-graph. Let us denote the event $B := \cup_{i\geq 1} B_i$. By definition the event $\mathcal{F} \in \mathcal{P}_K = A \cap B$. Then we can write

$$\begin{aligned}\mathbb{P}(\mathcal{F} \in \mathcal{P}_K \mid \mathcal{E}) &= \mathbb{P}(A \cap B \mid \mathcal{E}) = 1 - \mathbb{P}(A^c \cup B^c \mid \mathcal{E}) \\ &= 1 - \left(\mathbb{P}(\kappa = 0 \mid \mathcal{E}) + \mathbb{P}(\kappa = 1 \mid \mathcal{E}) + \sum_{k=2}^{\infty} \mathbb{P}(B_1^c \cap \cdots \cap B_k^c \mid \kappa = k, \mathcal{E})\mathbb{P}(\kappa = k \mid \mathcal{E})\right) \\ &= 1 - \mathbb{P}(\kappa = 0 \mid \mathcal{E}) - \mathbb{P}(\kappa = 1 \mid \mathcal{E}) - \sum_{k=2}^{\infty} (1-q)^k \mathbb{P}(\kappa = k \mid \mathcal{E}), \end{aligned} \tag{4.19}$$

where the last line followed since we assumed that B_i are independent $\mathrm{Ber}(q)$ conditioned on κ and $\mathcal{E}$. Noticing that the last sum is the generating function of $(\kappa \mid \mathcal{E})$ evaluated at $1-q$, except that the index starts from two instead of zero, we get

$$\begin{aligned}\mathbb{P}(\mathcal{F} \in \mathcal{P}_K \mid \mathcal{E}) &= 1 - \mathbb{P}(\kappa = 0 \mid \mathcal{E}) - \mathbb{P}(\kappa = 1 \mid \mathcal{E}) - G_{\kappa|\mathcal{E}}(1-q) + \sum_{k=0}^{1} (1-q)^k \mathbb{P}(\kappa = k \mid \mathcal{E}) \\ &= 1 - G_{\kappa|\mathcal{E}}(1-q) - q\mathbb{P}(\kappa = 1 \mid \mathcal{E}) \end{aligned} \tag{4.20}$$

□

Remark 4.3.1. *If we were interested in simply exterior vertices, using the same ideas, the expression in equation* (4.18) *would simplify to* $1 - G_{\kappa|\mathcal{E}}(1-q)$.

4.4 Proofs

In this section we prove Theorems 4.1.2, and 4.2.2–4.2.4.

4.4.1 Metric Dimension of Conditioned Galton-Watson Trees

Proof of Theorem 4.1.2. Combining Lemma 4.3.2 and Theorem 4.3.1, we have that

$$\frac{\mathrm{MD}(\mathcal{T}_n)}{n} \xrightarrow{p} \mathbb{P}(\mathcal{F} \in \mathcal{P}_L) - \mathbb{P}(\mathcal{F} \in \mathcal{P}_K), \tag{4.21}$$

where $\mathcal{F}$ is a Galton-Watson tree with offspring distribution ξ.

Clearly, $\mathbb{P}(\mathcal{F} \in \mathcal{P}_L) = p_0$. It remains to compute $\mathbb{P}(\mathcal{F} \in \mathcal{P}_K)$. Since the subtrees of each offspring in a Galton-Watson tree are independent the conditions of Lemma 4.3.3 are satisfied without conditioning.

We still need to find the value of $q = \mathbb{P}(B_v)$, which is the probability that $\mathcal{F}_v$ is a line-graph since the subtree $\mathcal{F}_v$ is independent of the degree of the root of $\mathcal{F}$. Vertex v can have (i) zero offspring, in which case $\mathcal{F}_v$ is a (trivial) line graph, (ii) one offspring, in which case $\mathcal{F}_v$ is a line with probability q, or (iii) more than one offspring, in which case $\mathcal{F}_v$ is not a line. Hence, we have the equation

$$q = p_0 + qp_1, \tag{4.22}$$

which gives $q = p_0/(1-p_1)$. Substituting equation (4.18) into equation (4.21) with $q = p_0/(1-p_1)$ we obtain the desired result. □

4.4.2 Metric Dimension of Binary Search Trees (Combinatorial Proof)

Proof of Theorem 4.2.2, $m = 2$. Combining Lemma 4.3.2 and Theorem 4.3.2, we obtain that

$$\frac{\mathrm{MD}(\mathcal{T}_n^{(2,-1)})}{n} \xrightarrow{p} \sum_{k=1}^{\infty} \frac{2}{(k+1)(k+2)} \mathbb{P}(\mathcal{T}_k^{(2,-1)} \in \mathcal{P}_L) - \sum_{k=1}^{\infty} \frac{2}{(k+1)(k+2)} \mathbb{P}(\mathcal{T}_k^{(2,-1)} \in \mathcal{P}_K) \tag{4.23}$$

Clearly $\mathbb{P}(\mathcal{T}_k^{(2,-1)} \in \mathcal{P}_L)$ equals 1 for $k = 1$ and 0 for $k > 1$, which implies that the first term in equation (4.23) is 1/3.

It remains to compute the second term in equation (4.23). Recall full and potential vertices from the description of binary search trees on page 2. Let $k' = k - 1$ and $S_k \in \{0, \dots, k'\}$ be the number of (full) vertices in the left subtree when the tree has k (full) vertices. Notice, that the number of potential vertices in the left and right subtrees follows a Pólya urn process with two urns initially with a single white and a single black ball, and that the number of full vertices is always one less than the number of potential vertices in each subtree. Elementary calculation using induction shows that S_k is then uniform over the set $\{0, \dots, k'\}$, or in other words $\mathbb{P}(S_k = \ell) = 1/(k'+1)$, see e.g. [112, Theorems 5.2, 5.3].

Since $S_k \in \{0, k'\}$ implies that the root has degree less than two $\mathbb{P}(\mathcal{T}_k \in \mathcal{P}_K | S_k \in \{0, k'\}) = 0$. By the law of total probability,

$$\begin{aligned}\mathbb{P}(\mathcal{T}_k^{(2,-1)} \in \mathcal{P}_K) &= \sum_{\ell=0}^{k'} \mathbb{P}(\mathcal{T}_k^{(2,-1)} \in \mathcal{P}_K | S_k = \ell)\mathbb{P}(S_k = \ell) \\ &= \frac{1}{k'+1} \sum_{\ell=1}^{k'-1} \mathbb{P}(\mathcal{T}_k^{(2,-1)} \in \mathcal{P}_K | S_k = \ell).\end{aligned} \tag{4.24}$$

Now we focus on the second condition of $\mathcal{P}_K$, the existence of a subtree that is a line. If a subtree has ℓ vertices, we argue that the probability that it is a line is

$$\prod_{i=3}^{\ell} \frac{2}{i} = \frac{2^{\ell-1}}{\ell!}.$$

Indeed, if the subtree has just one or two vertices, it must be a line. Thereafter, conditionally that the subtree is a line after having $i-1$ vertices, when we place the i^{th} vertex into the subtree, we have to sample from i possible places, only two of which keep the subtree a line. Namely, the children of the last vertex on the line. Here we use that the placement of vertices in the binary search tree is uniform over the possible locations, and conditioned that the vertex falls into the left (resp. right) subtree, its placement is uniform over the available locations within this subtree. To compute the probability that at least one of the subtrees is a line we apply an elementary inclusion-exclusion argument. For $1 \le \ell \le k'-1$, we have

$$\begin{aligned}\mathbb{P}(\mathcal{T}_k^{(2,-1)} \in \mathcal{P}_K | S_k = \ell) &= \mathbb{P}(B_1 | S_k = \ell) + \mathbb{P}(B_2 | S_k = \ell) - \mathbb{P}(B_1 \cap B_2 | S_k = \ell) \\ &= \frac{2^{\ell-1}}{\ell!} + \frac{2^{k'-\ell-1}}{(k'-\ell)!} - \frac{2^{k'-2}}{\ell!(k'-\ell)!},\end{aligned} \tag{4.25}$$

where in the last term we used that conditioned on their sizes, the left and right subtree evolve independently. Substituting the rhs back into equation (4.24) and using the basic identities of binomial coefficients, and recalling that $k' = k-1$, we obtain

$$\begin{aligned}\sum_{\ell=1}^{k'-1} \mathbb{P}(\mathcal{T}_k^{(2,-1)} \in \mathcal{P}_K | S_k = \ell) &= \sum_{\ell=1}^{k'-1} \frac{2^{\ell-1}}{\ell!} + \frac{2^{k'-\ell-1}}{(k'-\ell)!} - \frac{2^{k'-2}}{\ell!(k'-\ell)!} = \sum_{\ell=1}^{k'-1} \frac{2^{\ell}}{\ell!} - \frac{2^{k'-2}}{k'!} \sum_{\ell=1}^{k'-1} \binom{k'}{\ell} \\ &= \sum_{\ell=1}^{k'-1} \frac{2^{\ell}}{\ell!} - \frac{2^{k'-2}(2^{k'}-2)}{k'!}.\end{aligned} \tag{4.26}$$

Substituting (4.26) into (4.24) and then into (4.23) we obtain (with $k' = k-1$)

$$\frac{\mathrm{MD}(\mathcal{T}_n^{(2,-1)})}{n} \xrightarrow{p} \frac{1}{3} - \sum_{k=3}^{\infty} \frac{2}{(k+1)(k+2)(k'+1)} \left(\sum_{l=1}^{k'-1} \frac{2^l}{l!} - \frac{2^{k'-2}(2^{k'}-2)}{k'!} \right). \tag{4.27}$$

Getting a closed form expression for $\sum_{\ell=1}^{k'-1} 2^{\ell}/\ell!$ is difficult, but it is clearly bounded by e^2. Since the sum $\sum_{k=\ell+2}^{\infty} 2/(k(k+1)(k+2))$ is also bounded, we can swap the order of the sums to get the easier expression

$$\sum_{k=3}^{\infty} \frac{2}{k(k+1)(k+2)} \sum_{\ell=1}^{k-2} \frac{2^{\ell}}{\ell!} = \sum_{\ell=1}^{\infty} \frac{2^{\ell}}{\ell!} \sum_{k=\ell+2}^{\infty} \frac{2}{k(k+1)(k+2)}. \tag{4.28}$$

The sum $\sum_{k=\ell+2}^{\infty} 2/(k(k+1)(k+2))$ can be evaluated by elementary arithmetic operations and a telescopic sum. Indeed,

$$\sum_{k=\ell+2}^{\infty} \frac{2}{k(k+1)(k+2)} = \sum_{k=\ell+2}^{\infty} \left(\frac{1}{k(k+1)} - \frac{1}{(k+1)(k+2)} \right) = \frac{1}{(\ell+2)(\ell+3)}. \tag{4.29}$$

Substituting back into equation (4.28), elementary arithmetic operations give

$$\sum_{\ell=1}^{\infty}\frac{2^{\ell}}{\ell!}\frac{1}{(\ell+2)(\ell+3)}=\sum_{\ell=1}^{\infty}\frac{(\ell+1)2^{\ell}}{(\ell+3)!}=\sum_{\ell=1}^{\infty}\frac{(\ell+3)2^{\ell}}{(\ell+3)!}-\sum_{\ell=1}^{\infty}\frac{2\cdot 2^{\ell}}{(\ell+3)!}=\frac{1}{3}. \tag{4.30}$$

The last equality follows if we notice that the sum that we are subtracting is the same as the sum we are subtracting from, except it is shifted by one index. Hence, the result of the subtraction is the simply the first term of the sum. A similar compuation yields the following equalities,

$$\sum_{k=3}^{\infty}\frac{2}{k(k+1)(k+2)}\frac{2^{k'-2}(2^{k'}-2)}{k'!}=\sum_{k=3}^{\infty}\frac{2^{2k-3}}{(k+2)!}-\frac{2^{k-1}}{(k+2)!}=\frac{3e^4-48e^2+233}{384}. \tag{4.31}$$

Finally, substituting into equation (4.27) we obtain

$$\frac{\mathrm{MD}(\mathcal{T}_n^{(2,-1)})}{n}\xrightarrow{p}\frac{1}{3}-\frac{1}{3}+\frac{3e^4-48e^2+233}{384}=\frac{3e^4-48e^2+233}{384}, \tag{4.32}$$

which is the desired result. □

4.4.3 Metric Dimension of General Linear Preferential Attachment Trees (Proof Using Fringe Trees)

In this section we prove Theorems 4.2.2, 4.2.3 and 4.2.4. First, we state a few preliminary lemmas. We handle all values of (ρ,χ) together until the last step when we obtain the numerical values. Recall that Lemma 4.3.1 gives an embedding of $(\mathcal{T}_n^{(\rho,\chi)})_{n\geq 1}$ into a Crump-Mode-Jagers process with reproduction function $\Sigma_{\rho,\chi}$ given in Definition 4.3.2. Combining Lemma 4.3.2 and Theorem 4.3.3, we have that

$$\frac{\mathrm{MD}(\mathcal{T}_n^{(\rho,\chi)})}{n}\xrightarrow{a.s.}\mathbb{P}(\mathcal{F}\in\mathcal{P}_L)-\mathbb{P}(\mathcal{F}\in\mathcal{P}_K), \tag{4.33}$$

where $\mathcal{F}$ is a CMJ tree with offspring point process $\Sigma_{\rho,\chi}$ stopped at random time $\tau=\mathrm{Exp}(\rho+\chi)$.

By Definition 4.3.2, the time of the first offspring of the root of $\mathcal{F}$ is an $\mathrm{Exp}(\rho)$ random variable. To find $\mathbb{P}(\mathcal{F}\in\mathcal{P}_L)$ we need to compute the probability that the doomsday clock $\mathrm{Exp}(\rho+\chi)$ rings before the first offspring clock $\mathrm{Exp}(\rho)$. Hence,

$$\mathbb{P}(\mathcal{F}\in\mathcal{P}_L)=\frac{\rho+\chi}{2\rho+\chi}. \tag{4.34}$$

Next, we check that the conditions of Lemma 4.3.3 are satisfied, which will help us to find $\mathbb{P}(\mathcal{F}\in\mathcal{P}_K)$. Let $\Sigma_{\rho,\chi}=(\xi_1,\xi_2\ldots)$ be a linear preferential attachment reproduction process as described in Definition 4.3.2. We will apply Law of Total Probability with respect to the ringing time of the doomsday clock τ. So, for infinitesimal $\mathrm{d}x$, let us take $\mathcal{E}_x:=\{\tau\in(x,x+\mathrm{d}x)\}$ be the event that the doomsday clock τ rings in the interval $(x,x+\mathrm{d}x)$. Recall that we denote by κ the degree of the root of $\mathcal{F}$. Recall that we write κ for the number of children of the root in the limiting fringe tree $\mathcal{F}$.

Lemma 4.4.1. *Conditioned on $\mathcal{E}_x \cap \{\kappa = k\}$, the (unordered) set of times $\{\xi_1, \dots, \xi_k\}$ have the same distribution as k i.i.d. random variables with density*

$$g_x(y) = \frac{1}{Z_g(x)} \mathrm{e}^{\chi y} \tag{4.35}$$

supported on the interval $[0, x]$, with $Z_g(x) = \int_0^x \mathrm{e}^{\chi y}\,\mathrm{d}y$.

This statement is commonly known for $\chi = 0$, when $\Sigma_{(\rho,0)}$ is a Poisson point process (PPP) on $\mathbb{R}^+$ with intensity ρ. In this case, the lemma states that conditioned on the event that $\Sigma_{(\rho,0)}$ has k points on the interval $[0, x]$, the locations of these points have the same distribution as that of k i.i.d. uniform random variables on $[0, x]$.

Proof of Lemma 4.4.1. Recall the distribution of the consecutive birth times

$$\xi_j - \xi_{j-1} \overset{d}{=} \mathrm{Exp}(\rho + (j-1)\chi).$$

Conditioned on $\mathcal{E}_x$, the density that there are k children of the fringe-root, precisely born at ordered times $\underline{r} := (r_1, r_2, \dots, r_k)$, and the $(k+1)$-st child has $r_{k+1} > x$ is:

$$f_o(k, r_1, \dots, r_k \mid \mathcal{E}_x) := \rho \mathrm{e}^{-\rho r_1}(\rho+\chi)\mathrm{e}^{-(\rho+\chi)(r_2-r_1)} \cdot \dots \cdot (\rho+\chi(k-1))\mathrm{e}^{-(\rho+\chi(k-1))(r_k - r_{k-1})}\mathrm{e}^{-(\rho+\chi k)(x-r_k)}.$$

Observing that the coefficient of r_j in the exponent is χ, we see that

$$f_o(k, r_1, \dots, r_k \mid \mathcal{E}_x) = \frac{1}{Z_{f_o}(x)} \cdot \mathrm{e}^{\chi(r_1 + \dots + r_k)} = \frac{Z_g(x)^k}{Z_{f_o}(x)} \prod_{i=1}^{k} g_x(r_i), \tag{4.36}$$

where $Z_{f_o}(x) = \mathrm{e}^{-(\rho+\chi k)} / \prod_{i=0}^{k-1}(\rho + i\chi)$ is the normalizing factor independent of $\underline{r}$ (as long as $\underline{r}$ is really an ordered sequence, otherwise $f_o(k, r_1, \dots, r_k \mid x) = 0$). However, we are not interested in the density of the ordered set of times. The *unordered* set of times $\{\xi_1, \dots \xi_\kappa\}$ has density

$$f_u(k, r_1, \dots, r_k \mid \mathcal{E}_x) = \frac{1}{k!} f_o(k, r_1, \dots, r_k \mid \mathcal{E}_x) = \frac{Z_g(x)^k}{k! Z_{f_o}(x)} \prod_{i=1}^{k} g_x(r_i)$$

by the symmetry of the possible permutations of $r_1, \dots r_k$. Conditioning on k, by Bayes rule we know that

$$f_u(r_1, \dots, r_k \mid \mathcal{E}_x, \kappa = k) = \frac{1}{\mathbb{P}(\kappa = k \mid \mathcal{E}_x)} f_u(k, r_1, \dots, r_k \mid \mathcal{E}_x) = \frac{1}{Z_{f_u}(x)} \prod_{i=1}^{k} g_x(r_i),$$

where $Z_{f_u}(x)$ is the appropriate normalizing factor independent of $\{r_1, \dots, r_k\}$, that is $Z_{f_u}(x) = Z_g(x)^k$. Since the density $f_u(r_1, \dots, r_k \mid \mathcal{E}_x, \kappa = k)$ is the product of the densities $g(r_i)$, the random variables $\{\xi_1,\dots\xi_\kappa\}$ must be i.i.d., with density $g_x(y)$. □

The implication of this lemma is that conditioned on $\mathcal{E}_x$ and $\kappa = k$, the k subtrees of the fringe root are born independently at times following density $g_x(y)$, and evolve independently. Consequently, we can apply Lemma 4.3.3, and we proceed to computing the terms that appear in (4.18). Some of these terms can be simply deduced from a result of [116].

Lemma 4.4.2 (Theorem A.7. of [116])**.** *The offspring distribution of the root (denoted by κ) of a linear preferential attachment tree with parameters ρ and χ stopped at time x is given by* $\mathrm{NBin}(\rho,\mathrm{e}^x)$ *if* $\chi = 1$, $\mathrm{Poi}(\rho x)$ *if* $\chi = 0$ *and* $\mathrm{Bin}(\rho,1-\mathrm{e}^{-x})$ *if* $\chi = -1$, *where* NBin *denotes the negative binomial distribution,* Poi *denotes the Poisson distribution and* Bin *denotes the binomial distribution. In particular,*

$$G_{\kappa|\mathcal{E}_x}(z) = \begin{cases} (\mathrm{e}^{\chi x} + (1-\mathrm{e}^{\chi x})z)^{-\rho/\chi} & \textit{for } \chi = \pm 1 \\ \mathrm{e}^{-x(1-z)} & \textit{for } \chi = 0, \rho = 1, \end{cases} \tag{4.37}$$

and

$$\mathbb{P}(\kappa = 1 \mid \mathcal{E}_x) = \begin{cases} -\frac{\rho}{\chi}(1-\mathrm{e}^{\chi x})\mathrm{e}^{-x(\rho+\chi)} & \textit{for } \chi = \pm 1 \\ x\mathrm{e}^{-x} & \textit{for } \chi = 0, \rho = 1. \end{cases} \tag{4.38}$$

We refer the reader to [116] for a proof. The last unknown variable that we need to compute to apply Lemma 4.3.3 is $q = \mathbb{P}(B_v \mid \kappa = k, \mathcal{E}_x)$, the probability that a subtree $\mathcal{F}_v$ of a child v of $\mathrm{root}(\mathcal{F})$ is a line graph.

Definition 4.4.1. *For an offspring v of the root of $\mathcal{F}$, let us denote v by v_0, and v_j the first offspring of v_{j-1} for $j \geq 1$. In addition, let us denote by τ_{v_j} the birth time of v_j. Let $\tau_{v_j,2}$ denote the birth-time of the second offspring of individual v_j and let $\tau_2 = \min(\{\tau_{v_j,2}\})$.*

We condition on the doomsday clock to ring at time x (this the event $\mathcal{E}_x$). Since we assumed that $v = v_0$ is an offspring of a root, and v is alive before time x, by Lemma 4.4.1, the random variable τ_{v_0} has density $g_x(y)$ defined in equation (4.35). By definition, the event B_v holds if none of the v_j have two offspring until time x, hence, we must find $q = \mathbb{P}(\tau_2 > x)$. To describe τ_2, the following definition will be useful.

Definition 4.4.2. *Consider a Poisson point process $\Pi := \{0 = \pi_0, \pi_1, \pi_2, \dots\}$ on $\mathbb{R}^+$ with intensity $\lambda \in \mathbb{R}^+$ and let $(Y_j)_{j\geq 1}$ be an independent collection of exponential variables, independent of Π, with Y_j having parameter $j\nu \in \mathbb{R}^+$. Let $\zeta := \min\{j : Y_j \leq \pi_{j+1} - \pi_j\}$. Then, the exponential random variable with Poisson increasing rate is*

$$H_{\lambda,\nu} = \pi_\zeta + Y_\zeta. \tag{4.39}$$

Due to the memoryless property of exponential variables, we can think of $H_{\lambda,\nu}$ as a single exponential clock, that starts with initial rate 0 at time 0, and every time the governing Poisson point process Π has a new point, the rate of the clock increases by ν. The next lemma relates $H_{\lambda,\nu}$ to τ_2:

Lemma 4.4.3. *Recall that τ_{v_0} has density $g(y)$ defined in equation* (4.35)*, and let $H_{\rho,\rho+\chi}$ be an exponential random variable with Poisson increasing rate as defined in Definition 4.4.2 independent of τ_{v_0}. Then,*

$$\mathbb{P}(\tau_2 > x) = \mathbb{P}(H_{\rho,\rho+\chi} + \tau_{v_0} > x). \tag{4.40}$$

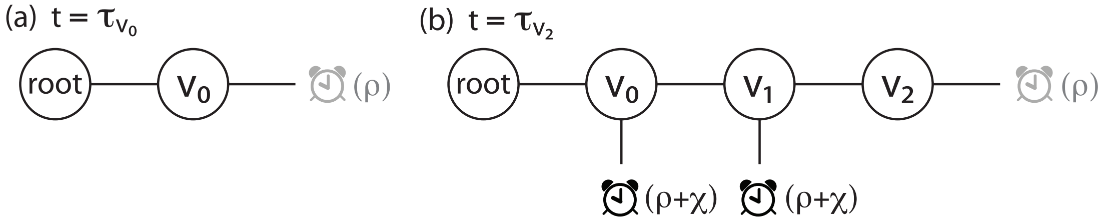


Figure 4.4: Illustration of the proof of Lemma 4.4.3. Part (a) shows the tree at time $t = \tau_{v_0}$, when only v_0 is born, and part (b) shows the tree at time $t = \tau_{v_2}$ assuming $\tau_2 > \tau_{v_2}$. If v_j is the last-born vertex at some time $t < \tau_2$, we have an (grey) exponential clock with intensity ρ to govern the Poisson point process $\tau_{v_1}, \tau_{v_2}, \ldots$, and j (black) exponential clocks with intensity $(\rho+\chi)$ that govern τ_2. If the grey clock rings, a new (black) exponential clock with intensity $(\rho+\chi)$ appears, and τ_2 is the time when the first black clock rings.

Proof. We are going to show that $\tau_2 - \tau_{v_0}$ and $H_{\rho,\rho+\chi}$ has the same distribution and both are independent of τ_{v_0}. First, the independence follows from the fact that differences between births of consecutive children in the Crump-Mode-Jagers tree are using independent exponential clocks, see Definition 4.3.2.

Next we show that $\tau_2 - \tau_{v_0} \overset{d}{=} H_{\rho,\rho+\chi}$. First we identify the underlying PPP. In the CMJ tree, by Definition 4.3.2, the first offspring of every vertex is governed by an exponential clock with rate ρ, hence $(\tau_{v_j} - \tau_{v_0})_{j\geq 0}$ has the same distribution as $(\pi_j)_{j\geq 1}$, a Poisson point process Π with intensity $\lambda = \rho$ in Definition 4.4.2. The first offspring form the line-graph emanating from $v = v_0$, see Figure 4.4.

The random variable τ_2 is defined as the first time any of the vertices $\{v_j \mid j \geq 0\}$ have degree at least three. The inequality $\tau_2 > \tau_{v_1}$ holds deterministically, because this is the first time any vertex (in this case, v_0) can have a second child within the subtree $\mathcal{F}_{v_0}$. This means that until $\tau_{v_1} = \pi_1$, τ_2 cannot happen. Indeed, $\zeta = 0$ cannot happen, since the rate of the exponential clock Y_0 is 0, hence $Y_0 \leq \pi_1 - 0$ happens with probability 0.

By Definition 4.3.2 again, the rate of arrival of the second child of any individual is $\rho+\chi$. For $j \geq 1$, let us look at a scenario when $v_0, v_1, \ldots, v_{j-1}, v_j$ are born and forming a line, i.e., they are born, none of them has a second child yet, and v_{j+1} has not been born yet. That is, we look at a time $t \in (\tau_{v_j}, \tau_{v_{j+1}})$. In this scenario, all of the vertices $v_0, v_1, \dots, v_{j-1}$ are waiting for their second offspring to be born, hence the total rate of arrival of the second offspring is governed by an exponential clock with parameter $j(\rho+\chi)$.

As a result, between τ_{v_j} and $\tau_{v_{j+1}}$ the random variable τ_2 can be described as an $\mathrm{Exp}(j(\rho+\chi))$ random variable (see Figure 4.4). With $\nu = \rho + \chi$, the random variable Y_j in Definition 4.4.2 is also an $\mathrm{Exp}(j(\rho+\chi))$ random variable.

By the memoryless property of exponential variables, conditioned that $\tau_2 > \tau_{v_j}$, τ_2 happens before $\tau_{v_{j+1}}$ if the $\mathrm{Exp}(j(\rho+\chi))$ variable is less than $\tau_{v_{j+1}} - \tau_{v_j}$. Since $\tau_{v_{j+1}} - \tau_{v_j} \overset{d}{=} \pi_{j+1} - \pi_j$, this inequality can be expressed as $Y_j \le \pi_{j+1} - \pi_j$.

In other words, if $Y_j \le \pi_{j+1} - \pi_j$, then j is the index of the last vertex v_j that is born before τ_2, and $\tau_2 = \tau_{v_j} + Y_j$. Otherwise, if $Y_j > \tau_{v_{j+1}} - \tau_{v_j}$, the value of Y_j is irrelevant, $\tau_{v_{j+1}}$ is born before any of the $v_0, v_1, \dots, v_{j-1}$ has a second child, and the rate of getting a second child on the line present goes up by $\rho + \chi$ since now v_j is also waiting for his second offspring to be born. By the memoryless property, we can restart the clocks and use new exponential variables for comparison. Hence, we move on to the next index $j+1$. The random variable ζ describes the first index j for which $Y_j \le \tau_{v_{j+1}} - \tau_{v_j}$, which is the first (and only) "relevant" index. Then,

$$\tau_2 - \tau_{v_0} = Y_\zeta + \tau_{v_\zeta} - \tau_{v_0} \overset{d}{=} Y_\zeta + \pi_\zeta,$$

which is precisely what we needed. □

Although for the proofs we will only need $H_{\rho,\rho+\chi}$, we find the tail distribution of $H_{\lambda,\nu}$ in a general form.

Lemma 4.4.4. *The tail distribution of $H_{\lambda,\nu}$ is given by*

$$\mathbb{P}(H_{\lambda,\nu} > t) = \exp\Big\{-\lambda t + \frac{\lambda}{\nu}(1 - \mathrm{e}^{-\nu t})\Big\}. \tag{4.41}$$

Proof of Lemma 4.4.4. Let us condition on the number of points in the Poisson point process $\pi_1, \pi_2, \dots$ before time t, which is just a Poisson random variable with intensity λt. We have

$$\mathbb{P}(H_{\lambda,\nu} > t) = \sum_{k=1}^{\infty} \frac{(\lambda t)^k \mathrm{e}^{-\lambda t}}{k!} \mathbb{E}\big[\mathrm{e}^{-\nu(\pi_2-\pi_1)} \mathrm{e}^{-2\nu(\pi_3-\pi_2)} \dots \mathrm{e}^{-(k-1)\nu(\pi_k-\pi_{k-1})} \mathrm{e}^{-k\nu(t-\pi_k)}\big], \tag{4.42}$$

where the expectation is over the random points $\pi_1, \dots, \pi_k$. By standard properties of the Poisson point process (in the spirit of Lemma 4.4.1 with $\rho = \lambda$ and $\chi = 0$), we can sample the points $\pi_1, \dots, \pi_k$ by sampling k points uniformly from interval $[0, t]$ and then indexing them such that $\pi_1 < \cdots < \pi_k$. Then, by a telescopic cancellation we obtain

$$\mathbb{P}(H_{\lambda,\nu} > t) = \sum_{k=1}^{\infty} \frac{(\lambda t)^k \mathrm{e}^{-\lambda t}}{k!} \mathbb{E}\big[\mathrm{e}^{\nu(\pi_1+\cdots+\pi_k)}\big] \mathrm{e}^{-k\nu t}. \tag{4.43}$$

Since each π_j appears exactly once in the sum, and we can forget about their ordering. Then, the π_j become independent uniform random variables on $[0,t]$, and we can simplify to

$$\begin{aligned}\mathbb{P}(H_{\lambda,\nu} > t) &= \sum_{k=1}^{\infty} \frac{(\lambda t)^k \mathrm{e}^{-\lambda t}}{k!} \mathbb{E}\big[\mathrm{e}^{\nu t U[0,1]}\big]^k \mathrm{e}^{-k\nu t} \\ &= \sum_{k=1}^{\infty} \frac{(\lambda t)^k \mathrm{e}^{-\lambda t}}{k!} \mathrm{e}^{-k\nu t} \left(\int_0^1 \mathrm{e}^{\nu t x} \mathrm{d}x \right)^k \\ &= \sum_{k=1}^{\infty} \frac{(\lambda t)^k \mathrm{e}^{-\lambda t}}{k!} \mathrm{e}^{-k\nu t} \frac{(\mathrm{e}^{\nu t}-1)^k}{(\nu t)^k}. \end{aligned} \tag{4.44}$$

Now simply cancelling the appropriate terms and factoring out the term not depending on k we reach the final result

$$\begin{aligned}\mathbb{P}(H_{\lambda,\nu} > t) &= \mathrm{e}^{-\lambda t} \sum_{k=1}^{\infty} \frac{\lambda^k \nu^{-k}}{k!} \left(\frac{\mathrm{e}^{\nu t}-1}{\mathrm{e}^{\nu t}} \right)^k \\ &= \exp\Big(-\lambda t + \frac{\lambda}{\nu}(1-\mathrm{e}^{-\nu t}) \Big). \end{aligned} \tag{4.45}$$

□

We proceed by computing $\mathbb{P}(\mathcal{F} \in \mathcal{P}_K)$ in (4.33). In order to do this, we make use of Lemma 4.3.3, that requires the conditional generating function of $\kappa \mid \mathcal{E}$, that we identified in Lemma 4.4.2 when we take $\mathcal{E}_x = \{\tau \in (x, x+\mathrm{d}x)\}$. It remains to calculate $1-q = \mathbb{P}(\tau_2 < x)$ that is needed as the argument of the generating function. So, we combine Lemmas 4.4.3 and 4.4.4 to find $q = \mathbb{P}(\tau_2 > x)$. By Lemma 4.4.3, we must compute the convolution of $H_{\rho,\rho+\chi}$ and the random variable with density $g_x(y)$ defined in equation (4.35), which gives

$$q = \mathbb{P}(\tau_2 > x) = \frac{1}{Z_g(x)} \int_0^x \mathrm{e}^{\chi(x-t)} \exp\Big(-\rho t + \frac{\rho}{\rho+\chi} - \frac{\rho \mathrm{e}^{-(\rho+\chi)t}}{\rho+\chi} \Big) \mathrm{d}t$$

We make the substitution $u = \frac{\rho}{\rho+\chi}\mathrm{e}^{-(\rho+\chi)t}$, which gives $t = -\log(\frac{\rho+\chi}{\rho}u)/(\rho+\chi)$ and $\mathrm{d}t = \frac{-1}{u(\rho+\chi)}\,\mathrm{d}u$, to get

$$\begin{aligned} q = \mathbb{P}(\tau_2 > x) &= \frac{\mathrm{e}^{\chi x + \frac{\rho}{\rho+\chi}}}{Z_g(x)} \int_{\frac{\rho}{\rho+\chi}}^{\frac{\rho}{\rho+\chi}\mathrm{e}^{-(\rho+\chi)x}} \frac{\rho+\chi}{\rho} u \mathrm{e}^{-u} \frac{-1}{u(\rho+\chi)} \mathrm{d}u \\ &= \frac{\mathrm{e}^{\chi x + \frac{\rho}{\rho+\chi}(1-\mathrm{e}^{-(\rho+\chi)x})} - \mathrm{e}^{\chi x}}{\rho Z_g(x)}, \end{aligned} \tag{4.46}$$

where $Z_g(x) = \int_0^x \mathrm{e}^{\chi y}\,\mathrm{d}y$ is from Lemma 4.4.1. Finally, we are ready to apply Lemma 4.3.3. Let us assume $\chi \neq 0$. The $\chi = 0$ case will be handled in Section 4.4.3 below.

Proof of Theorem 4.2.4. Substituting equations (4.37), (4.38) into (4.18) we obtain

$$\begin{aligned}\mathbb{P}(\mathcal{F}\in\mathcal{P}_K\mid\mathcal{E}_x) &= 1-G_{\kappa|\mathcal{E}}(1-q)-q\mathbb{P}(\kappa=1\mid\mathcal{E})\\ &= 1-(\mathrm{e}^{\chi x}+(1-\mathrm{e}^{\chi x})(1-q))^{-\rho/\chi}+q\frac{\rho}{\chi}(1-\mathrm{e}^{\chi x})\mathrm{e}^{-x(\rho+\chi)}\\ &= 1-(1-q(1-\mathrm{e}^{\chi x})))^{-\rho/\chi}+\frac{\rho}{\chi}q(1-\mathrm{e}^{\chi x})\mathrm{e}^{-x(\rho+\chi)}. \end{aligned}\tag{4.47}$$

Now, using the value q from (4.46), and by $Z_g(x)=(\mathrm{e}^{\chi x}-1)/\chi$, we have

$$q(1-\mathrm{e}^{\chi x}) = -\chi\frac{\mathrm{e}^{\chi x+\frac{\rho}{\rho+\chi}(1-\mathrm{e}^{-(\rho+\chi)x})}-\mathrm{e}^{\chi x}}{\rho}.\tag{4.48}$$

Substituting (4.48) into the second and third terms of equation (4.47), the formula becomes:

$$\mathbb{P}(\mathcal{F}\in\mathcal{P}_K\mid\mathcal{E}_x) = 1-\left(1+\chi\frac{\mathrm{e}^{\chi x+\frac{\rho}{\rho+\chi}(1-\mathrm{e}^{-(\rho+\chi)x})}-\mathrm{e}^{\chi x}}{\rho}\right)^{-\rho/\chi}-\mathrm{e}^{-\rho x+\frac{\rho}{\rho+\chi}(1-\mathrm{e}^{-(\rho+\chi)x})}+\mathrm{e}^{-\rho x}.$$

We apply the law of total probability with respect to the density of the doomsday clock τ with rate $\rho+\chi$ to compute

$$\begin{aligned}\mathbb{P}(\mathcal{F}\in\mathcal{P}_K) &= \int_0^\infty(\rho+\chi)\mathrm{e}^{-x(\rho+\chi)}\mathbb{P}(\mathcal{F}\in\mathcal{P}_K\mid\mathcal{E}_x)\,\mathrm{d}x\\ &= 1-\int_0^\infty(\rho+\chi)\mathrm{e}^{-x(\rho+\chi)}\left(1+\chi\frac{\mathrm{e}^{\chi x+\frac{\rho}{\rho+\chi}(1-\mathrm{e}^{-(\rho+\chi)x})}-\mathrm{e}^{\chi x}}{\rho}\right)^{-\rho/\chi}\mathrm{d}x\\ &\quad -\int_0^\infty(\rho+\chi)\mathrm{e}^{-x(\rho+\chi)}\mathrm{e}^{-\rho x+\frac{\rho}{\rho+\chi}(1-\mathrm{e}^{-(\rho+\chi)x})}\,\mathrm{d}x+\int_0^\infty(\rho+\chi)\mathrm{e}^{-x(\rho+\chi)}\mathrm{e}^{-\rho x}\,\mathrm{d}x.\end{aligned}\tag{4.49}$$

Let us denote the three integrals on the right hand side by I_1, I_2, I_3, respectively. The third integral can be computed explicitly as

$$I_3=\int_0^\infty(\rho+\chi)\mathrm{e}^{-x(\rho+\chi)}\mathrm{e}^{-\rho x}\,\mathrm{d}x=\frac{\rho+\chi}{2\rho+\chi},\tag{4.50}$$

and we observe that this term equals $\mathbb{P}(\mathcal{F}\in\mathcal{P}_L)$ in (4.34), and hence it cancels when substituted back into equation (4.33). So, for (4.33), we obtain the result for linear preferential attachment trees

$$\begin{aligned}\frac{\mathrm{MD}(\mathcal{T}_n^{(\rho,\chi)})}{n}\xrightarrow{a.s.} -1+\int_0^\infty(\rho+\chi)\mathrm{e}^{-x(\rho+\chi)}\left(1+\chi\frac{\mathrm{e}^{\chi x+\frac{\rho}{\rho+\chi}(1-\mathrm{e}^{-(\rho+\chi)x})}-\mathrm{e}^{\chi x}}{\rho}\right)^{-\rho/\chi}\mathrm{d}x\\ +\int_0^\infty(\rho+\chi)\mathrm{e}^{-x(\rho+\chi)}\mathrm{e}^{-\rho x+\frac{\rho}{\rho+\chi}(1-\mathrm{e}^{-(\rho+\chi)x})}\,\mathrm{d}x.\end{aligned}\tag{4.51}$$

This is the general formula for (ρ,χ) when $\chi \neq 0$. This also finishes the proof of Theorem 4.2.4, since the formula in (4.12) is recovered when $\chi = 1$. □

Now we evaluate this further for the special case $\chi = -1$, and obtain the metric dimension of m-ary increasing trees ($\rho = m \in \mathbb{N}$, and $\chi = -1$).

Proof of Theorem 4.2.2. When $\chi = -1$ and $\rho = m \in \mathbb{N}$, equation (4.51) simplifies to

$$\begin{aligned}\frac{\mathrm{MD}(\mathcal{T}_n^{(m,-1)})}{n} \xrightarrow{a.s.} -1 &+ \int_0^\infty (m-1)\mathrm{e}^{-x(m-1)} \left(1 - \frac{\mathrm{e}^{-x+\frac{m}{m-1}(1-\mathrm{e}^{-(m-1)x})} - \mathrm{e}^{-x}}{m}\right)^m \mathrm{d}x \\ &+ \int_0^\infty (m-1)\mathrm{e}^{-x(m-1)}\mathrm{e}^{-mx+\frac{m}{m-1}(1-\mathrm{e}^{-(m-1)x})}\,\mathrm{d}x.\end{aligned} \tag{4.52}$$

In the first row, the last bracket is of the form $(1-\mu+\nu)^m$, that we expand using the trinomial formula:

$$(1-\mu+\nu)^m = \sum_{(i,j)\in\mathbb{N}^2 : i+j\le m} \binom{m}{i,j}(-1)^i \mu^i \nu^j.$$

We apply this formula with $\mu = \mathrm{e}^{-x+\frac{m}{m-1}(1-\mathrm{e}^{-(m-1)x})}/m$ and $\nu = \mathrm{e}^{-x}/m$. After collecting terms, and taking into account that the integral in the last row of equation (4.52) can be merged with the term corresponding to $(i,j) = (1, m-1)$ of the expansion, changing the coefficient, we arrive at

$$\frac{\mathrm{MD}(\mathcal{T}_n^{(m,-1)})}{n} \xrightarrow{a.s.} \sum_{0\le i+j\le m} a_{i,j} \int_0^\infty \mathrm{e}^{-b_{i,j}x + i\frac{m}{m-1}(1-\mathrm{e}^{-(m-1)x})}\mathrm{d}x, \tag{4.53}$$

where

$$a_{1,m-1} = (m-1)\left(1 - \frac{m}{m^m}\right) \tag{4.54}$$

$$a_{i,j} = (m-1)\frac{(-1)^i}{m^{i+j}}\binom{m}{i,j} \qquad \text{if } (i,j) \neq (1,m-1) \tag{4.55}$$

$$b_{i,j} = (m-1) + i + j. \tag{4.56}$$

For $i = 0$, the coefficient $im/(m-1)$ of the doubly-exponential term in equation (4.53) in the exponent is 0, hence these terms simplify. We sum over the $i = 0$ terms in j, and perform the integration to obtain

$$\sum_{j=0}^m \frac{m-1}{m^j}\binom{m}{j}\int_0^\infty \mathrm{e}^{-(m-1+j)x}\mathrm{d}x = \sum_{j=0}^m \frac{m-1}{m^j}\binom{m}{j}\frac{1}{m-1+j}. \tag{4.57}$$

Observe that the $j = 0$ term is 1, and hence cancels the -1 in the first term of the right hand side of equation (4.52). For the integral indexed by $(i \neq 0, j)$ we can substitute $u = i\frac{m}{m-1}\mathrm{e}^{-(m-1)x}$

which gives $x = -\log(\frac{m-1}{im}u)/(m-1)$ and $\mathrm{d}x = \frac{-1}{u(m-1)}\,\mathrm{d}u$, to obtain

$$\begin{aligned}\sum_{\substack{0\le i+j\le m\\ i\neq 0}} a_{i,j}\int\limits_{0}^{\infty} \mathrm{e}^{-b_{i,j}x+\frac{im}{m-1}(1-\mathrm{e}^{-(m-1)x})}\,\mathrm{d}x &= \sum_{\substack{0\le i+j\le m\\ i\neq 0}} a'_{i,j}\int\limits_{\frac{im}{m-1}}^{0} u^{b'_{i,j}}\mathrm{e}^{-u}\,\mathrm{d}u\\ &= \sum_{\substack{0\le i+j\le m\\ i\neq 0}} -a'_{i,j}\gamma\left(b'_{i,j}+1,\frac{im}{m-1}\right), \qquad (4.58)\end{aligned}$$

where

$$a'_{i,j} = a_{i,j}\mathrm{e}^{\frac{im}{m-1}}\left(\frac{m-1}{im}\right)^{\frac{b_{i,j}}{m-1}}\frac{-1}{m-1} \qquad (4.59)$$

$$b'_{i,j} = \frac{b_{i,j}}{m-1}-1 = \frac{i+j}{m-1}. \qquad (4.60)$$

Combining equations (4.53)-(4.60) gives the formula

$$\frac{\mathrm{MD}(\mathcal{T}_n^{(m,-1)})}{n} \xrightarrow{a.s.} \sum_{j=1}^{m}\frac{m-1}{(m-1+j)m^j}\binom{m}{j} - \sum_{\substack{0\le i+j\le m\\ i\neq 0}} a'_{i,j}\gamma\left(\frac{i+j}{m-1}+1,\frac{im}{m-1}\right), \qquad (4.61)$$

which agrees with equation (4.7) in Theorem 4.2.2 with $A_{i,j} = -a'_{i,j}$.

For the binary search tree, that is, $m = 2$ we evaluate the coefficients in equations (4.57) and (4.59) numerically. Starting with equation (4.57), then proceeding to the coefficients $a'_{1,1}, a'_{1,0}, a'_{2,0}$, we get

$$\begin{aligned}\sum_{j=1}^{m}\frac{m-1}{m^j}\binom{m}{j}\frac{1}{m-1+j} &= \frac{1}{2}\cdot 2\cdot\frac{1}{2}+\frac{1}{3}\cdot\frac{1}{4} = \frac{7}{12}\\ -a'_{1,1} &= \frac{1}{2}\mathrm{e}^2\frac{1}{2^3} = \frac{\mathrm{e}^2}{2^4}\\ -a'_{1,0} &= -\frac{1}{2}2\mathrm{e}^2\frac{1}{2^2} = -\frac{\mathrm{e}^2}{2^2}\\ -a'_{2,0} &= \frac{1}{4}\mathrm{e}^4\frac{1}{4^2} = \frac{\mathrm{e}^2}{2^8}.\end{aligned}$$

Substituting these values into equation (4.61) gives

$$\frac{\mathrm{MD}(\mathcal{T}_n^{(2,-1)})}{n} \xrightarrow{a.s.} \frac{\mathrm{e}^2}{2^4}\gamma(3,2)-\frac{\mathrm{e}^2}{4}\gamma(2,2)+\frac{\mathrm{e}^4}{2^8}\gamma(3,4)+\frac{7}{12} = \frac{233-48\mathrm{e}^2+3\mathrm{e}^4}{384}. \qquad (4.62)$$

□

Next we proceed with the random recursive tree ($\chi = 0$ and $\rho = 1$). The proof is analogous to the proof of Theorem 4.2.4. We proceed from formula (4.46).

Proof of Theorem 4.2.3. In this case, equation (4.46) yields

$$q = \frac{\mathrm{e}^{\chi x + \frac{\rho}{\rho+\chi}(1-\mathrm{e}^{-(\rho+\chi)x})} - \mathrm{e}^{\chi x}}{\rho Z_g(x)} = \frac{\mathrm{e}^{(1-\mathrm{e}^{-x})} - 1}{x}. \tag{4.63}$$

Substituting equations (4.37), (4.38) for $\chi = 0$ and (4.63) into equation (4.18) now gives

$$\mathbb{P}(\mathcal{F} \in \mathcal{P}_K \mid \mathcal{E}_x) = 1 - \mathrm{e}^{-x(1-(1-q))} - qx\mathrm{e}^{-x} = 1 - \exp(1 - \mathrm{e}^{1-\mathrm{e}^{-x}}) - \mathrm{e}^{1-x-\mathrm{e}^{-x}} + \mathrm{e}^{-x}. \tag{4.64}$$

In this case, τ is exponential with rate $\rho = 1$. We apply the law of total probability to compute

$$\mathbb{P}(\mathcal{F} \in \mathcal{P}_K) = \int_0^\infty \mathbb{P}(\mathcal{F} \in \mathcal{P}_K \mid \mathcal{E}_x)\mathrm{e}^{-x}\,\mathrm{d}x = 1 - \int_0^\infty \mathrm{e}^{-x}\big(1 - \exp(\mathrm{e}^{1-\mathrm{e}^{-x}}) + \mathrm{e}^{1-x-\mathrm{e}^{-x}} - \mathrm{e}^{-x}\big)\,\mathrm{d}x. \tag{4.65}$$

We make the substitution $u = \mathrm{e}^{-x}$, which gives $x = -\log(u)$ and $\mathrm{d}x = -\frac{1}{u}\,\mathrm{d}u$, to get

$$\mathbb{P}(\mathcal{F} \in \mathcal{P}_K) = 1 + \int_1^0 (\mathrm{e}^{1-\mathrm{e}^{1-u}} + u\mathrm{e}^{1-u} - u)\,\mathrm{d}u = 1 - \int_0^1 \mathrm{e}^{1-\mathrm{e}^{1-u}}\,\mathrm{d}u - \mathrm{e}\gamma(2,1) + \frac{1}{2}, \tag{4.66}$$

where γ was defined in equation (4.5). Furthermore, we substitute $v = \mathrm{e}^{1-u}$ in the integral still remaining, which gives $u = 1 - \log(v)$ and $\mathrm{d}u = -\frac{1}{v}\,\mathrm{d}v$, to get

$$\int_0^1 \mathrm{e}^{1-\mathrm{e}^{1-u}}\,\mathrm{d}u = -\int_\mathrm{e}^1 v^{-1}\mathrm{e}^{1-v}\,\mathrm{d}v = \mathrm{e}\int_1^\mathrm{e} v^{-1}\mathrm{e}^{-v}\,\mathrm{d}v. \tag{4.67}$$

Substituting back into equation (4.33) we obtain the final result

$$\frac{\mathrm{MD}(\mathcal{T}_n^{(1,0)})}{n} \xrightarrow{a.s.} \frac{1}{2} - \left(1 - \mathrm{e}\int_1^\mathrm{e} v^{-1}\mathrm{e}^{-v}\,\mathrm{d}v - \mathrm{e}\gamma(2,1) + \frac{1}{2}\right) = \mathrm{e}\left(\int_1^\mathrm{e} v^{-1}\mathrm{e}^{-v}\,\mathrm{d}v + \gamma(2,1)\right) - 1. \tag{4.68}$$

This finishes the proof. □

5 Sequential Metric Dimension of Erdős-Rényi Random Graphs

In this chapter, we prove Theorem 3.4.1 by analyzing the sequential metric dimension (SMD) of Erdős-Rényi random graphs. See Chapter 3 for a general introduction about the SMD and its connection with source identification.

We start by reviewing the previous work related to this chapter in Section 5.1, after which we proceed to analyzing a simplified model involving random matrices instead of random graphs in Sections 5.2 and 5.3. The ideas from these *warmup* sections can be combined with the theory of Erdős-Rényi graphs (presented in Section 5.4) to prove our main result in Section 5.5.

This chapter is based on the publication [188] by Ódor and Thiran.

5.1 Related Work

Our analysis on the SMD of Erdős-Rényi random graphs is inspired by a similar analysis on the MD by [36]. The most important example is the *expansion properties* of connected Erdős-Rényi graphs, the main technique developed in [36]. According to this property, the distances $d(w_i, v)$ between any node v and a randomly picked (sensor) node w_i are dominated by one or two values from the set $\{D, D-1\}$, where D is the diameter of G (see Figure 5.2 and Table 5.1). Therefore, the information acquired in each step is essentially binary, which is a very important intuition to keep in mind. While the techniques of [36] are useful to study the SMD, we must also complement them with new techniques of our own. For instance in the SMD upper bound, we need to analyse an interactive game of possibly N steps instead of selecting the queries in a single round. In particular, the order in which we reveal the edges of the random graph is completely different in our analysis than in [36]. For the SMD lower bound, we could have split our proof in several cases, and for some of them we could have used the results of [36] directly. Instead, we introduce a coupling argument, which succeeds without case-work, and gives a clean alternative proof to the MD lower bound as well.

In [142], the authors assume a very similar model to ours, except that the queries are of the form (v, r), and the answers are binary indicating whether the target is in the ball around node v with radius r. Clearly, in Erdős-Rényi graphs, where distance queries happen to have essentially binary answers, the two models are very similar. Indeed, [142] independently recovers many of the results of [36]. In [142], the adaptive version of the problem is also introduced, which is very similar to the SMD, but they do not have any results on the adaptive version of the problem in Erdős-Rényi graphs. Since in the SMD we assume that we can use strictly more information than the binary model, our lower bounds are readily applicable to the binary model. The upper bounds are not readily applicable, but they could be extended with minimal modifications to the proof.

The binary nature of the answers to distance queries in Erdős-Rényi graphs suggests that our setup has close connections with Generalized Binary Search [184]. In a sense, our problem setup can be seen as the dual version of graph binary search introduced by [80], where the observations reveal the first edge in the shortest path instead of the path length. Although the two models share some similarities, we must point out that while [80] focuses on an algorithm for general graphs (with noisy but adaptive observations), our work provides asymptotically almost sure results on the sample complexity of an algorithm and a matching lower bound for all possible algorithms on Erdős-Rényi graphs (with noiseless observations), thus the authors aim for different goals. In terms of goals, the work most similar to ours is perhaps [77]. In their paper, the authors consider a version of the Cop and Robber game on Erdős-Rényi graphs of diameter two: The target can “move” between turns, and in order to locate this moving target, it is not the number of turns but the number of sensors that the player selects in each turn that we want to minimize. Recently, the results of [77] were extended by [76] to Erdős-Rényi graphs with a diameter larger than two, and they found that in that range, the number of sensors needed in the Cop and Robber game is strictly less then the SMD. Similarly to our proofs, the proofs in [76] make use of the expansion properties developed in [36].

The methods in this chapter connect several different ideas developed in different communities, which have not been connected before: in Section 5.2, we abstract out one of the key ideas of [36] and connect it with the Birthday Problem, and in Section 5.3, we connect the SMD with Generalized Binary Search [184].

5.2 Warmup1: Random Bernoulli Matrices with Pairwise Different Columns

In this section we consider an $M \times N$ random matrix A, with entries drawn independently from a Bernoulli distribution, and we are interested in the minimal M for which A still has pairwise different columns with high probability. This M can be viewed as the query complexity of binary search with random Bernoulli queries, where the i^{th} query can distinguish between targets j and k if $A_{ij} \neq A_{ik}$.

For notation, let us consider the binary matrix A with row indices $\mathcal{R} = [M]$ and column indices $\mathcal{C} = [N]$. For $R \subseteq \mathcal{R}$ and $W \subseteq \mathcal{C}$, let $A_{R,W}$ be the submatrix of A restricted to rows R and columns W.

Theorem 5.2.1. *Let $N \in \mathbb{N}$, let $0 < q(N) \leq 1/2$ and $M(N) \in \mathbb{N}$ be functions possibly depending on N, and let us define the random matrix $A \in \mathrm{Ber}(q)^{M \times N}$. Let $\mathcal{A}$ be the property that A has pairwise different columns. Then*

$$\hat{M}(N) = \frac{\log(N)}{\log\left(1/\sqrt{q^2 + (1-q)^2}\right)} \tag{5.1}$$

is the threshold function for $\mathcal{A}$. That is, for any $0 < q(N) \leq 1/2$ and $1 \gg \epsilon(N) \gg \frac{1}{\log(N)}$,

(i) if $M \geq (1 + \epsilon(N))\hat{M}$, then $\lim_{N\to\infty} P(A \in \mathcal{A}) = 1$

(ii) if $M \leq (1 - \epsilon(N))\hat{M}$, then $\lim_{N\to\infty} P(A \in \mathcal{A}) = 0$.

We could not find this particular theorem stated this way in the literature, however, there exist many related results. Computing the probability that an $N \times N$ random Bernoulli matrix is singular is a famous problem first proposed by Komlós in 1967 [148]. Clearly, if the matrix has two identical columns, then it is also singular, hence we obtain the lower bound

$$\mathbb{P}(A \notin \mathcal{A}) \leq \mathbb{P}(A \text{ is singular}).$$

Most of the research on the singularity of random Bernoulli matrices has been on the upper bound [133, 231], with the exception of [17]. In [17], the authors lower bound $\mathbb{P}(A \notin \mathcal{A})$ by using an inclusion-exclusion type argument. However, this bound is too loose in our case as we are interested in $P(A \in \mathcal{A})$ of an $M \times N$ matrix, where M is close to the threshold. Our analysis in this chapter could potentially be applied to tighten some of the bounds in [17], although the improvement would appear only in a high (5^{th}) order term of the bound.

Another well-studied problem related to $P(A \in \mathcal{A})$ is the Birthday Problem (BP). Indeed, when $q = 1/2$, we obtain the standard formulation of the BP with N people and 2^M days. For $0 < q < 1/2$, the columns (birthdays) are not equiprobable anymore, hence we obtain BP with heterogenous birthday probabilities. The non-coincidence probability of two birthdays in this case has been computed exactly using a recursive formula by [171]. Rigorous closed-form approximations for the constant qs were given by [18]. Intuitively, the events that two birthdays coincide are rare and almost independent, so the number of coincidences can be approximated by a Poisson random variable. However, Poisson approximation can work only as long as the number of pairwise collisions dominates the number of multi-collisions (i.e., collisions of ≥ 3 columns), which happens only for qs that do not decrease too fast with N. For fast-decaying qs, we need to use a different technique; we upper bound $P(A \in \mathcal{A})$ by the probability that A does not contain two identically zero columns. Indeed, the event that all columns are different implies that that no two identically zero columns can exist, hence it must have a smaller probability.

Theorem 5.2.1 can also be viewed as a simplified version of [36], which will become clear in Section 5.5. Our proofs also follow [36]: We chose to use a combination of the first moment method and Suen's inequality [228, 125] is used instead of Poisson approximation.

5.2.1 Proof of Theorem 5.2.1 Part (i)

Proof. Let $M = \frac{(1+\epsilon(N))\log(N)}{\log\left(1/\sqrt{q^2+(1-q)^2}\right)}$ with $\epsilon(N) \gg \frac{1}{\log(N)}$ and X be the number of pairs of columns in $\mathcal{C}$ which are identical (i.e., colliding). Let X_{xy} be the indicator of $A_{\mathcal{R},x} = A_{\mathcal{R},y}$ for $x \neq y \in \mathcal{C}$, and let P_k be the marginal that k fixed columns all collide. By the identity

$$\alpha^{\hat{M}} = \alpha^{\frac{\log(N)}{\log\left(1/\sqrt{q^2+(1-q)^2}\right)}} = N^{-\frac{2\log(\alpha)}{\log\left(q^2+(1-q)^2\right)}} \tag{5.2}$$

for all $\alpha \in \mathbb{R}$, we have by taking $\alpha = (q^2 + (1-q)^2)$

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[\sum_{x\neq y\in\mathcal{C}} X_{xy}] = \sum_{x\neq y\in\mathcal{C}} P_2 = \binom{N}{2}(q^2+(1-q)^2)^{(1+\epsilon(N))\hat{M}} \\ &= \frac{N(N-1)}{2}N^{-2(1+\epsilon(N))} < \frac{1}{2}N^{-2\epsilon(N)} \to 0 \end{aligned} \tag{5.3}$$

as $N \to \infty$. By the first moment method, it follows as $N \to \infty$ that

$$\mathbb{P}(A \notin \mathcal{A}) = \mathbb{P}(X > 0) \leq \mathbb{E}[X] \to 0 \tag{5.4}$$

□

5.2.2 Proof of Theorem 5.2.1 Part (ii)

The lower bound will be more involved. We are going to split our argument into two cases; Case 1: $q > \epsilon$ and Case 2: $q \leq \epsilon$. In Case 1, we will use Suen's inequality and in Case 2 we will bound the probability that A does not contain two identically zero columns.

Proof of Theorem 5.2.1 Part (ii), Case 1: $q > \epsilon$. To be able to apply Suen's inequality, we will need to show that pairwise collisions dominate three-way collisions. For this we must estimate P_2 and P_3. Using equation (5.2), with $\alpha = (q^2 + (1-q)^2)^{(1-\epsilon)}$

$$P_2 = (q^2 + (1-q)^2)^{(1-\epsilon)\hat{M}} = N^{-2(1-\epsilon)}, \tag{5.5}$$

and with $\alpha = (q^3 + (1-q)^3)^{(1-\epsilon)}$

$$P_3 = (q^3 + (1-q)^3)^{(1-\epsilon)\hat{M}} = N^{-2(1-\epsilon)\frac{\log(q^3+(1-q)^3)}{\log\left(q^2+(1-q)^2\right)}} \leq N^{-(1-\epsilon)(3+\frac{3}{2}q)}, \tag{5.6}$$

because one can check that for $0 \leq q \leq \frac{1}{2}$

$$\frac{\log(q^3 + (1-q)^3)}{\log\left(q^2 + (1-q)^2\right)} \geq \frac{3}{2} + \frac{3q}{4}. \tag{5.7}$$

Now we apply Suen's inequality [228, 125] to our setting. Let us define the index set $I = \{\{x,y\} \mid x \neq y \in \mathcal{C}\}$, which allows us to index X_{xy} as X_α for $\alpha \in I$. For each $\alpha \in I$ we define the "neighborhood of dependence" $B_\alpha = \{\beta \in I \mid \beta \cap \alpha \neq \emptyset\}$. Indeed, X_α independent of X_β if $\beta \notin B_\alpha$. Then, Suen's inequality implies

$$\mathbb{P}(A \in \mathcal{A}) = \mathbb{P}(X = 0) \leq \exp(-\lambda + \Delta \mathrm{e}^{2\delta}) \tag{5.8}$$

where, using $|I| = \binom{N}{2}$, $|B_\alpha| = (2N-3)$ and equations (5.5) and (5.6),

$$\lambda = \sum_{\alpha \in I} \mathbb{E}[X_\alpha] = \binom{N}{2} P_2 > \frac{1}{4} N^{2\epsilon} \tag{5.9}$$

$$\Delta = \sum_{\alpha \in I} \sum_{\alpha \neq \beta \in B_\alpha} \frac{1}{2} E[X_\alpha X_\beta] = \binom{N}{2} (2N-4) \frac{1}{2} P_3 \leq N^{3-(1-\epsilon)(3+\frac{3}{2}q)} \tag{5.10}$$

$$\delta = \max_{\alpha \in I} \sum_{\alpha \neq \beta \in B_\alpha} E[X_\beta] = (2N-4) P_2 < 2N^{-1+2\epsilon}. \tag{5.11}$$

We note that as $N \to \infty$ we have $1 < \mathrm{e}^{2\delta} < \mathrm{e}^{4N^{-1+2\epsilon}} \to 1$. Hence, for N large enough we have

$$-\lambda + \Delta \mathrm{e}^{2\delta} < -\lambda + 2\Delta < -\frac{1}{4} N^{2\epsilon} + 2N^{3-(1-\epsilon)(3+\frac{3}{2}q)} = N^{2\epsilon}\left(-\frac{1}{4} + 2N^{\epsilon - \frac{3}{2}q(1-\epsilon)}\right) \to -\infty \tag{5.12}$$

because when $q > \epsilon$ and $\frac{3(1-\epsilon)}{2} > 1$ we have $N^{\epsilon - \frac{3}{2}q(1-\epsilon)} \to 0$. By Equation (5.8) we can conclude $\mathbb{P}(A \in \mathcal{A}) \to 0$.

We see that for smaller qs such a Suen's inequality type analysis cannot work because the number of colliding triples (Δ), starts dominating the number of colliding pairs (λ). □

Proof of Theorem 5.2.1 Part (ii), Case 2: $q \leq \epsilon$. We would like to upper bound the probability that all columns are distinct by the probability that at most one column is identically 0. If we denote the number of identically 0 columns by Z, we want to show that $P(Z < 2) \to 0$. We start by proving $\mathbb{E}[Z] \to \infty$.

One can check that for $0 \leq q \leq \frac{1}{4}$ (which we may assume as $q \leq \epsilon \to 0$),

$$\frac{\log(1-q)}{\log\left(\sqrt{q^2+(1-q)^2}\right)} \leq 1 + \frac{9}{10}q. \tag{5.13}$$

Then, first using equation (5.2) with $M = (1-\epsilon)\hat{M}$ and $\alpha = 1-q$, and next applying inequality (5.13) and the assumptions $q \leq \epsilon$ and $\epsilon \gg \frac{1}{\log(N)}$, we have

$$\begin{aligned}\mathbb{E}[Z] &= N(1-q)^M = N^{1+(1-\epsilon)\frac{\log(1-q)}{\log\left(1/\sqrt{q^2+(1-q)^2}\right)}} \geq N^{1-(1-\epsilon)(1+\frac{9}{10}q)} \\ &= N^{\epsilon - \frac{9}{10}q(1-\epsilon)} \geq N^{\epsilon-\frac{9}{10}\epsilon} \to \infty \end{aligned} \tag{5.14}$$

We are going to use a standard concentration bound to finish this proof.

Lemma 5.2.1 (Chernoff bound)**.** *Let X be a binomial random variable. Then, for* $0 < \tau < 1$ *we have*

$$P(|X - \mathbb{E}[X]| \geq \tau E[X]) \leq 2\mathrm{e}^{-\frac{\tau^2 \mathbb{E}[X]}{3}}.$$

By Lemma 5.2.1, and since for N large enough we have $\mathbb{E}[Z] > 2$,

$$\begin{aligned}\mathbb{P}(A \in \mathcal{A}) &\leq \mathbb{P}(Z < 2) = \mathbb{P}\left(Z - \mathbb{E}[Z] < 2 - \mathbb{E}[Z]\right) \\ &\leq \mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq \left(1 - \frac{2}{\mathbb{E}[Z]}\right)\mathbb{E}[Z]\right) \\ &\leq 2\mathrm{e}^{-\frac{1}{3}\left(1-\frac{2}{\mathbb{E}[Z]}\right)^2\mathbb{E}[Z]} \leq 2\mathrm{e}^{-\frac{1}{3}\left(1-\frac{4}{\mathbb{E}[Z]}\right)\mathbb{E}[Z]} = 2\mathrm{e}^{\frac{4}{3}}\mathrm{e}^{-\frac{E[Z]}{3}} \to 0 \end{aligned} \tag{5.15}$$

□

5.3 Warmup2: Identifying Codes or Binary Search with Randomly Restricted Queries

In the previous section, we treated binary search with completely random entries. In this section, the queries will be selected by us, but we may only choose from a random subset of all queries (of size N), and we will only need to distinguish the columns for which there is no row selected with the same index. We start by adapting the definitions in Section 3.3 to the matrix setup used in Section 5.2.

Definition 5.3.1 (QC)**.** *Let $A \in \{0,1\}^{N\times N}$ be a binary matrix. Let the query complexity (*QC*) of A be the minimum $|R|$ such that $A_{R,\mathcal{C}\setminus R}$ has pairwise different columns.*

If the submatrix $A_{R,\mathcal{C}\setminus R}$ has pairwise different columns, the set R is also called an *identifying code* [137] of the graph which has adjacency matrix A (we must also allow self-edges to have equivalence, which is usually not part of the definition). Identifying codes are are closely related to resolving sets. If the graph has diameter two, the only difference between the two notions is that in the case of resolving sets we may receive three kinds of measurements (0, 1 and 2), not just two. However, we receive the 0 measurement only if we accidentally query the target, which can be ignored in many cases, hence the information we get is essentially binary for resolving sets as well.

Identifying codes of Erdős-Rényi graphs have been studied by [95]. In fact, [95] already featured some of the ideas that lead to characterizing the MD of Erdős-Rényi graphs in [36]. Part of the main theorem in this section (on the QC of Bernoulli random matrices, Theorem 5.3.1) has also appeared in [95] for a limited range of parameters, which we extend using the tools of [36]. Our proof of the QC of Bernoulli random matrices does not feature any new ideas, we only include it for the sake of completeness. We also define the adaptive version of the problem, the *sequential query complexity* (SQC), which will be similar to the SMD. The upper bound on the SQC will be algorithmic and will be quite different from the tools in [95] and [36].

Definition 5.3.2 (candidate targets)**.** *Given a set of queries R and target $v^\star$, the set of candidate targets for the matrix A is*

$$\mathcal{T}_R(A) = \begin{cases} \{v^\star\} & \text{if} v^\star \in R \\ \{v \in \mathcal{C}\setminus R \mid A_{R,\{v\}} = A_{R,\{v^\star\}}\} & \text{if} v^\star \not\in R. \end{cases}$$

Definition 5.3.3 (SQC)**.** *Let* $\mathrm{ALG}(G)$ *be the set of functions*

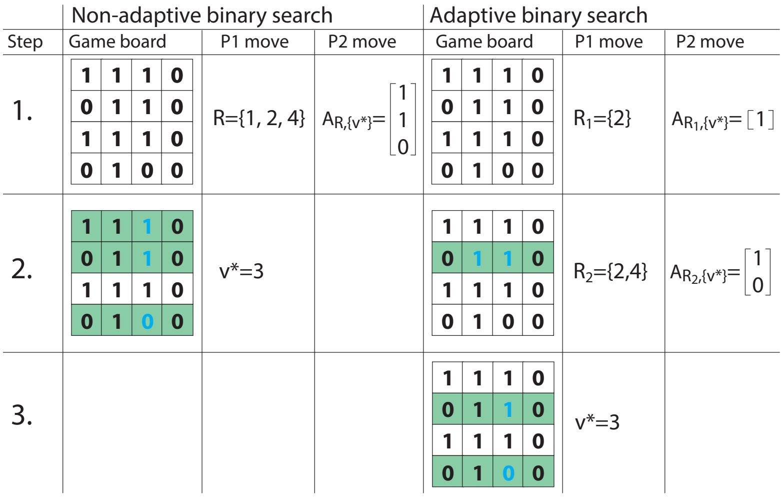
$$g : \{(A, R, A_{R,\{v^\star\}}) \mid R \subseteq \mathcal{R}, v^\star \in \mathcal{C}\} \to \mathcal{R}.$$

The sequential query complexity $\mathrm{SQC}(G)$ *is the minimum $r \in \mathbb{N}$ such that there is a query selection algorithm $g \in \mathrm{ALG}(G)$, for which if we let $R_0 = \emptyset$ and $R_{j+1} = R_j \cup g(G, R_j, A_{R_{j-1},\{v^\star\}}))$, then $v^\star \in R_r(A)$ for any $v^\star \in \mathcal{C}$.*

We show the difference between the QC and SQC on an example in Figure 5.1. The advantage studying the QC and SQC before the MD and SMD is that we have simpler results without the small dependencies that always arise in the graph setting.

Theorem 5.3.1. *Let $N \in \mathbb{N}$, let $0 < q < 1$, let $A \in \mathrm{Ber}(q)^{N\times N}$ and $\gamma_{sqc} = \max(q, 1-q)$. Then a.a.s,*

$$\mathrm{SQC}(A) = \min\left(N, (1+o(1))\frac{\log(N(1-\gamma_{sqc}))}{\log(1/\gamma_{sqc})}\right). \tag{5.16}$$



	Non-adaptive binary search			Adaptive binary search		
Step	Game board	P1 move	P2 move	Game board	P1 move	P2 move
1.	$\begin{matrix}1&1&1&0\\0&1&1&0\\1&1&1&0\\0&1&0&0\end{matrix}$	R={1, 2, 4}	$A_{R,\{v^*\}}=\begin{bmatrix}1\\1\\0\end{bmatrix}$	$\begin{matrix}1&1&1&0\\0&1&1&0\\1&1&1&0\\0&1&0&0\end{matrix}$	R_1={2}	$A_{R_1,\{v^*\}}=\begin{bmatrix}1\end{bmatrix}$
2.	$\begin{matrix}1&1&1&0\\0&1&1&0\\1&1&1&0\\0&1&0&0\end{matrix}$	v*=3		$\begin{matrix}1&1&1&0\\0&1&1&0\\1&1&1&0\\0&1&0&0\end{matrix}$	R_2={2,4}	$A_{R_2,\{v^*\}}=\begin{bmatrix}1\\0\end{bmatrix}$
3.				$\begin{matrix}1&1&1&0\\0&1&1&0\\1&1&1&0\\0&1&0&0\end{matrix}$	v*=3	

Figure 5.1: The process of non-adaptive and adaptive binary search with restricted queries with target $v^\star = 3$. The queries are marked with green and the observations are marked with blue.

The results for $\mathrm{QC}(A)$ *are of the same form, except that instead of* γ_{sqc} *we have* $\gamma_{qc} = \sqrt{q^2 + (1-q)^2}$.

Remark 5.3.1. *If* $1-\gamma_{sqc} = o(1)$ *(which is equivalent to* $1-\gamma_{qc} = o(1)$*), we have*

$$\frac{1}{\log(1/\gamma_{qc})} = (1+o(1))\frac{1}{\log(1/\gamma_{sqc})} = (1+o(1))\frac{1}{1-\gamma_{sqc}}, \tag{5.17}$$

In this case $\mathrm{SQC}(A) = (1+o(1))\mathrm{QC}(A) = \omega(\log(N))$, *so the SQC and the QC have the same asymptotic behavior.*

If $1-\gamma_{sqc} = \Theta(1)$, then the QC and the SQC are both logarithmic, and the SQC is a constant factor smaller than the QC.

5.3.1 Connection between Theorems 5.2.1 and 5.3.1

The main difference between the two theorems is that in Theorem 5.2.1 we sample M queries and use all of them, whereas in Theorem 5.3.1 we sample N queries and select (adaptively or non-adaptively) only a subset of them. Also, in Theorem 5.3.1 we only need to distinguish the columns $\mathcal{C} \setminus R$, for which there is no row selected with the same index. The subset we select is of size $\mathrm{SQC}(A)$ or $\mathrm{QC}(A)$.

Remark 5.3.2. *If* $\frac{\log(N)}{N} \gg 1-\gamma_{sqc}$, *then the first term (N) in the minimum in equation* (5.16) *is the dominant one, and* $\mathrm{SQC} = \Theta(N)$. *In this case, Theorem 5.2.1 implies that there are two identical columns in matrix A with high probability. Keep in mind that since in the QC and SQC we only need to distinguish between the columns of* $A_{R,\mathcal{C}\setminus R}$, *and therefore the condition becomes trivial for* $R = [N]$, *we must always have* $\mathrm{SQC} \le \mathrm{QC} \le N$.

Let us also give intuition about the new notation. On the range $q \in (0, \frac{1}{2}]$, the variable $1 - \gamma_{sqc}$ is just q, and it serves essentially the same purpose. The reason for introducing a new variable is that we can highlight the symmetry of the adaptive and non-adaptive case here and later in the text.

In the remainder of this section we sketch the proof of Theorem 5.3.1.

5.3.2 Proof of the QC Upper Bound of Theorem 5.3.1

Proof. For $1 - \gamma_{qc} = \Theta(1)$, a simple application of Theorem 5.2.1 (i) with $\epsilon = \log\log(N)/\log(N)$ shows that $A_{R_1, \mathcal{C} \setminus R_1}$ will have pairwise different columns with high probability, therefore step 2 is not even needed in this case. In the rest of the proof, we focus on the $1 - \gamma_{qc} \to 0$ case. For simplicity, let us assume that $q \le 1/2$, which then implies $q \to 0$.

For this upper bound, we provide an algorithm, which will select the row indices in two steps:

1. With $\epsilon_{R_1} \to 0$ slowly (to be specified at the end of the proof), we select
$$r = \min\left(N, (1 + \epsilon_{R_1})\frac{\log(Nq)}{\log(1/(1-q))}\right) \tag{5.18}$$
row indices uniformly at random to create set R_1.

2. Add all rows with index i to R_2 for which column i of $A_{R_1, \mathcal{C} \setminus R_1}$ is not unique, i.e., for which we have $A_{k,i} = A_{k,j}$ for some column index j and all $k \in R_1$.

Clearly, the set $A_{R_1 \cup R_2, \mathcal{C} \setminus (R_1 \cup R_2)}$ has pairwise different columns, because if any two columns i, j were identical in $A_{R_1, \mathcal{C} \setminus R_1}$, then rows i and j are added in step 2 to R_2, and therefore the column indices i and j are not in the set $\mathcal{C} \setminus (R_1 \cup R_2)$.

Since
$$\min\left(N, (1 + \epsilon_{R_1})\frac{\log(Nq)}{\log(1/(1-q))}\right) = \min\left(N, (1 + o(1))\frac{\log(N(1-\gamma_{qc}))}{\log(1/\gamma_{qc})}\right)$$
nodes are selected in step 1, to complete the proof we must show that we added $o(r)$ rows in step 2. In other words, if we let Y to be the number of non-unique columns after step 1 of the algorithm, it would be enough to show that $P(Y > r/\log(N)) = o(1)$.

Let Z be the number of pairs of non-unique columns after step 1. Clearly $Y \le \frac{Z}{2}$. It turns out that this upper bound is too loose in our case, because there are large subsets of columns which are all indistinguishable from each other, and the number of indistinguishable pairs in such a subset is quadratic in the number of columns. Since $q = o(1)$, the best candidate to find such a subset is among the columns that only contain entries with value 1. Let Y_0 be the number of non-unique columns after step 1 with all-1 entries in the rows indexed by R_1. Then, if we count the all-1 non-unique columns and the other non-unique columns separately, we have the upper bound $Y \le Y_0 + \frac{Z - \frac{Y_0(Y_0-1)}{2}}{2}$, since all $\frac{Y_0(Y_0-1)}{2}$ non-unique column pairs are counted into Z.

With our upper bound on Y using Y_0 and Z, our goal is to upper bound $\mathbb{E}[Y]$. By the linearity of expectation,

$$\mathbb{E}[Y] \le \mathbb{E}[Y_0] + \frac{\mathbb{E}[Z] - \frac{\mathbb{E}[Y_0(Y_0-1)]}{2}}{2}. \tag{5.19}$$

Since the probability of a column (with an index that is not in R_1) having no zeros is $(1-q)^r$ in the rows indexed by R_1, we have

$$\mathbb{E}[Y_0] = (N-r)(1-q)^r. \tag{5.20}$$

Moreover, if $Y_{0,j}$ is the indicator of column j being all-1, then

$$\mathbb{E}[Y_0(Y_0-1)] = \sum_{j=1}^{N-r} \mathbb{E}[Y_{0,j}^2] + \sum_{j\neq k=1}^{N-r} \mathbb{E}[Y_{0,j}Y_{0,k}] - \mathbb{E}[Y_0] = (N-r)(N-r-1)(1-q)^{2r}. \tag{5.21}$$

Finally, we can compute the expectation of the number non-unique column pairs as

$$\mathbb{E}[Z] = \frac{(N-r)(N-r-1)}{2}(q^2+(1-q)^2)^r. \tag{5.22}$$

Substituting equations (5.20), (5.21) and (5.22) into equation (5.19) yields

$$\begin{aligned}\mathbb{E}[Y] &\le (N-r)(1-q)^r + \frac{(N-r)(N-r-1)}{2}((q^2+(1-q^2))^r - ((1-q)^2)^r) \\ &< N(1-q)^r + N^2((q^2+(1-q^2))^r - ((1-q)^2)^r)\end{aligned} \tag{5.23}$$

Note that since x^r is a convex function in x for $r > 1$,

$$(q^2+(1-q)^2)^r - ((1-q)^2)^r \le rq^2(q^2+(1-q)^2)^{r-1}. \tag{5.24}$$

Also note that our choice of r satisfies

$$(1-q)^r = \frac{1}{(Nq)^{1+\epsilon_{R_1}}}, \tag{5.25}$$

and we also have

$$(q^2+(1-q)^2)^{r-1} = \frac{(1+o(1))}{(Nq)^{2(1+\epsilon_{R_1})}}. \tag{5.26}$$

Then, if we substitute back into (5.23), we get

$$\mathbb{E}[Y] < \frac{N}{(Nq)^{1+\epsilon_{R_1}}} + \frac{(1+o(1))N^2q^2r}{(Nq)^{2(1+\epsilon_{R_1})}} = O\left(\frac{r}{(Nq)^{2\epsilon_{R_1}}}\right). \tag{5.27}$$

Finally, choosing $\epsilon_{R_1} = \frac{\log\log(N)}{\log(Nq)}$, by Markov's inequality we have

$$\mathbb{P}(Y > r\log^{-1}(N)) < \frac{\mathbb{E}[Y]}{r\log^{-1}(N)} = O(\log^{-1}(N)) = o(1), \tag{5.28}$$

and the proof is completed.

□

5.3.3 Proof of the QC Lower Bound of Theorem 5.3.1

Proof. Here we only consider the $1-\gamma_{qc} = \Theta(1)$ case. By Remark 5.3.1, the $1-\gamma_{qc} = o(1)$ case will follow from the SQC lower bound. Let $r \le (1-\epsilon)\frac{\log(N)}{\log(1/\gamma_{qc})}$ and $\epsilon \gg \frac{\log\log(N)}{\log(N)}$ with $\epsilon \to 0$. Let Y be the number of subsets $W \subset \mathcal{R}$ with $|W| \le r$ for which $A_{W,\mathcal{C}\setminus W}$ has no repeated columns. For the lower bound to hold we must show that $Y = 0$ a.a.s.

Let us now select a set $R \subset \mathcal{R}$ of r rows in advance, and when A is revealed, let $\mathcal{A}_R$ be the event that $A_{R,\mathcal{C}\setminus R}$ has no repeated columns. Then

$$\mathbb{P}(Y > 0) \le \mathbb{E}[Y] \le N^r\mathbb{P}(A \in \mathcal{A}_R). \tag{5.29}$$

Using our result in equations (5.8) and (5.12) and because $1-\gamma_{qc} = \Theta(1)$ implies $\epsilon \ll q(1-\epsilon)$, for N large enough

$$\mathbb{P}(A \in \mathcal{A}_R) \le \exp(-\lambda + \Delta e^{2\delta}) < \exp\left(N^{2\epsilon}\left(-\frac{1}{4} + 2N^{\epsilon-\frac{3}{2}q(1-\epsilon)}\right)\right) < \exp\left(-\frac{1}{8}N^{2\epsilon}\right). \tag{5.30}$$

Then,

$$\mathbb{E}[Y] \le N^r \exp\left(-\frac{1}{8}N^{2\epsilon}\right) \le \exp\left((1-\epsilon)\frac{\log^2(N)}{\log(1/\gamma_{qc})} - \frac{1}{8}N^{2\epsilon}\right) \to 0 \tag{5.31}$$

since by assumption $\frac{1-\epsilon}{\log(1/\gamma_{qc})} = \Theta(1)$ and since $\epsilon \gg \frac{\log\log(N)}{\log(N)}$ implies $\log^2(N) \ll N^{2\epsilon}$. Finally, by equations (5.29) and (5.31), we have $Y = 0$ a.a.s. □

5.3.4 Proof of the SQC Upper Bound of Theorem 5.3.1

For $1-\gamma_{sqc} = o(1)$, the upper bound follows from the QC upper bound from Section 5.3.2, and we concentrate only on the $1-\gamma_{sqc} = \Theta(1)$ case in the current section. In order to prove this upper bound, we analyse the performance of a greedy query selection algorithm called MAX-GAIN.

Definition 5.3.4 (k-reducer). *For a query $w \in \mathcal{R}$ and an observation $l \in \{0,1\}$, let the targets agreeing with the pair (w,l) be denoted as*

$$\mathcal{S}_A(w,l) = \{v \in \mathcal{C}\setminus R \mid A_{w,v} = l\}. \tag{5.32}$$

Given an integer k and the triple $(A, R_j, A_{R_j,\{v^\star\}})$, a row w is called a k-reducer if after adding w to R_j, the worst case cardinality of R_{j+1} is upper bounded by k, that is

$$\max_{l\in\{0,1\}} |\mathcal{T}_{R_j} \cap \mathcal{S}_A(w,l)| \le k. \tag{5.33}$$

Definition 5.3.5 (MAX-GAIN)**.** *The MAX-GAIN algorithm finds the target by iteratively selecting as a query the k-reducer with the smallest k. That is,*

$$\text{MAXGAIN}(A, R_j, A_{R_j,\{v^\star\}}) = \operatorname*{argmin}_{w\in V\setminus R_j} \max_{l\in\{0,1\}} |\mathcal{T}_{R_j} \cap \mathcal{S}_A(w,l)|.$$

Note that the MAX-GAIN algorithm always finds the target in at most N steps by selecting all rows. Moreover, if we can always find better reducers, the number of steps decreases dramatically. Since each node is connected to a $\gamma_{sqc} > 1/2$ fraction of the nodes, it is reasonable to expect that we can find k-reducers with $k \approx |\mathcal{T}_{R_j}|\gamma_{sqc}$. The existence of such reducers would already imply the result we need.

Lemma 5.3.1. *If MAX-GAIN can select a $(|\mathcal{T}_{R_j}|\gamma_{sqc} + f(|\mathcal{T}_{R_j}|))$-reducer in the $(j+1)^{th}$ step of the algorithm with $f(n) = o\left(\frac{n}{\log(n)}\right)$ for any $j \in \mathbb{N}$ for which the candidate set size is $|\mathcal{T}_{R_j}| = \Omega\left(\frac{\log(N)}{\log(1/\gamma_{sqc})}\right)$, then the algorithm finds the source in $(1+o(1))\frac{\log(N)}{\log(1/\gamma_{sqc})}$ steps.*

The proof of Lemma 5.3.1 is included in Section B.1.1 of the appendix. For the SQC upper bound we will be able to prove the existence of a $(|\mathcal{T}_{R_j}|\gamma_{sqc}+f(|\mathcal{T}_{R_j}|))$-reducer for any candidate size, hence this condition of Lemma 5.3.1 may be ignored for the moment. The condition on the minimum candidate set size for which there exists a $(|\mathcal{T}_{R_j}|\gamma_{sqc} + f(|\mathcal{T}_{R_j}|))$-reducer will be important in Section 5.5.3.

Now we need to show the existence of such reducers. This will be a structural result on the matrix A which holds independently of the state of our algorithm, so we find it useful to define another notion quite similar to k-reducers.

Definition 5.3.6 (f-separator)**.** *Let $f(n) \in \mathbb{N} \to \mathbb{R}^+$ be a function, and γ_{sqc} as defined in Theorem 5.3.1. A set of columns $W \subseteq \mathcal{C}$, $|W| = n$ has an f-separator if there is a row $w \in \mathcal{R}$ such that*

$$\max_{l\in\{0,1\}} |W \cap \mathcal{S}_A(w,l)| \le n\gamma_{sqc} + f(n). \tag{5.34}$$

Remark 5.3.3. *An f-separator w for $\mathcal{T}_{R_j}$ is an $(|\mathcal{T}_{R_j}|\gamma_{sqc} + f(|\mathcal{T}_{R_j}|))$-reducer for the triple $(A, R_j, A_{R_j,\{v^\star\}})$. The difference between the two terms is that the term f-separator refers to a property of A and W, whereas the term k-reducer refers to a property of the state of an algorithm. The role of these two terms is also quite different. A k-reducer with a small k makes MAX-GAIN more efficient, whereas an f-separator with a small f is a typical separator, and its existence makes the analysis of this upper bound easier.*

Proof of the SQC *upper bound of Theorem 5.3.1.* To use Lemma 5.3.1, we have to show that for every $W \subseteq \mathcal{C}$ we have an f-separator with $f(n) = o\left(\frac{n}{\log(n)}\right)$. Let $X_w = |W \cap \mathcal{S}_A(w,1)|$ and note that $\mathbb{E}[X_w] = nq$. It is clear that if X_w is close to its expected value then v must be an f separator. Indeed, $|X_w - nq| \le f(n)$ implies

$$X_w - nq \le f(n) \Rightarrow X_w \le nq + f(n) \le n\gamma_{sqc} + f(n) \tag{5.35}$$

and

$$-(X_w - nq) - n + n \le f(n) \Rightarrow n - X_w \le n(1-q) + f(n) \le n\gamma_{sqc} + f(n), \tag{5.36}$$

hence w is an f separator. We first show that for any $v \in \mathcal{R}$ we have $|X_w - nq| \le f(n)$ with constant probability. Using Lemma 5.2.1 and substituting $f(n) = \sqrt{3n}$, we get

$$\mathbb{P}(|X_w - nq| \ge f(n)) = \mathbb{P}\left(|X_w - \mathbb{E}[X_w]| \ge \frac{f(n)}{nq} nq\right) \le 2\mathrm{e}^{\frac{-2nq\frac{f(n)^2}{n^2q^2}}{3}} = 2\mathrm{e}^{-\frac{6}{3q}} < \mathrm{e}^{-1}, \tag{5.37}$$

because $q \le 1$. Since the random variables X_w are mutually independent, the probability that none of the rows are f-separators for W is upper bounded by e^{-N}. Let Y be the number of subsets W that do not have a $\sqrt{3n}$-separator. Then,

$$\mathbb{E}[Y] < \sum_{W \subseteq \mathcal{C}} \mathrm{e}^{-N} \le 2^N \mathrm{e}^{-N} \to 0.$$

By the first moment method we can conclude that every $W \subseteq \mathcal{C}$ has a $\sqrt{3n}$-separator a.a.s. By Lemma 5.3.1 this concludes the proof. □

5.3.5 Proof of the SQC Lower Bound of Theorem 5.3.1

For this lower bound, we focus on the regime $\frac{\log(N)}{N} \ll 1 - \gamma_{sqc}$, and we look for columns with identical elements similarly to Case 2 of the Proof of Theorem 5.2.1, part (ii). Recall, that at the end of Section 3.3 we modelled the SMD as the number of steps in a two-player game, if both players play optimally. In this proof, we are essentially analyzing a strategy for Player 2, who does not decide the target in advance, and always provides observations 0 if $q \le \frac{1}{2}$ and 1 if $q > \frac{1}{2}$. By showing that with high probability any $r \times N$ submatrix of A has at least two columns with identically 0 or 1 elements, we assure that Player 2 can follow this simple strategy, and the size of the candidate set will be at least two after r queries, independently of the strategy of Player 1. We also note that deciding whether there exists a set of rows R so that each column of $\mathcal{A}_{R,\mathcal{C}\setminus R}$ has at least one zero element is equivalent to considering a bipartite meta-graph, where the two independent sets are the defined by the row and column indices of $\mathcal{A}$, and the edges are determined by the zero values of the matrix $\mathcal{A}$, and in this meta-graph our goal is to find a dominating set (vertex cover) among the row nodes, which dominate (cover) of the column nodes. Reassuringly, the domination number of a $\mathcal{G}(N,q)$ random graph is found to be $(1+o(1))\log(Nq)/q$ a.a.s. by [103], which agrees with Theorem 5.3.1 for $q \to 0$.

Proof of the SQC *lower bound of Theorem 5.3.1.* Similarly to Section 5.3.3, let Y be the number of subsets $W \subset \mathcal{R}$ with $|W| \leq r = (1-\epsilon)\frac{\log(N(1-\gamma_{sqc}))}{\log(1/\gamma_{sqc})}$ for which $A_{W,\mathcal{C}\setminus W}$ has at most one column with identically 0 (if $q \leq \frac{1}{2}$) or 1 (if $q > \frac{1}{2}$) elements. For the lower bound to hold we must show $Y = 0$ a.a.s. Let us now select a $R \subset \mathcal{R}$ of size r in advance, and when A is revealed let $\mathcal{A}_R$ be the event that $A_{R,\mathcal{C}\setminus R}$ has at most one column with identical elements. Then,

$$\mathbb{P}(Y > 0) \leq E[Y] \leq N^r \mathbb{P}(A \in \mathcal{A}_R). \tag{5.38}$$

Let Z_R be the number of identically 0 (or 1 if $q > \frac{1}{2}$) columns in $A_{R,\mathcal{C}\setminus R}$. By equation (5.15),

$$\mathbb{P}(A \in \mathcal{A}_R) \leq 2\mathrm{e}^{\frac{4}{3}}\mathrm{e}^{-\frac{\mathbb{E}[Z_R]}{3}} \tag{5.39}$$

Then, using the equation $\mathbb{E}[Z_R] = (N-r)\gamma_{sqc}^r$ and the definition of r,

$$\begin{aligned}
\mathbb{E}[Y] &\leq N^r 2\mathrm{e}^{\frac{4}{3}}\mathrm{e}^{-\frac{\mathbb{E}[Z_R]}{3}} \\
&= 2\mathrm{e}^{\frac{4}{3}} \exp\left(r\log(N) - \frac{1}{3}(N-r)\gamma_{sqc}^r\right) \\
&\leq 2\mathrm{e}^{\frac{4}{3}} \exp\left(r\log(N) - \frac{1}{3}(N-r)\gamma_{sqc}^{(1-\epsilon)\frac{\log(N(1-\gamma_{sqc}))}{\log(1/\gamma_{sqc})}}\right) \\
&= 2\mathrm{e}^{\frac{4}{3}} \exp\left(r(\log(N) + \frac{1}{3}N^{-(1-\epsilon)}) - \frac{1}{3}N(N(1-\gamma_{sqc}))^{\epsilon-1}\right) \\
&\leq 2\mathrm{e}^{\frac{4}{3}} \exp\left(\frac{\log(N)(\log(N)+1)}{\log(1/\gamma_{sqc})} - \frac{1}{3}\frac{N^\epsilon}{1-\gamma_{sqc}}\right) \\
&\leq 2\mathrm{e}^{\frac{4}{3}} \exp\left(\frac{\log^3(N) - \frac{1}{3}N^\epsilon}{\log(1/\gamma_{sqc})}\right) \to 0
\end{aligned} \tag{5.40}$$

since $\frac{1}{\log(1/\gamma_{sqc})} > 1$ and $\log^3(N) - \frac{1}{3}N^\epsilon \to -\infty$ as long as $\epsilon \gg \frac{\log\log(N)}{\log(N)}$. Finally, by equations (5.38) and (5.40) we have $Y = 0$ a.a.s. □

5.4 Expansion Properties of $\mathcal{G}(N, p)$

Before we proceed to our main results, we must establish some properties about the exponential growth of Erdős-Rényi graphs in the sizes of the level sets $\mathcal{S}_G(v, l)$ defined below. This exponential growth is depicted on Figure 5.2. Most statements in this section already appeared in [36] with a different notation, or can be easily derived from their results.

Definition 5.4.1 (level sets)**.** *For a graph $G = (V, E)$ and a node $v \in V$, let the level set of v be defined as $\mathcal{S}_G(v, l) = \{w \in V \mid d(v, w) = l\}$ for every $l \in \{0, \ldots, |V|\}$. The level sets from a set of nodes $V' \subseteq V$ is defined as $\mathcal{S}_G(V', l) = \bigcup_{v \in V'} \mathcal{S}_G(v, l)$.*

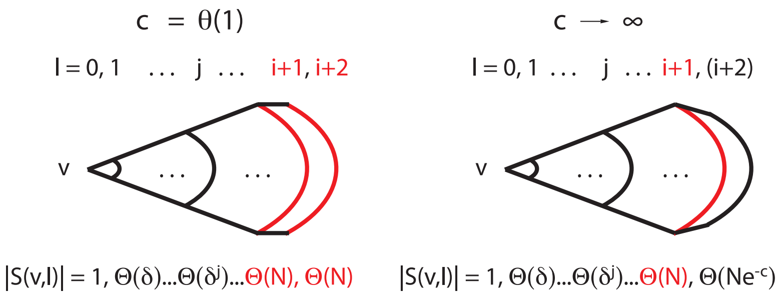

c = θ(1)
l = 0, 1 . . . j . . . i+1, i+2
v
. . .
. . .
|S(v,l)| = 1, Θ(δ)...Θ(δʲ)...Θ(N), Θ(N)
c → ∞
l = 0, 1 . . . j . . . i+1, (i+2)
v
. . .
. . .
|S(v,l)| = 1, Θ(δ)...Θ(δʲ)...Θ(N), Θ(Ne⁻ᶜ)

Figure 5.2: The arcs in the figures represent the level sets $\mathcal{S}_G(v,l)$ of $\mathcal{G}(N,p)$ for different ranges of c. The layers containing a constant fraction of the nodes marked red. In the $c = \Theta(1)$ case there are two such layers, whereas in the $c \to \infty$ there is only one. In this latter case the $(i+2)^{th}$ layer is in parenthesis because that layer may or may not exist depending on c.

$c(N)$	$c = \Theta(1)$		$1 \ll c$ $c \ll \log(\frac{N}{\delta^i})$	$\log(\frac{N}{\delta^i}) \ll c$ $c \ll 2\log(N)$	$2\log(N) \ll c$ $c \ll \delta$
D	$i+2$	$i+2$	$i+2$	$i+2$	$i+1$
$\lvert\mathcal{S}_G(v,i-1)\rvert$	δ^{i-1}	δ^{i-1}	δ^{i-1}	δ^{i-1}	δ^{i-1}
$\lvert\mathcal{S}_G(v,i)\rvert$	δ^i	δ^i	δ^i	δ^i	δ^i
$\lvert\mathcal{S}_G(v,i+1)\rvert$	$(1-\mathrm{e}^{-c})N$	$(1-\mathrm{e}^{-c})N$	$(1-\mathrm{e}^{-c})N$	$\left(1-\frac{\delta^i}{N}\right)N$	$\left(1-\frac{\delta^i}{N}\right)N$
$\lvert\mathcal{S}_G(v,i+2)\rvert$	$\mathrm{e}^{-c}N$	$\mathrm{e}^{-c}N$	$\mathrm{e}^{-c}N$	$\mathrm{e}^{-c}N$	0
MD ub	T3.1 case 1 [36]		(5.5.2), similar to the QC ub		
MD lb	T4.2 [36]		use SMD lb		
SMD ub	(5.5.3, 5.5.5), similar to the SQC ub		use MD ub		
SMD lb	(5.5.4), coupling and an analysis similar to the DQC lb				

Table 5.1: Overview of the main tools to prove Theorem 5.5.1. Each column corresponds to a different range of parameter c. The $c = \Theta(1)$ columns are split in two sub-columns: in the first, $\mathrm{e}^{-c} > 1-\mathrm{e}^{-c}$ and in the second, $\mathrm{e}^{-c} < 1-\mathrm{e}^{-c}$. Only the leading terms of the size of the level sets S are shown. The largest level set is colored in red, and the second largest is colored in pink. The last level set before one of the two dominating level sets is colored in grey. The bottom half of the table points to the proof of the upper/lower bound for each parameter range of Theorem 5.5.1, both in previous work and in this chapter.

We also define three functions δ, i and c (all depending on the parameters of an Erdős-Rényi graph $\mathcal{G}(N,p)$), which will be useful throughout the rest of this chapter.

Definition 5.4.2 (parameters δ, i and c of the expansion properties)**.** *Let $\delta = Np$, let $i \geq 0$ be the largest integer such that $\delta^i = o(N)$, and finally let $c = \frac{\delta^{i+1}}{N} = \delta^i p$.*

In this chapter we only consider connected Erdős-Rényi graphs with $\delta = Np \gg \log(N)$. We defer the interpretation of these definitions and introduce the main technical lemma that establishes the exponential growth of the level sets. This lemma also appeared in Lemma 2.1 of [36] (we replaced their $O(1/\sqrt{\omega}) + O(d^i/n)$ term with $O(\zeta)$ for simplicity), with an extra condition which we removed (proof in the Appendix).

Lemma 5.4.1 (Expansion property)**.** *With parameters i, c and $\delta \gg \log(N)$ as defined in Definition 5.4.2, let $\zeta = \zeta(N)$ be a function tending to slowly to zero with N such that*

$$\zeta \geq \max\left(\sqrt{\frac{\log(N)}{\delta}}, \frac{\delta^i}{N}\right). \tag{5.41}$$

For a node $v \in V$, let $\mathcal{E}(v, j)$ be the event that for every $l \leq j$

$$|\mathcal{S}_G(v,l)| = (1+O(\zeta))\,\delta^l, \tag{5.42}$$

and for two nodes $v \neq u \in V$, let $\mathcal{E}_2(u, v, j)$ be the event that for every $l \leq j$

$$|\mathcal{S}_G(\{u,v\},l)| = 2\,(1+O(\zeta))\,\delta^l. \tag{5.43}$$

For a subset $V' \subset V$ let $\mathcal{E}(V', j) = \bigcap_{v\in V'} \mathcal{E}(v, j)$ be the event that expansion properties hold for all nodes in V', and let $\mathcal{E}(j)$ be a shorthand for $\mathcal{E}(V, j)$. Similarly, let $\mathcal{E}_2(j) = \bigcap_{u\neq v\in V} \mathcal{E}_2(u, v, j)$. Then, for G sampled from $\mathcal{G}(N, p)$ the event $\mathcal{E}(i) \cap \mathcal{E}_2(i)$ holds a.a.s.

Corollary 5.4.1. *For every $v \in V$ we have $\sum_{l=1}^{i} |\mathcal{S}_G(v,l)| = (1+O(\zeta))\delta^i = o(N)$ a.a.s.*

Proof. By Lemma 5.4.1 by taking $l \leq i$ since $\zeta \gg 1/\delta$ and

$$\sum_{l=1}^{i} |\mathcal{S}_G(v,l)| = \sum_{l=1}^{i} (1+O(\zeta))\,\delta^l = (1+O(\zeta))\delta^i = o(N). \tag{5.44}$$

□

Parameter δ is essentially the expected degree of each node in $G \sim \mathcal{G}(N, p)$. We require $\delta \geq \log(N)/\zeta^2 \gg \log(N)$, so the graph is a.a.s. connected. The function ζ serves as the error term. Note that equation (5.41) implies that $\zeta \geq p$ for $i > 0$. Parameters i and c are both derived from $1/\log_N(\delta)$; parameter c is δ raised to one minus the fractional part of δ, and parameter i is the integer part of $1/\log_N(\delta)$, or more precisely the ceiling minus one. Qualitatively, the level set structure of $\mathcal{G}(N, p)$ has a periodic behavior as we tune p. As p decreases, in each such "period" the outmost layer gains more and more nodes until it is fully saturated and another layer appears. Roughly speaking, parameter i indicates the "period", and parameter c provides a fine-grain tuning of p within a "period". However, the "periods" and the appearance of new level sets are not exactly aligned. The next lemma tells us about how the diameter depends on δ and N.

Lemma 5.4.2 (Lemma 4.1 [36])**.** *Suppose that $\delta = pN \gg \log(N)$, and that for some integer $D \geq 1$*

$$\frac{\delta^{D-1}}{N} - 2\log(N) \to -\infty \qquad \textit{and} \qquad \frac{\delta^{D}}{N} - 2\log(N) \to \infty \tag{5.45}$$

Then the diameter of G sampled from $\mathcal{G}(N,p)$ is equal to D a.a.s.

Corollary 5.4.2. *Let $\mathcal{D}$ be the event that the diameter of G sampled from $\mathcal{G}(N,p)$ is either $i+1$ or $i+2$ with all parameters, including i, given by Definition 5.4.2, and as always $\delta \gg \log(N)$. Then $\mathcal{D}$ holds a.a.s.*

Proof. We distinguish three cases:

1. If $\delta^i/N - 2\log(N) \to -\infty$ and $\delta^{i+1}/N - 2\log(N) \to \infty$, then taking $D = i+1$ in Lemma 5.4.2 implies that the diameter is $i+1$ a.a.s.
2. If $\delta^{i+1}/N - 2\log(N) \to -\infty$ and $\delta^{i+2}/N - 2\log(N) \to \infty$, then taking $D = i+2$ in Lemma 5.4.2 implies that the diameter is $i+2$ a.a.s.
3. If $\delta^{i+1}/N - 2\log(N) = \Theta(1)$, then let us consider $G_1 = \mathcal{G}(N, p\omega)$ and $G_2 = \mathcal{G}(N, p/\omega)$ with $\omega \to \infty$ very slowly. Using Lemma 5.4.2, for ω growing slowly enough, the graphs G_1 and G_2 have diameter $i+1$ and $i+2$ respectively a.a.s. Since having diameter at least D is a monotone graph property, G must also have diameter $i+1$ or $i+2$ a.a.s.

There are no other cases than the three outlined above because the equations

$$\frac{\delta^{i}}{N} - 2\log(N) \to -\infty \tag{5.46}$$

and

$$\frac{\delta^{i+2}}{N} - 2\log(N) = c\delta - 2\log(N) \geq \left(\frac{c}{\zeta^2} - 2\right)\log(N) \to \infty. \tag{5.47}$$

must always hold by Definition 5.4.2 and the assumption $\delta \gg \log(N)$. □

The previous results shows that most of the nodes are either distance $i+1$ or $i+2$ away from any arbitrary node $v \in V$. We now extend Lemma 5.4.1 to these level sets too.

Lemma 5.4.3. *For every $v \in V$, let $\mathcal{E}(v, i+1)$ be the intersection of $\mathcal{E}(v,i)$ and of the event*

$$|\mathcal{S}_G(v,i+1)| = \begin{cases} \left(1 + O\left(\sqrt{\frac{\log(N)}{N}}\right)\right) Np & \textit{if } i = 0 \\ \left(1 - \left(\mathrm{e}^{-c} + \frac{\delta^i}{N}\right) + O\left(\zeta\left(\mathrm{e}^{-c} + \frac{\delta^i}{N}\right) + \sqrt{\frac{\log^2(N)}{N}}\right)\right) N & \textit{if } i > 0. \end{cases} \tag{5.48}$$

For a subset $V' \subset V$ let $\mathcal{E}(V', i+1) = \bigcap_{v \in V'} \mathcal{E}(v, i+1)$ be the event that expansion properties hold for all nodes in V', and let $\mathcal{E}(i+1)$ be a shorthand for $\mathcal{E}(V, i+1)$. Then event $\mathcal{E}(i+1)$ holds a.a.s.

Proof. For a fixed node $v \in V$, let us expose all of its edges (i.e., sample the edges of $\mathcal{G}(N,p)$ adjacent to v in any order and reveal them), and do the same for each of its neighbors recursively until depth i. This way of exposing edges also exposes all nodes in $W = \bigcup_{l \le i} \mathcal{S}_G(v,l)$.

After exposing these edges, we have for both $i = 0$ and $i > 0$ that

$$|\mathcal{S}_G(v,i+1)| = \text{Binom}\left(|V \setminus W|, 1-(1-p)^{|\mathcal{S}_G(v,i)|}\right), \tag{5.49}$$

because $\mathcal{S}_G(v,i+1) = \{w \in V \setminus W \mid \exists v' \in \mathcal{S}_G(v,i) \text{ s.t. } d(v',w) = 1\}$.

(i) In the $i > 0$ case, let us condition on the event $\mathcal{E}(\{v\},i)$. By Corollary 5.4.1, the set $V \setminus W$ has $N-(1+O(\zeta))\delta^i$ nodes. Then, we have

$$\begin{aligned}
\mathbb{E}[|\mathcal{S}_G(v,i+1)| \mid \mathcal{E}(\{v\},i)] &= (N-(1+O(\zeta))\delta^i)(1-(1-p)^{|\mathcal{S}_G(v,i)|}) \\
&\overset{(5.42)}{=} (N-(1+O(\zeta))\delta^i)(1-\mathrm{e}^{-(p+O(p^2))\delta^i(1+O(\zeta))}) \\
&= (N-\delta^i)(1-\mathrm{e}^{-c}(1+O(\zeta))) + O(\delta^i\zeta) \\
&= N\left(\left(1-\frac{\delta^i}{N}\right)(1-\mathrm{e}^{-c}) + O\left(\zeta\mathrm{e}^{-c} + \frac{\delta^i}{N}\zeta\right)\right) \\
&= N\left(1-(1+O(\zeta))\left(\mathrm{e}^{-c}+\frac{\delta^i}{N}\right)\right)
\end{aligned} \tag{5.50}$$

Let us denote $\mu = \mathbb{E}[|\mathcal{S}_G(v,i+1)| \mid \mathcal{E}(\{v\},i)]$. Then, by Lemma 5.2.1 with $\tau = \sqrt{6\log(N)/\mu}$ we have

$$\mathbb{P}(\left||\mathcal{S}_G(v,i+1)|-\mu\right| > \tau\mu \mid \mathcal{E}(\{v\},i)]) < \mathrm{e}^{-\frac{6\log(N)}{3}} = \frac{1}{N^2}. \tag{5.51}$$

Since equation (5.50) imples $N/\log(N) \ll \mu$, we have $\tau < \sqrt{6\log^2(N)/N}$. This, together with equation (5.51) implies that for any $v \in V$, with probability at least $1-\frac{1}{N^2}$, we have

$$\begin{aligned}
|\mathcal{S}_G(v,i+1)| &= \left(1+O\left(\sqrt{\frac{\log^2(N)}{N}}\right)\right)N\left(1-(1+O(\zeta))\left(\mathrm{e}^{-c}+\frac{\delta^i}{N}\right)\right) \\
&= \left(1-\left(\mathrm{e}^{-c}+\frac{\delta^i}{N}\right)+O\left(\zeta\left(\mathrm{e}^{-c}+\frac{\delta^i}{N}\right)+\sqrt{\frac{\log^2(N)}{N}}\right)\right)N \\
&= \left(1-(1+o(1))\left(\mathrm{e}^{-c}+\frac{\delta^i}{N}\right)\right),
\end{aligned} \tag{5.52}$$

because if $\sqrt{\frac{\log^2(N)}{N}} \gg \frac{\delta^i}{N}$, then $\mathrm{e}^{-c} \ge \mathrm{e}^{-o(1)\delta\sqrt{\frac{\log^2(N)}{N}}} > \mathrm{e}^{-o(1)\log(N)} \gg \sqrt{\frac{\log^2(N)}{N}}$. The desired result is implied by a union bound.

(ii) For $i = 0$ we have $\mathbb{E}[\mathcal{S}_G(v, i+1)] = (N-1)p$. In this case, by Lemma 5.2.1 with $\tau = \sqrt{6\log(N)/N}$ we get that with probability at least $1 - \frac{1}{N^2}$ we have

$$|\mathcal{S}_G(v,1)| = \left(1 + O\left(\sqrt{\frac{\log(N)}{N}}\right)\right) Np. \tag{5.53}$$

The desired result is again implied by a union bound.

□

In Lemma 5.4.3 we were quite precise about the error terms. A much weaker formulation of the same idea can be useful for interpreting Theorem 5.5.1.

Corollary 5.4.3. *In the same setting as in Lemma 5.4.1, the expected fraction of nodes in the level set with the largest expected size (conditioning on the expansion properties $\mathcal{E}(\{v\}, i)$) is*

$$\begin{cases} (1+o(1))p & \text{if } i = 0 \text{ and } (1+o(1))p \geq 1-p \\ 1-(1+o(1))p & \text{if } i = 0 \text{ and } p < 1-p \\ (1+o(1))\mathrm{e}^{-c} & \text{if } i > 0 \text{ and } \mathrm{e}^{-c} \geq 1 - \left(\mathrm{e}^{-c} + \frac{\delta^i}{N}\right) \\ 1-(1+o(1))\left(\mathrm{e}^{-c} + \frac{\delta^i}{N}\right) & \text{if } i > 0 \text{ and } \mathrm{e}^{-c} < 1 - \left(\mathrm{e}^{-c} + \frac{\delta^i}{N}\right). \end{cases}$$

Proof. The case $i = 0$ is obvious. The case $i > 0$ is a corollary of equation (5.50). Indeed, the level set with the largest expected size as N tends to infinity is $\mathcal{S}_G(v, i+2)$ if and only if $\mathrm{e}^{-c} \geq 1 - \left(\mathrm{e}^{-c} + \delta^i/N\right)$, in which case it contains a fraction

$$1 - \left(1 - (1+o(1))\left(\mathrm{e}^{-c} + \frac{\delta^i}{N}\right)\right) - o(1) = (1+o(1))\mathrm{e}^{-c}$$

of all nodes by Lemma 5.4.2 and equation (5.50). Otherwise, the level set with the largest expected size as N tends to infinity is $\mathcal{S}_G(v, i+1)$, and its expected size is computed in equation (5.50).

□

The results in this section are summarized in the first five rows of Table 5.1.

5.5 Main Results

We are ready to state our main theorem.

Theorem 5.5.1. *Let $N \in \mathbb{N}$ and $p \in [0,1]$ such that $\frac{\log^5(N)}{N} \ll p$ and $\frac{1}{\sqrt{N}} \ll 1-p$. With the parameters given in Definition 5.4.2, let*

$$\gamma_{smd} = \begin{cases} \max(p, 1-p) & \text{if } i = 0 \quad (\text{i.e., } p = \Theta(1)) \\ \max(\mathrm{e}^{-c}, 1-\mathrm{e}^{-c} - \frac{\delta^i}{N}) & \text{if } i > 0 \quad (\text{i.e., } p = o(1)). \end{cases} \tag{5.54}$$

and let

$$\gamma_{md} = \begin{cases} \sqrt{p^2 + (1-p)^2} & \text{if } i = 0 \\ \sqrt{(\mathrm{e}^{-c})^2 + (1-\mathrm{e}^{-c} - \frac{\delta^i}{N})^2} & \text{if } i > 0. \end{cases} \tag{5.55}$$

Finally, let $G = (V, E)$ *be a realization of a* $\mathcal{G}(N, p)$ *random graph. Then, the following assertion holds a.a.s.*

$$\mathrm{MD}(G) = (1 + o(1)) \frac{\log(N(1-\gamma_{md}))}{\log(1/\gamma_{md})}, \tag{5.56}$$

and

$$\mathrm{SMD}(G) = (1 + o(1)) \frac{\log(N(1-\gamma_{smd}))}{\log(1/\gamma_{smd})}. \tag{5.57}$$

Remark 5.5.1. *In case of* $c \to \infty$ *we have*

$$\frac{1}{\log(1/\gamma_{smd})} = (1 + o(1)) \frac{1}{1-\gamma_{smd}} = (1 + o(1)) \left(\mathrm{e}^{-c} + \frac{\delta^i}{N}\right)^{-1}. \tag{5.58}$$

The same equations also hold for γ_{md}.

Remark 5.5.2. *The term* F_γ *from Theorem 3.4.1 can now be expressed based on Theorem 5.5.1. With the definitions in equations* (5.54) *and* (5.55), *we can write*

$$\frac{\mathrm{SMD}(G)}{\mathrm{MD}(G)} = (1 + o(1)) \frac{\log(1/\gamma_{md})}{\log(1/\gamma_{smd})} \quad \text{if } \delta^{i+1} = \Theta(N), \tag{5.59}$$

hence we have

$$F_\gamma = \frac{\log(\gamma_{md})}{\log(\gamma_{smd})}. \tag{5.60}$$

Finally, we note that since $p \in (0, 1)$ *and* $\mathrm{e}^{-c} \in (0, 1)$, *elementary analysis can show that* F_γ *is a continuous function in* p *taking every value in the interval* $(1/2, 1)$.

5.5.1 Connection between Theorems 5.3.1 and 5.5.1

Clearly, there is a great deal of similarity between the MD/SMD in $\mathcal{G}(N,p)$ and the QC/SQC in $\mathrm{Ber}(q)^{N\times N}$. The final expression can always be written in the form

$$(1+o(1))\frac{\log(N(1-\gamma_\star))}{\log(1/\gamma_\star)},$$

where $\gamma_\star$ is a root mean square in the non-adaptive case and a maximum in the adaptive case. The main difference is that in binary search with randomly restricted queries, $\gamma_\star$ depends on parameter q through a simple direct relation, whereas in random graph binary search the dependence of $\gamma_\star$ on p is more complicated (see equation (5.54)).

We can understand the mapping from p to $\gamma_\star$, if we understand how we can map p to the parameter q in Theorem 5.3.1, which we can do based on our results in Section 5.4. When $p=\Theta(1)$, the mapping is just $p=q$. Indeed, since the diameter is 2 a.a.s, the vector $d(R,v)$ for $v\notin R$ is essentially $\mathrm{Ber}(p)+1$. When $p=o(1)$, the size of either one or two level sets dominates the size of the others (see Figure 5.2), hence the information we get is still basically a random Bernoulli vector, although here we must be more careful in the analysis. The mapping from p to q uses exactly the fraction of nodes in the largest level set established in Corollary 5.4.3.

Since different ranges of parameters require different proof techniques both in this chapter and in [36], an overview of the proofs is presented in Table 5.1 for better clarity.

5.5.2 Proof of Theorem 5.5.1 for the MD

The result on the MD in Theorem 5.5.1 has already been published in [36], with the exception of the upper bound in the case when $c=\Omega(\log(N))$. Indeed, the upper bound of [36]

$$\mathrm{MD}(G)\le(1+o(1))\frac{\log(N)}{\log(1/\gamma_{md})},$$

and does not match the lower bound

$$\mathrm{MD}(G)\ge(1+o(1))\frac{\log(N(1-\gamma_{md}))}{\log(1/\gamma_{md})}$$

unless $c=o(\log(N))$. Here, we present an improved algorithm (in [36] all sensors were selected uniformly at random), and a refined analysis based on the QC upper bound presented in Section 5.3.2 to match the lower bound for the MD for the remaining ranges of c.

Let us assume that $c\to\infty$ (which allows us to use Remark 5.5.1 repeatedly in the proof). The improved algorithm is defined as follows:

1. Select
$$r=\min\left(N,(1+\epsilon_{R_1})\frac{\log(N(1-\gamma_{md}))}{\log(1/\gamma_{md})}\right)\tag{5.61}$$

row nodes uniformly at random to create set R_1, where ϵ is a function slowly tending to zero (to be specified later.)

2. Add all v_j to R_2 for which $d(R, v_j)$ is not unique, i.e., for which we have $d(R, v_j) = d(R, v_k)$ for some node v_k.

The analysis of the algorithm is very similar to the analysis in Section 5.3.2. In the definition of the QC, our goal is to select a set of rows R so that $\mathcal{A}_{R,C\setminus R}$ has pairwise different columns. Notice that the MD can also be defined the same way if we use the distance matrix (where entry (j,k) equals $d(v_j, v_k)$) of the graph instead of matrix $\mathcal{A}$. Indeed, the nodes in R do not need to be distinguished anymore (they always have a unique signature), and the distance signatures $d(R, v)$ are precisely the columns of the submatrix of the distance matrix restricted to the rows in R (which need to be pairwise different for R to be a resolving set). Of course the distance matrix of a $\mathcal{G}(N, p)$ graph does not have the exact same distribution as a Bernoulli random matrix, let alone because the distance matrix of a $\mathcal{G}(N, p)$ graph does not need to be strictly binary. However, we established in Section 5.4 that when $c \to \infty$, most of the distances are concentrated on the value $(i+1)$, and indicators of each entry having the value $(i+1)$ are distributed almost like a Bernoulli random matrix.

Let us make this intuition more precise for an analysis similar to Section 5.3.2. Let us expose all edges in the $\mathcal{G}(N, p)$ graph. Then, we sample each node in R_1 uniformly at random from the nodes of the graph. With this way of sampling, it is possible that we sample the same node twice, however, this does not affect the correctness of the analysis (we can assume that in this case the second node is ignored). With such a sampling procedure, for each fixed node v, the indicator vector of the value $(i+1)$ appearing in $d(R, v)$ is a Bernoulli random vector with i.i.d. entries, with parameter (corresponding to parameter q in Section 5.3.2)

$$q_v = 1 - \frac{|\mathcal{S}(v, i+1)|}{N}. \tag{5.62}$$

For our proof we need to go beyond understanding $d(R, v)$ for a single node; we also need to understand pairwise interactions. Let us define the shorthand

$$s_0(u,v) = \sum_{l=1}^{i} |\mathcal{S}_G(\{u\}, l) \cap \mathcal{S}_G(\{v\}, l)| + |\mathcal{S}_G(\{u\}, i+2) \cap \mathcal{S}_G(\{v\}, i+2)|, \tag{5.63}$$

and

$$s_1(u,v) = |\mathcal{S}_G(\{u\}, i+1) \cap \mathcal{S}_G(\{v\}, i+1)|. \tag{5.64}$$

For q_v we have tight bounds from Lemma 5.4.3. However, to get tight bounds for $s_0(u,v)$ and $s_1(u,v)$ we are going to need even more involved statements about the expansion properties of Erdős-Rényi graphs; this time about the intersection of level sets.

Lemma 5.5.1 (Expansion property for intersections). *With parameter i, c and $\delta \gg \log(N)$ as defined in Definition 5.4.2, for two nodes $v \neq u \in V$, with $s_0(u,v)$ and $s_1(u,v)$ given in equations*

(5.63) *and* (5.64), *let* $\mathcal{E}_3(u,v)$ *be the event that*

$$s_0(u,v) = O\left(N(1-\gamma_{md})^2\right), \tag{5.65}$$

and

$$s_1(u,v) < N\left(1-(2+o(1))\left(\mathrm{e}^{-c}-\frac{\delta^i}{N}\right)\right), \tag{5.66}$$

holds. Let, $\mathcal{E}_3 = \bigcap_{u\neq v\in V}\mathcal{E}_3(u,v)$*. Then, for G sampled from* $\mathcal{G}(N,p)$ *the event* $\mathcal{E}_3$ *holds a.a.s.*

Proof. First, we are going to show inductively that for $l \leq i$ the event

$$|\mathcal{S}_G(\{u\},l)\cap\mathcal{S}_G(\{v\},l)| < (1+o(1))\max\left(\frac{\delta^{2l}}{N}, 9\delta^{l-1}\log(N)\right) \tag{5.67}$$

holds a.a.s. Clearly, we have $|\mathcal{S}_G(\{u\},0)\cap\mathcal{S}_G(\{v\},0)| = 0$, therefore the statement holds for the base step. Similarly to Section 5.4, in the induction step, we expose all level sets $\mathcal{S}_G(\{v\},l)$ and $\mathcal{S}_G(\{u\},l)$, we condition on equation (5.67) for l, and we aim to show the statement for $l+1$. The probability for a new node to be in $\mathcal{S}_G(\{u\},l+1)\cap\mathcal{S}_G(\{v\},l+1)$ is

$$\left(1-(1-p)^{|\mathcal{S}_G(\{v\},l)\setminus\mathcal{S}_G(\{u\},l)|}\right)\left(1-(1-p)^{|\mathcal{S}_G(\{u\},l)\setminus\mathcal{S}_G(\{v\},l)|}\right)(1-p)^{|\mathcal{S}_G(\{v\},l)\cap\mathcal{S}_G(\{u\},l)|} + \left(1-(1-p)^{|\mathcal{S}_G(\{v\},l)\cap\mathcal{S}_G(\{u\},l)|}\right)$$

which is clearly upper bounded (using Bernoulli's inequality) by

$$p|\mathcal{S}_G(\{v\},l)\setminus\mathcal{S}_G(\{u\},l)|\cdot p|\mathcal{S}_G(\{u\},l)\setminus\mathcal{S}_G(\{v\},l)| + p|\mathcal{S}_G(\{v\},l)\cap\mathcal{S}_G(\{u\},l)|.$$

Using Lemma 5.4.1 and the induction hypothesis,

$$\begin{aligned}\mathbb{E}[|\mathcal{S}_G(\{u\},l+1)\cap\mathcal{S}_G(\{v\},l+1)|] &< (1+o(1))N\left(p^2\delta^{2l} + p\cdot\max\left(\frac{\delta^{2l}}{N}, 9\delta^{l-1}\log(N)\right)\right)\\ &= (1+o(1))\max\left(\frac{\delta^{2l+2}}{N}, 9\delta^{l}\log(N)\right).\end{aligned} \tag{5.68}$$

By Lemma 5.2.1 (on Chernoff bounds) with $\tau = (1+o(1))$, equation (5.68) holds with probability $1-\frac{1}{N^3}$, even if we drop the expected value sign, because the expectation is larger than $9\log(N)$ for all $l \geq 0$.

We proceed to the case of $l = i+1$. Since by the induction hypothesis

$$|\mathcal{S}_G(\{u\},i)\cap\mathcal{S}_G(\{v\},i)| < (1+o(1))\max\left(\frac{\delta^{2i}}{N}, 9\delta^{i-1}\log(N)\right) = o\left(\delta^i\right) \tag{5.69}$$

we have that

$$\begin{aligned}
&\mathbb{E}[|\mathcal{S}_G(\{u\}, i+1) \cap \mathcal{S}_G(\{v\}, i+1)|] \\
&\qquad = \left(N - (2+o(1))\delta^i\right)\left(\left(1-(1-p)^{(1+o(1))\delta^i}\right)^2 (1-p)^{o(\delta^i)} + \left(1-(1-p)^{o(\delta^i)}\right)\right) \\
&\qquad = \left(N - (2+o(1))\delta^i\right)\left(\left(1-\mathrm{e}^{-c(1+o(1))}\right)^2 \mathrm{e}^{-o(c)} + \left(1-\mathrm{e}^{-o(c)}\right)\right) \\
&\qquad = N\left(1-(2+o(1))\left(\mathrm{e}^{-c(1+o(1))} - \frac{\delta^i}{N}\right)\right). \qquad (5.70)
\end{aligned}$$

Let us denote $\mu = \mathbb{E}[|\mathcal{S}_G(\{u\}, i+1) \cap \mathcal{S}_G(\{v\}, i+1)|]$. Then, by Lemma 5.2.1 with $\tau = \sqrt{9\log(N)/\mu}$, similarly to equations (5.51) and (5.52), we can prove that equation (5.70) holds with probability $1 - \frac{1}{N^3}$, even if we drop the expected value sign.

Finally for the level $i+2$,

$$\begin{aligned}
\mathbb{E}[|\mathcal{S}_G(\{u\}, i+2) \cap \mathcal{S}_G(\{v\}, i+2)|] &< (1+o(1))N(1-p)^{|\mathcal{S}_G(\{v\},l)|+|\mathcal{S}_G(\{u\},l)|-|\mathcal{S}_G(\{v\},l)\cap\mathcal{S}_G(\{u\},l)|} \\
&= (1+o(1))N\mathrm{e}^{-2c+o\left(p\delta^i\right)} \\
&= O\left(N\mathrm{e}^{-2c}\right). \qquad (5.71)
\end{aligned}$$

Now either $N\mathrm{e}^{-2c} = o(\log(N))$, in which case

$$\frac{\delta^{2i}}{N} > \frac{\delta^{i+1}}{N} = c = \Omega(\log(N)),$$

and the contribution of level $i+2$ in $s_0(u,v)$ is dominated by the contribution of level i a.a.s, or $N\mathrm{e}^{-2c} = \Omega(\log(N))$, in which case (by Lemma 5.2.1 with $\tau = O(1)$) equation (5.71) holds with probability $1 - \frac{1}{N^3}$, even if we drop the expected value sign. We focus on the latter case from now on (for the former case, the calculation is almost identical). We proved that a.a.s.,

$$\begin{aligned}
s_0(u,v) &= \sum_{l=1}^{i} |\mathcal{S}_G(\{u\}, l) \cap \mathcal{S}_G(\{v\}, l)| + |\mathcal{S}_G(\{u\}, i+2) \cap \mathcal{S}_G(\{v\}, i+2)| \\
&= \sum_{l=1}^{i} (1+o(1)) \max\left(\frac{\delta^{2l}}{N}, 9\delta^{l-1}\log(N)\right) + O\left(N\mathrm{e}^{-2c}\right). \\
&= O\left(\max\left(\frac{\delta^{2i}}{N}, 9\delta^{i-1}\log(N)\right) + N\mathrm{e}^{-2c}\right)
\end{aligned}$$

Notice that the term $9\delta^{i-1}\log(N)$ is always dominated by either δ^{2i}/N or e^{-2c}. Indeed, either $c = \delta^{i+1}/N = \Omega(\log(N))$, which implies $\delta^{2i}/N = \Omega(\delta^{i-1}\log(N))$, or $c = o(\log(N))$, which implies $N\mathrm{e}^{-2c} = \Theta(N) = \omega(\delta^{i-1}\log(N))$. Then

$$\begin{aligned} s_0(u,v) &= O\left(N\left(\frac{\delta^{2i}}{N^2} + \mathrm{e}^{-2c}\right)\right) \\ &= O\left(N\left(1-\gamma_{md}\right)^2\right). \end{aligned}$$

Since all equations held with probability $1-\frac{1}{N^3}$, a union bound completes the proof. □

Now we are ready to compute the most important quantities from the proof in Section 5.3.2. Let Y be the number of nodes v with non-unique distance vectors $d(R_1,v)$ after step 1 of the algorithm, and let Y_0 is the number of nodes v with distance vectors $d(R_1,v)$ being all-$(i+1)$. Then,

$$\mathbb{E}[Y_0] = \sum_{v\in V\setminus R_1} (1-q_v)^r. \tag{5.72}$$

Moreover, if $Y_{0,v}$ is the indicator of $d(R,v)$ being all-$(i+1)$, then

$$\mathbb{E}[Y_0(Y_0-1)] = \sum_{v\in V\setminus R_1} \mathbb{E}[Y_{0,v}^2] + \sum_{v\neq u\in V\setminus R_1} \mathbb{E}[Y_{0,v}Y_{0,u}] - \mathbb{E}[Y_0] = \sum_{u\neq v\in V\setminus R_1} \left(\frac{s_1(u,v)}{N}\right)^r. \tag{5.73}$$

Finally, we can upper bound the expectation of the number pairs of nodes $u \neq v \in V\setminus R_1$ with $d(R,u) = d(R,v)$ as

$$\mathbb{E}[Z] \leq \sum_{u\neq v\in V\setminus R_1} \left(\frac{s_0(u,v)}{N} + \frac{s_1(u,v)}{N}\right)^r. \tag{5.74}$$

Substituting equations (5.72), (5.73) and (5.74) into equation (5.19) yields

$$\mathbb{E}[Y] \leq \sum_{v\in V\setminus R_1} (1-q_v)^r + \sum_{u\neq v\in V\setminus R_1} \left(\left(\frac{s_0(u,v)}{N} + \frac{s_1(u,v)}{N}\right)^r - \left(\frac{s_1(u,v)}{N}\right)^r\right). \tag{5.75}$$

Note that since x^r is a convex function in x for $r > 1$,

$$\left(\frac{s_0(u,v)}{N} + \frac{s_1(u,v)}{N}\right)^r - \left(\frac{s_1(u,v)}{N}\right)^r \leq r\frac{s_0(u,v)}{N}\left(\frac{s_0(u,v)}{N} + \frac{s_1(u,v)}{N}\right)^{r-1}. \tag{5.76}$$

Note that for any constant $y = \Theta(1)$, and for $z \to 0$,

$$\frac{\log(1-(y+o(1))z)}{\log(1-z)} = (1+o(1))y. \tag{5.77}$$

Using the results of Lemma 5.5.1, and equation (5.77) with $y = 1$ and

$$z = 1-\gamma_{md} = 1-\sqrt{\mathrm{e}^{-2c} + \left(1-\mathrm{e}^{-c}-\frac{\delta^i}{N}\right)^2} = (1+o(1))\left(\mathrm{e}^{-c} + \frac{\delta^i}{N}\right),$$

our choice of r satisfies

$$\sum_{v\in V\setminus R_1}(1-q_v)^r < N\left(1-(1+o(1))\left(\mathrm{e}^{-c}+\frac{\delta^i}{N}\right)\right)^r = N\mathrm{e}^{(1+\epsilon_{R_1})\log(N(1-\gamma_{md}))\frac{\log\left(1-(1+o(1))\left(\mathrm{e}^{-c}+\frac{\delta^i}{N}\right)\right)}{\log(1/\gamma_{md})}}$$

$$\overset{(5.77)}{=} \frac{N}{(N(1-\gamma_{md}))^{(1+\epsilon_{R_1})(1+o(1))}}. \tag{5.78}$$

Similarly, using equation (5.77) with $y=2$, our choice of r satisfies

$$\sum_{\substack{u\neq v\in V\setminus R_1\\ k_{uv}=2}} r\frac{s_0(u,v)}{N}\left(\frac{s_0(u,v)}{N}+\frac{s_1(u,v)}{N}\right)^{r-1} \tag{5.79}$$

$$< N^2O\big((1-\gamma_{md})^2\big)\,r\left(1-(2+o(1))\left(\mathrm{e}^{-c}-\frac{\delta^i}{N}\right)\right)^r$$

$$= N^2O\big((1-\gamma_{md})^2\big)\,r\mathrm{e}^{(1+\epsilon_{R_1})\log(N(1-\gamma_{md}))\frac{\log\left(1-(2+o(1))\left(\mathrm{e}^{-c}+\frac{\delta^i}{N}\right)\right)}{\log(1/\gamma_{md})}}$$

$$\overset{(5.77)}{=} \frac{N^2O\big((1-\gamma_{md})^2\big)\,r}{(N(1-\gamma_{md}))^{2(1+o(1))(1+\epsilon_{R_1})}}. \tag{5.80}$$

Notice that we can choose ϵ_{R_1} to converge slowly enough so that the terms $(1+o(1))(1+\epsilon_{R_1})$ in the exponents are lower bounded by $1+\epsilon_{R_1}/2$. Then, if we substitute back (5.78) and (5.80) into (5.75), we get

$$\mathbb{E}[Y] < \frac{N}{(N(1-\gamma_{md}))^{1+\epsilon_{R_1}/2}} + \frac{N^2O\big((1-\gamma_{md})^2\big)\,r}{(N(1-\gamma_{md}))^{2(1+\epsilon_{R_1}/2)}} = O\left(\frac{r}{(N(1-\gamma_{md}))^{\epsilon_{R_1}}}\right). \tag{5.81}$$

Finally, choosing $\epsilon_{R_1} > \frac{2\log\log(N)}{\log(N(1-\gamma_{md}))}$, by Markov's inequality we arrive to

$$\mathbb{P}(Y > r\log^{-1}(N)) < \frac{\mathbb{E}[Y]}{r\log^{-1}(N)} = O(\log^{-1}(N)) = o(1), \tag{5.82}$$

and the proof is completed.

5.5.3 Proof of Theorem 5.5.1 for $p=\Theta(1)$, SMD Upper Bound

Proof. The $p=\Theta(1)$ case of Theorem 5.5.1 seems very similar to Theorem 5.3.1 because entries of the distance matrix of $G\sim\mathcal{G}(N,p)$ are essentially $\mathrm{Ber}(p)+1$ random variables. However, the matrix is always symmetric, which causes some complications in the proof.

We start by providing an analogous definition of an f-separator for graphs.

Definition 5.5.1 (f-separator). *Let $f(n)\in\mathbb{N}\to\mathbb{R}^+$ be a function. A set of nodes $W\subseteq V$, $|W|=n$*

has an f-separator if there is a node $w \in V$ such that

$$\max_{l \in \mathbb{N}} |W \cap \mathcal{S}_G(w,l)| \leq n\gamma_{smd} + f(n). \tag{5.83}$$

For the upper bound, the statement that for any set $W \subseteq V$, any node $w \in V$ can independently be an f-separator is not true anymore in contrast to the proof of Theorem 5.3.1, since the neighborhoods of the nodes in W are slightly correlated. However, the statement is still true for nodes $w \in V \setminus W$. Hence the proof will go on two steps. In step 1 we prove that V has a $2\sqrt{n}$-separator a.a.s., and next in step 2 we prove that any set W of cardinality at most $\gamma_{smd}N + 2\sqrt{N}$ has an f-separator with $f(n) = o\left(\frac{n}{\log(n)}\right)$.

For step 1, let us pull aside from V a random subset $F \subset V$ of cardinality $|F| = \log\log(N)$. By equation (5.37) with $q = p$ and $X_w = |V' \cup \mathcal{S}_G(w,1)|$, each $w \in F$ is not a $\sqrt{3n}$-separator of $V' = V \setminus F$ with probability at most e^{-1}. Since these events are independent, the probability that no node $w \in F$ is an $\sqrt{3n}$-separator of V' is then $\mathrm{e}^{-\log\log(N)} = \log(N)^{-1} \to 0$. On the other hand, a $\sqrt{3n}$-separator of V' is also a $2\sqrt{n}$-separator of V, since $\sqrt{3N} + \log\log(N) < 2\sqrt{N}$ for N large enough.

For step 2, we repeat the calculation in equation (5.37) with $f(n) = \sqrt{\frac{6}{1-\gamma_{smd}}}$. Let $X_w = |W \cap \mathcal{S}_G(w,1)|$, then

$$\begin{aligned}\mathbb{P}(|X_w - np| \geq f(n)) = \mathbb{P}\left(|X_w - \mathbb{E}[X_w]| \geq \frac{f(n)}{np}np\right) &\leq 2\mathrm{e}^{\frac{-2np\frac{f(n)^2}{n^2p^2}}{3}} \\ = 2\mathrm{e}^{-\frac{12}{3p(1-\gamma_{smd})}} &< \mathrm{e}^{-\frac{2}{1-\gamma_{smd}}},\end{aligned} \tag{5.84}$$

because $p \leq 1$. Note that by equations (5.35) and (5.36) with $q = p$ and $\gamma_{sqc} = \gamma_{smd}$, the event $|X_w - np| \geq f(n)$ implies that w is an f-separator for W.

Let Y be the number of subsets W of cardinality at most $\gamma_{smd}N + 2\sqrt{N}$ that do not have a $\sqrt{\frac{6}{1-\gamma_{smd}}}$-separator. Then,

$$\begin{aligned}\mathbb{P}(Y > 0) \leq \mathbb{E}[Y] &< \sum_{|W| \leq \gamma_{smd}N + 2\sqrt{N}} \mathrm{e}^{-\frac{2}{1-\gamma_{smd}}(N-|W|)} \\ &\leq 2^N \mathrm{e}^{-\frac{2}{1-\gamma_{smd}}(N-(\gamma_{smd}N+2\sqrt{N}))} \\ &< \mathrm{e}^{-N+\frac{4\sqrt{N}}{1-\gamma_{smd}}} \to 0,\end{aligned}$$

as long as $1 - \gamma_{smd} \gg \frac{1}{\sqrt{N}}$. The existence of a $2\sqrt{n}$-separator for V (step 1) and a $\sqrt{\frac{6}{1-\gamma_{smd}}}$-separator for all $W \subseteq V$ of cardinality at most $\gamma_{smd}N + 2\sqrt{N}$ (step 2) holds together a.a.s. by the union bound. Note that for $\gamma_{smd} \to 1$

$$\begin{aligned}
\sqrt{\frac{6}{1-\gamma_{smd}}} &\overset{(5.58)}{=} \sqrt{\frac{6\,(1+o\,(1))}{\log\left(\frac{1}{\gamma_{smd}}\right)}} \ll \frac{1}{\log\left(\frac{1}{\gamma_{smd}}\right)\left(1+\log\left(\frac{1}{\log\left(\frac{1}{\gamma_{smd}}\right)}\right)\right)} \\
&\ll \frac{\log\,(N)}{\log\left(\frac{1}{\gamma_{smd}}\right)\log\log(N)\left(\log\log\,(N)-\log\log\log(N)+\log\left(\frac{1}{\log\left(\frac{1}{\gamma_{smd}}\right)}\right)\right)} \\
&= \frac{n}{\log\,(n)} \qquad (5.85)
\end{aligned}$$

for $n = \frac{\log(N)}{\log\log(N)\log(1/\gamma_{smd})}$. Hence, $\sqrt{\frac{6}{1-\gamma_{smd}}} = o\left(\frac{n}{\log(n)}\right)$ for $n = \Omega\left(\frac{\log(N)}{\log(1/\gamma_{smd})}\right)$, which is the necessary condition on $f(n)$ to apply Lemma 5.3.1.

Since we were able to prove the existence of $(|\mathcal{T}_{R_j}|\gamma_{sqc} + f(|\mathcal{T}_{R_j}|))$-reducers for $j = 1$ (step 1 of the analysis), and for any $j > 1$ with candidate set size $|\mathcal{T}_{R_j}| = \Omega\left(\frac{\log(N)}{\log(1/\gamma_{smd})}\right)$ (step 2 of the analysis), Lemma 5.3.1 concludes the proof. □

5.5.4 Proof of Theorem 5.5.1 for $p = o(1)$, SMD Lower Bound

Proof. This proof is based on a coupling between the graph case and a simple stochastic process, which we can analyse similarly to the SQC lower bound. We start by upper bounding the probability that a random set is a resolving set by the probability that a certain survival process leaves at least two of the nodes alive. Since this survival process is still too complicated to analyse, we are going to introduce a second survival process later in the proof, which will give us the desired bound.

As usual, we first select queries $R = \{w_1, \ldots, w_r\}$ at random, with $|R| \leq r = (1-\epsilon)\frac{\log(N)}{\log(1/\gamma_{sqc})}$ and ϵ slowly decaying to zero. In the proof of the SQC lower bound, we had an explicit lower bound on how slowly ϵ must tend to zero, however, this time we do not provide such guarantees for the sake of simplicity.

Let $l^\star$ be the index of the largest level set in expectation. Using the results of Corollary 5.4.3,

$$l^\star = \begin{cases} i+1 & \text{if } \mathrm{e}^{-c} < 1-\mathrm{e}^{-c}-\frac{\delta^i}{N} \\ i+2 & \text{if } \mathrm{e}^{-c} \geq 1-\mathrm{e}^{-c}-\frac{\delta^i}{N}. \end{cases} \qquad (5.86)$$

Let $\mathcal{R}_R$ be the event that the randomly sampled set R of size $|R|$ is a resolving set in G, and $\mathcal{R}$ that there exists at least one resolving set. Similarly to Section 5.3.5 we want to upper bound $\mathbb{P}(\mathcal{R})$ by $N^r\mathbb{P}(\mathcal{R}_R)$, and $\mathbb{P}(\mathcal{R}_R)$ by the probability of the event that there are at least two distinct nodes $u \neq v \in V$ with $d(R,u) = d(R,v) = l^\star \mathbf{1}$.

Since R is uniformly random, we may sample it before any of the edges in G are exposed. Let us now expose the edges of G similarly to the proof of Lemma 5.4.3, except this time starting from the set R instead of a single node. Notice that before any of the graph is exposed, any of the nodes $v \in V \setminus R$ could possibly have $d(R,v) = l^\star \mathbf{1}$. Then, as more and more edges get exposed, many of the nodes lose this property. For instance the neighbors of the nodes in R cannot have $d(R,v) = l^\star \mathbf{1}$, because $l^\star > i \geq 1$. Hence focusing on the event $\mathcal{R}_R$, this exploration process of the graph can be seen as a survival process, where at least two nodes must survive.

Definition 5.5.2 (ESP). *In the exploration survival process (ESP) all nodes $v \in V \setminus R$ start out alive. In step $l < l^\star$, we expose all unexposed edges incident to the nodes in $\mathcal{S}_G(w_j, l)$ to expose $\mathcal{S}_G(w_j, l+1)$ for all $j \in \{1, \ldots, r\}$. Every node exposed this way dies. Then, if $l^\star = i+1$ we play an extra round, in which we expose all unexposed edges incident to the nodes in $\mathcal{S}_G(w_j, i+1)$, and a node that was still alive at the end of round $l^\star - 1$ survives this round if it connects to $\mathcal{S}_G(w_j, i)$ for all $j \in \{1, \ldots, r\}$. If $l^\star = i+2$, there is no extra round.*

Note that the ESP always takes $i+1$ steps, it is only the nature of the final step that depends on $l^\star$. The event that v survives the ESP is equivalent to $d(R,v) = l^\star \mathbf{1}$, unless $l^\star = i+2$ and event $\mathcal{D}$ in Corollary 5.4.2 does not hold (in this case node v surviving the ESP could have $d(w_j, v) > l^\star$). Since $\mathcal{D}$ holds a.a.s., we can assume it holds (formally, we may intersect all of our events with $\mathcal{D}$ and apply a union bound in the end).

In the first $l < l^\star$ rounds the probability of survival is $\rho_l^{(0)} = (1-p)^{|\mathcal{S}_G(R, l-1)|}$, and this probability is itself a random variable. When $l = l^\star = i+1$, we need each node to connect to each $\mathcal{S}_G(w_j, l-1)$, but these sets might have an intersection, so the exact value of the probability of survival (which we call $\rho_{l^\star}^{(0)}$) is complicated to write down. Fortunately, $\rho_{l^\star}^{(1)}$ can be lower bounded by $\prod_{j=1}^{r} \left(1 - (1-p)^{|\mathcal{S}_G(w_j, l^\star - 1)|}\right)$, since the events of connecting to $\mathcal{S}_G(w_j, l-1)$ for each $j \in \{1, \ldots, r\}$ are positively correlated. This motivates an alternative but still complicated survival process, which will serve as a bridge to the simple process we will finally analyse.

Definition 5.5.3 (CSP). *In the complex survival process (CSP) all nodes $v \in V \setminus R$ start out alive. In each of the $i+1$ rounds, each node survives with probability $\rho_l^{(1)}$, where*

$$\rho_l^{(1)} = \begin{cases} (1-p)^{|\mathcal{S}_G(R, l-1)|} & \text{if } l < l^\star \\ \prod_{j=1}^{r} \left(1 - (1-p)^{|\mathcal{S}_G(w_j, l^\star - 1)|}\right) & \text{if } l = l^\star \text{ and } l^\star = i+1, \end{cases} \tag{5.87}$$

where the sets $\mathcal{S}_G(R, l-1)$ are the same as in the ESP.

Let Y_0 (respectively, Y_1) be the indicator variable that at least two nodes survive the ESP (respectively, CSP). Then $Y_0 = 0$ is the same event as $\mathcal{R}_R$ and $\rho_l^{(1)} \leq \rho_l^{(0)}$, which implies $\mathbb{P}(\mathcal{R}_R) = \mathbb{P}(Y_0 = 0) \leq \mathbb{P}(Y_1 = 0)$. However, the CSP is still too difficult to analyse even with the lower bound in (5.87) because each term is itself a random variable depending on earlier levels. Instead, we study a "simple" survival process.

Corollary 5.5.1. *Let us denote the event $\mathcal{E}(R, l^\star - 1)$ by $\mathcal{E}_R$. Then, if $\mathcal{E}_R$ holds, there exists a constant C such that*

$$|\mathcal{S}_G(R, l-1)| \le r\delta^{l-1}(1 + C\zeta) \tag{5.88}$$

for all R and $l < l^\star$, and when $l^\star = i + 1$

$$|\mathcal{S}_G(w_j, l^\star - 1)| \ge \delta^{l^\star - 1}(1 - C\zeta) \tag{5.89}$$

for all w_j.

Proof. The existence of such a constant C is implied by equation (5.42) in Lemma 5.4.1 (for equation (5.88) since we need an upper bound, the intersection of the sets $\mathcal{S}_G(v, l-1)$ can be ignored). □

Definition 5.5.4 (SSP)**.** *In the simple survival process (SSP) all nodes $v \in V \setminus R$ start out alive. In each of the $i+1$ rounds, each node survives with probability $\rho_l^{(2)}$, where*

$$\rho_l^{(2)} = \begin{cases} (1-p)^{r\delta^{l-1}(1+C\zeta)} & \text{if } l < l^\star \\ \left(1 - (1-p)^{(1-C\zeta)}\right)^r & \text{if } l = l^\star \text{ and } l^\star = i+1, \end{cases} \tag{5.90}$$

and C is the constant in Corollary 5.5.1.

Let Y_2 be the indicator variable that at least two nodes survive the SSP. Equations (5.88) and (5.89) imply $\rho_l^{(1)} \ge \rho_l^{(2)}$, so it is in fact easier for nodes to survive the CSP than the SSP (the words "simple" and "complex" in the names of the terms SSP and CSP refer to the difficulty of analysis not the difficulty of survival). Hence, we should be able to prove $\mathbb{P}(Y_1 = 0, \mathcal{E}_R) \le \mathbb{P}(Y_2 = 0, \mathcal{E}_R)$, and we will prove it rigorously by coupling $(Y_1, \mathcal{E}_R)$ and $(Y_1, \mathcal{E}_R)$. Recall, that a coupling is a joint distribution $((\hat{Y}_1, \hat{\mathcal{E}}_R), (\hat{Y}_2, \hat{\mathcal{E}}_R))$ on $\{0,1\} \times \{0,1\}$ with the property that its first marginal is $(Y_1, \mathcal{E}_R)$ and the second marginal is $(Y_2, \mathcal{E}_R)$.

We define the joint distribution $(\hat{Y}_1, \hat{Y}_2, \hat{\mathcal{E}}_R)$ by specifying how to sample from it. We will simultaneously play the ESP, CSP and the SSP, and since all three processes can be simulated using Bernoulli trials with parameter p, we will use the same outcomes of these trials whenever possible. More precisely, the probability space of $(\hat{Y}_1, \hat{Y}_2, \hat{\mathcal{E}}_R)$ is the probability space of $N(N-r)(l^\star - 1 + r)$ coin flips where the probability of heads is p. The first $N(N-r)(l^\star - 1 + r)$ coin flips are first organized into $N - r$ buckets of size $N(l^\star - 1 + r)$ indexed by nodes $v \in V \setminus R$. Then, each of the $N - r$ buckets are further divided into $l^\star - 1$ blue and r red subbuckets, all of size N. The k^{th} coin flip in the l^{th} blue and respectively the j^{th} red subbucket of the bucket corresponding to node v is called flip$(v, "b", l, k)$ and respectively flip$(v, "r", j, k)$. Figure 5.3 explains the structure and function of these buckets through an example.

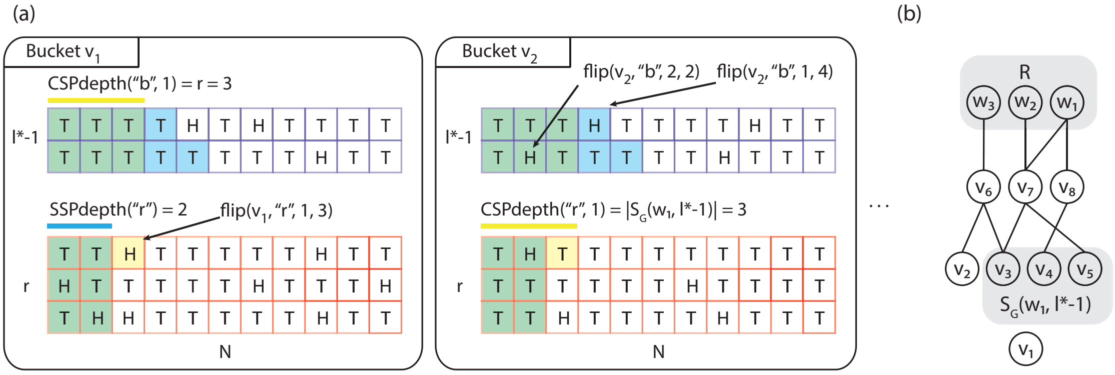


Figure 5.3: Part (a) of the figure shows a (partial) example sample of the elementary events of the coupled joint distribution. Here we set $N = 11, r = 3$ and $l^\star = i + 1 = 3$. These parameters might not actually correspond to any pair of parameters (N, p), we only use them for the example. The colors signify which survival processes use which coin flips; light blue is for the SSP, yellow is for the CSP and green is for the coin flips used by both of them. We should show $N - r = 8$ buckets in total; one for each node $v_i \in V \setminus R$, but we only show the bucket for two nodes v_1, v_2 in the interest of space. Node v_1 survives the first $l^\star - 1$ rounds in both processes, but survives the last round only in the CSP. In the SSP, it does not survive because the first red subbucket contains no head for this process (the CSP is saved by $\text{flip}(v_1, "r", 1, 3)$). Node v_2 dies in the first round of the SSP (because of $\text{flip}(v_2, "b", 1, 4)$) and in the second round of the CSP (because of $\text{flip}(v_2, "b", 2, 2)$).
Part (b) of the figure shows (a possible realisation of) the ESP corresponding the coin flips in part (a). The edges incident to nodes v_1 and v_2 correspond to the green coin flips in the blue subbuckets in part (a) of the figure. Only one such edge is present in this realization of the ESP; the edge between v_2 and v_6. This edge corresponds to $\text{flip}(v_2, "b", 2, 2)$, which is indeed the only green head in the blue subbuckets in part (a) of the figure. Part (b) of the figure also explains the values of $\text{CSPdepth}("r", j)$ in part (a). Indeed, we can check that $|\mathcal{S}_G(w_1, l^\star - 1)| = 3$ and $|\mathcal{S}_G(w_2, l^\star - 1)| = |\mathcal{S}_G(w_3, l^\star - 1)| = 2$. Similarly, we can check that $|\cup_{j=1}^{r} \mathcal{S}_G(w_j, 1)| = |\{v_6, v_7, v_8\}| = 3$, which corresponds to the value of $\text{CSPdepth}("b", 2)$ in part (a). Note that only coin flips in the blue subbuckets can correspond to edges in the ESP, as in the coupling we only simulate the ESP until round $l^\star - 1$.

To define $\hat{Y}_1$, we must explain how to simulate the CSP in Definition 5.5.3 using the coin flips in the blue and red subbuckets. To simulate the CSP we must know the level set sizes $\mathcal{S}_G(R, l - 1)$ for $l \le l^\star$, which requires simulating the ESP at least up to round $l^\star - 1$. Note that in the ESP, we only expose edges between one dead and one alive node (if we consider the set R dead at the start). For each such exposure of an edge between dead node u and alive node v in round $l < l^\star$, we use $\text{flip}(v, "b", l, k)$, where k is the lowest index for which $\text{flip}(v, "b", l, k)$ has not been used so far in the ESP. This way, the exact mapping between edges and coin flips can change depending on the order we sample the edges in each round, but this will not affect the coupling. Until round $l^\star - 1$, we are able to perfectly simulate the ESP, and fortunately we already have enough information to simulate the CSP. In round $l = l^\star$, we already know enough to finish simulating the CSP and we may ignore the ESP. We simply define $\hat{Y}_1$ to be the indicator of the event that there exists $V_{CSP} \subset V$ with $|V_{CSP}| \geq 2$, such that for any $v \in V_{CSP}$ we have no head in the set

$$\{\text{flip}(v, "b", l, k) \mid 1 \leq k \leq \text{CSPdepth}("b", l)\} \tag{5.91}$$

for each positive integer $l < l^\star$, where

$$\text{CSPdepth}("b", l) = \left| \bigcup_{j=1}^{r} \mathcal{S}_G(w_j, l-1) \right| \tag{5.92}$$

This set is exactly the "used" nodes of the ESP, since the ESP and the CSP are identical in rounds $l < l^\star$. In addition, if $l^\star = i+1$, we also need that there is at least one head in the set

$$\{\text{flip}(v, "r", j, k) \mid 1 \leq k \leq \text{CSPdepth}("r", j)\} \tag{5.93}$$

for each positive integer $j \leq r$, where

$$\text{CSPdepth}("r", j) = |\mathcal{S}_G(w_j, l^\star - 1)|. \tag{5.94}$$

It is clear that each bucket has at least as many coin flips as we need in each step, and that $\hat{Y}_1$ and Y_1 have the same distribution. Note that since we know the cardinality of the level sets $\mathcal{S}_G(R, l-1)$ for all $l \leq l^\star$, we can also determine $\hat{\mathcal{E}}_R$.

The random variable $\hat{Y}_2$ can simply be defined as the indicator of the event that there exists $V_{SSP} \subset V$ with $|V_{SSP}| \geq 2$, such that for any $v \in V_{SSP}$ we have no head in the set

$$\{\text{flip}(v, "b", l, k) \mid 1 \leq k \leq \text{SSPdepth}("b", l)\} \tag{5.95}$$

for each positive integer $l < l^\star$, where

$$\text{SSPdepth}("b", l) = r\delta^{l-1}(1 + C\zeta). \tag{5.96}$$

In addition, if $l^\star = i+1$, we also need that there is at least one head in the set

$$\{\text{flip}(v, "r", j, k) \mid 1 \leq k \leq \text{SSPdepth}("r")\} \tag{5.97}$$

for each positive integer $j \leq r$, where

$$\text{SSPdepth}("r") = (1 - C\zeta)\delta^{l^\star - 1} \tag{5.98}$$

(this depth does not depend on j, but it still needs to hold for r subbuckets). Clearly, $\hat{Y}_2$ and Y_2 have the same distribution.

By equation (5.88), if the event $\hat{\mathcal{E}}_R$ holds, each coin flip used by the CSP in rounds $\{1, \ldots, i+1\}$

is also used by the SSP. Hence if a node survives in SSP it must also survive in the CSP. When $l^\star = i+1$, the situation is reversed in the $(l^\star)^{th}$ round. By equation (5.89), each coin flip used by the SSP is used by the CSP. However, this time we need heads to survive, so again if a node survives in SSP it must also survive in CSP. Hence, $\hat{\mathbb{P}}(\hat{Y}_1, \hat{\mathcal{E}}_R) < \hat{\mathbb{P}}(\hat{Y}_2, \hat{\mathcal{E}}_R)$. We are now ready to use our coupling to bound the probability that there exists a resolving set.

$$\begin{aligned}
\mathbb{P}(\mathcal{R}) &\le \mathbb{P}(\mathcal{R}, \mathcal{E}) + \mathbb{P}(\overline{\mathcal{E}}) \\
&\le \sum_{|R| \le r} \mathbb{P}(\mathcal{R}_R | \mathcal{E}) + \mathbb{P}(\overline{\mathcal{E}}) \\
&\le \sum_{|R| \le r} \mathbb{P}(Y_1 = 0 | \mathcal{E}) + \mathbb{P}(\overline{\mathcal{E}}) \\
&\le N^r \frac{\mathbb{P}(Y_1 = 0, \mathcal{E}_R)}{\mathbb{P}(\mathcal{E})} + \mathbb{P}(\overline{\mathcal{E}}) = N^r \frac{\hat{\mathbb{P}}(\hat{Y}_1 = 0, \mathcal{E}_R)}{\mathbb{P}(\mathcal{E})} + \mathbb{P}(\overline{\mathcal{E}}) \\
&\le N^r \frac{\hat{\mathbb{P}}(\hat{Y}_2 = 0, \mathcal{E}_R)}{\mathbb{P}(\mathcal{E})} + \mathbb{P}(\overline{\mathcal{E}}) \\
&\le N^r \frac{\hat{\mathbb{P}}(\hat{Y}_2 = 0)}{\mathbb{P}(\mathcal{E})} + \mathbb{P}(\overline{\mathcal{E}}) = N^r \frac{\mathbb{P}(Y_2 = 0)}{\mathbb{P}(\mathcal{E})} + \mathbb{P}(\overline{\mathcal{E}}).
\end{aligned} \tag{5.99}$$

Now we proceed to upper bounding $\mathbb{P}(Y_2 = 0)$. Let Z_v be the indicator of the event that node $v \in V \setminus R$ survives in the SSP. We need to distinguish the two cases (i) $\mathrm{e}^{-c} \ge 1 - \mathrm{e}^{-c} - \frac{\delta^i}{N}$ and (ii) $\mathrm{e}^{-c} < 1 - \mathrm{e}^{-c} - \frac{\delta^i}{N}$.

(i) In the case $\mathrm{e}^{-c} < 1 - \mathrm{e}^{-c} - \frac{\delta^i}{N}$, since by equation (5.86) we have $l^\star = i+1$,

$$\begin{aligned}
\mathbb{P}(Z_v = 1) &\ge \left(\prod_{l=1}^{l^\star - 1} (1-p)^{r\delta^{l-1}(1+C\zeta)} \right) \left(1 - (1-p)^{(1-C\zeta)\delta^{l^\star - 1})} \right)^r \\
&\ge \mathrm{e}^{-p(1+o(1))r\delta^{i-1}} (1 - \mathrm{e}^{-p(1+o(1))\delta^i})^r \\
&\ge \left(1 - \frac{\delta^i}{N}\right)^{r(1+o(1))} (1 - \mathrm{e}^{-c(1+o(1))})^r \\
&\overset{(5.101)}{\ge} \left(1 - \frac{\delta^i}{N}\right)^{r(1+o(1))} (1 - \mathrm{e}^{-c})^{r(1+o(1))} \\
&\ge \left(1 - \frac{\delta^i}{N} - \mathrm{e}^{-c}\right)^{r(1+o(1))} \\
&\overset{(5.54)}{=} \gamma_{smd}^{r(1+o(1))}.
\end{aligned} \tag{5.100}$$

We used the fact that when $h_1(N) \to 0$, we have $\frac{1-\mathrm{e}^{-c(1+h_1(N))}}{1-\mathrm{e}^{-c}} \to 1$, which implies that there exists

$h_2(N) \to 0$ with $h_2(N) < \frac{1-\mathrm{e}^{-c(1+h_1(N))}}{1-\mathrm{e}^{-c}} - 1$. Then, taking $h_3(N) = \frac{\log(1+h_2(N))}{\log(1-\mathrm{e}^{-c})} \to 0$ we have

$$\begin{aligned}\frac{(1-\mathrm{e}^{-c(1+h_1(N))})^r}{(1-\mathrm{e}^{-c})^{r(1+h_3(N))}} &= \left(\frac{1-\mathrm{e}^{-c(1+h_1(N))}}{1-\mathrm{e}^{-c}}(1-\mathrm{e}^{-c})^{-h_3(N)}\right)^r \\ &> ((1+h_2(N))(1-\mathrm{e}^{-c})^{-h_3(N)})^r = 1. \qquad (5.101)\end{aligned}$$

(ii) In the case $\mathrm{e}^{-c} \geq 1-\mathrm{e}^{-c} - \frac{\delta^i}{N}$, since by equation (5.86) we have $l^\star = i+2$,

$$\begin{aligned}\mathbb{P}(Z_v = 1) &= \prod_{l=1}^{l^\star-1} (1-p)^{r\delta^{l-1}(1+C\zeta)} \\ &\geq \mathrm{e}^{-p(1+o(1))r\delta^i} \\ &\geq (\mathrm{e}^{-c})^{r(1+o(1))} \\ &\overset{(5.54)}{=} \gamma_{smd}^{r(1+o(1))}. \qquad (5.102)\end{aligned}$$

Combining equations (5.100) and (5.102) we can deduce that

$$\mathbb{P}(Z_v = 1) \geq \gamma_{smd}^{r(1+o(1))}. \qquad (5.103)$$

Let $Z = \sum_{v\in V\setminus R} Z_v$ be the number of survivors in the SSP. By equation (5.103) we have

$$\mathbb{E}[Z] \geq (N-r)\gamma_{smd}^{r(1+o(1))}. \qquad (5.104)$$

We finish with the computation similarly to equation (5.40) in Section 5.3.5. The $o(1)$ term will be swallowed by the ϵ term in r. In particular, we will need the inequality

$$(1+o(1))(1-\epsilon) < (1-\epsilon/2), \qquad (5.105)$$

which holds because we can choose an ϵ that tends to zero slower than the function hidden in the o(1) term. Then, putting it all together,

$$\begin{aligned}
(\mathbb{P}(\mathcal{R}) - \mathbb{P}(\overline{\mathcal{E}}))\mathbb{P}(\mathcal{E}) &\overset{(5.99)}{\le} N^r \mathbb{P}(Y_2 = 0) \\
&\overset{(5.15)}{\le} N^r 2\mathrm{e}^{\frac{4}{3}} \mathrm{e}^{-\frac{\mathbb{E}[Z]}{3}} \\
&\overset{(5.104)}{=} 2\mathrm{e}^{\frac{4}{3}} \exp\left(r\log(N) - \frac{1}{3}(N-r)\gamma_{smd}^{r(1+o(1))} \right) \\
&= 2\mathrm{e}^{\frac{4}{3}} \exp\left(r\log(N) - \frac{1}{3}(N-r)\gamma_{smd}^{(1+o(1))(1-\epsilon)\frac{\log(N(1-\gamma_{smd})}{\log(1/\gamma_{smd})}} \right) \\
&\overset{(5.105)}{\le} 2\mathrm{e}^{\frac{4}{3}} \exp\left(r\log(N) - \frac{1}{3}(N-r)\gamma_{smd}^{(1-\epsilon/2)\frac{\log(N(1-\gamma_{smd}))}{\log(1/\gamma_{smd})}} \right) \\
&= 2\mathrm{e}^{\frac{4}{3}} \exp\left(r\left(\log(N) + \frac{1}{3}N^{-(1-\epsilon/2)}\right) - \frac{1}{3}N(N(1-\gamma_{smd}))^{\epsilon-1} \right) \\
&\le 2\mathrm{e}^{\frac{4}{3}} \exp\left(\frac{2\log^2(N)}{\log(1/\gamma_{smd})} - \frac{1}{3}\frac{N^{\epsilon/2}}{1-\gamma_{smd}} \right) \\
&\le 2\mathrm{e}^{\frac{4}{3}} \exp\left(\frac{\log^3(N) - \frac{1}{3}N^{\epsilon/2}}{\log(1/\gamma_{smd})} \right) \to 0 \qquad (5.106)
\end{aligned}$$

since $\frac{1}{\log(1/\gamma_{smd})} > 1$ and $\log^3(N) - \frac{1}{3}N^{\epsilon} \to -\infty$ as long as $\epsilon \gg \frac{\log\log(N)}{\log(N)}$. Finally, since $\mathbb{P}(\mathcal{E}) \to 1$, we have that there exists no resolving set a.a.s. □

5.5.5 Proof of Theorem 5.5.1 for $p = o(1)$, Upper Bound

If $c \to \infty$, the theorem follows directly by SMD ≤ MD and the results of [36]. For the remainder of this proof, we assume $c = \Theta(1)$.

It would be ideal to reduce this proof to the SQC proof, similarly to the lower bound. However, in case $p = \Theta(N^{-\frac{i}{i+1}})$ with $i > 0$, the same approach as in Section 5.3.4 does not work. The events that two nodes v and w separate a set W are not independent anymore and we cannot show that every set W has an f-separator. Fortunately, we do not need such a powerful result for Theorem 5.5.1; it is enough to prove the existence of f-separators for the subsets that can be candidates set in MAX-GAIN. In order to know which subsets can be candidate sets, we expose most of the graph G, except that we reserve a small set F of $c_\gamma \log^2(N)$ nodes with $c_\gamma = 2/\log(1/\gamma_{smd})$, which we keep completely unexposed. The advantage of this is twofold. Now we can have a very good idea about which sets we need to separate, and we still have a large enough set of unexposed nodes that can independently separate the potential candidate sets (conditioned on the expansion properties of the exposed graph).

We have to make the claim that we have “a very good idea about which sets we need to separate” rigourous. Also, we must develop tools that allow us to reason about distances in the graph even when a small subset of the nodes are kept unexposed. We start showing that the unexposed nodes are a.a.s. far from each other.

Lemma 5.5.2. *In $G \sim \mathcal{G}(N,p)$, for a randomly selected $F \subset V$ with size $|F| = c_\gamma \log^2(N)$ with $c_\gamma = \frac{2}{\log(1/\gamma_{smd})}$ and for any two nodes $v, w \in F$, we have $d(v,w) \in \{i+1, i+2\}$ a.a.s., with i given in Definition 5.4.2.*

Proof of Lemma 5.5.2. We can sample the $c_\gamma \log^2(N)$ nodes one by one. Each time we sample one, let us also expose its i-neighborhood as done in the proof of Lemma 5.4.3. When we already sampled the first $j < \log^2(N)$ nodes, there are at most $c_\gamma \log^2(N)\delta^i(1+O(\zeta))$ nodes exposed (by Corollary 5.4.1), so the probability that we select an unexposed node is at least $1 - \frac{c_\gamma \log^2(N)\delta^i(1+O(\zeta))}{N}$. The probability that we always select an unexposed node is at least

$$\left(1 - \frac{c_\gamma \log^2(N)\delta^i(1+O(\zeta))}{N}\right)^{c_\gamma \log^2(N)} \to 1, \tag{5.107}$$

because $\delta \geq \log^5(N)$ and $cN = \delta^{i+1}$ implies $\frac{\log^2(N)\delta^i}{N} = \frac{c\log^2(N)}{\delta} \leq \frac{1}{\log^3(N)}$ for N large enough. Since unexposed nodes are always at least $i+1$ distance away from all previous nodes and not more than $i+2$ by Corollary 5.4.2, this completes the proof. □

We now introduce the necessary definitions to reason about distances in G with a small subset of the nodes F unexposed.

Definition 5.5.5 (exposed graph)**.** *Let $V' = V \setminus F$, let G' be the subgraph of G restricted to nodes V'. Let $N' = |V'|, \delta_2, c', i', \gamma'_{smd}, \zeta'$ be the parameters defined in Definition 5.4.2 and equation* (5.54) *for graph G'. Let $d'(u,v)$ be the length of shortest path between nodes $v \in F$ and $u \in V'$. For $v \in V'$, $\mathcal{S}_{G'}(v,l)$ is defined in Definition 5.4.1 for graph G'. For $v \in F$ let us extend the definition of $\mathcal{S}_{G'}(v,l)$ using the distance function d' instead of d.*

In the rest of the section we will mainly use parameters $N', \delta_2, c', i', \gamma'_{smd}, \zeta'$. We must keep in mind, that to prove Theorem 5.5.1 we must show

$$\text{SMD} \leq (1+o(1))\log(N)/\log(1/\gamma_{smd}).$$

Fortunately, we can show that $\gamma'_{smd} = \gamma_{smd}^{1+o(1)}$, which means that proving $\text{SMD} \leq (1+o(1))\log(N')/\log(1/\gamma'_{smd})$ is enough for the theorem to hold. The following lemma will also show that $i' = i$ (consequently, we will not use the notation i' after the following lemma).

Lemma 5.5.3. *With the definitions given in Definition 5.5.5 we have*

$$i' = i \quad \textit{and} \quad \gamma'_{smd} = \gamma_{smd}^{1+o(1)}. \tag{5.108}$$

Proof of Lemma 5.5.3. First, we show that for any constant $k \geq 1$, we have

$$\delta^k = (Np)^k \geq (N - c_\gamma \log^2(N))^k p^k = \delta_2^k \geq \delta^k \left(1 - \frac{kc_\gamma \log^2(N)}{N}\right). \tag{5.109}$$

Only the last inequality is not trivial. For the last inequality we note that since x^k is a convex function for $k > 1$,

$$\delta^k - \delta_2^k = (Np)^k - (N - c_\gamma \log^2(N))^k p^k \leq \left(c_\gamma \log^2(N) k N^{k-1}\right) p^k = \delta^k \frac{k c_\gamma \log^2(N)}{N}. \tag{5.110}$$

Equation (5.109) implies that $\delta^i / N = \Theta(1)$ if and only if $\delta_2^i / N' = \Theta(1)$, hence $i' = i$. Moreover,

$$\frac{N}{N'} c = \frac{\delta^{i+1}}{N'} > c' = \frac{\delta_2^{i+1}}{N'} \geq \frac{\delta^{i+1}}{N} \left(1 - \frac{i c_\gamma \log^2(N)}{N}\right) \geq c \left(1 - \frac{c_\gamma \log^3(N)}{N}\right), \tag{5.111}$$

since $i \leq \log(N)$. Equations (5.111) and (5.54) and imply equation (5.108). □

Definition 5.5.5 also introduces the distance function $d'(u,v)$ on G' (and one extra node from F). This function will be useful for us, first because it does not use any edge incident to $F \setminus \{v\}$, and therefore can be evaluated even if none of the edges incident to $F \setminus \{v\}$ are exposed, and second because we can prove that with high probability it is the same as the true distance. For the rest of this section, all expectations and probabilities are conditioned on the event that the expansion properties hold in the exposed graph G'.

Lemma 5.5.4. *In $\mathcal{G}(N, p)$, for a randomly selected $F \subset V$ with size $|F| = c_\gamma \log^2(N)$ with $c_\gamma = \frac{2}{\log(1/\gamma'_{smd})}$ for any two nodes $v \in F$ and $u \in V'$, we have $d'(u,v) = d(u,v)$ a.a.s.*

Proof of Lemma 5.5.4. Since both $d(u,v)$ and $d'(u,v)$ represent distances, and the only difference is that the former is the distance in G and latter is the distance in a subgraph of G, we must have $d(u,v) \leq d'(u,v)$. Suppose that for some $v \in F, u \in V'$ we have $d(u,v) < d'(u,v)$. This can happen only if there exists $w \in F \setminus \{v\}$ for which $d(v,w) + d(w,u) < d'(u,v)$. By Lemma 5.5.2 we have that $d(v,w) \geq i + 1$ a.a.s., and we also know that $d(u,w) \geq 1$ since $w \neq u$. If we could also show $d'(u,v) \leq i + 2$ a.a.s., the contradiction given by the inequality

$$1 + (i+1) \leq d(v,w) + d(w,u) < d'(u,v) \leq i + 2$$

would prove that no such w can exist a.a.s. Unfortunately, we cannot simply apply Corollary 5.4.2 to show $d'(u,v) \leq i+2$ a.a.s., since $d'(u,v)$ could be larger than $d(u,v)$. However, a simple analysis conditioning on the expansion properties suffices on the set V' suffices. Indeed,

$$\begin{aligned}
\mathbb{P}(\exists v \in F, u \in V' \text{ with } d'(u,v) > i+2) &\leq \sum_{u \in V'} \mathbb{P}(\exists v \in F \text{ with } d'(u,v) > i+2)) \\
&= \sum_{u \in V'} \left(1 - \mathbb{P}(\forall v \in F, v \in \bigcup_{l=1}^{i+2} \mathcal{S}_{G'}(u,l))\right) \\
&= \sum_{u \in V'} \left(1 - \left(1 - (1-p)^{|\bigcup_{l=1}^{i+1} \mathcal{S}_{G'}(u,l)|}\right)^{|F|}\right) \\
&\overset{(5.48)}{=} |V'| \left(1 - \left(1 - \mathrm{e}^{-p(1-\mathrm{e}^{-c'}+o(1))\delta_2^{i+1}}\right)^{c_\gamma \log^2(N)}\right).
\end{aligned} \tag{5.112}$$

By $1-x = \mathrm{e}^{-x(1+O(1))}$ for $x = o(1)$, and $p\delta_2^{i+1} = c'\delta_2$ we proceed to

$$\begin{aligned}
\mathbb{P}(\exists v \in F, u \in V' \text{ with } d'(u,v) > i+2) &\le N\left(1-\mathrm{e}^{-\mathrm{e}^{-c'\delta_2(1-\mathrm{e}^{-c'}+o(1))}(1+o(1))c_\gamma \log^2(N)}\right)\\
&\le N\mathrm{e}^{-c'\delta_2(1-\mathrm{e}^{-c'}+o(1))}(1+o(1))c_\gamma \log^2(N)\\
&\le N^{1-c'\log^3(N)(1-\mathrm{e}^{-c'}+o(1))}(1+o(1))c_\gamma \log^2(N)\\
&\to 0, \qquad (5.113)
\end{aligned}$$

where in the last inequality we used $\delta_2 = N'p \gg N'\log^5(N)/N \gg \log^4(N)$. □

We have proved that we may use $d'(u,v)$ instead of $d(u,v)$, but we still do not know which sets need to be separated. To determine which sets we need to separate, we will simulate the game on the exposed graph. In the j^{th} step, we assume that we have a candidate set, we expose a subset of the reserved nodes $F_j \subset F$ of size log(N), and we select the best reducer v from only the nodes F_j (unless the candidate set is small enough to query the whole set). Then, we consider all possible answers we could get if we selected v as a query and we continue our simulation for each possible scenario (Figure 5.4 (b)). This analysis is different from the proof in Section 5.3.4 where we first proved a structural result for every subset and then proved that the MAX-GAIN algorithm finds the target. This time, the structural argument and the simulation of the algorithm will be intertwined. We simulate all possible scenarios of MAX-GAIN before actually taking observations from Player 2, and we construct the function g from Definition 3.3.2 to form a "game plan" that we can follow later in real-time. To implement this analysis, we need to slightly extend our definitions.

From now on we will index our set of queries as $R_{j,\tilde{v}}$, since we must prepare for observations for any target $\tilde{v}$. We now have the property $|R_{j,\tilde{v}}| = j$ and $R_{j,\tilde{v}} \subset R_{j+1,\tilde{v}}$, for all $\tilde{v} \in V$. We will also define a new version of candidate targets that are indexed by $\tilde{v}$, and which in addition uses the new distance function that we defined above. In this new notion of candidate targets we assume that the target is in the exposed graph V' (the case when the target is in F will be handled at the end of the proof).

Definition 5.5.6. *Given a graph $G = (V,E)$ with unexposed nodes F and queries $R_{j,\tilde{v}}$, the set of pseudo-candidate targets*

$$\mathcal{T}'_{j,\tilde{v}} = \{v \in V' \mid d'(w,v) = d'(w,\tilde{v}) \textit{ for all } w \in R_{j,\tilde{v}}\}.$$

Remark 5.5.3. *Notice that $\tilde{v} \in \mathcal{T}'_{j,\tilde{v}}$ always holds. Also the sets $\mathcal{T}'_{j,\tilde{v}}$ define a partition on V', that is for any $\tilde{v} \in V'$,*

- $w \in \mathcal{T}'_{j,\tilde{v}} \Rightarrow \mathcal{T}'_{j,\tilde{v}} = \mathcal{T}'_{j,w}$
- $w \not\in \mathcal{T}'_{j,\tilde{v}} \Rightarrow \mathcal{T}'_{j,\tilde{v}} \cap \mathcal{T}'_{j,w} = \varnothing.$

This can be seen by an inductive argument. For $j = 0$, all sets $\mathcal{T}'_{0,\tilde{v}}$ coincide with V', as $R_{0,\tilde{v}} = \emptyset$. For step $j+1$, each equivalence class at step j is partitioned further by the new query.

We must also define an analogous notion to extend Definition 5.5.1 of f-separators.

Definition 5.5.7 (f-pseudo-separator)**.** *Let $f(n) \in \mathbb{N} \to \mathbb{R}^+$ be a function. A set of nodes $W \subseteq V'$, $|W| = n$ has an f-pseudo-separator if there is a node $w \in F$ such that*

$$\max_{l \in \mathbb{N}} |W \cap \mathcal{S}_{G'}(w,l)| \le n\gamma'_{smd} + f(n). \tag{5.114}$$

Algorithm 5.5.1: Simulating all scenarios of MAX-GAIN

1. We arbitrarily select $\log(N)$ disjoint sets $F_j \subset V$ of size $\log(N)$ and we let $F = \cup F_j$. We expose all edges of $V' = V \setminus F$.
2. In step $j \ge 0$, we expose the edges of nodes of F_j. For each $\mathcal{T}'_{j,\tilde{v}}$ we pick the best reducer $s_{j,\tilde{v}} \in F_j$ (possibly a different reducer one for each $\mathcal{T}'_{j,\tilde{v}}$) and add $s_{j,\tilde{v}}$ as a query to $R_{j,\tilde{v}}$. In the analysis, we prove that there always exists an f-separator (we define f later), so the new query will also be a $(|\mathcal{T}'_{j,\tilde{v}}|\gamma'_{smd} + f(|\mathcal{T}'_{j,\tilde{v}}|)$-reducer for $\mathcal{T}'_{j,\tilde{v}}$. Selecting it produces the new sets $\mathcal{T}'_{j+1,\tilde{v}}$.
3. When a set $\mathcal{T}'_{j,\tilde{v}}$ reaches size $o(\log(N)/\log(1/\gamma_{smd}))$, we query the entire set (see the proof of Lemma 5.3.1, for the base case).

j	$\mathcal{T}'_{j,v_1}$	$\mathcal{T}'_{j,v_2}$	$\mathcal{T}'_{j,v_3}$	$\mathcal{T}'_{j,v_4}$
0	V'	V'	V'	V'
1	$\{v_1, v_2\}$	$\{v_1, v_2\}$	$\{v_3, v_4\}$	$\{v_3, v_4\}$
2	$\{v_1\}$	$\{v_2\}$	$\{v_3\}$	$\{v_4\}$

j	R_{j,v_1}	R_{j,v_2}	R_{j,v_3}	R_{j,v_4}
0	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
1	$\{v_5\}$	$\{v_5\}$	$\{v_5\}$	$\{v_5\}$
2	$\{v_5, v_7\}$	$\{v_5, v_7\}$	$\{v_5, v_8\}$	$\{v_5, v_8\}$

Table 5.3: The query sets corresponding to each $\tilde{v}$ and j from the example in Figure 5.4.

With these definitions, the simulation of MAX-GAIN is defined in Algorithm 5.5.1. We must show that while constructing our “game plan”, it is in fact possible with probability tending to one to select an f-pseudo-separator $s_{j,\tilde{v}}$ in each step of the algorithm only from the F_j (we define f later). Then Lemma 5.5.4 implies that these f-pseudo-separators for the pseudo-candidate sets are in fact f-separators for the true candidate sets. This will allow us to use Lemma 5.3.1.

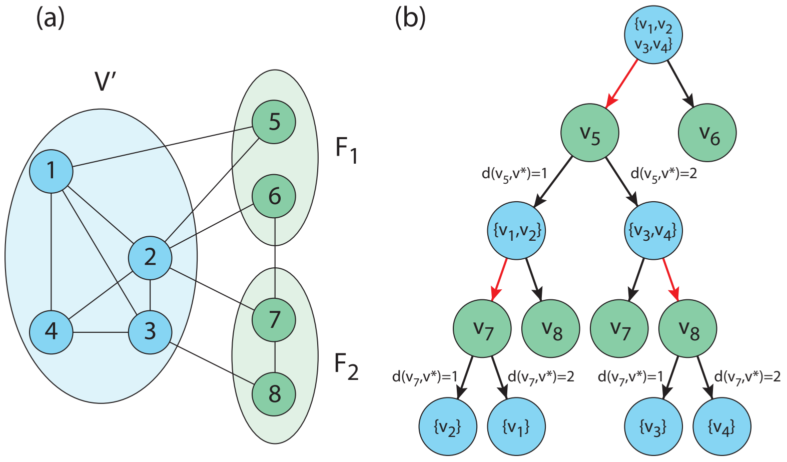


Figure 5.4: (a) A small example graph with V', F_1 and F_2. (b) The "game plan" corresponding to the graph in (a). The j^{th} blue layer ($j \geq 0$) shows the potential pseudo-candidate sets we might encounter in step j. Arrows exiting blue nodes point to the potential of queries F_j (green nodes on the j^{th} level). The red arrow marks the query we picked. Arrows exiting green nodes correspond to the potential observations provided by Player 2 in the actual game. Each scenario ends when the potential candidate set has exactly one element. Tables 5.2 and 5.3 show which sets $\mathcal{T}'_{j,\tilde{v}}$ and $R_{j,\tilde{v}}$ correspond to each step of this "game plan".

Lemma 5.5.5. *Let $n = |\mathcal{T}'_{j,\tilde{v}}|$ be the cardinality of the pseudo-candidate target set in step j of the game plan for target $\tilde{v}$. Let $v \in F_j$, let X_{vw} be the indicator of the event $v \in \mathcal{S}_{G'}(w, i+2)$ for $w \in \mathcal{T}'_{j,\tilde{v}}$, and let $X_v = \sum_{w \in \mathcal{T}'_{j,\tilde{v}}} X_{vw} = |\mathcal{S}_{G'}(v, i+2) \cap \mathcal{T}'_{j,\tilde{v}}|$. Then,*

$$\mathbb{E}[X_v] = n(\mathrm{e}^{-c'} + O(\zeta')). \tag{5.115}$$

Proof of Lemma 5.5.5. This is an analogous result to equation (5.50) in Lemma 5.4.3, except that now most of the graph is exposed. Consider $w \in \mathcal{T}'_{j,\tilde{v}}$, then

$$\begin{aligned}\mathbb{P}(X_{vw} = 1) &= (1-p)^{|\cup_{j=1}^{i} \mathcal{S}_{G'}(w,j)|} \\ &= \mathrm{e}^{(-p+O(p^2))(\delta_2^i(1+O(\zeta')))} \\ &= \mathrm{e}^{-c'}(1+O(\zeta'))) \\ &= \mathrm{e}^{-c'} + O(\zeta'). \end{aligned} \tag{5.116}$$

The result on the expectation follows immediately. □

The previous lemma only covered the expectation. Now we will establish a result on concentration.

Lemma 5.5.6. *Let v, X_v, X_{vw}, n be defined as in Lemma 5.5.5, and let ω be a function tending slowly to infinity. Then,*

$$\mathbb{P}\left(|X_v - n\mathrm{e}^{-c'}| > \frac{n}{\omega\log(n)}\right) \to 0 \tag{5.117}$$

as $N \to \infty$ independently of n.

Proof of Lemma 5.5.6. By Lemma 5.5.5 and Chebyshev's inequality,

$$\begin{aligned}
&\mathbb{P}\left(|X_v - n\mathrm{e}^{-c'}| > \frac{n}{\omega\log(n)}\right)\\
&\le \mathbb{P}\left(|X_v - n(\mathrm{e}^{-c'} + O(\zeta'))| > \frac{n}{2\omega\log(n)}\right)\\
&= \mathbb{P}\left(|X_v - \mathbb{E}[X_v]| > \frac{n}{2\omega\log(n)}\right) < \frac{4\omega^2\log^2(n)\mathrm{Var}[X_v]}{n^2}.
\end{aligned} \tag{5.118}$$

To compute $\mathrm{Var}[X_v]$ we will need

$$\begin{aligned}
\mathbb{E}[X_{vw}X_{vx}] &= \mathbb{P}(d(v,w) = i+2 \text{ and } d(v,x) = i+2)\\
&= (1-p)^{|\cup_{j=1}^{i}\mathcal{S}_{G'}(w,x,j)|}\\
&\overset{(5.43)}{=} \mathrm{e}^{(-p+O(p^2))(2\delta_2^i+O(\zeta'))}\\
&= \mathrm{e}^{-2c'}(1+O(\zeta'))\\
&= \mathrm{e}^{-2c'} + O(\zeta').
\end{aligned} \tag{5.119}$$

Then,

$$\begin{aligned}
\mathrm{Var}[X_v] &= \mathbb{E}[X_v^2] - \mathbb{E}^2[X_v]\\
&= \sum_{w\in\mathcal{T}'_{j,\tilde{v}}} \left(\mathbb{E}[X_{vw}^2] - \mathbb{E}^2[X_{vw}]\right)\\
&\qquad + \sum_{w,x} (\mathbb{E}[X_{vw}X_{vx}] - \mathbb{E}[X_{vw}]\mathbb{E}[X_{vx}])\\
&\le \mathbb{E}[X_v] + n^2(\mathrm{e}^{-2c'} - \mathrm{e}^{-c'}\mathrm{e}^{-c'} + O(\zeta'))\\
&= n\mathrm{e}^{-c'} + n^2 O(\zeta').
\end{aligned} \tag{5.120}$$

Recall that we assumed $c = \Theta(1)$ at beginning of the section, and that in equation (5.111) we proved that $c' = \Theta(1)$ must hold as well. Consequently, we have

$$\frac{\delta_2^i}{N'} = \Theta\left(\frac{1}{\delta_2}\right) = o\left(\sqrt{\frac{\log(N')}{\delta_2}}\right),$$

which implies that we can choose $\zeta' = \sqrt{\log(N')/\delta_2}$ in the version of equation (5.41) where we replace the parameters ζ, δ, N by the parameters ζ', δ_2, N'. Then, by the assumption $pN \gg$

$\log^5(N)$ in the statement of Theorem 5.5.1, we have

$$\zeta' = \sqrt{\frac{\log(N')}{\delta_2}} = o\left(\sqrt{\frac{N\log(N')}{N'\log^5(N)}}\right) = o\left(\frac{1}{\log^2(N)}\right). \tag{5.121}$$

Finally, substituting equation (5.120) into (5.118) yields

$$\begin{aligned}\mathbb{P}\left(|X_v - n\mathrm{e}^{-c'}| > \frac{n}{\omega\log(n)}\right) &< \frac{4\omega^2\log^2(n)\mathrm{Var}[X_v]}{n^2} \\ &\overset{(5.120)}{=} \frac{4\omega^2\log^2(n)n\mathrm{e}^{-c'}}{n^2} + \frac{4\omega^2\log^2(n)n^2O(\zeta')}{n^2} \\ &\overset{(5.121)}{=} o(1) + o\left(\frac{\omega^2\log^2(n)}{\log^2(N)}\right),\end{aligned} \tag{5.122}$$

which proves the desired result since $N \geq n$ and ω tends to infinity very slowly. □

Lemma 5.5.7. *Let Z be the indicator variable that we cannot select an f-pseudo-separator in some step of the simulation with $f(n) = 2n/(\omega\log(n))$. Then $\mathbb{P}(Z) \to 0$.*

Proof of Lemma 5.5.7. Let Z_j be the indicator variable that we cannot select an f-pseudo-separator in the j^{th} step of the simulation. Let us fix j. Let $Y_{j,\tilde{v}}$ be the indicator variable that we cannot find an f-pseudo-separator for the pseudo-candidate set $\mathcal{T}'_{j,\tilde{v}}$. Since by Remark 5.5.3 the pseudo-candidate sets partition V', some (for $j=1$ all) of the $Y_{j,\tilde{v}}$ can be identical, but this will not matter as in the end we will apply a union bound. Similarly to the proof of the SQC upper bound, finding an f-pseudo-separator is equivalent to finding an X_v close to its expectation. Indeed, $|X_v - n\mathrm{e}^{-c'}| \leq f(n)/2$ implies

$$X_v \leq n\mathrm{e}^{-c'} + \frac{f(n)}{2} \leq n\gamma'_{smd} + \frac{f(n)}{2} < n\gamma'_{smd} + f(n) \tag{5.123}$$

and

$$n - X_v \leq n(1-\mathrm{e}^{-c'}) + \frac{f(n)}{2} \leq n\gamma'_{smd} + n\frac{\delta_2^i}{N'} + \frac{f(n)}{2} < n\gamma'_{smd} + f(n), \tag{5.124}$$

because, as we saw in equation (5.121), we can choose $f(n)$ so that $\delta_2^i/N' < \zeta' < 1/\log^2(N) < f(n)/(2n)$. Thus, v is an f-pseudo-separator. The non-existence of an f-pseudo-separator implies the non-existence of an X_v close to its expectation, which means

$$\mathbb{P}(Y_{j,\tilde{v}}) \leq \mathbb{P}\left(|X_v - n\mathrm{e}^{-c'}| > \frac{f(n)}{2} \quad \forall v \in F_j\right). \tag{5.125}$$

Let us choose N large enough such that for $v \in F_j$

$$\mathbb{P}\left(|X_v - n\mathrm{e}^{-c'}| > \frac{f(n)}{2}\right) < \mathrm{e}^{-2} \tag{5.126}$$

(which can be done for any constant by Lemma 5.5.6 since $f(n)/2 = n/(\omega\log(n))$). Then,

$$\mathbb{P}(Y_{j,\tilde{v}}) \le \mathrm{e}^{-2|F_j|} = N^{-2}. \tag{5.127}$$

By union bound, since in every step we have at most $N' < N$ pseudo-candidate sets $\mathcal{T}'_{j,\tilde{v}}$ to separate,

$$\mathbb{P}(Z_j) = \mathbb{P}\left(\bigcup_{\tilde{v}\in|V'|} Y_{j,\tilde{v}}\right) \le N\mathbb{P}(Y_{j,\tilde{v}}) \le N^{-1}. \tag{5.128}$$

Finally, since we have $\frac{2\log N}{\log(1/\gamma_{smd})}$ sets F_j, another union bound shows that

$$\mathbb{P}(Z) = \mathbb{P}\left(\bigcup_{j=1}^{\frac{2\log N}{\log(1/\gamma_{smd})}} Z_j\right) \le \frac{2\log N}{N\log(1/\gamma_{smd})} = o(1). \tag{5.129}$$

□

Now we just need to put the pieces together to prove Theorem 5.5.1.

Proof of Theorem 5.5.1, upper bound. We perform Algorithm 5.5.1, with the modification that besides F, we also reserve another set F' with $c_\gamma \log^2\log^2(N)$ nodes. The modified algorithm runs in three steps.

(i) We run Algorithm 5.5.1 on $V' = V \setminus (F \cup F')$. The additional set F' slightly increases the size of the reserved nodes, but this $\log^2\log^2(N)$ term does not affect the analysis. Lemma 5.5.7 ensures that the algorithm can find an f-pseudo-separator for all F_j with probability tending to 1. Lemma 5.5.4 shows that the only candidate sets we might encounter in the MAX-GAIN algorithm are pseudo-candidate target sets in our game plan, and the f-pseudo-separator we found for the pseudo-candidate target sets are f-separators for the corresponding candidate target sets, unless the source was in the reserved nodes.

Since the f we used in Lemma 5.5.7 was $o(n/\log(n))$ we can apply Lemma 5.3.1, which shows that in each possible scenario we simulate, we find the source in

$$(1+o(1))\frac{\log(N')}{\log(1/\gamma'_{smd})}$$

steps. Therefore the number of steps we require is always is less than

$$\frac{2\log(N)}{\log(1/\gamma_{smd})},$$

the number of sets F_j we can use to find f-separators in Lemma 5.5.7. Thus, if the target was in V', the algorithm will find it. By Lemma 5.5.3, $\gamma'_{smd} = \gamma_{smd}^{1+o(1)}$, hence the number steps taken is upper bounded by the desired

$$(1+o(1))\frac{\log(N)}{\log(1/\gamma_{smd})}$$

steps.

(ii) We repeat the argument with candidate set F and reserved nodes F'.

(iii) Finally we query the entire F'.

In this last two steps, we selected only $o(\log(N))$ extra queries, which does not change the leading term of our upper bound. We ensured that no matter whether the source is in $V \setminus (F \cup F')$, F or F', we will be able to find it in the desired number of steps with probability tending to 1. Recall that in all of our calculations in Lemmas 5.5.4-5.5.7 we conditioned on the event that the expansion properties hold in the exposed graph. Since the expansion properties also hold with high probability, the upper bound in Theorem 5.5.1 holds also without conditioning. □

5.6 Discussion

In this chapter, we proved tight asymptotic results for the SMD in $\mathcal{G}(N,p)$. We found that a.a.s., the ratio between the SMD and the MD is a constant as N tends to infinity, and we conjecture that this constant is 1 except for $(pN)^i = \Theta(N)$ for $i \in \mathbb{N}$, where the constant term is found explicitly and is smaller than 1. On the one hand, considering the equivalence of binary search with adaptive and non-adaptive queries, it is interesting that there is any difference at all between the SMD and the MD. On the other hand, experimental results suggest that on other graph models (and especially real-world networks), the SMD is orders of magnitude smaller than the MD [222]. Hence, the Erdős-Rényi graphs are an intermediate regime, where the restriction on the queries does favor adaptive algorithms, but not by too much.

It would be interesting to study random graph models other than the $\mathcal{G}(N,p)$ model, where we expect the difference between the MD and the SMD to be significantly larger. Adding noise to the measurements would be another step towards more realistic scenarios, and in this case too, we expect a larger difference between the MD and the SMD. The noise can come from faulty observers similarly to [80], or the noise can be proportional to the distances observed which would model stochastic disease propagation in source localization [156, 223].

Stochastic Spreading Part III

6 The Power of Adaptivity on the Path

In this chapter, we prove rigorous upper and lower bounds on the number of queries required in source identification in the stochastic S2 framework introduced in Chapter 1, both in the non-adaptive and in the adaptive settings. See Chapter 1 for a general review on why the S2 framework is important in the field of source identification, and Section 6.6 for a review of related works in other fields. Since these are the first rigorous results on the query complexity of stochastic source identification with time queries, we focus on one of the most elementary graphs, the path graph, which already poses significant challenges, especially in the case of the adaptive lower bound.

This chapter is based on the publication [156] by Lecomte, Ódor and Thiran.

6.1 Model

We consider the S2 source identification framework, which was introduced together with the S1 framework in Chapter 1. We have several reasons to focus on the S2 framework instead of the S1, including that S2 is easier to define and it is theoretically more appealing, as pointed out by several papers in the field [248, 188], and that there is little difference between the number of queries required in the two frameworks [226]. We further discuss the differences between the two frameworks and how our results can be extended to S1 in Appendix C.2.

Although the S2 framework was already defined in Chapter 1, we repeat the definitions here for convenience, and we introduce a few new notations that will be useful in this chapter (see Figure 6.1 for an illustration). In this chapter, we treat the case when the underlying graph $G = (V, E)$ is an n-node path, with nodes numbered from 1 to n (so $V = \{1, \ldots, n\}$). We will assume $n \geq 3$ for convenience. We will often say that a node u is to the "left" (respectively, "right") of a node v if $u < v$ (resp., $u > v$).

First, a node $v^* \in V$ is picked uniformly at random, which is called the *source*. Then for each edge $e \in E$, a weight $w(e)$ is drawn independently from some distribution $\mathcal{W}$. Both v^* and the weights $w(e)$ are hidden from the identification algorithm. Once they are drawn, the algorithm

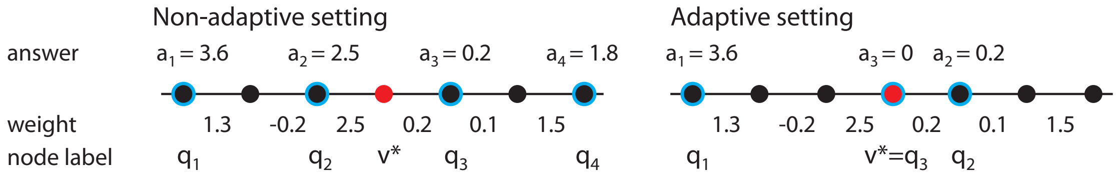


Figure 6.1: An illustration of the stochastic S2 model on the path graph. Query nodes are marked blue and the source is marked red.

will start making queries to identify v^*. To perform a query, the algorithm chooses a query node $q \in V$, and receives an *answer* with value $\mathrm{ans}_w(v^*, q)$: the shortest distance between v^* and q in graph G with edges weighted by w.

We distinguish between two settings. In the *non-adaptive* setting, the algorithm has to submit all of its queries in one batch, then receives all answers, and has to make a prediction. In the *adaptive* setting, the algorithm can make queries one by one, and adapt the choice of the next query based on previous answers. In both settings, the weights $w(e)$ are only drawn once, at the very beginning, and will not change between queries. As we will see, the difference between these two settings will have a huge impact on the number of queries that the algorithm needs, because in the adaptive setting the algorithm will be able to quickly zero in on the source v^* and receive progressively more refined information.

For the weight distribution, we choose $\mathcal{W} = \mathcal{N}(1, \sigma^2)$: a normal of mean 1 and variance $\sigma^2 > 0$, where σ is a parameter of the model. We choose a normal distribution because (i) by the Central Limit Theorem, the distances between faraway nodes converge to a normal distribution for most edge-delay distributions $\mathcal{W}$, while close-by nodes can be searched via exhaustive search (we discuss how to extend our results to these other distributions in Appendix C.1), (ii) there are several properties (e.g. additivity, tight concentration) of the normal distribution that simplify our calculations. One should note that the weights can take negative values, especially when σ is large, in contradiction with the non-negativity of propagation delays. Letting weights take negative values further accounts for the randomness in the incubation and reporting times. It makes source identification more challenging because of the absence of a deterministic, monotone dependence between the time of infection of a query node and its distance to the source. We discuss how to extend our results to propagation delay distributions that only take positive values, in Appendix C.1.

6.2 Results and Discussion

6.2.1 Non-Adaptive Setting

We present matching upper and lower bounds for the non-adaptive setting.

Theorem 6.2.1. *For any failure probability* $0 < \delta < 1/2$, *there is a deterministic algorithm for* *non-adaptive source identification on the n-node path which asks* $\min(O(1 + n\sigma^2 \log(1/\delta)), n)$ *queries and identifies v^* correctly with probability at least* $1-\delta$, *even if v^* is chosen adversarially instead of drawn uniformly at random.*

Theorem 6.2.2. *For any success probability $p > 1/n$, any (potentially randomized[1]) algorithm for non-adaptive source identification on the n-node path must ask* $\Omega(1 + \min(p^3 n\sigma^2, p^2 n)) = \Omega_p(1 + \min(n\sigma^2, n))$ *queries to identify v^* correctly with probability at least p when v^* is drawn uniformly at random.*

We can interpret the results in the following way. Intuitively, answers are "accurate" up to distance roughly $1/\sigma^2$: indeed, for a query node at distance d from the source, the mean of the received answer is d and the variance is $d\sigma^2$, so if $d = \omega(1/\sigma^2)$, the variance becomes $\omega(1)$ and it becomes impossible to deduce the real distance with constant probability. Therefore, instead of thinking of receiving stochastic answers, we can imagine that we receive the exact distance, but only if this distance is $\leq 1/\sigma^2$ (and otherwise they would give no answer at all). That is, we think of the queries as effective within a "limited range" $1/\sigma^2$. In that model, it is clear that $\Theta(1 + \min(n\sigma^2, n))$ queries are necessary and sufficient, which is exactly what we find in Theorems 6.2.1 and 6.2.2.

Our proofs also build on the intuition of query nodes with limited range. In the proof of Theorem 6.2.1, we show that if the query nodes are spaced equally (and deterministically), every v^* is in the "range" of the closest two query nodes, meaning that once rounded to the closest integer, they give the correct answer with high probability. This probability is based only on the randomness of the edge weights $w(\cdot)$, and is high no matter where v^* ends up, hence we can identify v^* even without assuming any prior on its distribution. This extra guarantee was not required by the model, but it comes naturally without any additional cost.

In contrast, in Theorem 6.2.2 we show that any algorithm that succeeds with constant probability must use $\Omega(1 + \min(n\sigma^2, n))$ queries even if the algorithm is allowed to take advantage of the assumption that v^* is uniformly distributed over the nodes V. The proof works by showing that if one uses fewer queries, then most of the nodes are so far away from the closest query node that they are indistinguishable from other close-by nodes.

While in this chapter we only consider the path graph, these results indicate that limited range query nodes might be a good proxy for non-adaptive source identification in other graphs as well, which has not been thoroughly explored in the source identification literature.

[1] Since the distribution for the identity of the source is fixed, rather than adversarial, randomness in the algorithm is not useful (as long as we are not considering running time): the algorithm should simply choose the set of queries that maximizes the probability of finding the source, and output the likeliest source given the answers it receives.

6.2.2 Adaptive Setting

The adaptive setting is more complex and more interesting than the non-adaptive case. For instance, it is not obvious anymore how the queries should be selected. We may consider an algorithm that, at each decision, selects a query node based on the posterior probabilities that the source is at some node: we call *posterior* at a node v the probability that the source is v, conditioned on the answers made so far. However, those posteriors might be hard to compute, and might not be well-behaved as a function of v (for example, they might not be unimodal). Fortunately, as long as the variance of the edge-delays is relatively low, the answers that we see are concentrated around their expected value and we can form a fairly good idea about what the posteriors might look like. This inspires the following algorithm, which (for intuition) can be seen as a procedure that computes at each step the posteriors approximately, and selects the next query node close to the node with the highest posterior (the node that is most likely to be the source).

Theorem 6.2.3. *For any failure probability* $0 < \delta < 1/2$*, there is a deterministic algorithm for adaptive source identification on the* n*-node path which uses*

$$\begin{cases} O(\log(1+\log_{1/\sigma} n) + \operatorname{polylog}(1/\delta)) \text{ queries if } \sigma^2 \leq 1/2 \\ O(\log\log n + \sigma^2 \cdot \operatorname{polylog}(\sigma, 1/\delta)) \text{ queries if } \sigma^2 \geq 1/2 \end{cases}$$

and identifies v^* *correctly with probability at least* $1-\delta$*, even if* v^* *is chosen adversarially instead of drawn uniformly at random.*

To show optimality, as we did in the non-adaptive case, we show that no algorithm can succeed without asking a large number of queries, even under the assumption that v^* is uniformly distributed over V.

Theorem 6.2.4. *For any success probability* $p > 1/n$ *and any* $n \geq \Theta_p(\max(\sigma^3, 1))$*, any (potentially randomized) algorithm for adaptive source identification on the* n*-node path must use*

$$\begin{cases} \Omega_p(1+\log(1+\log_{1/\sigma} n)) \text{ queries if } \sigma^2 \leq 1/2 \\ \Omega_p(\log\log n) \text{ queries if } \sigma^2 \geq 1/2 \end{cases}$$

to identify v^* *correctly with probability at least* p *when* v^* *is drawn uniformly at random.*

The proof of Theorem 6.2.4 is the most challenging proof we present. Since the algorithm is allowed to adapt each query based on the results of the previous ones, we have to somehow quantify the progress it has made towards identifying v^*. However, this is made delicate by the fact that the edge weights $w(e)$ are drawn only once at the onset, which means that the algorithm does not only accumulate information about v^*, but also about the edge weights $w(e)$. In particular, proof approaches that try to "fool" the algorithm by giving it answers from a modified distribution tend to fail because the algorithm can test for consistency across queries.

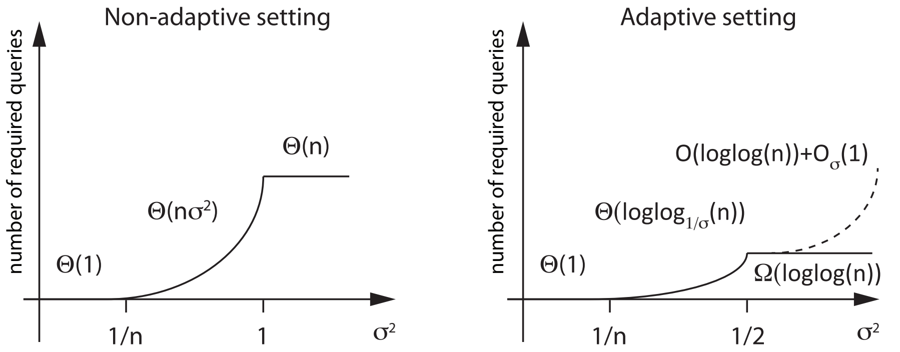


Figure 6.2: A sketch of a linear-log plot of the optimal number of queries in the non-adaptive and adaptive cases as a function of σ.

Instead, the proof that we present builds on a detailed understanding what the posteriors can look like in each step.

The upper and lower bounds in Theorems 6.2.3 and 6.2.4 match for σ up to $\tilde{\Theta}(\log\log n)$. For $\sigma \gg \log\log(n)$, our upper and lower bounds are separated by a $\sigma^2 \cdot \mathrm{polylog}(\sigma)$ term, which comes from the final steps, when the algorithm has gotten so close to the source that the variance of the edge-delays is about as big as the expected value of the answers, at which point the algorithm simply queries every node.

This $\sigma \gg \log\log(n)$ regime is a difficult regime to analyse, because for larger σ we lose the concentration of the answers, and cannot control the shape of the posteriors anymore. Moreover, we believe that in the high-σ regime, any asymptotically tight results for the Gaussian case would not carry over to other edge-delay distributions $\mathcal{W}$, because the $O_{\sigma,\delta}(1)$ term becomes sensitive to the specific $\mathcal{W}$ we pick. As an example, consider $\mathcal{W}$ to be a Gaussian distribution with a very large σ (say, larger than n), truncated at 0 (to prevent negative edge weights w). With this $\mathcal{W}$, we can always figure out which direction the source is from node v by simply querying a neighbor of v, and checking which answer is larger. Hence we can find the source with binary search in $\log_2(n)$ rounds, however, if $\mathcal{W}$ is a non-truncated Gaussian random variable and σ is large enough, we clearly have to query every node to find the source.

Finally, as an additional strength of our results, we note that the lower bounds (Theorems 6.2.2 and 6.2.4) continue to apply even if the algorithm were to also receive direction information (i.e. whether the source is on the left or on the right of the query node), whereas the upper bounds (Theorems 6.2.1 and 6.2.3) work well even without using this information.

6.3 Preliminaries

We denote a normal distribution with mean μ and variance σ^2 as $\mathcal{N}(\mu,\sigma^2)$. We occasionally call variables distributed according to this a normal distribution "Gaussians". We will often use the following basic facts about the normal distribution.

Fact 6.3.1. *If $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent Gaussians, then $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

Fact 6.3.2. *If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{P}[X \notin \mu \pm a] \leq \mathrm{e}^{-\frac{a^2}{2\sigma^2}}$.*

Proof. First, using the probability density function of the normal distribution and the change of variables $z = \frac{x-\mu}{\sigma}$ we have

$$\mathbb{P}[X \notin \mu \pm a] = 2 \times \int_{\mu+a}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \sqrt{\frac{2}{\pi}} \int_{a/\sigma}^{\infty} \mathrm{e}^{-z^2/2} dz.$$

It only remains is to prove that for all $b \geq 0$, $\sqrt{\frac{2}{\pi}} \int_b^\infty \mathrm{e}^{-z^2/2} dz \leq \mathrm{e}^{-b^2/2}$.

We separate into two cases.

- If $b \geq 1$, then we have

$$\begin{aligned} \sqrt{\frac{2}{\pi}} \int_b^\infty \mathrm{e}^{-z^2/2} dz &\leq \sqrt{\frac{2}{\pi}} \int_b^\infty \frac{z}{b} \mathrm{e}^{-z^2/2} dz && (z \geq b) \\ &= \sqrt{\frac{2}{\pi}} \frac{1}{b} \mathrm{e}^{-b^2/2} && \left(\tfrac{d}{dz} \mathrm{e}^{-z^2/2} = -z\mathrm{e}^{z^2/2}\right) \\ &\leq \mathrm{e}^{-b^2/2}. && \left(b \geq 1 \geq \sqrt{\tfrac{2}{\pi}}\right) \end{aligned}$$

- One can easily check that on interval $[0, 1]$, $\sqrt{\frac{2}{\pi}} \int_b^\infty \mathrm{e}^{-z^2/2} dz$ is convex while $\sqrt{\frac{2}{\pi}} \int_b^\infty \mathrm{e}^{-z^2/2} dz$ is concave (by computing their second derivatives), and in addition we have

$$\sqrt{\frac{2}{\pi}} \int_0^\infty \mathrm{e}^{-z^2/2} dz = 1 = \mathrm{e}^{-0^2/2} \quad \text{and} \quad \sqrt{\frac{2}{\pi}} \int_1^\infty \mathrm{e}^{-z^2/2} dz < 0.318 < 0.606 < \mathrm{e}^{-1^2/2}.$$

  Therefore, $\sqrt{\frac{2}{\pi}} \int_b^\infty \mathrm{e}^{-z^2/2} dz \leq \mathrm{e}^{-b^2/2}$ on the whole interval. □

6.4 Proofs for the Non-Adaptive Setting

6.4.1 Upper Bound

Proof of Theorem 6.2.1. First of all, observe that one can always find the source with probability 1 if one is willing to use n queries: just query each node. The query at node v^* will produce answer 0, while the other queries will almost surely produce nonzero answers.

Now it suffices to show that the source can be identified with $O(1 + n\sigma^2 \log(1/\delta))$ queries. The strategy is natural: query $\sim n/d$ nodes along the path at fixed intervals of some length d, where d is small enough to ensure that that the query nodes nearest to the source v^* return an answer that is very close to the real distance (and in particular, that will be exactly equal to it once rounded to the nearest integer).

What makes things a bit more complex is that:

(a) even if all the weights are positive, it may not be easy to determine between which two query nodes v^* is identified;

(b) since the weights may be negative, it is possible that a query node q_1 gives a smaller answer than a query node q_2 even though q_2 lies between q_1 and v^* (in particular, the answers do not necessarily form a unimodal sequence when read from left to right).

Concretely, the algorithm will query nodes $1, d+1, 2d+1, \ldots, 1+\lfloor\frac{n-1}{d}\rfloor d$. It will then find the query node with the smallest answer, which we call q_{smallest}, and the next query node to its left $q_{\text{left}} := q_{\text{smallest}} - d$ (let's assume for now that $q_{\text{smallest}} \neq 1$, so that q_{left} exists). Let $a_{\text{smallest}} := \text{ans}_w(v^*, q_{\text{smallest}})$ and $a_{\text{left}} := \text{ans}_w(v^*, q_{\text{left}})$ be the corresponding answers. Then the algorithm just assumes that both of them are correct (equal to the real distance) once rounded to the nearest integer (that is, $\lfloor a_{\text{smallest}} \rceil = |v^* - q_{\text{smallest}}|$ and $\lfloor a_{\text{left}} \rceil = |v^* - q_{\text{left}}|$), and computes v^* as

$$\begin{cases} q_{\text{smallest}} + \lfloor a_{\text{smallest}} \rceil & \text{if } \lfloor a_{\text{left}} \rceil \geq d \\ q_{\text{smallest}} - \lfloor a_{\text{smallest}} \rceil & \text{otherwise.}^2 \end{cases}$$

For this strategy to work, it is enough if the following statements hold simultaneously:

(a) among the query nodes located at or to the left of v^*, the closest one is the one with the smallest answer;

(b) among the query nodes located at or to the right of v^*, the closest one is the one with the smallest answer;

(c) the two closest query nodes to v^* on its left side and the closest query node on its right side all give a correct answer once rounded to the nearest integer.

Indeed, if this is true, then q_{smallest} will be the closest query node to v^* on either its left or right side, and thus both q_{smallest} and q_{left} will be among the three query nodes that are guaranteed by point (c) to give the correct result once rounded.

The following claim, which is purely technical and easily obtained from concentration bounds, is proved in Appendix C.4.

[2] If q_{left} does not exist, which happens only when $q_{\text{smallest}} = 1$, then the algorithm can simply compute v^* as $1 + \lfloor a_{\text{smallest}} \rceil$, again assuming that a_{smallest} is correct once rounded to the nearest integer.

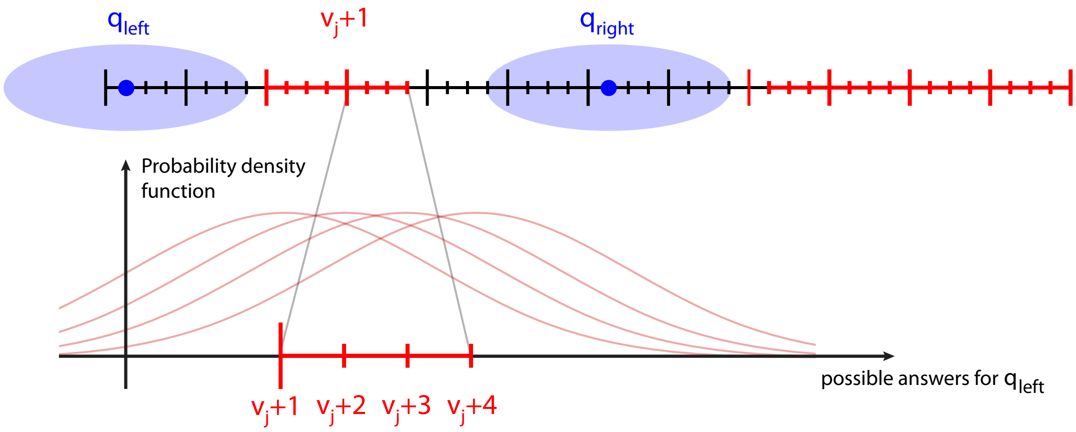


Figure 6.3: An illustration for the proof of Theorem 6.2.2 with $d = 6$ and $k = 4$. At the top of the figure, the graph G is shown with the set Q (the query nodes) and the set $C(d)$ (the "covered" nodes) marked blue, and the set $K(d,k)$ marked red. At the bottom of the figure, the probability density functions of the answers recorded by q_{left} are shown for each candidate source in the highlighted segment. Intuitively, the union of the areas under the red curves corresponds to the probability of success of the optimal source identification algorithm (this is made concrete in Equation (6.2), although we need to also consider the answer from q_{right}, so instead of a single integral we get a double integral). In the proof, we show that if too few nodes are queried, then the red segments will be far from the closest query node, and therefore the red curves will have a large overlap, and their union will be small.

Claim 6.4.1. *For some $d = \Omega\left(\frac{1}{\sigma^2 \log(1/\delta)}\right)$, all of (a), (b), (c) hold simultaneously with probability $\geq 1-\delta$.*

This means that the number of queries used is $\lfloor \frac{n-1}{d} \rfloor + 1 = O(1 + n/d) = O(1 + n\sigma^2 \log(1/\delta))$. □

6.4.2 Lower Bound

Proof of Theorem 6.2.2. Since $p > 1/n$, it is clear that at least one query is necessary (otherwise one could not do better than randomly guessing the source, which gives $p = 1/n$). In the rest of the proof, we show that one needs $\Omega(\min(p^3 n\sigma^2, p^2 n))$ queries to identify the source.

Let us introduce some notation. Let Q be the set of query nodes that the algorithm chooses, let $\text{ans}_w(v^*, Q) := \{\text{ans}_w(v^*, q)\}_{q \in Q}$ be the answers it receives from each query node, and let f be the function that takes in these answers and returns a prediction for v^*. Since we are not bounding the running time of the algorithm, we can assume that both Q and f are deterministic (the algorithm can simply choose the values of Q and f that give the best chance of finding the source), while the answers $\text{ans}_w(v^*, q)$ are random variables depending on both v^* and w (recall that v^* is drawn uniformly in $V = [n]$). The overall success probability of the algorithm is given by $p = \mathbb{P}_{v^*,w}[f(\text{ans}_w(v^*, Q)) = v^*]$. For a fixed node v, let $p(v) :=$

$\mathbb{P}_w[f(\text{ans}_w(v,Q)) = v]$ be the probability that the algorithm will output v conditioned on $v^* = v$. Clearly, $p = \frac{1}{n}\sum_{v\in V} p(v)$.

As discussed before in Section 6.2, our proofs in the non-adaptive case build on the intuition of query nodes with limited range: roughly speaking, we will show that most nodes are further than a distance $1/\sigma^2$ away from the closest query node, and therefore they will be hard to distinguish from their neighbors. Figure 6.3 sketches some of the key points used in the proof.

Let $d > 0$ be an integer which we will fix later in equation (6.6), representing the "range" of the query nodes. Intuitively, nodes outside the range d of any query node (the "uncovered" nodes) might be hard to distinguish, contrary to nodes within the range of a query node (the "covered" nodes). Let $C(d) \subset V$ be the set of nodes that are within distance d of some query node in Q.

In addition, let us subdivide the first $k\lfloor n/k\rfloor$ nodes of V into $\lfloor n/k\rfloor$ segments of length k, where $k > 0$ is an integer that we will fix later, and let $K(d,k) \subseteq V \setminus C(d)$ be the set of nodes contained in the segments that are entirely included in $V \setminus C(d)$ (we will call such segments "uncovered"). Our goal in defining these segments is to show that there are few "covered" segments, and that the source identification problem is hard to solve on "uncovered" segments.

More precisely, to demonstrate that Q needs to be large, we will split the probability of success $p = \frac{1}{n}\sum_{v\in V} p(v)$ into two parts:

- the part due to $v \in V \setminus K(d,k)$ (the "covered" segments), which will be small whenever Q, d and k are small (simply because the set $V \setminus K(d,k)$ will be small);
- the part due to $v \in K(d,k)$ (the "uncovered" segments), which will be small ($\leq p/2$) whenever d and k are large enough.

Concretely,

$$\begin{aligned}
pn &= \sum_{v\in V} p(v)\\
&= \sum_{v\in V\setminus K(d,k)} p(v) + \sum_{v\in K(d,k)} p(v)\\
&\leq |V \setminus K(d,k)| + \sum_{v\in K(d,k)} p(v). && (p(v) \text{ is a probability, so } p(v) \leq 1)\\
&\leq (2d+2k-1)|Q| + (k-1) + \sum_{v\in K(d,k)} p(v) && (6.1)
\end{aligned}$$

where the factor $(2d+2k-1)$ is because each query node covers $\leq 2d+1$ nodes directly, and can affect $\leq 2(k-1)$ more nodes by touching their segment; also, the $+(k-1)$ comes from the $< k$ nodes that were not within the first $k\lfloor n/k\rfloor$ nodes and therefore are not in a segment.

Let us now prove that the sum $\sum_{v\in K(d,k)} p(v)$ is small when d and k are large. This makes intuitive sense: the nodes in $K(d,k)$ are far from the closest query node, so they will be hard to distinguish from each other. To do this, we will use the following lemma.

Claim 6.4.2. *Let $\{v+1,\dots v+k\}$ be a set of k adjacent nodes. Let $q_{left} \in Q$ be the closest query node at or to the left of $v+1$, and let $q_{right} \in Q$ be the closest query node at or to the right of $v+k$. Assume that there are no query nodes between q_{left} and q_{right}, and that $q_{left} \le v-d$ and $q_{right} \ge v+k+d$. Then*

$$\sum_{i=1}^{k} p(v+i) < \frac{2(d+k)\mathrm{e}^{\frac{k^2}{2(d+k)\sigma^2}}}{d}.$$

Proof. Let us consider a scenario where the source is sampled uniformly from $\{v+1,\dots,v+k\}$, and let $f' := \arg\max_f \sum_{i=1}^{k} \mathbb{P}_w[f(\mathrm{ans}_w(v+i,Q)) = v+i]$ be the algorithm that maximizes the success probability in this scenario. Let $p'(v+i) := \mathbb{P}_w[f'(\mathrm{ans}_w(v+i,Q)) = v+i]$, then clearly $\sum_{i=1}^{k} p(v+i) \le \sum_{i=1}^{k} p'(v+i)$ by definition of f'. Also observe that f' will only depend on the answers at q_{left} and q_{right}, since the algorithm already knows that $v^* \in [q_{\mathrm{left}}, q_{\mathrm{right}}]$, and the other query nodes outside $[q_{\mathrm{left}}, q_{\mathrm{right}}]$ do not carry relevant information. Indeed, any answer outside of q_{left} and q_{right} is just the sum of the answer at q_{left} or q_{right} plus some extra term that does not say anything about the identity of v^*. More formally, one could recreate the other answers from just the answers at q_{left} and q_{right} in a way that exactly replicates the original distribution, so we can transform any algorithm that uses all the answers into an algorithm that uses only the answers at q_{left} and q_{right} with the same performance.

For similar reasons, we can assume that $q_{\mathrm{left}} = v-d$ and $q_{\mathrm{right}} = v+k+d$. If the query nodes were any further, that would be more difficult for the algorithm f', because we can simulate that case using the answers at $v-d$ and $v+k+d$.

Since f' maximizes the success probability of estimating the hidden parameter v^* with both the likelihood function and the prior distribution being completely known, the optimal f' computes the posterior distribution using Bayes rule, and picks the v^* that maximizes it (this is called Maximum A Posteriori or MAP estimation) [198]. In our case we have a uniform prior, which implies that f' is simply the Maximum Likelihood Estimator, i.e., for any answer $(a_{\mathrm{left}}, a_{\mathrm{right}})$,

$$f'(a_{\mathrm{left}}, a_{\mathrm{right}}) = \underset{i\in\{1,\dots,k\}}{\arg\max}(g_{v+i}(a_{\mathrm{left}}, a_{\mathrm{right}})),$$

where $g_{v+i}(x,y)$ denotes the probability density function of $\mathcal{W}_{v+i} = (\mathcal{N}(d+i,(d+i)\sigma^2), \mathcal{N}(d+k-i,(d+k-i)\sigma^2)$, the distribution of the answers at q_{left} and q_{right} (note that $\mathcal{W}_{v+i}$ is a pair of *independent* normal distributions). Consequently,

$$\sum_{i=1}^{k} p'(v+i) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \max_{i\in\{1,\dots,k\}}(g_{v+i}(x,y))\,\mathrm{d}x\,\mathrm{d}y. \tag{6.2}$$

Next, we provide the following upper bound to g_{v+i} for every $i \in \{1, \dots, k\}$:

$$\begin{aligned} g_{v+i}(x,y) &= \frac{1}{2\pi\sigma^2\sqrt{(d+i)(d+k-i)}} \exp\left(-\frac{(x-(d+i))^2}{2(d+i)\sigma^2} - \frac{(y-(d+k-i))^2}{2(d+k-i)\sigma^2}\right) \\ &< \frac{1}{2\pi d\sigma^2} \exp\left(-\frac{(x-(d+i))^2 + (y-(d+k-i))^2}{2(d+k)\sigma^2}\right) \end{aligned} \tag{6.3}$$

Notice that if we consider a triangle ABC with $A = (x, y)$, $B = (d+i, d+k-i)$ and $C = (d, d)$, and we denote the side lengths opposite of each point by a, b and c, then the numerator of the exponent in equation (6.3) equals c^2. The following lower bound holds for c^2 based on the law of cosines and elementary algebra:

$$c^2 = a^2 - 2ab\cos(\measuredangle ACB) + b^2 \geq a^2 - 2ab + b^2 \geq \frac{a^2}{2} - b^2.$$

The last inequality can be confirmed if we move all terms to the left side and find the expression $(a/\sqrt{2} - \sqrt{2}b)^2 \geq 0$. After substituting back into a, b and c, since the maximum distance between points $(d+i, d+k-i)$ and (d, d) is k for any $i \in \{1, \dots, k\}$, we get

$$(x-(d+i))^2 + (y-(d+k-i))^2 \geq \frac{1}{2}((x-d)^2 + (y-d)^2) - k^2. \tag{6.4}$$

Substituting equation (6.4) back into equation (6.3) yields

$$\begin{aligned} g_{v+i}(x,y) &< \frac{1}{2\pi d\sigma^2} \exp\left(-\frac{(x-d)^2 + (y-d)^2 - 2k^2}{4(d+k)\sigma^2}\right) \\ &= \frac{2(d+k)\mathrm{e}^{\frac{k^2}{(d+k)2\sigma^2}}}{d} \cdot \frac{1}{2\pi(d+k)2\sigma^2} \exp\left(-\frac{(x-d)^2 + (y-d)^2}{2(d+k)2\sigma^2}\right). \end{aligned} \tag{6.5}$$

Notice that the last line of equation (6.5) can be written as

$$\frac{2(d+k)\mathrm{e}^{\frac{k^2}{2(d+k)\sigma^2}}}{d} g(x,y),$$

where $g(x, y)$ is the probability density function of two independent copies of $\mathcal{N}(d, (d+k)2\sigma^2)$, so its double integral must sum to 1. Thus, plugging this upper bound into equation (6.2), we get

$$\sum_{i=1}^{k} p'(v+i) < \frac{2(d+k)\mathrm{e}^{\frac{k^2}{2(d+k)\sigma^2}}}{d} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\,\mathrm{d}x\,\mathrm{d}y = \frac{2(d+k)\mathrm{e}^{\frac{k^2}{2(d+k)\sigma^2}}}{d}.$$

Since $\sum_{i=1}^{k} p'(v+i)$ is an upper bound on $\sum_{i=1}^{k} p(v+i)$, the proof is completed. □

Since each segment in $K(d, k)$ contains k consecutive nodes that are all a distance d away from the closest query node, we can apply Claim 6.4.2 to each of them, and get

$$\begin{aligned}\sum_{v\in K(d,k)} p(v) &\le \frac{|K(d,k)|}{k}\cdot\frac{2(d+k)\mathrm{e}^{\frac{k^2}{2(d+k)\sigma^2}}}{d}\\ &\le \frac{n}{k}\cdot\frac{2(d+k)\mathrm{e}^{\frac{k^2}{2(d+k)\sigma^2}}}{d}.\end{aligned}$$

In order to make this $\le pn/2$, let us set

$$k := \lceil 16\mathrm{e}/p\rceil \qquad\text{and}\qquad d := \max\left(\left\lceil\frac{k^2}{2\sigma^2\ln(pk/8)}\right\rceil, k\right) \tag{6.6}$$

(this value of k is chosen so that the $\ln(\cdot)$ in the definition of d is positive[3]). Indeed, these choices give

$$\begin{aligned}\sum_{v\in K(d,k)} p(v) &\le \frac{n}{k}\cdot\frac{2(d+k)\mathrm{e}^{\frac{k^2}{2(d+k)\sigma^2}}}{d}\\ &\le \frac{n}{k}\cdot\frac{4d}{d}\mathrm{e}^{\frac{k^2}{2d\sigma^2}} && (k\le d \text{ by (6.6)})\\ &\le \frac{n}{k}\cdot 4\mathrm{e}^{\ln(pk/8)} && (d \ge \tfrac{k^2}{2\sigma^2\ln(pk/8)} \text{ by (6.6)})\\ &= pn/2. && (6.7)\end{aligned}$$

Combining (6.1) with (6.7), we finally get

$$pn \le (2d+2k-1)|Q| + (k-1) + pn/2$$

which further implies

$$\begin{aligned}|Q| &\ge \frac{pn/2-k}{2d+2k}\\ &\ge \frac{pn-2k}{8d}. && (k\le d)\end{aligned}$$

Let us assume $k \le pn/4$ (otherwise we have $p^2n < 4pk \le 4p\left\lceil\frac{16\mathrm{e}}{p}\right\rceil = O(1)$, which means the

[3] This value of k also makes sense intuitively: we are showing that no algorithm can solve the source identification problem with probability better than $p/2$ in an uncovered interval, so we definitely need at least $k \ge 2/p$ to rule out random guessing.

$\Omega(\min(p^3 n\sigma^2, p^2 n))$ bound we are trying to prove becomes a trivial $\Omega(1))$. Then we get

$$\begin{aligned}
|Q| &\geq \frac{pn-2k}{8d} \\
&\geq \frac{pn}{16d} && (k \leq pn/4) \\
&\overset{(6.6)}{=} \frac{pn}{16\max\left(\left\lceil \frac{k^2}{2\sigma^2 \ln(pk/8)} \right\rceil, k\right)} && \text{(by definition of } d\text{)} \\
&\overset{(6.6)}{=} \frac{pn}{O\left(\max\left(\frac{1}{p^2\sigma^2}, \frac{1}{p}\right)\right)} && \text{(by definition of } k\text{)} \\
&= \Omega(\min(p^3 n\sigma^2, p^2 n)). && \square
\end{aligned}$$

6.5 Proofs for the Adaptive Setting

6.5.1 Upper Bound

Proof of Theorem 6.2.3. The algorithm crucially uses the following result on the concentration of the answers at large distances. We prove it in Section C.5.

Lemma 6.5.1. *For any probability* $0 < \delta < 1/2$, *there is some constant* $C(\delta) = O\left(\sqrt{\log(1/\delta)}\right)$ *such that for any* n, σ *and any source* $v^* \in V$, *we have*

$$\mathbb{P}_w[\forall q \in V, \text{ans}_w(v^*, q) \in |v^* - q| \pm C(\delta)\cdot\sigma\sqrt{|v^* - q|}\ln(1 + |v^* - q|)] \geq 1 - \delta. \tag{6.8}$$

That is, the concentration bound $|v^* - q| \pm C(\delta)\cdot\sigma\sqrt{|v^* - q|}\ln(1+|v^* - q|)$ *holds simultaneously for all nodes* q *with probability at least* $1-\delta$ *over the choice of the weights* w.

With this concentration result in hand, the algorithm follows a natural "iterative refining" strategy: start by obtaining a rough estimate of the identity of v^*, then progressively refine it by querying nodes closer and closer to v^*. After k steps (where k is defined by Claim 6.5.1), only a few possible candidate sources will remain, and the algorithm will switch to testing them one by one.

Concretely, let us assume that Lemma 6.5.1 holds with the desired probability of failure δ. Then the algorithm will maintain a shrinking interval $[l_i, r_i]$ which contains v^*. Initially $l_0 = 1$ and $r_0 = n$. At each step, the algorithm will query node l_i. Let d_i be equal to $r_i - l_i$. v^* has to be to the right of l_i and at distance at most d_i from l_i, so

$$\begin{aligned}
\text{ans}_w(v^*, l_i) &\in |v^* - l_i| \pm C(\delta)\cdot\sigma\sqrt{|v^* - l_i|}\ln(1 + |v^* - l_i|), && \text{(by (6.8))} \\
&\subseteq v^* - l_i \pm C(\delta)\cdot\sigma\sqrt{d_i}\ln(1 + d_i) && (v^* \geq l_i \text{ and } |v^* - l_i| \leq d_i)
\end{aligned}$$

and thus given answer $\text{ans}_w(v^*, l_i)$ the algorithm knows that v^* must be in the interval

$$l_i + \text{ans}_w(v^*, l_i) \pm C(\delta) \cdot \sigma \sqrt{d_i} \ln(1 + d_i). \tag{6.9}$$

Therefore, it shrinks its interval as follows:

$$\begin{cases} l_{i+1} = \max(l_i, l_i + \lceil \text{ans}_w(v^*, l_i) - C(\delta) \cdot \sigma \sqrt{d_i} \ln(1 + d_i) \rceil) \\ r_{i+1} = \min(r_i, l_i + \lfloor \text{ans}_w(v^*, l_i) + C(\delta) \cdot \sigma \sqrt{d_i} \ln(1 + d_i) \rfloor) \end{cases}$$

The resulting interval has length $d_{i+1} = r_{i+1} - l_{i+1} \leq 2C(\delta) \cdot \sigma \sqrt{d_i} \ln(1 + d_i)$.

Now what remains to do is to figure out how fast this interval shrinks, and when we should switch to testing the remaining candidates one by one. To get a rough initial intuition of the speed at which it shrinks, let us imagine that $d_{i+1} = \sigma \sqrt{d_i}$. Then, the sequence would decrease very fast at the onset, when d_i is still large, then decrease slower and slower. We would observe that $\log(d_{i+1}/\sigma^2) = \log(\sqrt{d_i}/\sigma) = \frac{1}{2}\log(d_i/\sigma^2)$: the *logarithm* of the ratio of d_i to σ^2 is divided by 2 at each step. So it would be reasonable to assume that d_i will approach σ^2 in a doubly-logarithmic number of steps. This is made rigorous in the following claim, which is proved in Section C.6.

Claim 6.5.1. *Assume $d_0 \leq n$, and $d_{i+1} \leq C \cdot \sigma \sqrt{d_i} \ln(1 + d_i)$ for some value $C > 0$. Then*

- *if $\sigma^2 \leq 1/2$, there exists $k = O(\log\log_{1/\sigma} n)$ such that $d_k = \text{poly}(C)$;*
- *if $\sigma^2 \geq 1/2$, there exists $k = O(\log\log n)$ such that $d_k = \sigma^2 \cdot \text{poly}(C, \log(1 + \sigma^2))$.*

We instantiate Claim 6.5.1 with $C := 2C(\delta)$. After the first k steps, we simply go through the $d_k + 1$ remaining possible positions for node v^* in $[l_k, r_k]$, and check them all with one query each.[4] With probability 1, v^* will be the only one to give 0 as an answer . Thus, overall, this algorithm will succeed with probability at least $1 - \delta$. The total number of queries used is $k + d_k + 1$, which by Claim 6.5.1 gives the desired bounds in both the $\sigma^2 \leq 1/2$ case and the $\sigma^2 \geq 1/2$ case. □

6.5.2 Lower Bound

For your reading convenience, here is a quick reference of the notations that are used throughout the proof.

[4] As we explain in Section C.1, this theorem can be extended to apply to many other edge weight distributions. If the distribution's support is positive, a binary search can be used instead.

Notations cheatsheet (not exhaustive)

- p: the desired probability of identifying the true source.
- R: the internal randomness of the algorithm (see Definition 6.5.1).
- q_j: the j^{th} node queried by the algorithm.
- a_j: the j^{th} answer the algorithm receives (it will take the value $\text{ans}_w(v^*, q_j)$).
- T: shorthand for $\text{Typical}_{p/2}(v^*, w)$, i.e. the event that the concentration bounds from Definition 6.5.3 hold.
- C: shorthand for $C(p/2)$, a constant (in n and σ) factor involved in the concentration bounds of Definition 6.5.3 (see Definition C.7.1 for its precise value).
- D: shorthand for $D(\sigma, p/2)$, the minimum distance at which the concentration bounds of Definition 6.5.3 hold (see Definition C.7.1 for its precise value).
- l_j, r_j: the step counters at which the closest query nodes to v^* have been placed so far at step j; i.e. by the time the j^{th} query has been asked, v^* lies between query nodes q_{l_j} and q_{r_j}.
- μ_j: the minimum of the answers to query nodes q_{l_j} and q_{r_j}. Intuitively, it is a proxy for the smallest answer made so far (after asking j queries).
- $\text{reduce}_{n,\sigma}(x)$: a function $\mathbb{R} \to \mathbb{R}$ that models the fastest decrease of μ_j an algorithm can hope for: most of the time $\mu_{j+1} \geq \text{reduce}_{n,\sigma}(\mu_j)$ (see Definition 6.5.6).
- λ_i: a lower bound on μ_j with high probability, and therefore a limit on the progress that the algorithm can make (see Definition 6.5.7).
- K_j: a random variable representing all the information that the algorithm has at its disposal after asking the first j queries (see Definition 6.5.8).
- A_j: the event that $\mu_j \geq \lambda_j$; intuitively, the event that at step j, the algorithm has not queried any nodes very close to v^*.
- B_j: informally, the event that even based on everything the algorithm knows at step j, no node is particularly likely to be the source (see Definition 6.5.10).
- j_{stop} (also, j_{min}): a lower bound on the number of steps that an algorithm needs to find the source with probability p (see Definition 6.5.11).

Definition 6.5.1 (R)**.** *Let R be a random variable denoting the internal randomness of the algorithm. One can for example think of R as drawn uniformly from interval* $[0,1]$*, as this puts no limitation on the amount of randomness the algorithm can use.*

Definition 6.5.2 (q_j, a_j)**.** *Let q_j be the j^{th} query node selected by the algorithm, and let a_j be* *the answer that it gets to query q_j (i.e. $a_j := \mathrm{ans}_w(v^*, q_j)$). Both q_j and a_j are random variables that can depend on v^*, w and the internal randomness of the algorithm.*

Note that under the hypotheses of Theorem 6.2.4, any algorithm needs to query at least one node: otherwise it would succeed with probability at most $1/n < p$.

To simplify the proof, we make the following adaptations to the model, which only give the algorithm more power to identify the source, and therefore hold without loss of generality:

(a) Before the algorithm starts, two initial query nodes $q_{-1} = 1$ and $q_0 = n$ are already selected, resulting in answers a_{-1} and a_0 at no cost to the algorithm. The first query that is actually chosen by the algorithm is q_1.

(b) When querying node q_j, in addition to the answer a_j, the algorithm is told on which side of q_j the source v^* is identified.[5]

(c) Once the algorithm is ready to guess the position of v^*, it should query it.[6] If at any point the algorithm queries v^*, it immediately terminates and the identification is considered successful. More precisely, the number of queries that the algorithm uses is defined as the first positive integer j such that $q_j = v^*$.

The details of the proof are at times technically heavy, so we first give a general outline to provide the gist of the proof. It proceeds in the following 8 steps. We will cover each of them in detail in the next 8 subsections (Section 6.5.2.x corresponds to step x).

1. We define a random event T (over v^* and w) which has probability $\geq 1 - p/2$, guarantees that v^* is not too close to the ends of the path, and gives some concentration bounds for the answers when the query node is at least some distance D away from v^* (D will be defined in Definition 6.5.4). T represents a "typical situation": the role of this event is to exclude some extreme cases (e.g. $v^* = 1$ or $v^* = n$) that would derail the proof.

2. We define a sequence of random variables μ_j that describe how close the algorithm is to find v^* after it has asked j queries. μ_j is (roughly) the distance between v^* and the query node closest to v^*, and tends to decrease as j increases. We also define a corresponding *deterministic* sequence $\lambda_0 > \lambda_1 > \cdots$ where for each j, $\mu_j \geq \lambda_j$ with high probability.

3. We define the following events which (basically) imply each other in alternation (i.e. $A_j \Rightarrow B_j \Rightarrow A_{j+1}$), and will help us bound the progress of the algorithm:

[5] Note that this gives the algorithm the ability to perform a binary search, which is not necessarily easy when the weight distribution is not positive.

[6] This extra query does not affect the asymptotics because as noted in the previous paragraph the algorithm always needs to query at least one node anyway.

- A_j is the event that $\mu_j \geq \lambda_j$; it intuitively means "none of the j first query nodes are too close to v^*";
- B_j will be defined later, and intuitively means "even after asking j queries, the algorithm has only a vague idea where v^* is", or a bit more precisely, "even conditioned on all the answers gathered by the algorithm during the first j steps, none of the nodes have a high probability of being the source".

4. We define j_{stop} be the largest j such that $\lambda_j \geq D$ (recall that D is the distance above which event T gives concentration bounds on answers). Our goal will be to prove that with high probability, the algorithm needs to ask at least j_{stop} queries.

5. We prove two key lemmas, which show that in most cases, $A_j \Rightarrow B_j$ and $B_j \Rightarrow A_{j+1}$. They state that for $j < j_{\text{stop}}$,

   - A_j implies B_j (Lemma 6.5.3);
   - with probability $1 - \frac{1}{\log n}$, $T \wedge A_j \wedge B_j$ implies A_{j+1} (Lemma 6.5.4).

   This is the core technical part of the proof.

6. We chain the above lemmas by induction and use the fact that $\mathbb{P}[\neg T] \leq p/2$ to obtain $\mathbb{P}[A_{j_{\text{stop}}}] \geq 1 - p$.

7. We prove that $j_{\text{stop}} = \Omega(\log\log_{\max(1/\sigma,2)} n)$, the desired lower bound.

8. We observe that event $A_{j_{\text{stop}}}$ implies that the algorithm has not found v^* after asking j_{stop} queries, which using 6 and 7 completes the proof.

Typical instances: event T

In our model, there are no hard guarantees on how far away the answer $\text{ans}_w(v^*, q)$ might be from the real distance $|v^* - q|$. For example, $\text{ans}_w(v^*, v^*+1) \sim \mathcal{N}(1, \sigma^2)$ might be as large as 1000, even if $\sigma = 1$ (though with very low probability). While such extreme events are intuitively disadvantageous for the algorithm, they also make it harder to prove lower bounds. Therefore, we need to make basic assumptions on the range of $\text{ans}_w(v^*, q)$ at high distances.

To do this, we will need to use the notion of a "typical" instance: a choice of v^* and w for which some reasonable concentration results hold. Note that part (i) below is very similar to Lemma 6.5.1, which we used for the upper bound. Let $C(\delta)$ and $D(\sigma,\delta)$ be two values (defined later in Definition C.7.1) such that

$$\max(\sigma^2, \mathrm{e}^2) \leq D(\sigma,\delta) = o_\delta(\max(\sigma^2 \log\sigma, 1)). \tag{6.10}$$

Definition 6.5.3 (Typical$_\delta(v^*, w)$)**.** *For any probability $\delta > 0$, let* Typical$_\delta(v^*, w)$ *be the event that the following holds:*

(a) $\min(\text{ans}_w(v^*,1),\text{ans}_w(v^*,n)) \geq \frac{n}{C(\delta)}$;

(b) for all q *with* $d_q := |v^* - q| \geq D(\sigma,\delta)$,

(i) $\text{ans}_w(v^*,q) \in d_q \pm \sigma\sqrt{d_q}\ln d_q$

(ii) $\text{ans}_w(v^*,q) \in d_q \pm d_q/4 = [\frac{3}{4}d_q, \frac{5}{4}d_q]$.

Part (a) means that the two answers from the query nodes at either end of the path are not too much smaller than their expectation $\Omega(n)$*, and part (b) means that above a certain distance threshold* $D(\sigma,\delta)$*, all answers are concentrated around their mean.*

As the name indicates, most instances are typical (the proof is given in Section C.7).

Lemma 6.5.2. *For any probability* $\delta > 0$ *and any* $n \geq \Theta_\delta(\max(\sigma^2 \ln\sigma, 1))$, $\mathbb{P}_{v^*,w}[\text{Typical}_\delta(v^*, w)] \geq 1-\delta$.

We will apply Lemma 6.5.2 with $\delta := p/2$. We will use the following shorthands.

Definition 6.5.4 (T, C, D)**.** *Let* $T := \text{Typical}_{p/2}(v^*, w)$, $C := C(p/2)$ *and* $D := D(\sigma, p/2)$.

Corollary 6.5.1. $\mathbb{P}[T] \geq 1 - p/2$.

Measure of progress μ_j and benchmark λ_j

It turns out that the right metric of progress to look at is (roughly speaking) the *smallest answer value seen so far.* More precisely, suppose that the algorithm has asked j queries so far (and hence is at step j). Then we define the quantity μ_j as follows.

Definition 6.5.5 (l_j, r_j, μ_j)**.** *Let* $l_j := \arg\max_{i\leq j, q_i\leq v^*}(q_i)$ *and* $r_j := \arg\min_{i\leq j, q_i \geq v^*}(q_i)$*, which means that* q_{l_j} *(resp.* q_{r_j}*) is the closest query node at or to the left (resp. right) of* v^* *placed so far. Then* $\mu_j := \min(a_{l_j}, a_{r_j})$*, the smaller of the corresponding answers.*

Note in passing that by simplifying assumption (b) in the beginning of Section 6.5.2, the algorithm knows l_j and r_j. Also, if $\mu_j > 0$, then the algorithm has not found v^* yet (otherwise we would have $q_{l_j} = q_{r_j} = v^*$ and thus $a_{l_j} = a_{r_j} = 0$).

We want to show that, with high probability, μ_j cannot decrease too fast with j. To make it formal, we define an analogous *deterministic* sequence λ_j, which we will show is a lower bound for μ_j with high probability. We call λ_j a “benchmark” because it is a point of comparison to determine whether the algorithm is making fast progress or not. It decreases with j according to the following function.

Definition 6.5.6. *Let* $\text{reduce}_{n,\sigma}(x) = \frac{\sigma\sqrt{x}}{400\ln x \log n}$.

Definition 6.5.7. *Let* $\lambda_0 := n/C$ *and* $\lambda_{j+1} := \mathrm{reduce}_{n,\sigma}(\lambda_j)$.[7]

Observe that by point (a) in Definition 6.5.3, T implies $\mu_0 \geq n/C = \lambda_0$. Our goal will be to prove that $\mu_j \geq \lambda_j$ will likely continue to hold as j increases.

Events A_j and B_j

Informally, at step j, A_j is the event that the algorithm has not queried any nodes very close to v^*, and B_j is the event that the algorithm has only a vague idea of where v^* is (or more precisely, that even conditioned on all the answers so far, no node has a high probability of being the source). As we will see in Section 6.5.2), intuitively,

- A_j implies B_j because if the algorithm does not have any query nodes close to v^*, then the answers it got are all very noisy, and thus its confidence interval for v^* is wide (Lemma 6.5.3);
- B_j implies A_{j+1} because if all nodes are very unlikely to be the source v^*, then wherever it decides to query the next node, it is unlikely to be very close to v^* (Lemma 6.5.4).

As we will see, both events depend only on information that is available to the algorithm at step j. For convenience, we define random variable K_j, which describes all the knowledge of the algorithm up to step j.

Definition 6.5.8 (K_j)**.** *Let* $\{q_i\}_{\leq j} := (q_{-1},\ldots,q_j)$ *and*
$:= (a_{-1},\ldots,a_j)$ *be the query nodes and answers available at step* j. *Then let* $K_j = (\{q_i\}_{\leq j},\{a_i\}_{\leq j},l_j,r_j)$. *This encodes the locations of all query nodes, the answers received from them, as well as the identity of the two query nodes between which* v^* *lies.*

A_j is the event that μ_j is greater than the benchmark λ_j.

Definition 6.5.9. *Let* A_j *be the event that* $\mu_j \geq \lambda_j$.

B_j is the event that the posterior of $v^* = v$ given K_j is "diluted".

Definition 6.5.10. *Let* B_j *be the event that for all nodes* $v \in V$,

$$\mathbb{P}[T \wedge (v^* = v) \mid K_j] \leq \frac{1}{\left(\frac{8}{3}\lambda_{j+1}+1\right)\log n}.$$

Note that $\mathbb{P}[T \wedge (v^* = v) \mid K_j]$ itself is a random variable since it depends on K_j, so B_j is still a random event even though it is a statement about a probability. An equivalent way to define

[7] If $\lambda_j \leq 0$, we define $\lambda_{j+1} := 0$. However, we will never use such values.

B_j is to first define random variable

$$P_j := \max_{v \in V} \mathbb{P}[T \wedge (v^* = v) \mid K_j],$$

then to let B_j be the event that $P_j \le \frac{1}{\left(\frac{8}{3}\lambda_{j+1}+1\right)\log n}$.

Stopping step j_{stop}

Our goal is to show that for a high value of j, we have $\mu_j > 0$ with high probability, and therefore the algorithm has failed to find v^* using only j queries. We now define that value of j.

Definition 6.5.11 ($j_{\min}, j_{\text{stop}}$)**.** *Let $j_{\min}$ be the smallest integer $j \ge 0$ such that $\lambda_j < D$. Then*

$$j_{\text{stop}} := \min\left(j_{\min} - 1, \left\lfloor \frac{p \log n}{2} \right\rfloor\right).$$

This means that at step $j \le j_{\text{stop}}$, $\lambda_j \ge D$ is still big enough for the concentration bounds of event T to hold. The second argument of the $\min(\cdot,\cdot)$ is just for convenience of the proof, and will not matter if n is large enough. We will also use the following easily believable fact, proved in Section C.8.

Fact 6.5.1. $\lambda_0 > \lambda_1 > \cdots > \lambda_{j_{\text{stop}}} > \lambda_{j_{\text{stop}}+1}$.

Key lemmas

We now state our two main lemmas, which constitute the core technical part of the proof. The proof of Lemma 6.5.3 is very technical and not particularly enlightening, so it is deferred to C.9. The proof of Lemma 6.5.4, on the other hand, is much more straightforward, and we include it here.

Lemma 6.5.3. *If $j < j_{\text{stop}}$, then $A_j \Rightarrow B_j$.*

Lemma 6.5.4. *If $j \le j_{\text{stop}}$, then $\mathbb{P}[\neg T \vee \neg A_j \vee \neg B_j \vee A_{j+1}] \ge 1 - \frac{1}{\log n}$.*

Note that "$\neg T \vee \neg A_j \vee \neg B_j \vee A_{j+1}$" is logically equivalent to "$(T \wedge A_j \wedge B_j) \Rightarrow A_{j+1}$". Intuitively, if B_j holds, then the probability of $v^* = v$ (conditioned on the answers so far) is low for any v, which means that whatever the algorithm picks as its next query q_{j+1}, the probability that q_{j+1} is within some distance d of v^* is upper bounded by the sum of those probabilities over $v \in [q_{j+1} - d, q_{j+1} + d]$. Therefore, with high probability, a_{j+1} will not be too small, and the same holds for μ_{j+1}.

Proof of Lemma 6.5.4. We will show equivalently that $\mathbb{P}[T \wedge A_j \wedge B_j \wedge \neg A_{j+1}] \le \frac{1}{\log n}$.

At step j, the algorithm queries node q_{j+1} based on the information K_j it has so far and its internal randomness R, then receives answer a_{j+1}. The the only way for both A_j and $\neg A_{j+1}$ to hold is for the new answer a_{j+1} to be smaller than λ_{j+1}.[8]

Let $d := |v^* - q_{j+1}|$. If we had $d \geq \frac{4}{3}\lambda_{j+1} \geq D$, then if T occurs, by concentration bound (ii) we would have $a_{j+1} \geq \frac{3}{4}d \geq \lambda_{j+1}$. Let $I := V \cap (q_{j+1} \pm \frac{4}{3}\lambda_{j+1})$ (I also depends on (K_j, R)). Then the only way to have $a_{j+1} < \lambda_{j+1}$ is for v^* to be in I, which implies

$$\mathbb{P}[T \wedge A_j \wedge B_j \wedge \neg A_{j+1}] \leq \mathbb{P}[T \wedge v^* \in I \wedge B_j]. \tag{6.11}$$

Now, for any assignment (k_j, r) of random variables (K_j, R), we have

$$\begin{aligned}
&\mathbb{P}[T \wedge v^* \in I \wedge B_j \mid K_j = k_j \wedge R = r] \\
&\quad = \sum_{v \in I} \mathbb{P}[T \wedge (v^* = v) \wedge B_j \mid K_j = k_j \wedge R = r] && (I \text{ is fixed by } (K_j, R)) \\
&\quad = \sum_{v \in I} \mathbb{P}[T \wedge (v^* = v) \wedge B_j \mid K_j = k_j]. \\
&&& (T, v^* \text{ are independent from } R, \text{ and } B_j \text{ is fixed by } K_j)
\end{aligned}$$

If B_j is false given $K_j = k_j$, then the above sum has probability 0. If on the other hand B_j is true given $K_j = k_j$, then by definition of B_j,

$$\begin{aligned}
\sum_{v \in I} \mathbb{P}[T \wedge (v^* = v) \wedge B_j \mid K_j = k_j] &= \sum_{v \in I(k_j, r)} \mathbb{P}[T \wedge (v^* = v) \mid K_j = k_j] \\
&\leq \frac{|I|}{\left(\frac{8}{3}\lambda_{j+1} + 1\right)\log n} \leq \frac{1}{\log n}.
\end{aligned}$$

Therefore, in either case, $\mathbb{P}[T \wedge v^* \in I \wedge B_j] \leq \frac{1}{\log n}$, which, combined with (6.11), completes the proof. □

Induction on j

Lemmas 6.5.3 and 6.5.4 can now be chained to obtain the following result.

Lemma 6.5.5. $\mathbb{P}[A_{j_{\text{stop}}}] \geq 1 - p$.

Proof. First, as already noted at the end of Section 6.5.2, T implies $\mu_0 \geq \lambda_0$, which means that $T \Rightarrow A_0$ (by definition of A_0). Also, by Lemma 6.5.3, $A_j \Rightarrow B_j$ for $0 \leq j < j_{\text{stop}}$. In addition, by Corollary 6.5.1, we have $\mathbb{P}[T] \geq 1 - p/2$, and by Lemma 6.5.4, for $0 \leq j < j_{\text{stop}}$, we have $\mathbb{P}[\neg T \vee \neg A_j \vee \neg B_j \vee A_{j+1}] \geq 1 - \frac{1}{\log n}$. Therefore by a union bound, both T and "$\neg T \vee \neg A_j \vee \neg B_j \vee A_{j+1}$

[8] Formally, if $A_j \wedge \neg A_{j+1}$, then using Fact 6.5.1 we have $\mu_{j+1} < \lambda_{j+1} < \lambda_j \leq \mu_j$. Therefore, $\mu_{j+1} = a_{j+1}$, and thus $a_{j+1} < \lambda_{j+1}$.

for $0 \le j < j_{\text{stop}}$" simultaneously hold with probability at least

$$1 - p/2 - \frac{j_{\text{stop}}}{\log n} \ge 1 - p/2 - \frac{\left\lfloor \frac{p \log n}{2} \right\rfloor}{\log n} \ge 1 - p. \qquad \text{(by Definition 6.5.11)}$$

If they do hold, then the following logical statements are all true: "T", "$T \Rightarrow A_0$", "$A_j \Rightarrow B_j$" ($\forall\, j < j_{\text{stop}}$), and "$(T \wedge A_j \wedge B_j) \Rightarrow A_{j+1}$" ($\forall\, j < j_{\text{stop}}$). It is easy to see that, chained together, they imply $A_{j_{\text{stop}}}$. □

Asymptotics of j_{stop}

The following lemma gives us an asymptotic lower bound on j_{stop}. We prove it in Section C.10.

Lemma 6.5.6. *For $n \ge \Theta_p(\max(\sigma^3, 1))$, we have*

$$j_{\text{stop}} + 1 = \begin{cases} \Omega_p(1 + \log(1 + \log_{1/\sigma} n)) \text{ if } \sigma^2 \le 1/2 \\ \Omega_p(\log\log n) \text{ if } \sigma^2 \ge 1/2. \end{cases}$$

Proof of Theorem 6.2.4

All that is left to do is to conclude.

Proof of Theorem 6.2.4. If $A_{j_{\text{stop}}}$ holds, then $\mu_{j_{\text{stop}}} > 0$, which means the algorithm has not found v^* after asking j_{stop} queries (recall our assumption from the beginning of Section 6.5.2 that, without loss of generality, the algorithm must query the source in order to make its guess). By Lemma 6.5.5, this happens with probability at least $1 - p$. Therefore, any algorithm that finds v^* with probability at least p must use at least $j_{\text{stop}} + 1$ queries. The theorem then follows from Lemma 6.5.6. □

6.6 Related Work

We throughly reviewed the related work in source identification in Chapter 1. In this section we review the more distant related work in Information Theory and Theoretical Computer Science.

6.6.1 Related Work in Information Theory

The role of adaptivity is a central question in several fields in computer science, including property testing [50], information theory [58, 154] and learning theory [213]. The most well-known example is perhaps binary search on a line, where being adaptive reduces the number of queries from n to $\log_2(n)$. Such a significant decrease in the query complexity of standard binary search is possible because the queries are very constrained; we can only ask whether the target is to the left or to the right of the queried vertex. If instead we are allowed to query any subset for containment of the target without any noise, then there is no difference between the adaptive and the non-adaptive query complexities (this is the well-known BarKochba or 20 Questions Game between two players, where the first player comes up with an item that the other player must identify by asking (in principle up to 20) yes-no questions). Indeed, $\log_2(n)$ questions are necessary because every answer carries only binary information, and the target can be found by $\log_2(n)$ non-adaptive questions by querying each digit of the binary representation of the index of the target vertex. One way to reintroduce a difference between the adaptive and non-adaptive cases in the 20 Questions Game is to corrupt the answers by a query dependent noise, which was proposed initially by Rényi [204], and has been studied by several follow-up works, including [58, 154, 255].

The problem setup of [154] has a close resemblance to our setup. In both cases, the search is done on a line, and the answers are corrupted by Gaussian noise, the variance of which depends on how close the query was to finding the target. The notable differences between [154] and our setup are that:

(i) we have more restrictive queries (one query in our setup is a single node (hence there are n possible queries), whereas one query in the setup of [154] is a subset of the nodes (hence there are 2^n possible queries))

(ii) we receive more information (we receive a noisy version of the distance between the queried vertex and the source, whereas in [154] they receive a noisy binary answer for belonging to the query set)

(iii) in our case the noise that corrupts the answers is not independent between queries.

Because of these differences, our proof techniques and our results are also different from [154]. The main tool in [154] for the adaptive upper bound is the *posterior matching scheme*. Roughly speaking, posterior matching produces queries that split the line into two approximately equal-weight subsets weighed by the posterior. In particular, there is no restriction on the queries produced by the posterior matching scheme, and therefore it is not applicable in our case (see (i) above). We also note, that as opposed to our setup, in [154], the geometry of the search space does not play an important role; any subset of the vertices of the line can be a query and the answers are insensitive to the distances between the queried vertices and the target vertex. For this reason, the usual geometry-insensitive information-theoretic notions (such as the entropy of the posterior) that work well in [154], cannot be used in our setup (see Section 6.5.2 for our notion of "progress"). In terms of results, for constant σ, both the non-adaptive and adaptive query complexities are found to be $\Theta(\log(n))$ in [154], which is in sharp contrast with our finding of $\Theta(n)$ and $\Theta(\log\log n)$ in the non-adaptive and adaptive settings, respectively. Finally, we note that the paper [154] features results about the expected query complexity of the search algorithms, whereas we give query complexity bounds that hold with any constant failure (or success) probability.

6.6.2 Related Work in Theoretical Computer Science

Extensions of binary search to graphs have been proposed on numerous occasions [80, 92, 136, 190]. Of these, perhaps [80] has the closest connections with source identification with time queries. In this extension, a target vertex at an unknown position in a general graph is to be identified by adaptively querying vertices. A queried vertex can only respond whether it is the target or not, and if not, it indicates the edge on a shortest path between itself and the target. In the noiseless setting, queries always report the correct answer, whereas in the noisy setting, queries report a correct answer independently with probability $1/2 < p < 1$. In a sense, noisy binary search is an adaptive version of the source identification model proposed by [198] where we would keep the "who infected me" information and drop the time information instead, with the notable difference that in noisy binary search the noise that corrupts the answers is independent between queries. Since the information that a queried node can provide is its distance and/or its direction towards the source, adaptive source identification and noisy binary search on a line can be seen as "duals" of each other, in the sense that the former collects a noisy estimate of the distance whereas the latter collects a noisy estimate of the direction to the source. In the latter case, the adaptive query complexity is found to be $\Theta(\log n)$ for constant p in [80, 154]. Comparing this result with our result of $\Theta(\log\log n)$ for the number of required queries in the adaptive case indicates that the distance to the target is far more informative than the direction, at least on the path graph. On different graphs, notably on star graphs, the distance is expected to be less informative than the direction. We limited the study of stochastic source identification in this chapter to the path topology because of the complexity of the computations, and we leave the study in other graph topologies for further work.

6.7 Discussion and Future Work

We presented the first mathematical study of source identification with time queries in a non-deterministic diffusion process. We considered both the setting when the queries are selected adaptively and non-adaptively. We found that when the edge-delay distribution has constant variance, the number of required queries is $\Theta(\log\log n)$ in the adaptive setting, and $\Theta(n)$ in the non-adaptive setting. Our results are in sharp contrast with similar problems, such as measurement dependent noisy search on a line [154], or probabilistic binary search in graphs [80], where the query complexities were found to be $\Theta(\log n)$ in both cases.

The main open question is of course what happens in other graphs. Extending our results to certain classes of trees might be feasible with the methods presented in this chapter, however, an extension to graphs with cycles seems very challenging. Still, we hope that our results can inspire some, potentially more heuristic, ideas for treating graphs with cycles as well. Based on preliminary simulation results, the extension of our results seems most feasible in graphs with strong geometric properties. In Figure 6.4 we show the number of sensors required by a state of the art source identification algorithm in various network models as a function of

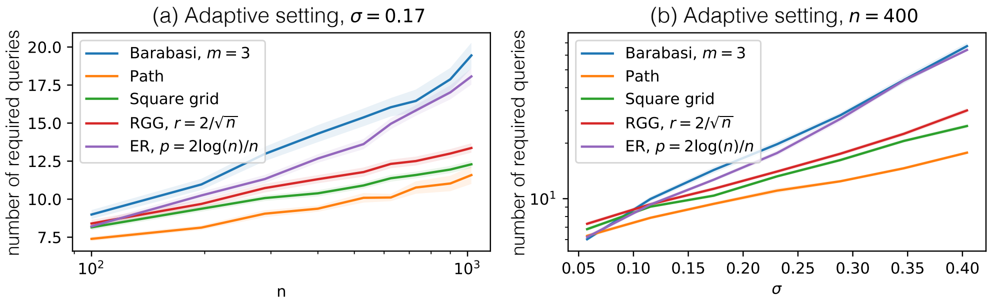


Figure 6.4: The number of queries required by the adaptive source identification algorithm called Max-Gain [224] in the S1 source identification model as a function of (a) the network size n, and (b) the standard deviation of the edge-delays σ in the following network models: Barabási-Albert network [23] with average degree $2m$, path graph, square grid, random geometric graphs [196] in the unit square with connection radius r, Erdős-Rényi graphs [82] with connection probability p. The network parameters, controlling the number of edges in the random networks, were chosen slightly above the connectivity threshold, so that all networks in the simulation were connected. Each datapoint is an average of 192 simulations and the confidence intervals are computed using the Student's t distribution-test. See Appendix C.3 for more details about how the Max-Gain was run to produce these simulation results.

(a) the network size n, and (b) the standard deviation σ of the propagation delay distribution $\mathcal{W}$. Both plots seem to suggest that the number of required sensors behave similarly to the path graph in square grids and random regular graphs [196], and rather differently in the Barabási-Albert [23] and the Erdős-Rényi [82] small world networks. Of course, it is difficult to read off anything conclusive from such simulation plots; for instance, from Figure 6.4 alone, we would not have been able to read off the $\Theta(\log\log(n))$ dependence of the query complexity on the size of the path graph, which we proved rigorously in this chapter.

While we do not consider this scenario, given the sensitive nature of health information, it would be interesting to study source identification with time queries in the context of privacy preserving learning. In a scenario where an adversary is watching our queries, but not the responses, a recent line of work characterized the tradeoff between query complexity and privacy in adaptive binary search on a line [235, 245]. The model has been extended to the case when the answers we receive are noisy in a follow-up work by [244]. It would be interesting to combine the methods presented in this chapter with the methods of [235, 245, 244] for new results in privacy preserving source identification.

Imperfect Knowledge of the Network Part IV

7 Robustness of the Metric Dimension to Adding a Single Edge

In this chapter, we study the increase of the metric dimension (MD) on adding a single edge to the graph. See Chapter 3 for a general introduction about the MD and its connection with source identification. Specifically, we reviewed related results about the robustness of the MD to edge additions and edge deletions in Section 3.2

This chapter is based on the publication [173] by Mashkaria, Ódor and Thiran.

7.1 Summary of Results

In a previous work Eroh et. al. [83], found that the increase of the MD not bounded by any constant in general graphs on a single edge addition. Their statement is supported by an example graph, where the addition of a single edge doubles the MD. In Section 7.3.1, we show an example graph where adding a particular edge increases the MD from $\Theta(\log(N))$ to $\Theta(N)$, which is a much larger increase than in the example of [83], where the MD only doubles. For a result in the opposite direction, in Section 7.3.2 we provide an upper bound on the MD of the graph with the extra edge in terms of the MD of two subgraphs of the original graph. We believe that this result can be used in several graph families to show that the exponential increase in Section 7.3.1 only happens for very special (in a sense very heterogeneous) graphs, and that in most cases the MD at most doubles. We prove this doubling upper bound for d-dimensional grid graphs in Section 7.4.1, and finally, we perform an even more refined analysis for the case of $d = 2$ in Section 7.4.2.

For the case $d = 2$, we conjecture that the limiting distribution of the MD after a uniformly random edge is added is $3 + \text{Ber}(8/27)$, where Ber is the Bernoulli distribution. The only part missing in proving this conjecture is a lower bound on the MD when the extra edge is in a specific configuration. Such lower bound proofs are especially tedious, since one must show that no set of landmark nodes of a certain size can distinguish every pair of nodes, which often leads to a long case-by-case analysis. Instead, we proved as much as we could reasonably write down, and state the rest of our results as a conjecture at the end of the chapter (Conjecture

7.4.1). A similar approach was used in [170] when determining the MD of torus graphs.

Our proofs rely on careful combinatorial analysis, and a detailed description of how the shortest paths change in a graph after adding an edge. In particular, when adding an edge to the graph, we study the set of node pairs between which the shortest paths are changed and unchanged. These sets depend on the extra edge, and they are highly structured. We are not aware whether this structure (described in Section 7.2) has been previously studied in the literature, but we believe that it could bring insight into different problems where the addition of a single edge is studied (i.e. wormhole attacks [118] and the dynamic all pairs shortest paths problem in data structures [67, 11]).

7.2 Changes in the All-Pairs Shortest Paths After Adding an Edge

In this section we will develop tools to understand how the shortest paths change in a graph after adding an extra edge.

Let $G = (V, E_G)$ be a connected simple graph, with vertex set V (we use the word vertex, node and point interchangeably) and edge set E_G. We add an edge e between two non-adjacent vertices E and F to obtain a graph $G' = (V, E_G \cup \{e\})$. Let $d_H(A,B)$ denote the length of the shortest path between vertices A and B in graph H. For simplicity, we will use the notation $d_G(A,B) = AB$.

Remark 7.2.1. *If we want to reach vertex B from vertex A, there are three options: Either we do not use e at all, or we use e from E to F or we use e from F to E. Hence,*

$$d_{G'}(A,B) = \min(AB, AE+1+FB, AF+1+EB). \tag{7.1}$$

Clearly, we cannot increase the distance between two vertices by adding an edge, or in other words either $d_{G'}(A,B) \le AB$. Next, we describe the pairs of vertices whose distance decreased after adding the edge.

Definition 7.2.1 (special region)**.** *For any vertex A, $R_A = \{Z \in V \mid d_{G'}(Z,A) < ZA\}$. We will refer R_A as the special region of A.*

The special region contains the vertices which will "use" the extra edge e to reach A. Formally, we can write this as $Z \in R_A$ is equivalent with $d_{G'}(A,Z) = \min(AE+1+FZ, AF+1+EZ) < ZA$.

Definition 7.2.2 (normal region, normal vertex)**.** *$N_A = V \setminus R_A$ will be referred as the normal region of A. We call the intersection of all normal regions as simply the normal region and we denote it by N. A vertex in the normal region is called a normal vertex.*

The normal region can be succinctly expressed as $N = \{Z \in V \mid R_Z = \varnothing\}$. For a normal vertex $Z \in N$ we have $d_{G'}(A,Z) = AZ$ for every vertex A, that is distances from or to these vertices Z

are unchanged after adding edge e, which makes normal vertices the simplest type of vertices from the point of view of our analysis. The following claim helps us to characterize the normal region for any graph.

Claim 7.2.1. *The set of vertices V can be partitioned to the following three sets,*

$$\begin{aligned} R_E &= \{A \in V \mid AE - AF > 1\} \\ N &= \{A \in V \mid |AE - AF| \leq 1\} \\ R_F &= \{A \in V \mid AE - AF < -1\}. \end{aligned}$$

The intuition for Claim 7.2.1 is that if we are trying to reach A from some other node, we may want to use e in the EF direction if F is closer to A, we may want to use e in the FE direction if E is closer to A, and there is no gain in using e if E and F are almost equidistant to A. The three regions are illustrated in Figure 7.1.

Proof. First assume that $|AE - AF| \leq 1$. For an arbitrary vertex B in the graph, using triangular inequality,

$$AB \leq AE + EB \leq AF + 1 + EB.$$

Similarly,

$$AB \leq AE + 1 + FB.$$

Hence, by Remark 7.2.1, we have that $d_{G'}(A,B) = AB$. As this is true for any vertex $B \in V$, we must have $A \in N$.

Next assume $AE - AF > 1$. Then, by Remark 7.2.1,

$$d_{G'}(A,E) = \min(AE, AE + 1 + AF, AF + 1) = AF + 1,$$

which implies that $A \in R_E$. The $AE - AF < -1$ case follows analogously. □

The usefulness of partitioning the vertices into R_E, N and R_F goes beyond just characterizing the normal region. Note that R_E collects the vertices that use the edge in the FE direction (because F is closer to them), and R_F collects the vertices that use the edge in the EF direction. There are no nodes that use the extra edge in both directions. Hence, if the two nodes are in the same special region R_E or R_F, they are using the extra edge in the same direction, and they cannot use the extra edge to reduce the distance between themselves. We formalize this intuition in the next claim.

Claim 7.2.2. *If two vertices A and B lie in the same special region R_E or R_F, then $d_{G'}(A,B) = d_G(A,B)$, or equivalently $B \notin R_A$ and $A \notin R_B$.*

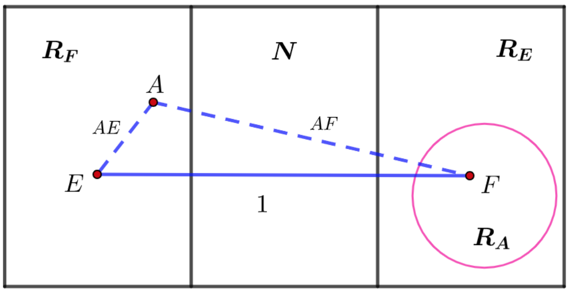


Figure 7.1: Graph G' partitioned into three regions: R_E, N and R_F.

Proof. Without loss of generality, let $A, B \in R_E$. Then, we have $AE - AF > 1$ and $BE - BF > 1$ by Claim 7.2.1. Therefore,

$$d_{G'}(A,B) = \min(AB, AE + 1 + FB, AF + 1 + EB) = AB,$$

because

$$AE + 1 + FB > AF + 2 + FB \geq AB + 2,$$

and

$$AF + 1 + EB > AF + 2 + FB \geq AB + 2$$

by the triangle inequality. □

Remark 7.2.2. *Containment in special regions defines an anti-reflexive, symmetric and anti-transitive (never transitive) relation between pairs of vertices. Containment in normal regions defines a reflexive, symmetric and intransitive (not necessarily transitive) relation between pairs of vertices.*

Proof. For both special and normal regions (anti-)reflexivity follows from the definition and symmetry follows from the symmetry of distances in both G and G'. The anti-transitivity of special regions follows from Claim 7.2.2. Indeed, if $A \in R_B$ and $B \in R_C$, then the pairs (A, B) and (B, C) are in different special regions R_E or R_F, which implies that A and C must be both in R_E or R_F and we cannot have $A \in R_C$. □

We are now ready to justify the illustration in Figure 7.1.

Remark 7.2.3. *For a vertex $A \in R_F$ we have*

$$F \in R_A \subseteq R_E,$$

and similarly, for a vertex $B \in R_E$ we have

$$E \in R_B \subseteq R_F.$$

Proof. For a vertex $A \in R_F$, the statement $F \in R_A$ follows by the symmetric nature of special regions (Remark 7.2.2). The $R_A \subseteq R_E$ is a simple consequence of anti-transitivity. Indeed, $R_A \cap N$ is empty by definition, and $R_A \cap R_F$ is empty because we cannot have $Z \in R_F$, $Z \in R_A$ and $A \in R_F$ all hold at the same time. □

Next, we use the anti-transitivity property to make equation (7.1) more explicit.

Claim 7.2.3. *For any $A, B \in V$, we have*

$$d_{G'}(A,B) = \begin{cases} AE + 1 + FB & \text{if } A \in R_B \text{ and } A \in R_F \\ AF + 1 + EB & \text{if } A \in R_B \text{ and } A \in R_E \\ AB & \text{otherwise, i.e., } A \in V \setminus R_B = N_B. \end{cases} \tag{7.2}$$

Proof. We consider only the case $A \in R_B$ and $A \in R_F$; the second case is symmetric, and the third holds by definition. By the definition of special regions, $A \in R_F$ is equivalent with

$$d_{G'}(A,F) = \min(AF + 1 + EF, AE + 1 + FF) < AF,$$

which further implies $AE + 1 < AF$.

By the anti-transitivity of special regions, $A \in R_B$ and $A \in R_F$ together imply $B \in R_E$, which is equivalent with

$$d_{G'}(B,E) = \min(BE + 1 + FE, BF + 1 + EE) < BE,$$

which further implies $BF + 1 < BE$.

Finally, $A \in R_B$ is equivalent with

$$d_{G'}(A,B) = \min(AE + 1 + FB, AF + 1 + EB),$$

which reduces to $d_{G'}(A,B) = AE + 1 + FB$ since $AE < AF$ and $BF < BE$. □

We already used the intuition that vertices in special regions "gain" from the addition of the extra edge. We formalize this intuition in the next definition.

Definition 7.2.3 (Gain, $\text{Gain}_{\max}$)**.** *Let the decrease in the distance between two vertices due to edge e be denoted as*

$$\text{Gain}(A,B) = AB - d_{G'}(A,B).$$

Let the maximum gain associated to a node A be denoted as

$$\text{Gain}_{\max}(A) = \max_X(\text{Gain}(A,X)).$$

Remark 7.2.4. *For vertex $A \in R_E$, vertex E gets the maximum benefit of the extra edge to reach A, that is,* $\text{Gain}_{\max}(A) = \text{Gain}(A, E) = AE - (1 + AF)$*. More generally, for any $A \in V$, we have*

$$\text{Gain}_{\max}(A) = \max(0, |AF - AE| - 1).$$

We can also observe that, by Claim 7.2.1, $A \in N$ if and only if $\text{Gain}_{\max}(A) = 0$*. A similar statement hold for vertex F instead of E.*

Proof. Suppose for contradiction that there is a node $B \in V$ for which $\text{Gain}(A, B) > \text{Gain}(A, E)$. Since $A \in R_E$, this node B must be in R_F, otherwise by Claim 7.2.2 we have $\text{Gain}(A, B) = 0$. Then, the following inequalities must hold:

$$\begin{aligned} \text{Gain}(A, B) &> \text{Gain}(A, E) \\ AB - d_{G'}(A, B) &> AE - d_{G'}(A, E) \\ AB - (BE + 1 + AF) &> AE - (1 + AF) \\ AB &> AE + BE. \end{aligned}$$

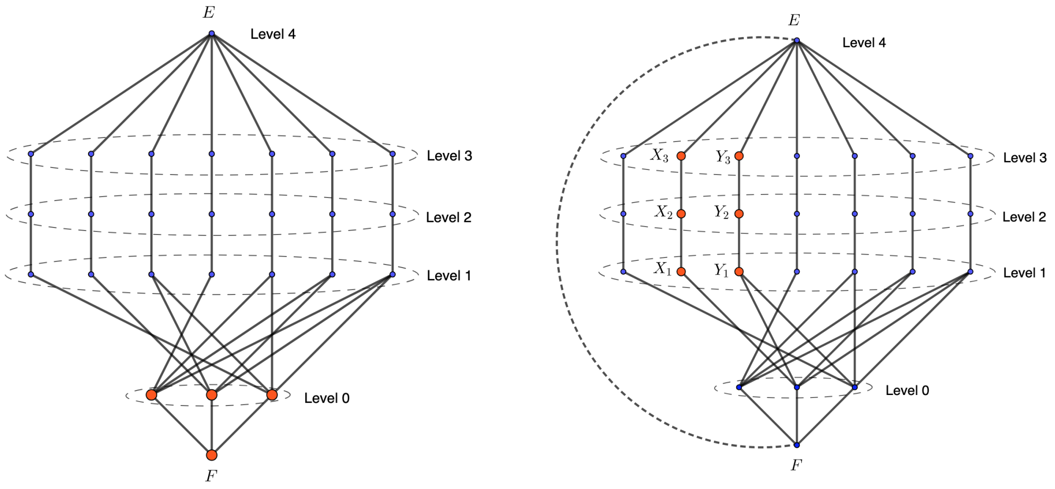
The last inequality above contradicts the triangle inequality, and the proof is completed. □

7.3 General Graphs

7.3.1 An Example with an Exponential Increase in the Metric Dimension

In this section, we give a construction for a graph $G^\star$ on $3n + \lceil \log_2(n) \rceil - 1$ nodes with $\text{MD}(G^\star) \leq \lceil \log_2(n) \rceil + 1$, in which the increase in the metric dimension is at least $n - \lceil \log_2(n) \rceil - 3$ on adding a single (specific) edge. The idea is that in $G^\star$, the vertices of R_F can be efficiently distinguished only by some vertices in R_E (but not by vertices in R_F). Then, after adding edge e, the vertices in R_E can reach R_F on new shortest paths, and they will not distinguish vertices in R_F anymore. Hence R_F will have to be distinguished by vertices in R_F, which will require significantly more nodes. The construction is shown in Figure 7.2 for $n = 8$.

$G^\star$ has 6 levels indexed by $l \in \{-1, \dots, 4\}$. Levels 1-3 each contain $n - 1$ vertices, which are indexed by i for each level. Level 0 contains $\lceil \log_2(n) \rceil$ vertices indexed by j. Levels -1 and 4 contain the single vertices $F = v_1^{(-1)}$ and $E = v_1^{(4)}$. We connect all of the vertices of level 0 and 3 to F and E, respectively. We connect the vertices of level 1 (respectively, level 2) to the vertices of level 2 (resp., level 3) if and only if the vertices of both levels 1-2 (resp., 2-3) have the same index. Finally, we connect a vertex labeled i in level 1 to a vertex labeled j in level 0 if and only if the j^{th} bit in the binary representation of i is one. For example, $v_1^{(1)}$ is connected only to $v_{\lceil \log_2(n) \rceil}^{(0)}$ because the binary representation of 1 is $0 \dots 01$. This construction leads therefore to the following definition.



(a) G

(b) G'

Figure 7.2: Example where MD increases by a large amount (for $n = 8$)

Definition 7.3.1 ($G^\star$)**.** *For $n > 1$, let $G^\star = (V^\star, E_{G^\star})$, with*

$$V^\star = \left\{v_j^{(0)} \mid j \in \{1, \dots, \lceil \log_2(n) \rceil\}\right\} \cup \left\{v_i^{(l)} \mid l \in \{1,2,3\}, i \in \{1, \dots, n-1\}\right\} \cup \{E, F\},$$
$$E_{G^\star} = \left\{(F, v_j^{(0)})\right\} \cup \left\{(v_j^{(0)}, v_i^{(1)}) \mid \text{bin}(i)_j = 1\right\} \cup \left\{(v_i^{(l)}, v_i^{(l+1)}) \mid l \in \{1,2\}\right\} \cup \left\{(v_i^{(3)}, E)\right\},$$

where $\text{bin}(i)_j$ *denotes the j^{th} bit of the binary representation of the number i.*

Claim 7.3.1. *The set $S^\star = \left\{v_j^{(0)}\right\} \cup F$ resolves $G^\star$. Consequently,* $\text{MD}(G^\star) \leq \lceil \log_2(n) \rceil + 1$.

Proof. We need to show that any pair of vertices in $V^\star \setminus S^\star$ are distinguished. There are two possibilities for any pair of distinct vertices: either they are in different levels or in the same level. If they are on different levels, vertex F will distinguish them, because for any $v_i^{(l)}$ with $l \in \{1,2,3,4\}$ we have $d_{G^\star}(v_i^{(l)}, F) = l + 1$. If they are on the same level, the binary representations of their index i will differ at at least one position. Let the j^{th} bit of both labels be different. Then, vertex $v_j^{(0)}$ will distinguish them, because its distance to the vertex whose label has the j^{th} bit equal to 1 is two hops shorter than its distance to the vertex whose label has the j^{th} bit equal to 0. Therefore all pairs of points are distinguished, which completes the proof. □

Now we add an edge e between vertices E and F. The resulting graph $G^{\star\prime}$ is shown in Figure 7.2b.

Claim 7.3.2. *The metric dimension of graph $G^{\star\prime}$ is at least $n - 2$.*

Proof. Notice that the set of nodes that can distinguish $v_j^{(3)}$ and $v_k^{(3)}$ is

$$\left\{v_i^{(l)} \mid l \in \{1,2,3\}, i \in \{j,k\}\right\}.$$

This is because all other nodes can reach both $v_j^{(3)}$ and $v_k^{(3)}$ through E on their shortest path and E cannot distinguish any pair of nodes on level 3. Hence, distinguishing nodes on level 3 is equivalent to resolving a star graph, and the metric dimension of $G^{\star\prime}$ is at least $n-2$. □

Combining Claims 7.3.1 and 7.3.2, we observe that the increase in the metric dimension of $G^\star$ on adding e is at least $n - \lceil \log_2(n) \rceil - 3$.

7.3.2 Bounds on the Change of the Metric Dimension

It has been shown in [83] that if G' is obtained from G by adding an extra edge, then $\mathrm{MD}(G') \geq \mathrm{MD}(G) - 2$, and if there are no even cycles in G', then $\mathrm{MD}(G') \leq \mathrm{MD}(G) + 1$. However, in the previous section we saw an example where $\mathrm{MD}(G')$ was exponentially larger than $\mathrm{MD}(G)$. In this section we provide an upper bound on $\mathrm{MD}(G')$ in terms of the MD of the subgraphs of G, which holds for all graphs G'.

Lemma 7.3.1. *Let $G = (V, E)$ be a connected graph, and let G' be the graph obtained by adding edge e between vertices E and F as before. Let $V_1 = \{U \in V \mid d_G(U,E) \leq d_G(U,F)\}$ and $V_2 = \{U \in V \mid d_G(U,E) \geq d_G(U,F)\}$. Let G_1 and G_2 be subgraphs of G induced on vertex sets V_1 and V_2, respectively. Then,*

$$\mathrm{MD}(G') \leq \mathrm{MD}(G_1) + \mathrm{MD}(G_2) + 2.$$

Proof. Let S_1 and S_2 be the resolving sets of minimum size of graphs G_1 and G_2, respectively. We prove that $S = S_1 \cup S_2 \cup \{E, F\}$ is a resolving set of G'. Let $N_1 = \{U \in V \mid 0 \leq d_G(U,F) - d_G(U,E) \leq 1\}$ and $N_2 = \{U \in V \mid 0 \leq d_G(U,E) - d_G(U,F) \leq 1\}$. By Claim 7.2.1, $N_1, N_2 \subseteq N$, where N is the normal region. Consider two vertices X and Y. There are two cases:

Case 1: $X, Y \in V_1$ or $X, Y \in V_2$.

Without loss of generality, let $X, Y \in V_1$. Let $A \in S_1$ be the vertex which resolves X and Y in G_1. Since $V_1 = R_F \cup N_1$ and $V_2 = R_E \cup N_2$, by Claim 7.2.2, $d_G(X, A) = d_{G'}(X, A)$ and $d_G(Y, A) = d_{G'}(Y, A)$, hence X and Y are resolved by A in G', too.

Case 2: $X \in V_1 \setminus V_2$ and $Y \in V_2 \setminus V_1$ or vice and versa.

Without loss of generality, let $X \in V_1 \setminus V_2$ and $Y \in V_2 \setminus V_1$. Note that by definition, $V_1 \setminus V_2$ and $V_2 \setminus V_1$ contain the nodes that are closer to E and F, respectively. Hence, we can always go through e when going from $V_1 \setminus V_2$ to F or $V_2 \setminus V_1$ to E on a shortest path, that is

$$d_{G'}(X,F) = 1 + d_{G'}(X,E) \tag{7.3}$$

$$d_{G'}(Y,E) = 1 + d_{G'}(Y,F). \tag{7.4}$$

Assume for contradiction that none of E and F distinguish X and Y. This implies that $d_{G'}(X,E) = d_{G'}(Y,E)$ and $d_{G'}(X,F) = d_{G'}(Y,F)$. Adding both equations gives

$$d_{G'}(Y,E) + d_{G'}(X,F) = d_{G'}(Y,F) + d_{G'}(X,E).$$

Substituting values from (7.3) and (7.4) gives a contradiction. Hence, either E or F will distinguish these two vertices.

For every possible pair of vertices we showed a distinguishing vertex in S. Finally,

$$\mathrm{MD}(G') \le |S| = |S_1| + |S_2| + 2 = \mathrm{MD}(G_1) + \mathrm{MD}(G_2) + 2.$$

□

Next we present a graph $G^{\star\star}$ for which the upper bound of Lemma 7.3.1 is achieved. The graph has 74 vertices and it is drawn in Figure 7.3. The four solid black nodes labeled as E in the figure represent a single vertex E in the graph $G^{\star\star}$. Similarly, the four solid black nodes labeled F represent vertex F. All other nodes shown in the figure represent distinct nodes. The graph $G^{\star\star\prime}$ is obtained by adding an edge between vertices E and F. In this setting, $G_1^{\star\star}$, defined in Lemma 7.3.1, will be the sub-graph induced by nodes having green and yellow outlines and $G_2^{\star\star}$ will be the sub-graph induced by nodes having orange and red outlines.

Claim 7.3.3. $\mathrm{MD}(G_1^{\star\star}) = \mathrm{MD}(G_2^{\star\star}) = 8$ *and* $\mathrm{MD}(G^{\star\star\prime}) = 18$.

Proof. Notice that $G_1^{\star\star}$ and $G_2^{\star\star}$ are isomorphic, hence their metric dimensions must be equal as well. First we show $\mathrm{MD}(G_1^{\star\star}) \le 8$. Indeed we have 8 triangles in $G_1^{\star\star}$, and selecting one degree 2 vertex in each triangle is enough to distinguish any two vertices. To show $\mathrm{MD}(G_1^{\star\star}) \ge 8$, observe that we need to select one vertex from each of the triangles. The equality $\mathrm{MD}(G_1^{\star\star}) = \mathrm{MD}(G_2^{\star\star}) = 8$ together with Lemma 7.3.1 proves $\mathrm{MD}(G^{\star\star\prime}) \le 18$.

Next, we show that $\mathrm{MD}(G^{\star\star\prime}) \ge 18$. Again, notice that $G^{\star\star\prime}$ contains 16 triangles, and we must select a vertex in each of them. Notice that even after we selected these 16 nodes, the solid colored pairs in Figure 7.3 are not distinguished. Moreover, it is not possible to distinguish all 4 of these solid colored pairs by adding a single vertex to the set. Indeed, if any of the green stroked nodes are selected, the green solid pair is not distinguished. A similar argument holds for all other colors. This shows that we must add at least two nodes to the initial 16, and the metric dimension of $G^{\star\star\prime}$ is at least 18, which completes the proof. □

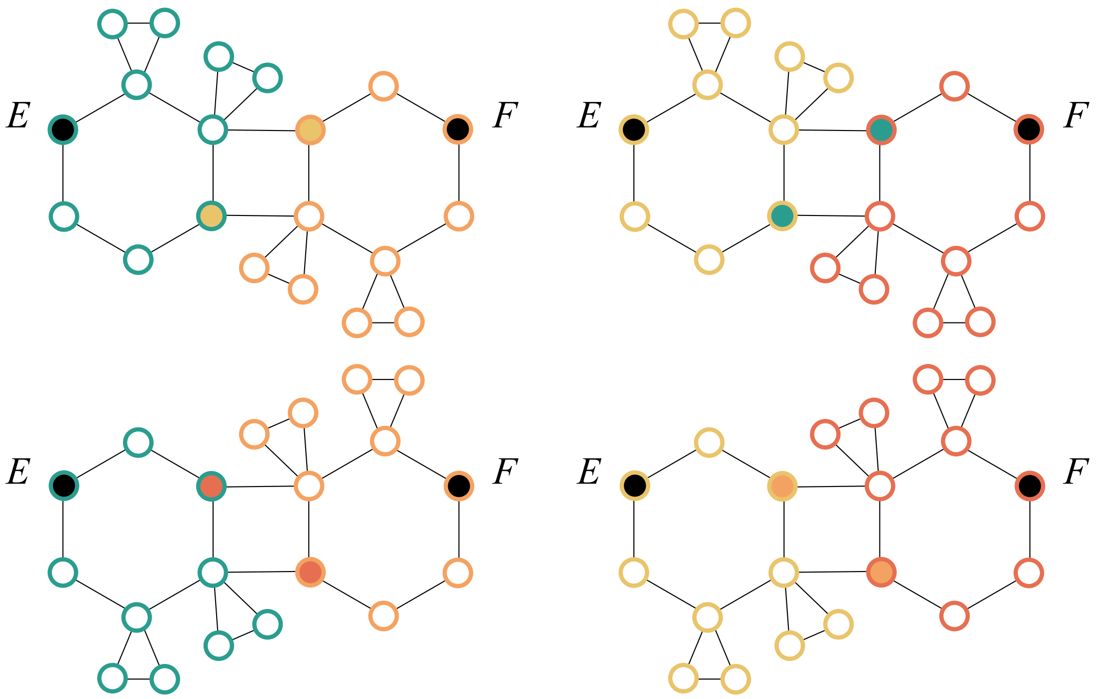


Figure 7.3: Graph $G^{\star\star}$ and points E, F for which upper bound is achieved

7.4 Grid Graph

The main technical result of this chapter is on the metric dimension of the grid graph augmented with one edge.

Definition 7.4.1. *Let the d-dimensional grid graph with side lengths $(n_1, n_2, \dots, n_d)$ be the Cartesian product of d paths indexed by i with length n_i.*

Let us represent each vertex A of the grid in a d-dimensional space as $(x_A^{(1)}, x_A^{(2)}, ..., x_A^{(d)})$ where $1 \le x_A^{(i)} \le n_i$ for $i \in \{1, ..., d\}$. For grid G and vertices A and B, we denote the distance

$$d_G(A, B) = AB = \sum_{i=1}^{d} |x_A^{(i)} - x_B^{(i)}|.$$

We state and prove the general result for d-dimensional grid graphs in Section 7.4.1, and we focus on the case of the 2-dimensional grid for more precise results in Section 7.4.2.

7.4.1 The d-Dimensional Grid

We start by understanding the MD of the d-dimensional grid without any extra edges. The paper [141] claims that the MD of a d-dimensional grid is d, however, [49] shows by computer search that this statement is false for hypercubes of dimensions $5 \le d \le 8$. It is not difficult to show that d is an upper bound, but it is believed asymptotically not to be tight when the side lengths are small. The paper [211] claims without proof that if all side lengths are $n_i = n$, then

$$\limsup_{d\to\infty} \frac{\mathrm{MD}(G)\log_n(d)}{d} \leq 2, \tag{7.5}$$

and they also prove

$$\liminf_{d\to\infty} \frac{\mathrm{MD}(G)\log_n(d)}{d} \geq 1. \tag{7.6}$$

However, when the side length n is large, then the MD of d-dimensional grid is exactly d, which was shown in [100]. Before stating this lower bound on n for the MD to be exactly d, we include a non-asymptotic lower bound on the MD for grids with general side lengths.

Lemma 7.4.1. *Let G be a grid of dimension d with side lengths* $(n_1, n_2, \dots, n_d)$*, and let us denote* $N_\Sigma = \sum_i n_i$ *and* $N_\Pi = \prod_i n_i$*. Then*

$$\mathrm{MD}(G) \geq \frac{\log(N_\Pi)}{\log(N_\Sigma - d + 1)}. \tag{7.7}$$

Proof. The distances in G range from 0 to $\sum_i(n_i - 1)$, which implies a total number of $(N_\Sigma - d + 1)^{\mathrm{MD}(G)}$ possible distinct distance vectors. Since the distance vectors must be unique, the number of possible distinct vectors must be at least as large as the total number of vertices in G, or formally $(N_\Sigma - d + 1)^{\mathrm{MD}(G)} \geq N_\Pi$. Taking the logarithm of both sides and rearranging the terms gives the desired result. □

The lower bound on n for the MD to be exactly d can be found as a corollary of Lemma 7.4.1.

Corollary 7.4.1 (Theorem 5.1 [100])**.** *Let G be a grid of dimension d with equal side lengths* $(n, n, \dots, n)$*. If* $n \geq d^{d-1}$*, then* $\mathrm{MD}(G) = d$*.*

Proof. The assumption $n \geq d^{d-1}$ is equivalent to $n^{\frac{d}{d-1}} \geq nd$, which, by taking the logarithm of both sides, gives

$$\frac{d}{d-1}\log(n) \geq \log(nd). \tag{7.8}$$

Combining inequalities (7.7) and (7.8) gives

$$\mathrm{MD}(G) \overset{(7.7)}{\geq} \frac{\log(N_\Pi)}{\log(N_\Sigma - d + 1)} > \frac{\log(N_\Pi)}{\log(N_\Sigma)} = \frac{d\log(n)}{\log(nd)} \overset{(7.8)}{\geq} d - 1. \tag{7.9}$$

Since it is well established that the MD of the d-dimensional grid is upper bounded by d, the proof is completed. □

We need a slightly more technical lemma before stating our main results on the MD of the grid with an extra edge.

Lemma 7.4.2. *Let $G = (V, E_G)$ be a grid graph of dimension d with side lengths $(n_1, n_2, ..., n_d)$. Let E and F be the endpoints of the extra edge e. As defined in Lemma 7.3.1, let $V_1 = \{U \in V \mid UE \leq UF\}$. Let G_1 be the subgraph of G induced on V_1. Then* $\mathrm{MD}(G_1) \leq d$.

We defer the proof to the end of the section, and we state and prove our main theorem for d-dimensional grid graphs.

Theorem 7.4.1. *Let $G = (V, E_G)$ be a grid graph of dimension d. For an edge e between any two vertices E and F in V, let $G' = (V, E_G \cup \{e\})$. Then,* $\mathrm{MD}(G') \leq 2d + 2$. *Moreover, the lower bound* (7.7) *in Lemma 7.4.1 holds for G' as well.*

Proof of Theorem 7.4.1. Let $V_1 = \{U \in V \mid UE \leq UF\}$ and $V_2 = \{U \in V \mid UE \geq UF\}$. Let G_1 and G_2 be the subgraphs of G induced on V_1 and V_2, respectively. Lemma 7.4.2 implies that $\mathrm{MD}(G_1) \leq d$, and $\mathrm{MD}(G_2) \leq d$ holds by symmetry. Finally, we apply Lemma 7.3.1 to arrive to

$$\mathrm{MD}(G') \leq \mathrm{MD}(G_1) + \mathrm{MD}(G_2) + 2 \leq 2d + 2.$$

For the lower bound, since adding an edge only decreases the distances in the graph, the same proof as in Lemma 7.4.1 applies. □

It is an interesting question, whether the upper bound in Theorem 7.4.1 can be improved to $2d$ by simply not including the two endpoints of the extra edge into the resolving set when applying Lemma 7.3.1. We saw in Claim 7.3.3, that the two endpoints are needed for general graphs, but we will see in the next section, that they are not needed for the 2-dimensional grid. We believe that the upper bound can be improved to $2d$, but the proof is not straightforward. In the proof of the 2-dimensional case, we rely heavily on the observation that the normal region has a specific shape no matter where the extra edge is added. We show in Figure 7.4 that this is not true anymore even for $d = 3$. Indeed, the shape of the normal regions (and thus of sets V_1 and V_2) can be quite different for different configurations of the extra edge, which suggests that the number of cases can explode.

We conclude the section by providing a proof for Lemma 7.4.2.

Proof of Lemma 7.4.2. The proof will consist of three parts. In the first part of the proof, we define our coordinate system so that the extra edge is oriented in a specific way. This part essentially breaks the symmetries of the grid, which will reduce the number of cases we need to inspect later in the proof. In the second part, we show that a set of d corners in V_1, which we denote by O, resolves the grid G. Finally, in the third part of the proof, we show that the distance between any vertex $X \in V_1$ and any corner in O is the same in both G and G_1. Hence,

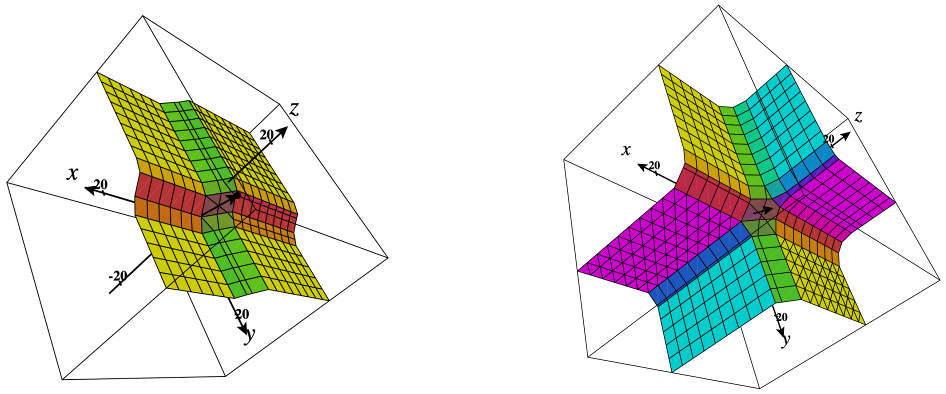


Figure 7.4: The 3D surfaces show the normal region in the 3-dimensional grid for two different configurations of the extra edge. The extra edges are marked with a black vector in the middle of the cube.

O will be a resolving set of G_1 as well, which proves that the MD of G_1 is upper bounded by d and completes the proof of the lemma.

Part 1: Without loss of generality, we can label the dimensions such that $|x_E^{(1)} - x_F^{(1)}| = \max_i(|x_E^{(i)} - x_F^{(i)}|)$, i.e., the distance between E and F along the first dimension is the maximum among distances along all the dimensions. Now, again without loss of generality, we also assume that $x_E^{(i)} \leq x_F^{(i)}$ for all i. We can assume that because if $x_E^{(j)} > x_F^{(j)}$ for any dimension j, we can reflect the grid along that dimension so that $x_E^{(j)}$ becomes less than $x_F^{(j)}$. Basically, this reflection will map coordinates $x_X^{(j)}$ to $n_j - x_X^{(j)}$, keeping all other coordinates unchanged. We summarize these assumptions, taken without loss of generality, below.

Assumption 7.4.1 (symmetry breaking)**.** *Without loss of generality, we assume that E and F satisfy*

$$x_E^{(i)} \leq x_F^{(i)} \quad \textit{for all } i \in \{1, \ldots, d\}, \tag{7.10}$$

and

$$x_F^{(1)} - x_E^{(1)} \geq x_F^{(i)} - x_E^{(i)} \quad \textit{for all } i \in \{2, \ldots, d\}. \tag{7.11}$$

The d-dimensional grid has $2^d d!$ symmetries for choosing a coordinate system (which form the hyperoctahedral group). Note that even after Assumption 7.4.1, we still have $(d-1)!$ ways of choosing the coordinates (each equation in (7.10) removes a factor of two, and equations (7.11) remove a factor of d). This is because we only require that $|x_F^{(i)} - x_E^{(i)}|$ takes (one of) its maximum value(s) for $i = 1$, and we have no constraint on the order of the values for the other indices. Thus, Assumption 7.4.1 does not break all symmetries of the grid, only the ones necessary for the proof. This also means that although we exhibit only a single resolving set O, there are multiple sets of d corners in V_1 that resolve G.

Part 2: In this part of the proof, we show there there exists a set of d corners in V_1 that resolves the grid G. Let us define

$$O = \left\{ \begin{array}{l} O_1 = (1,1,1,\ldots,1), \\ O_2 = (1,n_2,1,\ldots,1), \\ O_3 = (1,1,n_3,\ldots,1), \\ \ldots, \\ O_d = (1,1,1,\ldots,n_d) \end{array} \right\},$$

where O_1 is the all-ones vector of dimension d, and for $j > 1$ we get O_j from O_1 by changing its j^{th} entry to n_j. Khuller et al. show that the set of the d corners of O form a resolving set of G, and we only need to show that all d corners of O belong to V_1, that is $O_jE \le O_jF$ holds for all j. Because of equations (7.10),

$$O_1E = \sum_{i=1}^{d}(x_E^{(i)} - 1) \le \sum_{i=1}^{d}(x_F^{(i)} - 1) = O_1F.$$

Next, we consider the corners O_j for $j > 1$. Because of Assumption 7.4.1, we have

$$x_F^{(j)} - x_E^{(j)} \overset{(7.11)}{\le} x_F^{(1)} - x_E^{(1)} \overset{(7.10)}{\le} \sum_{i=1, i\neq j}^{d}(x_F^{(i)} - x_E^{(i)}).$$

Reorganizing the terms and then adding $n_j - d + 1$ to both sides of the inequality yields

$$-x_E^{(j)} + \sum_{i=1,i\neq j}^{d} x_E^{(i)} \le -x_F^{(j)} + \sum_{i=1,i\neq j}^{d} x_F^{(i)}$$

$$(n_j - x_E^{(j)}) + \sum_{i=1,i\neq j}^{d}(x_E^{(i)} - 1) \le (n_j - x_F^{(j)}) + \sum_{i=1,i\neq j}^{d}(x_F^{(i)} - 1)$$

$$O_jE \le O_jF.$$

Thus, all the corners in O lie inside V_1.

Part 3: In this part of the proof, we show that $d_G(X,O_j) = d_{G_1}(X,O_j)$ for all $X \in V_1$ and $O_j \in O$. We show this by exhibiting a shortest path between X and O_j in G such that all vertices on that path belong to V_1. This will show that $d_G(X,O_j) \ge d_{G_1}(X,O_j)$. The inequality in the opposite direction is trivial because G_1 is a subgraph of G, which means that we must have $d_G(X,O_j) = d_{G_1}(X,O_j)$.

For $j > 1$, the shortest path between X and O_j that we exhibit will have the following two parts:

1. decrease all the co-ordinates (in any order), except j, to 1 to reach $X_1 = (1,...,x_X^{(j)},...,1)$.
2. increase the j^{th} coordinate from $x_X^{(j)}$ to n_j in order to reach O_j.

For $j = 1$, we simply decrease all the coordinates (in any order) to 1 to reach O_1. Clearly, these define valid shortest paths in a grid graph, and next, we prove that we stay inside V_1 both throughout the first part (from X to X_1) and the second part (from X_1 to O_j) of the path.

First, we show that if $X = (x_X^{(1)}, x_X^{(2)}, ..., x_X^{(d)}) \in V_1$ with $x_X^{(1)} > 1$, then $X_0 = (x_X^{(1)} - 1, x_X^{(2)}, ..., x_X^{(d)}) \in V_1$ as well. We distinguish two cases based on the ordering of $x_E^{(1)}, x_F^{(1)}$ and $x_X^{(1)}$. On the one hand, if $x_X^{(1)} > x_F^{(1)} \geq x_E^{(1)}$ or $x_X^{(1)} \leq x_E^{(1)} \leq x_F^{(1)}$, then

$$X_0F - X_0E = |x_X^{(1)} - 1 - x_F^{(1)}| - |x_X^{(1)} - 1 - x_E^{(1)}| = |x_X^{(1)} - x_F^{(1)}| - |x_X^{(1)} - x_E^{(1)}| = XF - XE \geq 0,$$

since the terms inside the absolute values have the same sign. On the other hand, if $x_E^{(1)} < x_X^{(1)} \leq x_F^{(1)}$, then

$$X_0F - X_0E = (x_F^{(1)} - x_X^{(1)} + 1) - (x_X^{(1)} - 1 - x_E^{(1)}) = (x_F^{(1)} - x_X^{(1)}) - (x_X^{(1)} - x_E^{(1)}) + 2 = XF - XE + 2 \geq 2.$$

Since there are no other cases by Assumption 7.4.1, the inequality $X_0F - X_0E \geq 0$ must always hold, which implies $X_0 \in V_1$. Therefore, we showed that decrementing the first coordinate does not lead outside of V_1, and the same argument works for any of the d coordinates.

Next, we show for the second part of the shortest path, that each vertex in the path from X_1 to O_j with $j > 1$ belongs to V_1. Let $X_y = (1, ..., y, .., 1)$ be a vertex with $x_{X_y}^{(i)} = 1$ for $i \neq j$, and $x_X^{(j)} \leq y = x_{X_y}^{(j)} \leq n_j$. Clearly, X_y describes all intermediate vertices on the path between X_1 to O_j. Then, since $j > 1$,

$$\begin{aligned} X_yF - X_yE &= |y - x_F^{(j)}| + \sum_{i=1, i\neq j}^{d} (x_F^{(i)} - 1) - |y - x_E^{(j)}| - \sum_{i=1, i\neq j}^{d} (x_E^{(i)} - 1) \\ &= |y - x_F^{(j)}| - |y - x_E^{(j)}| + \sum_{i=1, i\neq j}^{d} (x_F^{(i)} - x_E^{(i)}) \\ &\overset{(7.11)}{\geq} |y - x_F^{(j)}| - |y - x_E^{(j)}| + (x_F^{(j)} - x_E^{(j)}). \end{aligned} \tag{7.12}$$

Finally, by applying the triangle inequality to the right hand side of equation (7.12), we arrive to

$$X_yF - X_yE \geq 0,$$

which implies that all the vertices X_y in the path from X_1 from O_j with $j > 1$ belong to V_1. This concludes the proof of the lemma. □

7.4.2 The 2-Dimensional Grid

For the sake of simplicity, we slightly adjust our notation to the $d = 2$ case. Let $G = (V, E_G)$ be a two-dimensional rectangle grid graph with m rows and n columns. Let the tuple (i, j) denote the vertex in i^{th} column and j^{th} row. The upper left, upper right, bottom right, bottom left corners are labeled as

$$P=(1,1), \qquad Q=(n,1), \qquad R=(n,m), \qquad S=(m,1),$$

respectively (see Figure 7.5). Let e be the edge between vertices $E=(x_E,y_E)$ and $F=(x_F,y_F)$ with $x_E,x_F \in \{1,\ldots,n\}$, $y_E,y_F \in \{1,\ldots,m\}$, with the assumption that $EF \geq 2$. Let $G'=(V,E_G \cup \{e\})$ be the 2-dimensional grid augmented with one edge.

Assumption 7.4.2 (symmetry breaking for $d=2$)**.** *We assume that*

1. $x_F \leq x_E$
2. $y_E \leq y_F$
3. $x_E - x_F \leq y_F - y_E$.

Assumption 7.4.2 is just a special case of Assumption 7.4.1 for $d=2$. Geometrically, it means that the edge is tilted right, F is below and to the left of E, and the angle between the edge and the horizontal axis is between 45 and 90 degrees (see Figure 7.5). As argued in the proof of Lemma 7.4.2, if the edge is in any other orientation, we can flip or rotate the grid horizontally and/or vertically to bring the edge in this orientation, hence Assumption 7.4.2 can be made without loss of generality.

Adversarial Setting

Theorem 7.4.2. *Let $G=(V,E)$ be a rectangle grid graph with m rows and n columns. For an edge e between any two nodes in V, let $G'=(V,E\cup\{e\})$. Then, the set of all 4 corners of the original grid is a resolving set for G', and consequently* $\mathrm{MD}(G') \leq 4$.

Proof. We start by making observations about which special regions the four corners P,Q,R,S belong to. First, notice that

$$QF - QE = (n-x_F)+(y_F-1)-(n-x_E)-(y_E-1) = EF \geq 2,$$

Hence by Claim 7.2.1, $Q \in R_F$. Similarly, $S \in R_E$.

Then, notice that

$$PF - PE = (x_F-1)+(y_F-1)-(x_E-1)-(y_E-1) = (y_F-y_E)-(x_E-x_F) \geq 0$$

where the last inequality holds by Assumption 7.4.2. Claim 7.2.1 implies therefore that P belongs to either R_F or N. Similarly, R belongs to either R_E or N. In any case, by Claim 7.2.2, it can be deduced that

$$(R_P \cup R_Q) \cap (R_S \cup R_R) = \varnothing \tag{7.13}$$

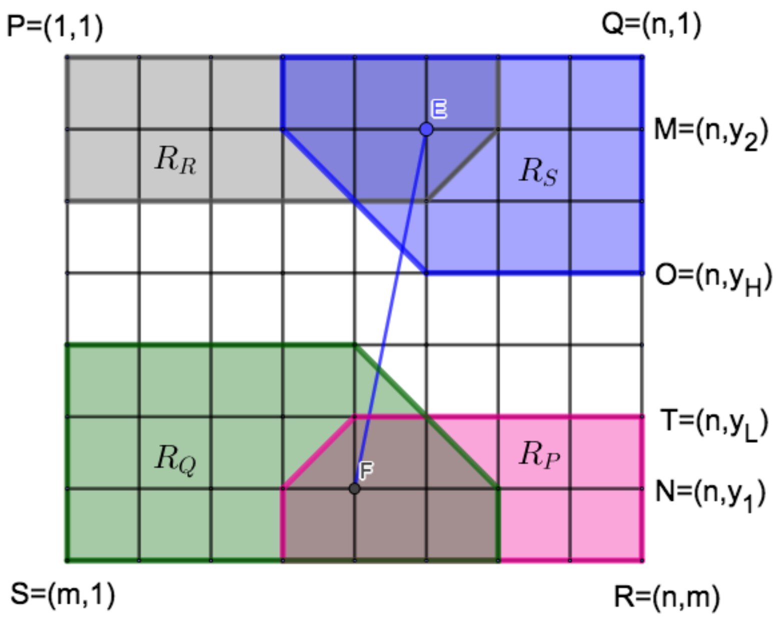


Figure 7.5: The sets R_P, R_Q, R_R, R_S, R_W are colored grey, blue, pink, green and white, respectively. Vertices on the boundary of coloured regions are included in the respective coloured region.

In fact, it turns out that $R_P \cup R_Q = R_E$ and $R_R \cup R_S = R_F$, but we are not showing this because it is not needed in this proof. Instead, let $R_W = V \setminus \{R_P \cup R_Q \cup R_R \cup R_S\}$ (the white region in Figure 7.5), and we note that the sets $R_P \cup R_Q$, $R_S \cup R_R$ and R_W partition the set of nodes V.

To prove the theorem, for any pair of nodes A, B, we are going to assign two of the corners $\{P, Q, R, S\}$ in the resolving set, and we are going to show that one of the two must distinguish A and B. The assignment will depend on whether A and B belong to $R_S \cup R_R$, $R_P \cup R_Q$ or R_W. Moreover, we further divide the region $R_S \cup R_R$ to $R_R \setminus R_S, R_R \cap R_S$ and $R_S \setminus R_S$, and the region $R_P \cup R_Q$ to $R_Q \setminus R_P, R_Q \cap R_P$ and $R_P \setminus R_Q$, and we treat each subregion separately.

This would mean treating $7 \cdot 7 = 49$ cases, but we make some simplifications. Let us suppose that the first point A is in R_W or in R_Q. The cases when A falls in R_P, R_R or R_S are very similar. We make no assumptions on where B falls, but combine similar cases. Finally, we arrive to 8 cases, which are presented in Table 7.1. The table shows the various possibilities of regions where A and B can belong to (denoted by R_1 and R_2), the corresponding pair of corners which distinguish A and B, and the claim which proves this.

R_1	R_2	Distinguishing Corners	Claim used
R_W	$R_W \cup R_P \cup R_Q$	R, S	7.4.2
R_W	$R_R \cup R_S$	P, Q	7.4.2
$R_Q \setminus R_P$	$R_W \cup R_Q \cup R_P$	R, S	7.4.2
$R_Q \setminus R_P$	R_S	Q, S	7.4.1
$R_Q \setminus R_P$	$R_R \setminus R_S$	Q, R	7.4.2
$R_Q \cap R_P$	$R_W \cup R_Q \cup R_P$	R, S	7.4.2
$R_Q \cap R_P$	R_S	Q, S	7.4.1
$R_Q \cap R_P$	R_R	P, R	7.4.1

Table 7.1: The assignment of corners to the pair A, B, when $A \in R_1$ and $B \in R_2$.

We conclude the proof by stating and proving Claims 7.4.1 and 7.4.2. □

Claim 7.4.1. *If $A \in R_Q$ and $B \in R_S$ then $d_{G'}(A,Q) \neq d_{G'}(B,Q)$ or $d_{G'}(A,S) \neq d_{G'}(B,S)$, i.e., A and B are distinguished by the opposite corners Q and S. Similarly, if $A \in R_P$ and $B \in R_R$, then they are distinguished by P and R.*

Proof. Suppose for contradiction that $d_{G'}(A,Q) = d_{G'}(B,Q)$ and $d_{G'}(A,S) = d_{G'}(B,S)$.

Since $A \in R_Q$, $A \not\in R_S$, $B \in R_S$ and $B \not\in R_Q$, we have

$$BQ = d_{G'}(B,Q) = d_{G'}(A,Q) = AF + 1 + EQ$$
$$AS = d_{G'}(A,S) = d_{G'}(B,S) = BE + 1 + FS.$$

Adding these equations gives

$$BQ + AS = AF + EQ + BE + FS + 2. \tag{7.14}$$

Applying the triangle inequality to points B, E, Q and A, F, S and adding both the inequalities, we get

$$BQ + AS \leq BE + EQ + AF + FS,$$

which contradicts (7.14). A similar proof holds for $A \in R_P$ and $B \in R_R$ with corners P and R. □

Claim 7.4.2. *If two vertices A, B are outside of the union of the special regions of two adjacent corners, then they are distinguished by those two corners. For example, if $A, B \in V \setminus \{R_P \cup R_Q\}$ then P and Q distinguish A and B.*

Proof. The distances from A, B to P, Q in G' are same as that in G, and we know [174] that the set of two adjacent corners is a resolving set of a rectangle grid. □

Random Setting

Theorem 7.4.2 tells us that the MD of a grid and one extra edge must take a value from the set $\{2,3,4\}$, and in fact, all three values can occur. In Conjecture 7.4.1, we present a set of conditions, which we believe completely characterize the MD of a grid and one extra edge, but proving this conjecture seems tedious. Instead, we are interested in a probabilistic approach: what is the distribution of the MD when a uniformly randomly selected edge is added?

First we define some quantities which will be useful for the remaining section.

Definition 7.4.2 (Gain$'$)**.** *Let*

$$\text{Gain} = \text{Gain}(E,F) = |y_F - y_E| + |x_E - x_F| - 1$$

as in Definition 7.2.3, and let

$$\text{Gain}' = \max(0, ||y_F - y_E| - |x_E - x_F|| - 1) = \text{Gain}(F,(x_F, y_E)).$$

The notion of Gain captures the maximum gain for any pair of nodes. The pair of vertices (E,F) obviously have maximum gain, however, there can be other pairs which have the same gain. For two vertices X, Y, let us denote by $\text{Rec}(X,Y)$ the rectangle that has opposite corners X and Y, and sides parallel to the sides of the grid. Then, the pairs (A,B), with $A \in \text{Rec}(E,Q)$, and $B \in \text{Rec}(F,S)$ also have $\text{Gain}(A,B) = \text{Gain}$, since there is a shortest path between A and B in G that passes through both E and F. The notion of Gain$'$ has a very similar interpretation as Gain. We defined Gain$'$ as the gain between vertices F and (x_F, y_E). Notice that (x_F, y_E) is also a corner of $\text{Rec}(E,F)$. Therefore, while Gain is about the gain between the opposite corners, Gain$'$ is about the gain between the adjacent corners of the same rectangle (by symmetry the gain between E and (x_E, y_F) is also Gain$'$). Similarly to Gain, there are many other pairs of vertex pairs (A,B) with $\text{Gain}(A,B) = \text{Gain}'$. These are the pairs (A,B) with $A \in \text{Rec}(P,(x_F, y_E))$ and $B \in \text{Rec}(F,R)$, and symmetrically the pairs with $A \in \text{Rec}(R,(x_E, y_F))$ and $B \in \text{Rec}(E,P)$. Roughly speaking, we could thus say that Gain is useful if we want to measure the distance between vertex pairs with one vertex close to S and the other close to Q, while Gain$'$ is useful if we want to measure the distance between vertex pairs with one vertex close to P and the other close to R.

One of the key steps of the main proof in this section will be about treating the case when Gain$'$ is very small. This is the case when the extra edge has (or is close to having) a 45 degree angle with the sides of the grid, and $\text{Rec}(E,F)$ is (or is close to being) a square. In the extreme case, when $\text{Gain}' = 0$, no vertex pairs close to P and R use the extra edge, and the structure of the special and normal regions are different from the case when $\text{Gain}' \geq 1$. When $\text{Gain}' = 1$, there are still some subtle but inconvenient structural differences compared to the $\text{Gain}' \geq 2$ case. Fortunately, since we are adopting a probabilistic framework, in the proof we will be able to ignore the cases with $\text{Gain}' \leq 1$, as these cases have a vanishing probability of occurring.

Definition 7.4.3. *Let $\mathbb{P}_n$ be the probability distribution over potential extra edges*

$$e_n = ((x_E, y_E), (x_F, y_F))$$

that we can add to G_n, where (x_E, y_E) and (x_F, y_F) are two uniformly random vertices of G_n.

Theorem 7.4.3. *Let G_n be the $n \times n$ grid and let $G'_n = G_n \cup \{e_n\}$ with e_n sampled from distribution $\mathbb{P}_n$. Then, the following results hold:*

$$\lim_{n\to\infty} \mathbb{P}_n(\mathrm{MD}(G'_n) \in \{3,4\}) = 1 \tag{7.15}$$

$$\lim_{n\to\infty} \mathbb{P}_n\left(\mathrm{MD}(G'_n) = 3 \,\middle|\, \mathrm{Gain}' \textit{ is odd or} \min(|x_E - x_F|, |y_E - y_F|) < \frac{\mathrm{Gain}'}{2} + 2\right) = 1 \tag{7.16}$$

$$\lim_{n\to\infty} \mathbb{P}_n\left(\mathrm{Gain}' \textit{ is odd or} \min(|x_E - x_F|, |y_E - y_F|) < \frac{\mathrm{Gain}'}{2} + 2\right) = \frac{19}{27}. \tag{7.17}$$

According to Theorem 7.4.3, the asymptotic probability that the MD of the square grid with an extra edge is three is at least 19/27. We believe that it is also true that the MD is at least four when Gain' is even and $x_E - x_F \geq \mathrm{Gain}'/2 + 2$. If we could prove this, we could state that the asymptotic probability of $\mathrm{MD}(G')$ being three is *exactly* 19/27, and $\mathrm{MD}(G') \to \mathrm{Ber}(8/27) + 3$ in probability, where $\mathrm{Ber}(q)$ is a Bernoulli random variable with parameter q. We believe that a brute-force approach similar to the proof of Theorem 7.4.2 can work, but it requires a tedious case-by-case analysis that is out of scope of this dissertation.

The probabilistic formulation of Theorem 7.4.3 allows us to ignore the edge-cases that would be too tedious to check individually, but it introduces new challenges as well. To address these new challenges and we reduce equations (7.15)-(7.17) to technical Lemmas D.3.1, D.3.2 and D.3.3, which are of deterministic nature. We give the proof of Theorem 7.4.3 at the end of this section, but we defer the proof of the technical lemmas to Section D.3.

The specific edge-cases that we ignore using the probabilistic formulation are given in Assumption 7.4.3.

Assumption 7.4.3 (edge-case removal)**.** *We assume that*

1. $x_F \neq x_E$
2. $\mathrm{Gain}' \geq 2$
3. *none of E and F lie on the boundary of the grid.*

In addition to Assumption 7.4.3, we are also going to make use of Assumption 7.4.2 as we did in the proof of Theorem 7.4.2. Assumptions 7.4.2 and 7.4.3 applied together have some additional implications.

Remark 7.4.1. *Assumption 7.4.2 and 7.4.3 together imply that*

1. $x_F < x_E$
2. $y_E < y_F$.
3. $x_E - x_F < y_F - y_E$

Using Assumption 7.4.2 in the probabilistic formulation is not as straightforward anymore, as symmetry breaking can also break the uniformity of the sampling of the extra edge. Indeed, sampling a random edge that satisfies Assumption 7.4.2 is not the same as sampling an edge from $\mathbb{P}_n$ and rotating and reflecting it so that Assumption 7.4.2 is satisfied. In Claims 7.4.3 and 7.4.5, we are going to show that after removing only $O(n^3)$ edges from $V \times V$, and thus slightly changing the distribution $\mathbb{P}_n$, the symmetry breaking will not violate the uniformity of the sampling anymore.

Definition 7.4.4 ($\mathcal{P}$,$\mathcal{Q}$,$\tilde{\mathbb{P}}_n$,$\mathbf{Q}_n$)**.** *Let $\mathcal{P}$ be the set of extra edges $((x_E, y_E), (x_F, y_F))$ that satisfy Assumption 7.4.3, and let $\mathcal{Q}$ the set of extra edges that satisfy both Assumptions 7.4.2 and 7.4.3. Let $\tilde{\mathbb{P}}_n$ and $\mathbf{Q}_n$ be the uniform probability distribution over $\mathcal{P}$ and $\mathcal{Q}$, respectively.*

In Claim, 7.4.3 we show that $\mathbb{P}_n$ is close to $\tilde{\mathbb{P}}_n$, and in Claim 7.4.5 we show that $\tilde{\mathbb{P}}_n$ is close to $\mathbf{Q}_n$. These two claims allow us to use $\mathbf{Q}_n$ instead of $\mathbb{P}_n$ in the proof of Theorem 7.4.3.

Claim 7.4.3. *For $\mathbb{P}_n$ and $\tilde{\mathbb{P}}_n$ given in Definitions 7.4.3 and 7.4.4,*

$$\lim_{n\to\infty} \|\mathbb{P}_n - \tilde{\mathbb{P}}_n\|_{TV} = 0.$$

Proof of Claim 7.4.3. The support of $\mathbb{P}_n$ is $V \times V$, and $|V \times V| = n^4$ because each of the four coordinates x_E, y_E, x_F and y_F can take four values. Recall, that $\mathcal{P} \subset (V \times V)$, and the set $(V \times V) \setminus \mathcal{P}$ consists of the edges that do not satisfy Assumption 7.4.3. Therefore, to upper bound the cardinality of $(V \times V) \setminus \mathcal{P}$, it is enough to upper bound the number of edges violating each of the conditions in Assumption 7.4.3. It is clear that the number of edges that violate the first condition is n^3; the coordinates x_E, y_E, y_F can be chose arbitrarily in n^3 different ways, and then setting $x_F = x_E$ gives exactly one unique edge that violates the first condition. For a more insightful but less precise explanation, notice that the original set $V \times V$ had four degrees of freedom, and we lost one to violating the condition, hence we are left with three degrees of freedom and $O(n^3)$ edges. It is not hard to see that we lose one degree of freedom to violate the second and third conditions as well, and therefore the number of edges violating these conditions are also $O(n^3)$. We conclude that the number of edges in $(V \times V) \setminus \mathcal{P}$ are also of order $O(n^3)$.

Then,

$$\begin{aligned}
2\|\mathbb{P}_n - \tilde{\mathbb{P}}_n\|_{TV} &= \sum_{e\in\mathcal{P}} |\mathbb{P}_n(e) - \tilde{\mathbb{P}}_n(e)| + \sum_{e\in V\times V\setminus\mathcal{P}} \mathbb{P}_n(e) \\
&= |\mathcal{P}|\left|\frac{1}{|V\times V|} - \frac{1}{|\mathcal{P}|}\right| + \frac{|(V\times V)\setminus\mathcal{P}|}{|V\times V|} \\
&= (n^4 + O(n^3))\left|\frac{1}{n^4} - \frac{1}{n^4+O(n^3)}\right| + \frac{O(n^3)}{n^4} \\
&= O\left(\frac{1}{n}\right).
\end{aligned}$$

□

Definition 7.4.5 (H)**.** *Let us consider the following actions on the extra edges of the grid:*

1. *by h_1 the reflection along the vertical line through the midpoints of sides PQ and SR,*
2. *by h_2 the reflection along the horizontal line through the midpoints of sides PS and QR,*
3. *and by h_3 switching the two endpoints of the edge.*

Let H be the group generated by h_1, h_2 and h_3 acting on the edges.

Notice that the group H acting on the edges is isomorphic to the $\mathbb{Z}_2^3$ group. Indeed, all three actions have order two and commute with each other. Thus, H can be described as $\{h_1^i h_2^j h_3^k \mid i,j,k \in \{0,1\}\}$. Also, notice that for $e \in \mathcal{Q}$, applying h_1, h_2 and h_3 flips the inequality labeled with the same index in Remark 7.4.1, and keeps the other two inequalities unchanged.

Definition 7.4.6. *Let h be a map, which for each edge $e \in \mathcal{Q}$ returns the set of edges that we get by applying the elements of H to e.*

The sets $h(e)$ can be seen as orbits of the edges under the action of H.

Claim 7.4.4. *With $\mathcal{P}$, $\mathcal{Q}$ and h given in Definitions 7.4.4 and 7.4.6, the following three statements must hold:*

1. *$|h(e)| = 8$ for every $e \in \mathcal{Q}$*
2. *the orbits of the edges in $\mathcal{Q}$ are disjoint, i.e., $h(e_1) \cap h(e_2) = \emptyset$ for $e_1 \neq e_2 \in \mathcal{Q}$*
3. *for every $e \in \mathcal{P}$, there is an $e_2 \in \mathcal{Q}$ with $e \in h(e_2)$.*

Proof of Claim 7.4.4. Statement 1 follows from the observation that every non-trivial group action in H flips a different subset of the inequalities in Remark 7.4.1, and two edges cannot coincide if they satisfy different sets of inequalities. For statement 2, since H is a group, if two orbits $h(e_1), h(e_2)$ have a non-empty intersection, we must have $e_1 \in h(e_2)$. However, every non-trivial group action in H flips at least one of the inequalities of Remark 7.4.1, which implies that if we apply a non-trivial group action, the image of $e_2 \in \mathcal{Q}$ cannot be in $\mathcal{Q}$. For statement 3, for edge $e \in \mathcal{P}$, let $\mathbf{v}(e) \in \{0,1\}^3$ be a binary vector, whose i^{th} entry indicates that e violates inequality i in Remark 7.4.1. Then $h_1^{\mathbf{v}(e)_1} h_2^{\mathbf{v}(e)_2} h_3^{\mathbf{v}(e)_3}$ is a group action that flips exactly the inequalities that are violated by e, and thus maps e into $\mathcal{Q}$. Let the image of e under this action be e_2, and then indeed, $e \in h(e_2)$. □

Claim 7.4.5. *Let $\mathcal{A}_n$ be a sequence of events defined on graph G'_n that are closed under the action of H. Then,*

$$\lim_{n\to\infty} |\mathbb{P}_n(\mathcal{A}_n) - \mathbf{Q}_n(\mathcal{A}_n)| = 0.$$

Proof of Claim 7.4.5. The three statements of Claim 7.4.4 together imply that the orbits $h(e)$ of $e \in \mathcal{Q}$ partition $\mathcal{P}$ into sets of cardinality 8. A simple corollary is that $|\mathcal{P}| = 8|\mathcal{Q}|$.

Let us suppose that event $\mathcal{A}_n$ is closed under the action of H, or formally as $e \in \mathcal{A}_n$ implies $h(e) \subset \mathcal{A}_n$. This closedness property, combined with Claim 7.4.4 implies that the edges in $\mathcal{A}_n$ can also be counted as 8 times the number of edges in $\mathcal{A}_n \cap \mathcal{Q}$. Then,

$$\tilde{\mathbb{P}}_n(\mathcal{A}_n) = \frac{|\mathcal{A}_n|}{|\mathcal{P}|} = \frac{8|\mathcal{A}_n \cap \mathcal{Q}|}{8|\mathcal{Q}|} = \mathbf{Q}_n(\mathcal{A}_n). \tag{7.18}$$

Finally, we combine equation (7.18) with Claim 7.4.3 as

$$\lim_{n\to\infty} |\mathbb{P}_n(\mathcal{A}_n) - \mathbf{Q}_n(\mathcal{A}_n)| = \lim_{n\to\infty} |\mathbb{P}_n(\mathcal{A}_n) - \tilde{\mathbb{P}}_n(\mathcal{A}_n)| \le \lim_{n\to\infty} \|\mathbb{P}_n(\mathcal{A}_n) - \tilde{\mathbb{P}}_n(\mathcal{A}_n)\|_{TV} = 0,$$

and the proof is completed. □

Now we have all the ingredients to prove Theorem 7.4.3.

Proof of Theorem 7.4.3. Since all events in the statement of Theorem 7.4.3 are closed under the action of H on the square grid, Remark 7.4.5 shows that it is enough to prove equations (7.15)-(7.17) for distribution $\mathbf{Q}_n$. Note that because of statements 1 and 3 of Remark 7.4.1, $\min(|x_E - x_F|, |y_E - y_F|) = x_E - x_F$ for edges in $\mathcal{Q}$. Hence, the second condition in (7.16) and (7.17) reduces to $x_E - x_F < \text{Gain}'/2 + 2$ for distribution $\mathbf{Q}_n$.

The rest of the proof relies on Lemmas D.3.1-D.3.3 given in Section D.3, which have purely deterministic nature. Lemma D.3.1 shows that for extra edges in $\mathcal{Q}$ (that is edges satisfying Assumption 7.4.2 and 7.4.3), the metric dimension of G' will be at least three deterministically, which, combined with Theorem 7.4.2, gives equation (7.15). Lemma D.3.2 shows that there exists a resolving set of cardinality three for every extra edge in $\mathcal{Q}$ with an odd Gain$'$. For the extra edges in $\mathcal{Q}$ with an even Gain$'$ and with $x_E - x_F < \text{Gain}'/2 + 2$, there exist different resolving sets of cardinality three, which is proved in Lemma D.3.3. Thus, Lemmas D.3.1, D.3.2 and D.3.3 combined imply equation (7.16).

Finally, we show equation (7.17). Let us denote by C the subset of vertex pairs in $\mathcal{Q}$ that satisfy the condition in equation (7.17), i.e.,

$$C = \left\{ (E, F) \in \mathcal{Q} \,\middle|\, \text{Gain}' \text{ is odd or } x_E - x_F < \frac{\text{Gain}'}{2} + 2 \right\}.$$

Let the complement of C be

$$\bar{C} = (V \times V) \setminus C = \left\{ (E, F) \in \mathcal{Q} \,\middle|\, \text{Gain}' \text{ is even and } x_E - x_F \geq \frac{\text{Gain}'}{2} + 2 \right\}.$$

Next, we calculate $|\bar{C}|/|\mathcal{Q}|$. Let $x_E - x_F = a$ and $y_F - y_E = b$, which together with Assumption 7.4.3 gives

$$b - a - 1 = \text{Gain}'.$$

Then, the conditions on a, b that need to be satisfied for an edge to be in $\bar{C}$ can be reformulated as :

1. $b - a$ is odd (equivalent to Gain$'$ is even)
2. $b - a \geq 3$ (equivalent to Gain$' \geq 2$)
3. $a \geq \frac{b}{3} + 1$ (equivalent to $x_E - x_F \geq \frac{\text{Gain}'}{2} + 2$)
4. $1 \leq a, b \leq n - 2$, as the extra edge is not horizontal nor vertical, and does not touch the boundary of the grid.

Let $b - a = 2i + 1$ with $i \geq 1$. With this parameterization, the first two conditions are already obviously satisfied. Substituting $a = b - 2i - 1$ into $a \geq b/3 + 1$ gives $b \geq 3(i + 1)$. Hence, for a fixed i, b can have values from $3(i + 1)$ to $n - 2$, and consequently, the maximum value that i can take is $\lfloor (n - 5)/2 \rfloor$. Note that for a given pair (a,b), there are $(n - a - 1)(n - b - 1)$ possible edges in G_n which do not touch the boundary. Therefore,

$$|\bar{C}| = \sum_{i=1}^{\lfloor \frac{n-5}{3} \rfloor} \sum_{b=3(i+1)}^{n-2} (n - b + 2i)(n - b - 1),$$

which reduces asymptotically to

$$|\bar{C}| = \frac{1}{27} n^4 + O(n^3).$$

Therefore,

$$\mathbf{Q}_n(\bar{C}) = \frac{|\bar{C}|}{|\mathcal{Q}|} = \frac{\frac{1}{27} n^4 + O(n^3)}{\frac{1}{8} n^4 + O(n^3)} = \frac{8}{27} + O\left(\frac{1}{n}\right), \tag{7.19}$$

Hence, $\mathbf{Q}_n(C) = 1 - \mathbf{Q}_n(\bar{C}) \to 19/27$, which shows equation (7.17) and completes proof of the theorem. □

Precise Conjecture

Finally, we present our precise conjecture which completely characterizes metric dimension for any 2-dimensional grid graph augmented with one edge. We believe this can be proved by rigorous case-wise analysis but it is out of the scope of this dissertation. We have verified this conjecture for square grids with sizes up to 15 × 15 using simple C++ programs available at [172]. Note that the conjecture is stated not only for square grids but also for $m \times n$ rectangular grids, but for these graphs we only verified the conjecture for a few parameter values due to the increased number of cases.

Conjecture 7.4.1. *Let G be a 2-dimensional grid graph with m rows and n columns. Let e be the edge between vertices $F = (x_F, y_F)$ and $E = (x_E, y_E)$ with $x_F, x_E \in \{1, ..., n\}$, $y_F, y_E \in \{1, ..., m\}$, with the assumption that $EF \geq 2$. Let $G' = (V, E_G \cup \{e\})$ be the grid augmented with one edge. Let* $\mathrm{Gain} = |y_E - y_F| + |x_F - x_E| - 1$ *and* $\mathrm{Gain}' = ||y_F - y_E| - |x_F - x_E|| - 1$.

- $\mathrm{MD}(G') = 4$ *if all of the following conditions are satisfied:*
  - *None of the endpoints of e is a corner of the grid. i.e.,*

    $$(x_E, y_E), (x_F, y_F) \notin \{(1,1), (n,1), (1,m), (n,m)\}$$

  - Gain' *is positive and even.*
  - $\min(|x_F - x_E|, |y_F - y_E|) \geq \frac{\mathrm{Gain}'}{2} + 2$
- $\mathrm{MD}(G') = 2$ *if any of the following conditions is satisfied:*
  - $\mathrm{Gain} = 1$
  - $\mathrm{Gain}' \leq 1$, Gain *is odd and one of the endpoints is a corner of the grid.*
  - $\mathrm{Gain}' \geq 3$, Gain *is odd,* $\mathrm{Gain} - \mathrm{Gain}' \leq 2$ *and one of the endpoints is a corner of the grid.*
  - Gain *is odd and both endpoints are corners of the grid.*
- $\mathrm{MD}(G') = 3$ *for all other cases.*

8 Source Identification via Contact Tracing

In this chapter, we develop a new framework for source identification, that is intended to improve upon the limitations of previous frameworks as discussed in Chapter 1. Since we are introducing a new framework, the beginning of the current chapter is focused on building intuition (Section 8.1), and it is more introductory in nature compared to the beginning of the previous chapters. An overview of the rest of the chapter is presented at the end of Section 8.1.

This chapter is based on the publication [189] by Ódor, Vuckovic, Ndoye and Thiran.

8.1 Rationale and overview

During the COVID-19 pandemic, we have seen a revolution of the contact tracing technology, which helped track and contain the epidemic [39, 150]. Some contact tracing programs were conducted by governmental/health agencies [194], while others relied on decentralized approaches [234]. Most contact tracing approaches work by notifying people who could have received the infection from known infectious patients, i.e., they trace "forward" in time. However, some advocate that a "bidirectional" tracing, where the past history of the infection is also tracked, can be more effective [38, 81, 146]. Our goal in this section is to improve on the common limitations of source identification algorithms reviewed in Section 1.3 based on ideas inspired from these recent advances in contact tracing.

We propose a new framework for source identification, which we call Source Identification via Contact Tracing Framework (SICTF). In SICTF, algorithms can have two types of queries: contact queries, which can be used to explore the network, and sensor (test) queries, after which agents reveal their symptom onset time as before. The goal of the algorithm is to find the source as accurately as possible, while minimizing the number of contact and sensor queries. The SICTF is a way to formalize the source identification task; it determines the goal of the algorithm and how information can be gained about the epidemic, but it does not specify the underlying epidemic and mobility data models (simulated or real). In this chapter, we analyse different algorithms in the SICTF with various epidemic and mobility models.

Besides specifying the possible queries that algorithms can make, the SICTF also determines the way the outbreak is detected, which marks the starting time of the source identification task. In the S1/S2 models introduced in Chapter 1, the source identification task often starts long after the outbreak, when essentially all agents in the network are infected [198], which can be seen as an additional limitation of source identification frameworks. The SICTF is also closely related to contact tracing frameworks, where it is standard to assign a probability that each node spontaneously self-reports after developing symptoms, which triggers the activation of contact tracing algorithms [150, 38]. In the SICTF, we adopt the idea of self-reporting with a slight modification. We believe that the most interesting time to perform the source identification task is when a new disease (or a new mutation of the disease) appears, and therefore we tie these self-reporting events to hospitalizations, where infections are properly diagnosed by healthcare professionals. In particular, this means that the SICTF can only be applied to epidemic data (and models) where hospitalizations are well-defined. In this chapter, we use the datasets generated by the Data-driven COVID Simulator (DCS) introduced in [162], which is one of the most realistic toolboxes that generate datasets modelling COVID-19, which we are aware of (notably, hospitalizations are part of the model). We also propose synthetic approximations for the epidemic and mobility models in the DCS; the Deterministically Developing Epidemic model and the Household Network Model, which improve the interpretability of our results since they have fewer parameters.

We propose a simple algorithm called LocalSearch (LS), which adaptively traces back the transmission path from the first hospitalized patient to the source. The LS algorithm is quite efficient at finding the source; the number of contact and sensor queries that it uses does not depend on the size of the network, but only on the local neighborhood of the source. Moreover, the LS algorithm provably finds the source with 100% accuracy, because of our assumption that every contact and sensor query is answered without noise. However, it is well-known that data-availability is a major issue in contact tracing [27], either because the agents do not comply with contact tracing efforts, or possibly (and in particular in the current COVID-19 epidemic) because they do not develop symptoms, and are unaware that they have the disease. In this chapter, we model the effect of asymptomatic agents. When queried and tested, these agents do not reveal their time of infection, only whether they have or had the disease at some point. We show that the accuracy of the LS algorithm drops in the presence of asymptomatic agents, because the algorithm can get stuck while tracing back the transmission path from the first hospitalized patient to the source. Therefore, we propose an improved version of LS called LS+, which accounts for the presence of asymptomatic agents by placing more sensors. We are not aware of any previous work in the source identification literature that models the effect of asymptomatic patients, but the resulting model can be seen as a mix between the snapshot and the S1/S2 models. We mention that non-complying agents or agents who provide noisy observations have been studied by [14, 110, 163]. Non-complying agents could also be included in our framework by treating them as asymptomatic agents (even though in this case we have no information about whether the agent had the disease or not), without jeopardizing the correctness of our algorithms.

We benchmark the LS and LS+ algorithms in both our data-driven and our synthetic epidemic and mobility models, and we compare them to state-of-the-art adaptive [224] and non-adaptive [130, 161] algorithms originally developed for the S1/S2 models, tailored to the SICTF. We find that both LS and LS+ outperform these baseline algorithms in accuracy (probability of finding the correct source).

While the LS/LS+ are designed to be simple algorithms, their theoretical analysis is quite challenging. Nevertheless, we are able to provide rigorous results about the success probability of both algorithms after a series of simplifications to the epidemic and mobility models, by extending some recent results on the theory of exponential random trees [93, 167], which have previously not been connected to the source identification literature. We present these theoretical results in Section 8.4, after formally introducing the SICTF, our models and the LS/LS+ algorithms in Section 8.3. By simulations, we show that our analytic results approximate the accuracy of the algorithms well, even in the most realistic setting in Section 8.5. Our analytic results provide additional insight into how the parameters of the epidemic and mobility models affect the performance of the algorithms. We discuss these insights along with some non-rigorous computations that mirror our main proof ideas in Section 8.2. Reading Section 8.2 before Sections 8.3-8.5 is useful to build intuition, but is not necessary to understand the chapter.

8.2 Warmup Results

8.2.1 A Simple Network and Epidemic Model and a Simple Algorithm

Let us consider a time-dependent network model, where each agent meets d new agents each day in such a way that the contact network is an infinite tree (ignoring the label of the edges giving the propagation time along the edge). This network models homogeneous mixing in a very large population; we consider more realistic network models in Section 8.3. On this network, we consider an epidemic model that starts at $t = 0$ with one infected agent, and then progresses as infected agents infect their d susceptible contacts each independently with probability p_i each day. Since our goal is to study the epidemic process, it is sufficient to track only the agents who are already infectious (also called *internal nodes*), and the agents who are in contact with infectious agents at time t (also called *external nodes*), as shown in Figure 8.1 (a)-(c). For $d = 1$, the spread of the infection is then equivalent to the growth a random tree $\mathcal{T}_t$ rooted at the source of the infection, known under the name of Random Exponential Recursive Tree (RERT) and recently introduced in [167]. Because of the similarities of the models, we refer to the model with general d as RERT in the remaining of this section. We point out that the standard literature on elementary branching processes such as Galton-Watson trees or random recursive trees [75] is not applicable in our scenario, because these branching processes have no notion of global time (i.e., a node in such processes becomes infectious immediately after receiving the infection), whereas nodes in diseases commonly go through an exposed, non-infectious period before becoming infectious, which is well captured by the

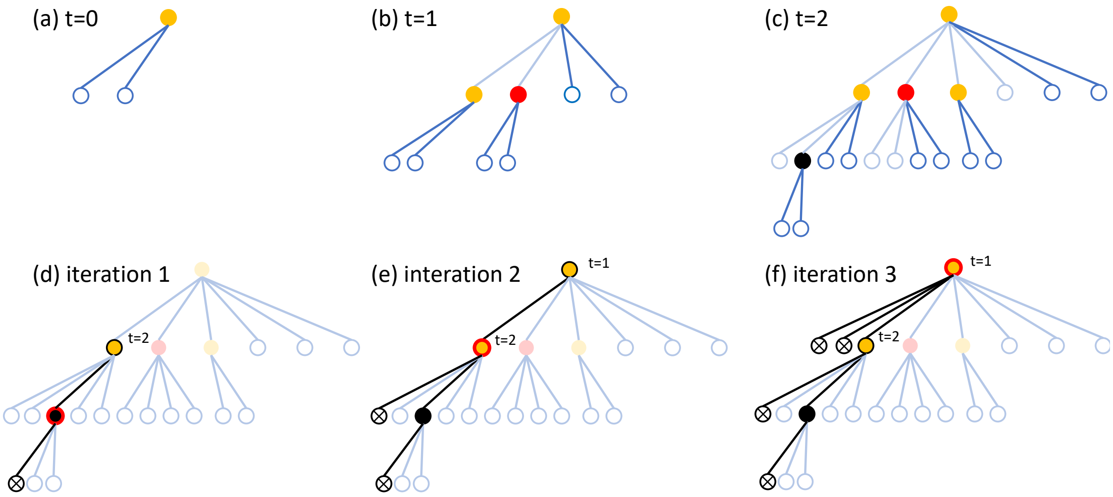


Figure 8.1: (a)-(c) shows the spread of the infection in the model considered in Section 8.2.1, which is equivalent to the growth of the RERT, with $d = 2$. Dark blue edges show the contacts on day t, and light blue edges show contacts present on previous days (and thus subfigures). Orange (resp., red; black) nodes mark symptomatic non-hospitalized (resp, asymptomatic; symptomatic hospitalized) nodes. (d)-(f) shows the LS source identification algorithm introduced in Section 8.2.2, which succeeds in this example because there are no asymptomatic nodes on the transmission path between the first hospitalized node and the source. Black edges show the queried edges, and black stroke marks nodes already discovered by the algorithm. A node with black X marks a negative test result, and red stroked node marks the node currently maintained as source candidate by the LS algorithm.

RERT model. We mention that there is literature on more advanced branching processes that do have a notion of global time, e.g. Crump-Mode-Jagers trees [124], however we opt for the RERT because of its simple definition.

After a node (patient) becomes infected, the disease can take three courses (which for now do not affect $\mathcal{T}_t$): with probability p_a the patient is asymptomatic, with probability $(1-p_a)p_h$ the patient is hospitalized, and with probability $(1-p_a)(1-p_h)$ the patient recovers without hospitalization. The governmental/health agency learns about the outbreak when the first hospitalization occurs (see Figure 8.1 (c)) and starts the source identification process right away. It can inquire about the contacts of each agent and it can test the agents. From patients that were symptomatic (at any point in time in the past), the agency learns about their symptom onset time (which, in this simple model, is always one day after the infection time), but from asymptomatic patients it only learns that they had (or have) the disease at some point when they are tested. The framework introduced in this paragraph (including both the identification of the outbreak through the first hospitalization, and the possible actions the agency can take) is a simplified version of the SICTF (Source Identification via Contact Tracing Framework), introduced in Section 8.3.3.

The network and epidemic models introduced in this section have four parameters: d, p_i, p_a, p_h, and it is important to understand how each of them affects the difficulty of source identification in the SICTF. We distinguish two important factors. First, if the outbreak is not detected rapidly enough, the length of the transmission path to the first hospitalized agent is long, and source identification becomes then difficult, because a lot of information needs to be recovered. Therefore, a low p_i, a low p_h and/or a high p_a parameter can hinder source identification (recall that the probability of hospitalization was $p_h(1-p_a)$). The second factor is related to the difficulty of recovering information about the transmission path. If p_a is high, then there are a lot of nodes who are asymptotic and therefore do not reveal their symptom onset time, making source identification very difficult. Since p_a affects both the length of the transmission path and the amount of collected information, it is safe to expect that, of all parameters, p_a has the largest effect on the difficulty of source identification. The parameter d is interesting, because a large d can reduce the length of the transmission path, but it also makes the information about the transmission path less accessible as more agents need to be tested. Since in this chapter we do not set a hard constraint on the total number of available tests, the advantage of a shorter path takes over the drawback of additional tests and a large d increases the success probability.

To say anything quantitative about source identification in the SICTF, we must discuss specific algorithms that solve the source identification task. In this chapter we propose a simple algorithm called LocalSearch (LS), shown in Figure 8.1 (d)-(f). The LS algorithm maintains one candidate node s_c at each iteration (initially, the first hospitalized node), which is always symptomatic, and it updates it in a greedy way: at the time of the infection of s_c, all its d incident edges are queried, and all its d neighbors are tested. Then the agent with the lowest reported infection time will be the new candidate s_c. The algorithm stops when s_c does not change anymore between two consecutive iterations. For simplicity, we assume that the infection does not spread any further during these iterations, however, this assumption does not affect the ability of the algorithm to find the source or not. Indeed, it is not difficult to see that on tree networks, LS succeeds if and only if there are no asymptomatic nodes on the transmission path from the source to the first hospitalized agent. This observation leads us to enhance the LS algorithm by also searching within the neighbors of asymptomatic nodes; we explore this idea in the LS+ algorithm introduced in Section 8.3.5. We are not aware of this simple greedy algorithm being studied in the context of source identification, although similar ideas were implemented for non-adaptive source identification to lower the runtime of the algorithms [192].

8.2.2 Back of the Envelope Calculation

Now, we have all the tools to estimate the probability of success of the LS algorithm. First we condition on the course of the disease in the source. With probability p_a, the source is asymptomatic and LS can never succeed. With probability $(1-p_a)p_h$, the source itself becomes hospitalized, and LS always succeeds. Finally, with probability $(1-p_a)(1-p_h)$ the source is symptomatic but not hospitalized, which we call event $\mathcal{A}$. If event $\mathcal{A}$ happens, then LS may or may not succeed depending on whether there are any asymptomatic nodes on the transmission path. More precisely, conditioned on event $\mathcal{A}$ and on the transmission path having length l, the probability of success is $(1-p_a)^{l-1}$ (since there are $l-1$ nodes on the path which can be asymptomatic), which implies

$$\mathbb{P}(\text{success}) = (1-p_a)p_h+(1-p_a)(1-p_h)\left(\sum_{l=1}^{t}\mathbb{P}\left(\text{transmission path has length } l \mid \mathcal{A}\right)(1-p_a)^{l-1}\right). \tag{8.1}$$

The difficult part is to compute the distribution of the transmission path conditioned on event $\mathcal{A}$; indeed we already saw that all four parameters d, p_i, p_a, p_h affect this distribution in a non-trivial way. Let us perform a back of the envelope computation to get more insight into the effect of these parameters. The exact structure of the infection tree will not matter for this computation, only its *profile* does. It is denoted by $\mathcal{T}_t(l)$ and defined as the number of (internal) nodes at level l (i.e., at distance l from the source of the infection). Remember that by definition the RERT has $d\cdot\mathcal{T}_{t-1}(l-1)$ external nodes on level l, and that at time t each external node is promoted to be internal with probability p_i to form $\mathcal{T}_t$. Consequently, the level of a node h added at time $t>0$ has the same distribution (conditioned on the tree $\mathcal{T}_{t-1}$ at the previous step) as the size (number of internal nodes) of the profile $\mathcal{T}_{t-1}(l-1)$, that is,

$$\mathbb{P}(\text{level}(h) = l \mid \mathcal{T}_{t-1} = T_{t-1}) = \frac{T_{t-1}(l-1)}{|T_{t-1}|}. \tag{8.2}$$

Working on the RERT directly can be a daunting task, therefore we propose to approximate the numerator and the denominator of equation (8.2) by $\mathbb{E}[\mathcal{T}_{t-1}(l-1)]$ and $\mathbb{E}|[\mathcal{T}_{t-1}|]$, respectively.

It can be shown by a simple inductive argument, or by generating functions as in [167], that for RERTs we have $\mathbb{E}[\mathcal{T}_t(l)] = \binom{t}{l}(dp_i)^l$ and $\mathbb{E}[|\mathcal{T}_t|] = (1+dp_i)^t$, which suggests a binomial distribution for the level of h. And indeed, we can approximate the distribution of the level of a node h added at time t as

$$\begin{aligned}\mathbb{P}(\text{level}(h) = l) &\approx \frac{\mathbb{E}[\mathcal{T}_{t-1}(l-1)]}{\mathbb{E}[|\mathcal{T}_{t-1}|]}\\ &= \frac{\binom{t-1}{l-1}(dp_i)^{l-1}}{(1+dp_i)^{t-1}}\\ &= \binom{t-1}{l-1}\left(\frac{dp_i}{1+dp_i}\right)^{l-1}\left(1-\frac{dp_i}{1+dp_i}\right)^{t-l}\\ &= \mathbb{P}(\text{Bin}(t-1, q) = l-1),\end{aligned}$$

with $q = dp_i/(1+dp_i)$.

One of the main challenges of this calculation is that we do not know the day of the first hospitalization t conditioned on event $\mathcal{A}$, we only know that each node is hospitalized with probability $(1-p_a)p_h$, which means that the index of the first hospitalized node follows a geometric distribution with mean $1/((1-p_a)p_h)$. We approximate $t-1$ by the first time that the expected size of the infection tree (excluding the source since we condition on event $\mathcal{A}$) exceeds the expected index of the first hospitalized node. Therefore we solve

$$\mathbb{E}[|\mathcal{T}_{t-1}|-1] = (1+dp_i)^{t-1} - 1 = \frac{1}{(1-p_a)p_h} = \mathbb{E}[\text{index of the first hospitalized node}]$$

for t (relaxing the constraint that t is an integer), which gives

$$t-1 = \frac{\log\left(1+\frac{1}{(1-p_a)p_h}\right)}{\log(1+dp_i)}.$$

Consequently, we approximate $\mathbb{P}\left(\text{transmission path has length } l \mid \mathcal{A}\right)$ by $\mathbb{P}(\text{Bin}(t-1,q) = l-1)$. Continuing equation (8.1), and using the well-known expression of the probability generating function of the binomial distribution, we get

$$\begin{aligned}\mathbb{P}(\text{success}) &\approx (1-p_a)p_h + (1-p_a)(1-p_h)\left(\sum_{l=1}^{t}\mathbb{P}\left(\text{Bin}\left(t-1,q\right) = l-1\right)(1-p_a)^{l-1}\right)\\ &= (1-p_a)\left(p_h + (1-p_h)\left((1-p_a)\frac{dp_i}{1+dp_i} + 1 - \frac{dp_i}{1+dp_i}\right)^{\frac{\log\left(1+\frac{1}{(1-p_a)p_h}\right)}{\log(1+dp_i)}}\right). \qquad (8.3)\end{aligned}$$

One can check that this expression agrees with our qualitative intuition. However, it is not at all clear whether it is valid because of the strong approximations made in some steps of the above computation. In Section 8.4, we prove a rigorous upper bound on the success probability, and we also provide much more careful approximations by proving exact theorems about the simplified models that we use. Then, in Section 8.5 we compare our results with simulation results on synthetic data, as well as with data generated by the DCS model.

8.3 Models, Methods, Algorithms

8.3.1 Epidemic Models

The DCS Model

We call DCS the model implemented by [162]. The DCS model is fairly complex, and we only give a brief overview.

Each agent in the agent set V can be in one of 8 states: susceptible, exposed, asymptomatic infectious, pre-symptomatic infectious, symptomatic infectious, hospitalized, recovered or dead. Transitions between different states are characterized by counting processes described by stochastic differential equations with jumps. The most important, and also most complicated of these counting processes is the exposure counting process $N_i(t)$, which is modeled by a Hawkes process for each agent i. Hawkes processes are point processes with a time-dependent, self-exciting conditional intensity function $\lambda_i^*(t)$.

$$\lambda_i^*(t) = \beta \sum_{j \in V \setminus \{i\}} \int_{t-\delta}^{t} K_{i,j}(\tau)\, \gamma \mathrm{e}^{-\gamma(t-\tau)}\, d\tau \tag{8.4}$$

where the kernel $K_{i,j}(\tau)$ indicates whether j has been at time τ at the same site where i is at time t, and whether j is in the infectious state. Parameters γ and δ are the decay of infectiousness at sites and the non-contact contamination window, respectively, and they account for the fact that j can infect i even if they are never at the same site, as j can leave some pathogens behind (airborne for instance). Parameter β is the transmission rate for symptomatic and asymptomatic individuals, and it comes in two versions: β_c accounts for infections outside the household and β_h accounts for infection in the household. Parameters β_c and β_h are fitted to the COVID-19 infection data of Tubingen from 12/03/2020 to 03/05/2020 using Bayesian Optimization. The model also has a parameter for the relative asymptomatic transmission rate built into the function $K_{i,j}(\tau)$, which scales down the infectiousness of asymptomatic agents (to 55% of the infectiousness of symptomatic agents by default).

Once a susceptible agent becomes infected, the disease can take three possible courses (see Figure 8.2 (a)). With probability p_a, the agent becomes asymptomatic infectious after time T_E, and then recovers after time T_I. With probability $1 - p_a$, the agent becomes pre-symptomatic infectious after time T_E, next symptomatic infectious after time T_P, and then recovers with probability $1 - p_h$ after time $T_I - T_P$, or becomes hospitalized with probability p_h after time T_H. Agents in the DCS are also assigned age values based on demographic data, and the hospitalization probability p_h of each agent is determined based on its age (following COVID-

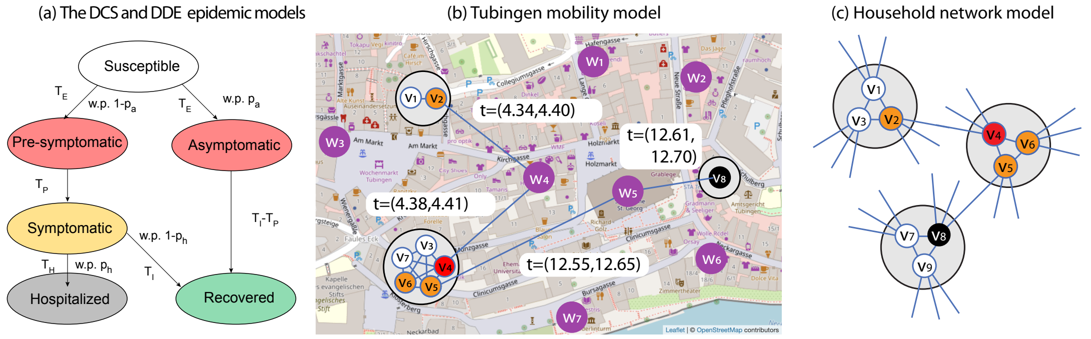


Figure 8.2: (a) The flow diagram of the DCS and DDE epidemic models. (b) A possible epidemic outbreak in the Tubingen mobility model, and (c) the Household network model. The large grey circles mark households, and the purple nodes mark places, otherwise we use the same coloring as in (a). In both cases (b) and (c), the transmission paths are (v_2, v_4, v_5, v_8).

19 infection data). The times T_E, T_P, T_I and T_H are drawn from an appropriately parametrized (using values from the COVID-19 literature) lognormal distribution as shown in Table 8.1.

The DDE Model

We start by taking the DCS model [162], which we simplify to enable its theoretical analysis. In the Deterministically Developing Epidemic (DDE) model, continuous time (used in DCS) is replaced by discrete time-steps: we refer to one time-step in the DDE as one day. Instead of modelling the infection propagation as a Hawkes process, an infectious agent (symptomatic or asymptomatic) can infect its susceptible neighbor with probability p_i each day. Thereafter, the disease progresses the same way as in the DCS, except that in the DDE model the transition times are deterministic (the infection events and the severity of the disease (i.e., the (a)symptomatic and hospitalized states) are still determined randomly), and we have a single parameter p_h for the hospitalization probability (agents in this model do not have an age parameter). We discuss how we set the parameters of the DDE model in Section 8.3.4.

8.3.2 Simulating Mobility

Tubingen Mobility Model

We briefly review the mobility model introduced in [162], and illustrated in Figure 8.2 (b). The population is partitioned into households of possibly varying size (usually between 1 and 5). The households are assigned a location, and we also place some external sites (shops, offices, schools, transport stations, recreating sites) on the map, which the agents may visit. The location of the households and the number of agents in them is sampled randomly based on demographic datasets. Initially, each agent is assigned a few favorite sites (randomly based on distance), and will only visit these throughout the simulation. Each agent decides to leave home after some exponentially distributed time, visits one of its (randomly chosen) favorite sites, and comes back home after another (usually much shorter) exponentially distributed time. If two agents visit the same site at the same time, or within some time δ, we record them as a contact, which gives an opportunity for the infection to propagate. We denote the Tubingen mobility model as TU, and the DCS epidemic model that runs on the TU mobility model as DCS+TU.

Household Network Model

The Household network model (HNM) was inspired by [162], however we note that similar models have been studied in the theoretical community by [22]. As in the Tubingen mobility model, in HNM N nodes are assigned into households, but of constant size $d_h + 1$. Every pair of nodes in the same household are connected by an edge, forming therefore cliques of size $d_h + 1$. Additionally, each node is assigned d_c half edges, which are paired uniformly at random with other half-edges in the beginning. Some half-edge pairings can result in self-loops or multi-edges, which are discarded. This construction defines a random graph generated by a configuration model, which shares a lot of similarities with Random Regular Graphs (RRG) [242]. In fact, if we join nodes in the same household into a single node in the HNM (which we refer to as the *network of households* of the HNM), then the resulting graph is equivalent to the *pairing model* of RRGs with degree $d_c(d_h + 1)$. It is well-known that in the pairing model of RGGs of degree d, the local neighborhood (of constant radius, as the number of nodes tends to infinity) of a uniformly randomly chosen vertex is a d-regular tree (with probability tending to 1), which implies that locally there are asymptotically almost surely no self-loops, multi-edges or any cycles in the graph. This result has various names; in random graph theory the result is usually proved by subgraph counting [242], in probability theory it is the basis of branching process approximations [22], and in graph limit theory it is called the local convergence to the infinite d-regular tree [28]. In our theoretical analysis, this result motivates the approximation of the neighborhood of the source in the network of households of the HNM by an infinite $d_c(d_h + 1)$-regular tree. The HNM itself is then approximated by replacing each (household) node of the infinite $d_c(d_h + 1)$-regular tree of households by a $(d_h + 1)$-clique, and by setting the edges so that each (individual) node has degree exactly $d_c + d_h$, while keeping the connection between cliques unchanged (see Figure 8.2 (c) for a visualization).

Since the HNM is a time-independent graph, we adopt the standard notations from graph theory. Formally, the HNM is given by the set of nodes and edges $G = (V, E)$. Let us denote by $H(v)$ the set of nodes that are in the same household as node v. The distance between two nodes $u, v \in V$ (denoted by $d(u, v)$) is defined as a number of edges of the shortest path between u and v. We denote the DDE epidemic model that runs on the HNM network as DDE+HNM.

8.3.3 The Source Identification via Contact Tracing Framework

We present the Source Identification via Contact Tracing Framework (SICTF), which can be applied to both epidemic and mobility models presented so far. The framework determines how the government/health agency, which conducts the source identification task, learns about the outbreak, and how it can gather further information to locate the source. In the SICTF, as in Section 8.2.1, the agency learns about the outbreak when the first hospitalization occurs, and it also learns the identity of nodes when they become hospitalized (including the identity of the first hospitalized node).

After the outbreak is detected, the agency can make three types of queries. The first type of query, the household query with parameter v, reveals the agents that live in the same household as v. The household query works the same way in both the TU and the HNM models, and we do not limit the number of times it can be called (these queries are considered as cheap in the SICTF). The second type of query, the contact query, works differently in the TU and the HNM models. For the TU model, a contact query has two parameters: an agent v and a time window $[t_1, t_2]$. As a result, all agents that have been in contact with v (and therefore could have infected v or could have been infected by v) at an external site between t_1 and t_2 are revealed. In the HNM, no time window is needed for the contact query (which we also call edge query), and all neighbors of v in graph G are revealed. Contact (and edge) queries are considered expensive in the SICTF. While in this chapter we do not limit the number of available queries, we track the number of contacts and edges that are revealed as the algorithm runs. Note that in the TU model if two agents v_1 and v_2 have been in contact during the time window $[t_1, t_2]$ and also during a different time window $[t_3, t_4]$, then those are counted as separate contacts, whereas in the HNM an edge between v_1 and v_2 is only counted once. Although contact queries are considered expensive, both household and contact queries are answered instantly in the SICTF.

The third kind of query is the test query with parameter v, which reveals information about the course of the disease in the queried agent (see Figure 8.2 (a)). Symptomatic patients reveal the time of their symptom onset (which exactly determines their time of infection in the DDE due to the deterministic transition times) if they are past the pre-symptomatic state (i.e., if they are either infectious or recovered). Asymptomatic and pre-symptomatic patients do not reveal any information about their infection time; they just reveal that they have the disease or had the disease at some point and have recovered. For all algorithms we assume that asymptomatic patients do not reveal whether they have the infection at the time they are queried. Finally, agents who have not been exposed, or are still in their exposed state, give a negative test result. Test queries are again considered expensive in the SICTF, we even limit the population that can be tested on any given day to at most 1% of the total population, due to the capacity of testing facilities. However, since in this chapter we do not limit the number of days that the algorithm can use to locate the source, the limit on the number of tests does not play an important role. As opposed to household and contact queries (and the model in Section 8.2.1), test results are only answered the next day in the SICTF, which means that the

Interpretation	Parameter	DCS+TU	DDE+HNM
Exposed time	T_E	Lognormal distribution with $\mu = 3.22$, $\sigma = 2.3$	3
Pre-symptomatic time	T_P	Lognormal distribution with $\mu = 2.3$ and $\sigma = 1$	2
Infectious time	T_I	Lognormal distribution with $\mu = 14.0$ and $\sigma = 1$	14
Hospitalization time	T_H	Lognormal distribution with $\mu = 7.0$ and $\sigma = 1$	7
Probability of asymptomatic	p_a	0.4	0.4
Probability of hospitalization	p_h	Age dependent (mean is 0.0817)	0.083
Probability of infection	p_i	Hawkes process with various parameters on average ≈ 0.02 for a contact	0.1
External contacts	d_c	From mobility simulation, on average 15 each day	3
Number of external infections caused by a single agent each day	$d_c p_i$	On average around 0.3	0.3
Household contacts	d_h	From data, on average 1.51	2
Number of nodes (agents)	N	9054	400 or 1000

Table 8.1: Default values for the infection parameters in the DCS+TU and the DDE+HNM models

algorithms must operate in “real-time”, while the epidemic keeps propagating.

8.3.4 Parameters

The DCS+TU model has many parameters, most of which are fitted to COVID-19 datasets of Tubingen from 12/03/2020 to 03/05/2020 by [162] (we show the most relevant parameters in Table 8.1). We determined the parameters of the DDE+HNM model so that they fit the parameters of the DCS+TU as closely as possible (see the precise values in Table 8.1). We determine the values of T_E, T_P, T_I in the DDE+HNM by rounding the expected value of the corresponding distribution in the DCS+TU to the nearest integer. Since p_a is simply a constant in both models, we keep the same numerical value in the DDE+HNM. The parameter p_h is more complicated, because in the DCS+TU model there is a different hospitalization probability for each age group. We take the average hospitalization probability across the population to be p_h. The most complicated parameter to fit is p_i, because in the DCS+TU model, infections are modelled by a Hawkes process, which depends on many parameters, including whether the infectious agent is symptomatic or asymptomatic, the length of the visit, the site where the infection happens, etc (see equation (8.4)). We empirically observe the probability of infection in every contact in several simulations, and we find that an agent has on average 15 contacts outside the household each day, and that the average probability of infection during such a contact is around 0.02. However, since we use smaller networks for the DDE+HNM ($N = 400$ or 1000, because running the baselines on larger networks is not feasible) than the DCS ($N = 9054$), setting d_c to be as high as 15 would violate the assumption that the network of households of the HNM can be locally approximated by a tree (see Section 8.3.2). Therefore we chose $d_c = 3$ for the HNM and we scale p_i so that $d_c p_i$ (the expected number of external infections caused by a single agent each day) is the same in the DCS+TU and the DDE+HNM models. Finally, we choose d_h in the DDE+HNM by rounding the average household connections in the DCS+TU. Note that the average number of household connections is not the same as the average number of household members, because the number of connections grows quadratically in the size of the households, and thus fitting to the number of connections results in a higher d_c (due to the Quadratic Mean-Arithmetic Mean inequality).

Finding the default values for the parameters is useful to create a realistic model. However, we are also interested in the effect of each of the parameters on the performance of our algorithms. Therefore, in the DDE+HNM, we vary the parameters p_a, p_h, p_i, d_h and d_c, while keeping the other ones unchanged. For the DCS+TU model, we also keep the mobility model fixed and we focus on varying the parameters p_a, p_h and p_i. As noted above, there is no single parameter p_h or p_i in the DCS+TU model, therefore we change all hospitalization probabilities and all intensities of the Hawkes processes so that the hospitalization probability averaged across the population and the infection probability averaged across contacts equal the desired values.

8.3.5 The LocalSearch Algorithms LS and LS+

The LS algorithm finds patient zero by local greedy search. It keeps track of a candidate node, which is always the node with the earliest reported symptom onset time. We denote the candidate of the algorithm at iteration $i > 0$ by $s_{c,i}$. We think of s_c as a list, which is updated in each iteration of the algorithm, and we use the notation $s_{c,-1}$ for the last element of the list (i.e., the current candidate). In each iteration of the algorithm, we compute a new candidate denoted by s'_c, and we append it at the end of the list s_c at the beginning of the next iteration, unless $s'_c = s_{c,-1}$, in which case the algorithm terminates.

```
Algorithm 1: The LS and the LS+ algorithms
$s_c \leftarrow []$;
$s'_c \leftarrow$ first hospitalized node ;
while $s_{c,-1} \neq s'_c$ do
    $s_c$.append($s'_c$);
    (a): Add household members and backwards contacts of $s_{c,-1}$ to the test queue;
    while the test queue is non-empty do
        (b): Test nodes of the test queue, which were untested for the current $s_{c,-1}$;
        for $v$ in test results do
            if $v$ is symptomatic and $t_v < t_{s'_c}$ then
                (c): $s'_c \leftarrow v$;
            if (LS+) and ($v$ is asymptomatic) then
                if $v \in H(s_{c,-1})$ then
                    (d): Add backwards contacts of $v$ to test queue;
                else
                    (e) or (f) Add household members of $v$ to test queue;
return $s_{c,-1}$;
```

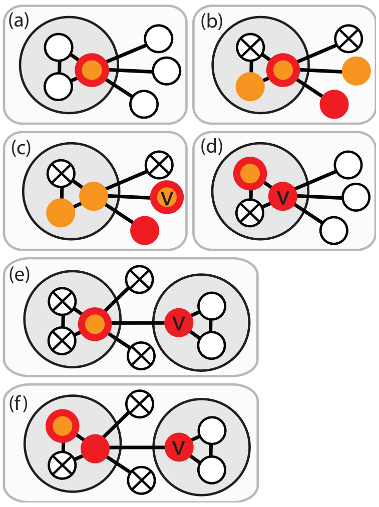


Figure 8.3: Pseudocode and graphical explanation for the LS and LS+ algorithms. We use the same coloring as in Figure 8.2 (a). Black edges show the queried edges, a node with black X marks a negative test result, and red stroked node marks the node currently maintained as source candidate by the LS algorithm. We denote by t_v the symptom onset time of symptomatic node v and by $H(v)$ the household of a node v similarly to the main text.

Since we consider the SICTF, the outbreak is detected when the first hospitalized case is reported. At that time, s'_c is initialized to be the hospitalized patient, the test queue is initialized to be empty, and the algorithm is started. In the beginning of an iteration, if the test queue is empty, the household members and the "backward" contacts of the current candidate $s_{c,-1}$ are queried and are added to the test queue (see Figure 8.3 (a)). We define "backward" contacts as the set of nodes that have been in contact with $s_{c,-1}$ in the interval $[t_{s_{c,-1}} - (T_E + T_P) - (\sigma_E + \sigma_P), t_{s_{c,-1}} - (T_E + T_P) + (\sigma_E + \sigma_P)]$, where $t_{s_{c,-1}}$ is the symptom onset time of current candidate $s_{c,-1}$. The terms σ_E and σ_P model the standard deviation of the transition times, and they are set to zero for the DDE and to $\sigma_E = 2$ and $\sigma_P = 1$ for the DCS based on Table 8.1. We note that the notion of "backward" contacts is only meaningful in the case of time-dependent network models; for the HNM, all neighbors are counted as backward contacts.

After the test queue is initialized, the agents inside the queue are tested (see Figure 8.3 (b)). Not all nodes can be tested on the same day because of the limitation on the number of tests available per day in the SICTF, however, this has little effect because we do not proceed to the next iteration until the test queue becomes empty. Once the test results come back to the agency, if any of the (symptomatic) nodes v reports an earlier symptom onset time than the current candidate $s_{c,-1}$, then we update our next candidate s'_c to be v (see Figure 8.3 (c)). We note that the iteration does not stop immediately after s'_c is first updated; the iteration runs until the test queue becomes empty, and until then, s'_c can be updated multiple times. This is important in the theoretical results to prevent the algorithm from getting sidetracked (see Figure E.1). We also experimented with a version of the LS and LS+ algorithms where the iteration stops immediately once s'_c is updated; we call these algorithms LSv2 and LS+v2.

The main drawback of the LS algorithm is that is gets stuck very easily if there is even one asymptomatic node on the transmission path. For this reason, we introduce the LS+ algorithm, in which we enter the backward contacts of the asymptomatic household members of $s_{c,-1}$, and the household members of any asymptomatic node into the testing queue (see Figure 8.3 (d)-(f)). Since the symptom onset times of asymptomatic nodes v are not revealed, we define backward contact in this case as any contact in the time window $[t_{s_{c,-1}} - (T_P + 2T_E + T_I), t_{s_{c,-1}} - (T_P + 2T_E)]$, where $t_{s_{c,-1}}$ is still the symptom onset time of the current candidate $s_{c,-1}$. Indeed, in the DDE model, since $s_{c,-1}$ was infected at $t_{s_{c,-1}} - (T_P + T_E)$, if v infected $s_{c,-1}$, agent v must have been infectious at that time, which implies that v could not have been infected later than $t_{s_{c,-1}} - (T_P + 2T_E)$ or earlier than $t_{s_{c,-1}} - (T_P + 2T_E + T_I)$. In the DCS model, the terms σ_E and σ_P can be subtracted and added to the two ends of the queried time window to account for the randomness in the transition times.

Both algorithms stop if the testing queue becomes empty before a node with an earlier symptom onset time than $s_{c,-1}$ is discovered, and both algorithm return $s_{c,-1}$ as their inferred source. The high level pseudocode and an illustration of the LS and LS+ algorithms are given in Figure 8.3.

8.4 Theoretical Results

In this section we present theoretical results for the LS and LS+ algorithms described in Section 8.3.5. We follow a similar approach as in the non-rigorous computation in Section 8.2.2, which is useful but not necessary for understanding this section. All the statements are rigorously established, and whenever we reach a point where the computations would become intractable, we propose a simpler approximate model to study. One of the main contributions of this chapter is to identify which computations can be done on more general models, and which computations need more simplified ones (see Figure 8.4 for an overview of the different models used for the computations in this section).

We compute the success probability of the LS and LS+ algorithms in two steps. We first assume the length of the transmission path known in Section 8.4.1 . This computation is then made possible by a tree approximation of the HNM, called the Red-Blue (RB) tree (defined in Section 8.4.1), and a slightly modified version of the DDE model called DDE_{NR} (defined in Section 8.4.1). The RB tree preserves some of the household structure in the HNM, and therefore allows us gain insight into the difference between the LS and LS+ algorithms, which would be difficult to obtain if we had worked on trees without taking the household structure in account.

For the second step, we would need to compute the distribution of the transmission path on the RB tree. However, finding a closed form expression is intractable. Instead, we combine the network and epidemic models into a growing random tree model, and we consider a d-ary Random Exponential Tree (RET). The d-ary RET model has only been studied for $d = 2$ [93]; we extend the results on their expected profile for general d in Section 8.4.2. Nevertheless, working on d-ary RETs still remains difficult, and therefore, in our last modeling step, we introduce a Deterministic Exponential Tree (DET) model, whose profile is close to the expected profile of the RET, and we compute the distribution of the transmission path on this model in Section 8.4.2.

To summarize all models considered in this chapter, we have a data-driven and a synthetic model for simulations (DCS+TU and HNM+DDE), an analytically tractable model (RB-tree+ tDDE_{NR}) where we can compute the success probability if the length of the transmission path is known. In a second stage, we compute the distribution of a transmission path on a deterministic tree (DET), which has a similar profile as a random tree (RET) that approximates our analytically tractable model. We visualize these five different models in Figure 8.4 (a), and we show by simulations in Figure 8.4 (b) that the distribution of the transmission path is similar in all of the considered models with appropriately scaled parameters. We compare our analytic results on the success probabilities of the LS and LS+ algorithms with our simulation results in Section 8.5.3 in Figure 8.6.

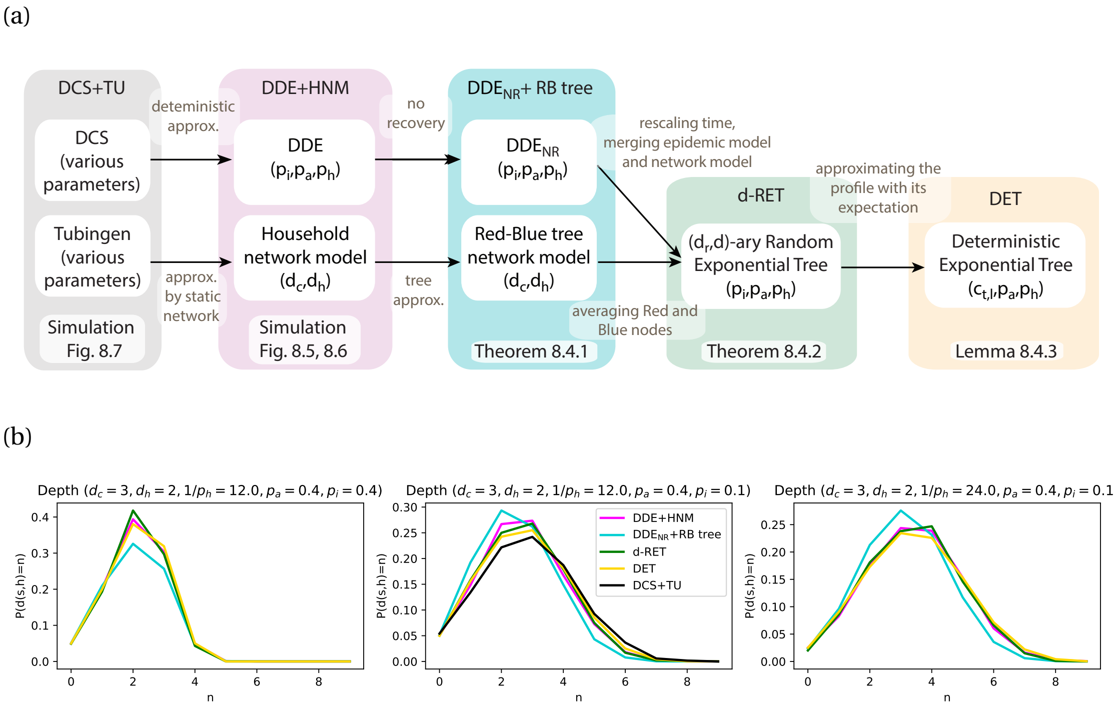

(a)
DCS+TU
DCS (various parameters)
Tubingen (various parameters)
Simulation Fig. 8.7
deteministic approx.
approx. by static network
DDE+HNM
DDE (p_i,p_a,p_h)
Household network model (d_c,d_h)
Simulation Fig. 8.5, 8.6
no recovery
tree approx.
DDE_{NR}+ RB tree
DDE_{NR} (p_i,p_a,p_h)
Red-Blue tree network model (d_c,d_h)
Theorem 8.4.1
rescaling time, merging epidemic model and network model
averaging Red and Blue nodes
d-RET
(d_r,d)-ary Random Exponential Tree (p_i,p_a,p_h)
Theorem 8.4.2
approximating the profile with its expectation
DET
Deterministic Exponential Tree ($c_{t,l}$,p_a,p_h)
Lemma 8.4.3
(b)
Depth ($d_c = 3, d_h = 2, 1/p_h = 12.0, p_a = 0.4, p_i = 0.4$)
Depth ($d_c = 3, d_h = 2, 1/p_h = 12.0, p_a = 0.4, p_i = 0.1$)
Depth ($d_c = 3, d_h = 2, 1/p_h = 24.0, p_a = 0.4, p_i = 0.1$
P(d(s,h)=n)
n
DDE+HNM
DDE_{NR}+RB tree
d-RET
DET
DCS+TU

Figure 8.4: The different approximation methods (a) and the distribution of the length of transmission path in the different models (b) proposed in Section 8.4. Panel (b) also shows the length of the transmission path in the DCS model on the TU dynamics, to highlight the fit of our model.

8.4.1 Success Probability of LS and LS+ Algorithms on the RB Tree

In this section we introduce the Red-Blue (RB) tree model (which is a tree approximation to the HNM), and we calculate the exact probability that the LS and LS+ algorithms succeed, if the length of the transmission path is known.

Red-Blue tree models

In short, a RB tree is a two-type branching process with a deterministic offspring distribution that depends on d_h and d_c. The lack of randomness in this distribution makes us adopt the formalism of deterministic rooted trees.

Definition 8.4.1. *Let a rooted tree, denoted by $G(s)$, be a tree graph with a distinguished node root node s. Let u and v be two nodes connected by an edge in $G(s)$. If $d(u,s) < d(v,s)$, we say that u is a parent of v, otherwise u is a child of v. Moreover, if $d(s,v) = l$ we say that v is on level l. An RB tree with parameters (d_c, d_h) is an infinite rooted tree, such that the nodes also have an additional color property. The root is always colored red and the rest of the nodes are colored red or blue. The root has d_c red and d_h blue children. Every other red node has $d_c - 1$ red and d_h blue children, and every blue node has d_c red children and no blue children. Red nodes and their d_h blue children partition the nodes of the RB tree $G(s)$ into subsets of size $d_h + 1$, which we call households.*

Remark 8.4.1. *In the RB tree, each blue node has degree $d_c + 1$, and each red node has degree $d_c + d_h$, including the root of the tree s (which is the source of the epidemic, when the RB tree is combined with an epidemic model).*

The RB tree can be seen as a local tree approximation of the HNM. Let $G = (V, E)$ be an HNM with parameters (d_c, d_h), and let $s \in V$ be the distinguished source node. In Section 8.3.2 we noted that the HNM can be approximated locally around the source node by replacing each node of an infinite $d_c(d_h + 1)$-regular tree by a $(d_h + 1)$-clique, and setting the edges so that each node has degree exactly $d_c + d_h$, while keeping the connection between cliques unchanged. Let us call this infinite graph G^*. Although G^* is not a tree, all cycles in G^* must be contained entirely inside the households, which implies that in each household there exists exactly one node that has the minimal distance to the source. We will refer to these nodes with minimal distance to the source as the red nodes, and we color the rest of the nodes blue. In other words, the red nodes will be the first ones in their households to be infected. Let us now delete the edges between the blue nodes in G^* to obtain graph G'. We claim that G' is isomorphic to the RB tree $G(s)$ rooted at the source s. Indeed, since the edges between blue nodes have been deleted in G^* to form G', each blue node has $d_c + 1$ red neighbors and no blue neighbor, and since the edges incident to red nodes have been unchanged, each red node has d_c red and d_h blue neighbors, exactly as in the definition of RB tree above.

Note that a household in G^* is completely characterized by only specifying the colors of the nodes: a household always consists of one red node and of its d_h blue children. We use this characterization as a definition for households in the RB tree G', because it does not depend on the edges from G that are deleted in G^*, whereas this deletion makes the original definition of a household as a clique in G unusable.

Next, we make some important observations the behavior of the LS and the LS+ algorithms on RB trees, which we prove in Appendix E.1.1. We start by formalizing the notion of transmission path.

Definition 8.4.2. *Let h be the first hospitalized node and s be the source. We call the path $(s = v_0, v_1, ... v_l = h)$, where v_i is the infector of v_{i+1} for $0 \leq i < l$, the* transmission path. *Also we call the path $(v_l, v_{l-1}, ... v_1)$ the* reverse transmission path.

Remark 8.4.2. *Note that in an RB tree, each household traversed by a transmission path shares one (the red node in the household) or two (the red node of the household and one of its d_h children in the household) nodes with this path. Moreover, the red node of a household traversed by a transmission path is followed by another red node on the path (in another household) if it is the only node of that household on the transmission path, whereas it is followed by a blue node (in the same household) if two nodes of that household are on the transmission path.*

Lemma 8.4.1. *In the RB tree network, the LS algorithm succeeds if and only if all nodes on the transmission path are symptomatic, and the LS+ algorithm succeeds if among the nodes of the transmission path, there exists a symptomatic node in each household, and the source is symptomatic.*

Remark 8.4.3. *We note that the statement for LS+ in Lemma 8.4.1 cannot be reversed, i.e., it is possible that LS+ succeeds even if among the nodes of the transmission path, there is a household with no symptomatic node (see Figure E.1 (a)). Also, the proof of Lemma 8.4.1 does not hold if the LS+ algorithm proceeds to the next iteration at the first time s_c' is updated (see Figure E.1 (b)). Finally, in the proof of Lemma 8.4.1, we do not make any assumptions about asymptomatic patients having had the disease previously or not, which implies that we could treat non-complying agents as asymptomatic patients without jeopardizing the correctness of the algorithms.*

The $\mathrm{DDE_{NR}}$ Model

Focusing on tree networks is an important step towards making our models tractable for theoretical analysis, but it will not be enough; we will make two minor simplifications to the DDE model as well: we eliminate (i) the pre-symptomatic state and (ii) the recovered state, and we call the new model $\mathrm{DDE_{NR}}$ (where NR stands for No Recovery). (i) The first assumption can be made without loss of generality, because the pre-symptomatic state does not have any effect on the disease propagation, nor on the success of the source identification algorithm. Indeed, according to Lemma 8.4.1, the success of the LS and LS+ algorithms depends only on the information gained about the transmission path, and by the time of the first hospitalization, every node on the transmission path must have left the pre-symptomatic state (since we always have $T_P < T_E + T_H$), even if we include it in the model. (ii) The second assumption on the absence of recovery states amounts to take $T_I \to \infty$, which does have a small effect on the disease propagation, however, this effect is minimal because $T_I = 14$ is already quite large, and because only the very early phase of the infection is interesting for computing the success probabilities of the algorithms. Finally, this last assumption has no effect on the information gained by the algorithm since we assumed that recovered patients (who were symptomatic) can remember and reveal their symptom onset time in the same way as symptomatic infectious patients.

Success Probability of LS

Assuming that the distribution of length of the transmission path is provided for us (we give an approximation in Section 8.4.2), the success probability of LS can be computed succinctly. We need a short definition before stating our result.

Definition 8.4.3. *Let p be the probability that a node is asymptomatic conditioned on the event that it is not hospitalized.*

A simple computation shows that

$$p = \mathbb{P}(v \text{ is asy} \mid v \text{ is not hosp}) = \frac{p_a}{p_a + (1-p_a)(1-p_h)}. \tag{8.5}$$

Lemma 8.4.2. *For the* $\mathrm{DDE}_{\mathrm{NR}}$ *epidemic model with parameters* (p_i, p_a, p_h) *on the RB tree with parameters* (d_c, d_h)*, and with p computed in equation* (8.5)*, we have*

$$\mathbb{P}(\textit{LS succeeds}) = \sum_{n=0}^{\infty} \left(1-p\right)^n \mathbb{P}(d(s,h) = n). \tag{8.6}$$

Proof. Let us reveal the randomness that generates the epidemic in a slightly modified way than in the definition (Sections 8.3.1 and 8.4.1). As before, at the beginning only the source is infectious, and depending on course of the disease, the source can be symptomatic and hospitalized, symptomatic but not hospitalized, or asymptomatic with probabilities $(1-p_a)p_h, (1-p_a)(1-p_h), p_a$, respectively. In each moment, each infectious node infects each of its susceptible neighbors with probability p_i. If a node is infected, we reveal the information whether it will become hospitalized (which happens with the probability $(1-p_a)p_h$), but if it does not become hospitalized, we do not reveal whether the node is asymptomatic or symptomatic yet. Indeed, this information is not necessary for continuing the simulation of the epidemic since we assumed that there is no difference between the infection probabilities of symptomatic and asymptomatic nodes. Thereafter, when the first hospitalized case occurs, we reveal for each infected node v on the transmission path (except the last node, which we know is hospitalised; see Definition 8.4.2) whether it is asymptomatic or not. The only information we have about these nodes is that they are not hospitalized, which implies that the probability that a node is revealed to be asymptomatic on the transmission path is exactly the probability p from Definition 8.4.3 computed in (8.5).

By Lemma 8.4.1, LS succeeds if and only if each node on the transmission path is symptomatic. Conditioning on the length of the transmission path, we can compute the probability of each node being symptomatic by equation (8.5) as

$$\mathbb{P}(LS \text{ suceeds} | d(s,h) = n) = \left(1 - \mathbb{P}(v \text{ is asy} | v \text{ is not hosp})\right)^n = \left(1-p\right)^n, \tag{8.7}$$

from which (8.6) follows immediately. □

Success Probability of LS+

Computing the success probability of the LS+ algorithm is far more challenging compared to the LS algorithm, even if the distribution of the length of the transmission path is provided to us. Indeed, since the LS+ algorithm does further testing on the contacts and household members of asymptomatic nodes, it is essential to have additional information about the number of households on the transmission path. We give our main result on the LS+ in the next theorem, which we prove in Appendix E.1.2.

Theorem 8.4.1. *Let p be as in* (8.5) *and let $\mathcal{S}(n,\alpha,\beta)$ be the set of k integer values such that k and n have different parity and $n+1-2(\alpha+\beta) \geq k \geq 2-(\alpha+\beta)$. Then, for the* $\mathrm{DDE_{NR}}$ *epidemic model with parameters (p_i, p_a, p_h) on the RB tree with parameters (d_c, d_h), we have*

$$\mathbb{P}(LS+ \text{ succeeds}) \geq \mathbb{P}(d(s,h)=0) + (1-p)\mathbb{P}(d(s,h)=1) +$$
$$\sum_{n=2}^{\infty} \sum_{\substack{\alpha,\beta\in\{0,1\} \\ k\in\mathcal{S}(n,\alpha,\beta)}} \binom{\frac{n+k-3}{2}}{k-2+\alpha+\beta} \frac{(d_h(1-p))^{\frac{n+k-1}{2}} (d_c(1+p))^{\frac{n-k+1}{2}-\alpha-\beta} d_c(d_c-1)^{k+\alpha+\beta-2}}{\lambda_1 \left(\frac{d_c-1+D}{2}\right)^n + \lambda_2 \left(\frac{d_c-1-D}{2}\right)^n} \mathbb{P}(d(s,h)=n), \tag{8.8}$$

where

$$D = \sqrt{(d_c-1)^2 + 4d_c d_h} \tag{8.9}$$
$$\lambda_1 = \frac{(d_c+1+D)(2d_h+d_c-1+D)}{2D(d_c-1+D)} \tag{8.10}$$
$$\lambda_2 = \frac{(D-d_c-1)(2d_h+d_c-1-D)}{2D(d_c-1-D)}. \tag{8.11}$$

8.4.2 Approximating the Depth of the Path to the First Hospitalized Node

Section 8.4.1 was dedicated to the success probability of the LS and LS+ algorithms, however, in these results, we are still missing the distribution of the transmission path length. In this subsection we address this problem by introducing simpler approximate models.

(d_r, d)-ary Random Exponential Tree

When we introduced the $\mathrm{DDE}_{\mathrm{NR}}$ model in Section 8.4.1, we removed both parameters T_P and T_I from the DDE model (by removing the presymptomatic and the recovered states, respectively), but we kept the parameter T_E. In this step we will rescale the time parameter to make $T'_E = 1$ by changing p'_i to be $1-(1-p_i)^{T_E}$. Since we had $T_E = 3$ by default, using T'_E and p'_i instead of T_E and p_i means that we choose 3 days to be our time unit, and the probability of infection is scaled to be the probability that the infection is passed in at least one of three days (since the RB tree is time-independent, if two nodes are connected, the infection can spread on it every day). We drop the prime from p'_i and T'_E for ease of notation. As a second approximation, instead of keeping track of two types of nodes (red and blue) as it is done in the RB tree, we propose to change our network model to an infinite d-regular tree, where d is set to be the average degree of an RB tree.

By making these two changes (tracking time at a coarser scale and simplifying the network topology to a d-regular tree), the growth of the epidemic becomes equivalent to a known model, the d-ary Random Exponential Tree (d-RET). Binary RETs have been introduced in [93]. We give the definition below for completeness.

Definition 8.4.4. *A d-ary Random Exponential Tree (d-RET) with parameters d, p_i at time day t, denoted by $G_t(s)$, is a random tree rooted at node s. At day 0, the tree $G_t(s)$ only has its root node s. Let $\bar{G}_t(s)$ be the closure of $G_t(s)$, which is obtained by attaching external nodes to $G_t(s)$ until every internal node (a node that was already present in $G_t(s)$) has degree exactly d in the graph $\bar{G}_t(s)$. Then, $G_{t+1}(s)$ is obtained from $\bar{G}_t(s)$ by retaining each external node with probability p_i, and dropping the remaining external nodes.*

Indeed, each node of a d-RET infects a new node with probability p_i each day, and after a sufficiently long time, the d-RET becomes close to a large d-ary tree. Of course, we do not want to let the d-RET grow for a very long time, we only want it to grow until the first hospitalization occurs. So far we have not talked about the course of the disease of the nodes in the d-RET model because we could define the spread of the infection without it. Since we still need to do one final simplification to compute the distribution of the transmission path, we defer the discussion about hospitalizations, and how the parameters p_a and p_h are part of the model, to Section 8.4.2. Note that by considering the d-RET, we deviate from the idea of separating the epidemic and the network models; we only have a randomly growing tree, which is stopped at some time, when the tree is still almost surely finite.

So far we only did simplifications to the model, which resulted in further and further deviations from the original version. Now we will make a small modification that brings our model back closer to the RB tree, without complicating the computations too much. We still make almost all maximum degrees of the RET uniform d, but we make an exception with the root, which will have maximum degree $d_r = d_c + d_h$. This makes the maximum degree of the root the same as the degree of the root of the RB tree. We call the resulting model a (d_r, d)-RET with parameter p_i. Since the close neighborhood of the source has a high impact on the success probability, we found that this solution gives the best results while keeping the computations tractable.

In our computations, only the profile the infection tree will be important, which motivates the next definition.

Definition 8.4.5. *In the (d_r, d)-RET model with parameter p_i, let $A_{t,l}$ be the number of nodes during day t at level l, and let $a_{t,l} = \mathbb{E}[A_{t,l}]$. Moreover, we define the random variable*

$$A_t = \sum_{t=0}^{+\infty} A_{t,l} \tag{8.12}$$

with $A_{-1,l} = 0$ for all l, and its expectation $a_t = \mathbb{E}[A_t]$.

As noted earlier, the d-RET model has only been analyzed for $d = 2$ to this date. We provide the expected number $a_{t,l}$ of nodes at level l in day t for the general case in the next theorem and corollary, which we prove in Appendices E.1.3 and E.1.4.

Theorem 8.4.2. *In the (d_r, d)-RET with parameter p_i, let $a_{t,l}$ be as in Definition 8.4.5. Then*

$$a_{t,0} = 1 \tag{8.13}$$

$$a_{t,l} = d_r p_i \sum_{m=l-1}^{t-1} \binom{m}{l-1} (1-p_i)^{m-l+1} d^{l-1} p_i^{l-1}, \text{for } t \geq l \geq 1 \tag{8.14}$$

$$a_{t,l} = 0, \text{for } l > t. \tag{8.15}$$

Corollary 8.4.1. *In the RET(p_i, d_r, d), let a_t be the expectation of* (8.12)*, as in Definition 8.4.5. For $t \geq 0$,*

$$a_t = 1 + d_r \frac{(1 - p_i + dp_i)^t - 1}{d-1}. \tag{8.16}$$

Deterministic Exponential Tree with Parameters p_a, p_h and $(c_{t,l})_{t,l\in\mathbb{N}}$

In the (d_r, d)-RET model it is still complicated to calculate the distribution of the depth of the first hospitalized node. For this reason, we approximate the RET model by a deterministic time-dependent tree with a prescribed profile.

Definition 8.4.6. *Let $(c_{t,l})_{t\in\mathbb{N}\bigcup\{-1\},l\in\mathbb{N}}$ be a two-dimensional array with $c_{t,l} = 0$ for $t \in \{-1, 0\}$ and $l \in \mathbb{N}$, except for $c_{0,0} = 1$, and with $c_{t,l} \geq c_{t,l-1}$ for any t and any $l \geq 1$. Additionally, if we define $c_t = \sum_l c_{t,l}$, then the array $(c_{t,l})$ must satisfy $c_t > c_{t-1}$ for $t \geq 0$. Then, we define the Deterministic Exponential Tree (DET) with parameter $(c_{t,l})_{t\in\mathbb{N}\bigcup\{-1\},l\in\mathbb{N}}$, as a time-dependent rooted tree, that has exactly $c_{t,l}$ nodes on level l at time t. The edges between the adjacent levels are drawn arbitrarily so that the tree structure is preserved.*

The formal assumptions on the array $(c_{t,l})$ are simply made to ensure that the DET starts with a single node at $t = 0$, that it never shrinks on any level $(c_{t,l} \geq c_{t,l-1})$, and that it grows by at least one node in each time step $(c_t > c_{t-1})$.

We have defined the DET at any given time t, however, to determine the length of the transmission path, we are not interested in the DET at any given time, but only when the first hospitalization occurs. To compute the distribution of the first hospitalized node, we would like to have an absolute order on the times when the nodes are added, which we do by randomization. We say that on day t, nodes are added one by one to the DET, their order given by a uniformly random permutation, and each node is hospitalized with probability $(1-p_a)p_h$ (as in the original DDE model). When the first hospitalization occurs, we stop growing the tree, and we call the resulting (now random) model a stopped DET with parameters $(c_{t,l}), p_a, p_h$. We find the transmission path length distribution on the stopped DET in the next lemma, which we prove in Appendix E.1.5.

Lemma 8.4.3. *Let us consider the stopped DET model with parameters $(c_{t,l}), p_a, p_h$, and let h denote the first hospitalized node. Then*

$$\mathbb{P}(d(s,h) = l) = \sum_{t=0}^{+\infty} \frac{c_{t,l} - c_{t-1,l}}{c_t - c_{t-1}} (1-(1-p_a)p_h)^{c_{t-1}} \left(1 - (1-(1-p_a)p_h)^{c_t - c_{t-1}}\right). \tag{8.17}$$

We would like to set $c_{t,l}$ so that the DET is close to the RET described in Section 8.4.2. For equation (8.17) to make sense, we should substitute integer values for $c_{t,l}$, however, for an approximation the equation can also be evaluated for fractional values as well.

Remark 8.4.4. *If we substitute $c_{t,l} = a_{t,l}$ and $c_t = a_t$ in equation* (8.17)*, where $a_{t,l}$ is given in Theorem 8.4.2 and a_t is computed in Corollary 8.4.1, then we get the expression*

$$d_r p_i^{l-1} d^{l-1} \sum_{t=l}^{+\infty} \frac{\binom{t-1}{l-1}(1-p_i)^{t-l}}{(1-p_i+dp_i)^{t-1}} (1-(1-p_a)p_h)^{1+d_r \frac{(1-p_i+dp_i)^{t-1}-1}{d-1}} \left(1-(1-(1-p_a)p_h)^{d_r(1-p_i+dp_i)^{t-1}}\right), \tag{8.18}$$

which approximates the distribution of the transmission path length in the (d_r, d)-ary RET stopped at the first hospitalization.

8.5 Simulation Results

8.5.1 Baseline Algorithms

Non-adaptive Baseline: Dynamic Message Passing

There are few source identification algorithms that are compatible with time-varying networks in the literature [120, 130, 87, 55]. The most promising one among these algorithms [130] has a close resemblance to the a previous work of [161] on Dynamic Message Passing (DMP) algorithms. Given the initial conditions on the identity of the source node and its time of infection, the DMP algorithm approximates the marginal distribution of the outcome of an epidemic at some later time t. The algorithm is exact on tree networks, and it computes a good approximation when there are not too many short cycles in the network. Therefore, the DMP algorithm can be used to approximate the likelihood of the observed symptom onset times for any (source,time) pair. Due to its flexibility, we were able to adapt the DMP algorithm to the SICTF (see Appendix E.2 for more details).

Originally, the DMP was applied to the source identification problem by computing the likelihood values for all possible (source,time) pairs, and then choosing the source node from the most likely pair as the estimate [161]. However, testing all (source,time) pairs increases the time complexity of the algorithms potentially by a factor of N^2, which makes the algorithm intractable in many applications. Jiang et. al. [130] proposed a very similar algorithm to the DMP equations (which is unfortunately not exact even on trees), and solved the issue of intractability by a heuristic preprocessing step to the DMP algorithm. This preprocessing step, identifies a few candidate (source,time) pairs, by spreading the disease backward from the observations in a deterministic way (called reverse dissemination). Since we already approximate our data-driven model (DCS) by an epidemic model with deterministic transition times (DDE), it is natural for us to also implement the deterministic preprocessing step proposed by [130]. We produce 5 (source,time) pairs which are feasible for the 5 earliest symptom onset time observations (see Appendix E.2.3 for more details). It would have been ideal to run the algorithms for more than 5 pairs, but this was made impossible by the runtimes becoming very high. We run therefore our implementation of the DMP algorithm with the previously computed feasible (source,time) pairs as initial conditions to find the most likely source candidate.

The source estimation algorithms developed using the DMP algorithm do not specify how the sensors should be selected, and therefore place these non-adaptive sensors randomly. We refer to the resulting algorithm as random+DMP. The number of sensors is set so that it always exceeds the number that LS/LS+ would use. The simulation results are shown in Figure 8.5 for the DDE+HNM model. Importantly, the deterministic preprocessing step of [130] is compatible with time-varying networks, which allows us to run the algorithm for the DCS+TU model as well (see Figure 8.7).

Adaptive Baseline: Size-Gain

The Size-Gain (SG) algorithm was developed for epidemics which spread deterministically [249], and has been later extended to stochastic epidemics [224]. It works by narrowing a candidate set based on a deterministic constraint. If v_1, v_2 are symptomatic observations, then s_c is in the candidate set of SG if and only if

$$|(t_{v_2} - t_{v_1}) - (d(v_2, s_c) - d(v_1, s_c))| < \sigma(d(v_2, s_c) + d(v_1, s_c)), \tag{8.19}$$

where σ is the standard deviation of the infection time of a susceptible contact. If one of the observations, say v_2, is negative, then SG uses a condition almost identical to equation (8.19), except that the absolute value is dropped, since a negative observation at time t_{v_2} is only a lower bound on the true symptom onset time of v_2. These deterministic conditions are checked for every symptomatic-symptomatic or symptomatic-negative pair (v_1, v_2) to determine if s_c can be part of the candidate set. Next, SG places the next sensor adaptively at the node which reduces the candidate set by the largest amount in expectation (assuming a uniform prior on the source and its infection time), and it terminates when the candidate set shrinks to a single node. Note that the SG algorithm can fail if at least one of the deterministic conditions in equation (8.19) is violated for some (v_1, v_2) because of the randomness of the epidemic.

We use the existing implementation of the SG algorithm by [224], and adapt it to the SICTF. We incorporate asymptomatic-symptomatic and asymptomatic-negative observations (v_1, v_2) the same way as symptomatic-negative are incorporated; we drop the absolute value sign in equation (8.19), because an asymptomatic observation at time t_{v_1} is only an upper bound on the true symptom onset time of v_1. We impose the same daily limit to the number of sensors that can be placed by the SG algorithm in a single day as for the LS/LS+ algorithm, and if the candidate set size does not shrink to one on the day when both LS and LS+ have already provided their estimates, then the SG algorithm must make a uniformly random choice from the current candidate set as its source estimate. The simulation results are shown in Figure 8.5 for the DDE+HNM model. We do not implement the SG algorithm for the DCS+TU model, because its runtime is too high, and because it is not clear how it should be implemented for time-varying networks.

8.5.2 Comparison with Baselines

We show our simulation results comparing the random+DMP, SG, LS and LS+ algorithms in Figure 8.5. In the first row of Figure 8.5, we show the accuracy of the algorithms with solid curves. Since the LS/LS+ algorithms cannot identify an asymptomatic source, we also show what the accuracy would look like if the goal of the SICTF was to identify the first symptomatic agent with dashed lines. It is clear that in both metrics and across a wide range of parameters, the LS+ algorithm performs best, followed by LS, next random+DMP, and finally SG. The only exception is for high values of p_i, where SG performs best. The good performance of SG for these parameters is expected, because SG was originally developed for deterministically spreading epidemics (i.e., $p_i = 1$).

In the second row of Figure 8.5, we show the number of test/sensor queries used by the algorithms. LS uses the fewest tests, followed by LS+. The random+DMP and SG algorithms always use more tests than LS/LS+, except for large values of p_i. Finally, in the last row of Figure 8.5 we show the number of contact (or in this case edge) queries used by the algorithms. Again, LS uses fewer queries than LS+, while both the random+DMP and SG algorithms query essentially the entire network.

Figure 8.5 shows that the LS/LS+ algorithms are fairly robust to changes in the parameters of the model, except for the parameter p_a. Indeed, if there are many asymptomatic nodes in the network, then source identifyion becomes very challenging. It may be surprising that as p_a grows, the number of tests that LS uses decreases, contrary to LS+. This is because as p_a grows, the LS algorithm gets stuck more rapidly, while the LS+ algorithm compensates for the presence of asymptomatic nodes by using more test/sensor queries.

8.5.3 Comparison of Simulations and Theoretical Results

The analytic results from Section 8.4 are in good agreement with the simulation results in Figure 8.6. We also experiment with changing the parameters d_h, d_c while keeping all the parameters fixed, and with changing d_c while keeping the product $d_c p_i$ fixed. We observe that LS is not affected by the parameter d_h, whereas LS+ performs better with a higher d_c, which is expected because LS+ leverages the household structure of the network to improve over LS. Somewhat surprisingly, we also observe that a higher d_c also improves the performance of both algorithms. This can be explained by the fact that a larger d_c implies that there are more nodes in the close neighborhood of the source, which results in shorter transmission paths, making source identification less challenging. Finally, if we increase d_c but keep $d_c p_i$ fixed, the performance of the algorithms does not change as much, which confirms the intuition that it is the number of infections caused by an infectious node in a single day that matters the most (as we discussed in Section 8.2).

8.5.4 Simulations on the DCS Model

We show our simulation results on our most realistic DCS+TU model in Figure 8.7. We make very similar observations on this model as the ones that we have made on the DDE+HNM model in Sections 8.5.2 and 8.5.3, which shows that the LS/LS+ algorithms and our analysis of their performance is robust to changes in the epidemic and network models.

In the DCS+TU model, we used a fixed limit on the number of sensors that the random+DMP model selects, instead of setting the limit based on the LS+ algorithm. As a result, for a few parameters the LS+ algorithm used more tests than the random+DMP model. However, we note that by updating the candidate node immediately after an earlier symptom onset time is revealed (see Section 8.3.5), we can essentially cut the number of required tests for the LS+ algorithm by half (LSv2 and LS+v2), without sacrificing the performance of the algorithms.

8.6 Discussion

We have introduced the LS and LS+ algorithms in the SICTF, and we have used a sequence of models on which we can compute their accuracy (probability of finding the correct source) rigorously. We find that both LS and LS+ outperform baseline algorithms, even though the baselines essentially query all contacts on a transmission path between agents, while LS and LS+ query only a small neighborhood of the source. One could argue that LS and LS+ beat the baseline algorithms only because we benchmark them in our own framework, which is different from the framework for which the baseline algorithms were developed. However, we argue that the LS/LS+ algorithms are robust to changes in the framework due to their simplicity, and we support our argument by simulation results. The runtimes of the LS/LS+ algorithms are also much lower than the baselines and do not depend on the network size since they are local algorithms - as opposed to the baselines, which have quadratic or even larger dependence on the network size. The "low-tech" approach in the design of the LS/LS+ algorithms increases their potential to be implemented in real-world scenarios, possibly even in a decentralized way, similarly to contact tracing smart phone applications [234], which is an interesting direction for future work.

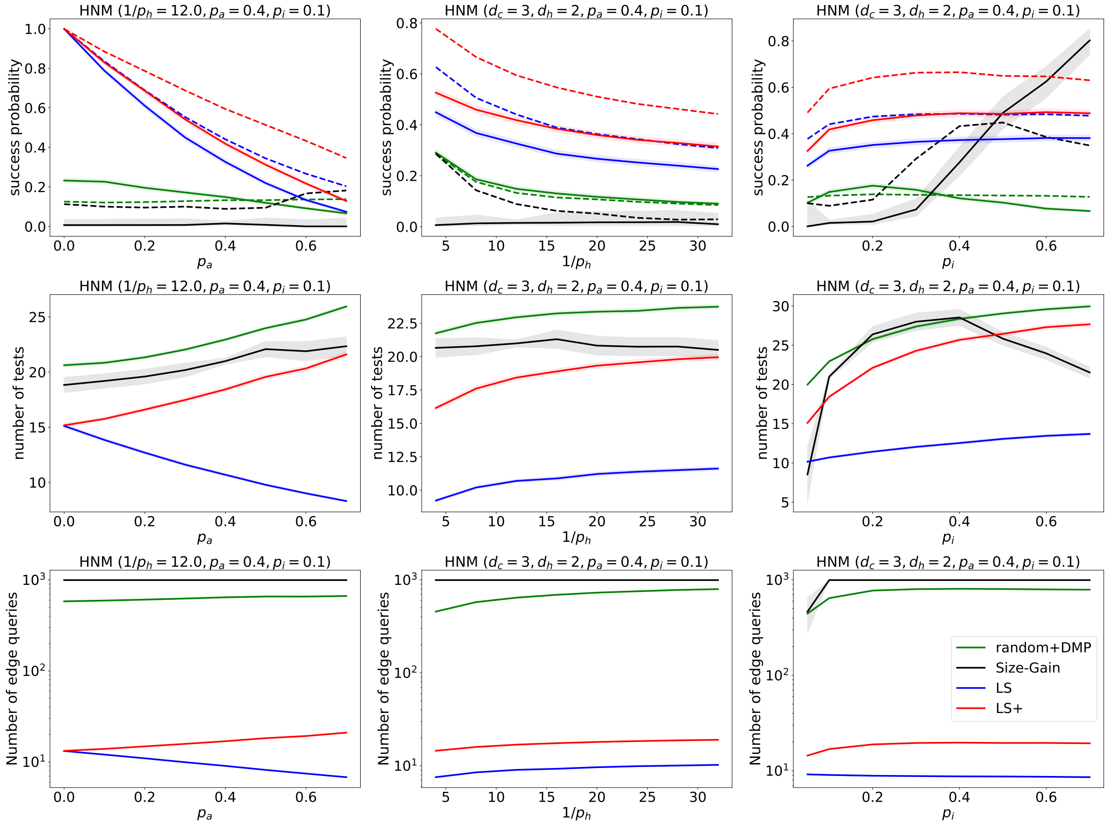


Figure 8.5: The performance of the algorithms LS, LS+, R and SG if the metric is the probability of finding the source (solid curves) or the first symptomatic patient (dashed curves). The simulations were computed on a population of $n = 400$ individuals in the DDE model on the HNM, and each datapoint is the average of 4800 independent realizations except for the SG algorithm, which was run with 192 independent realizations. The confidence intervals for the success probabilities are computed using the Wilson score interval method, and for the tests and the queries using the Student's t-distribution.

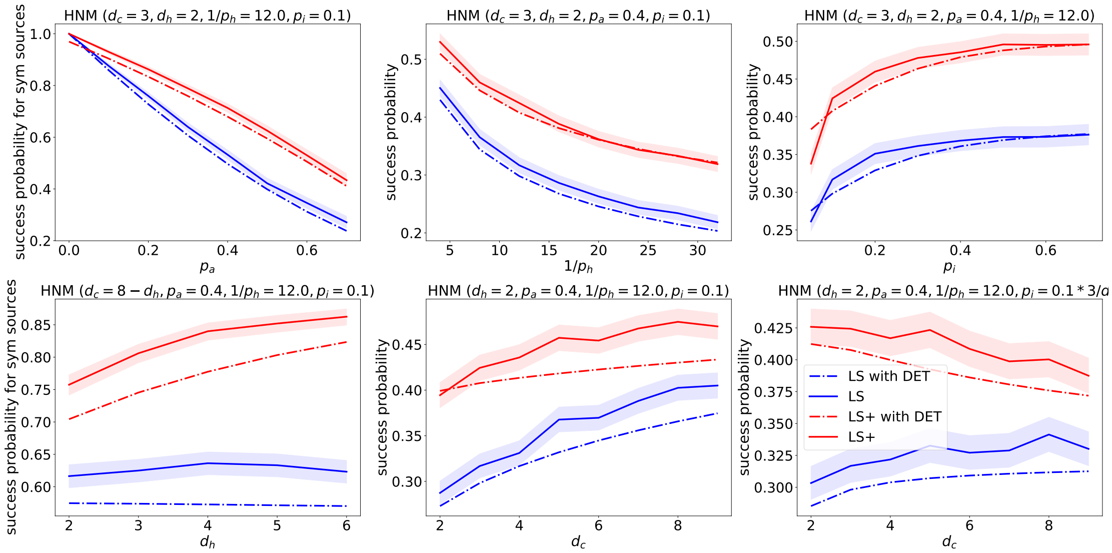


Figure 8.6: The probability of success of the LS and LS+ algorithms (solid curves) and their theoretical estimate (dash-dotted curves) with the success probabilities computed in Lemma 8.4.2 and Theorem 8.4.1, while the transmission path distribution computed in equation (8.18). The simulation results were generated using the DDE model on HNM networks of size $n = 1000$ with 4800 independent samples. The 95% confidence intervals are computed using the Wilson score interval method.

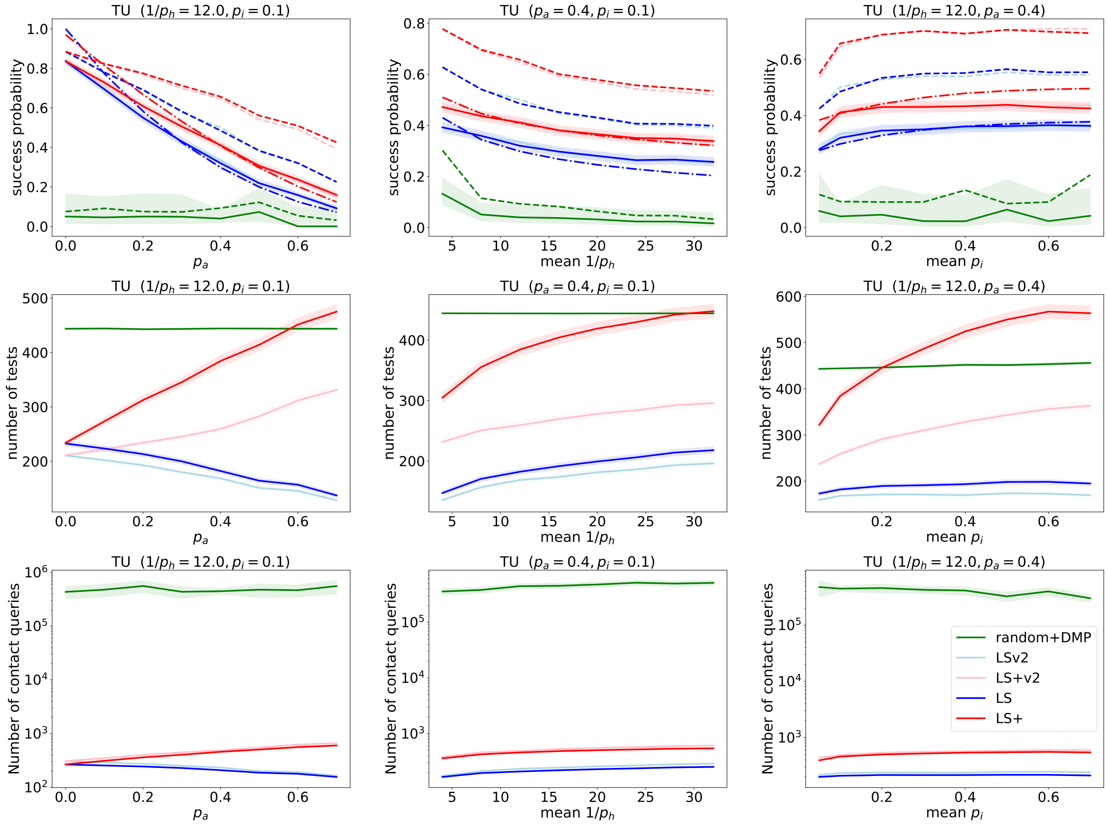


Figure 8.7: The performance of the algorithms LS, LS+ and random+DMP on the DCS model with the Tubingen dynamics if the metric is the probability of finding the source (solid curves) or the first symptomatic patient (dashed curves), together with the theoretical results (dash-dotted lines), as shown in Figure 8.6. The simulations were computed on a population of $n = 9054$ individuals, and each datapoint is the average of 2400 independent realizations for the LS/LS+/LSv2/LS+v2 algorithms, and 48 independent realizations for the random+DMP algorithm. The default population and infection parameters were selected to match the population and COVID-19 infection datasets of Tubingen. The confidence intervals for the success probabilities are computed using the Wilson score interval method, and for the tests and the queries using the Student's t-distribution.

Conclusion Part V

9 Conclusion

Our goal in this dissertation was twofold: we aimed to relax the limitations of source identification frameworks discussed in Section 1.3 by introducing adaptivity to the assumptions, and we aimed to provide rigorous mathematical results, as previous results in source identification with time queries were mostly based on simulations.

Chapter 2 mainly served as motivation for the source identification problem; we showed through data-driven simulations and rigorous proofs on real and modelled geometric metapopulation networks that the location of the sources can have a counter-intuitive effect on the outcome of an epidemic. We identified the so called *swithcover phenomenon*, according to which epidemics started from the central nodes of a geometric metapopulation network can reach more individuals if the basic reproduction number is small, but if the epidemic is more infectious, it reaches a larger population when seeded from uniformly selected nodes.

In Chapter 3, we reviewed the most important results about the role of adaptivity in the case of deterministic epidemic models, where the number of required sensors (queires) is equivalent to the metric dimension (MD) in the non-adaptive case, and to the sequential metric dimension (SMD) in the adaptive case. At the end of Chapter 3, we compared the known results about the MD and the SMD of random graph models, and drew important conclusions about how the network structure impacts the role of adaptivity in deterministic epidemics.

In Chapter 4, we extended previous mathematical results on the MD to a large class of random trees, including critical Galton-Watson trees conditioned to have a certain size, and growing general linear preferential attachment trees. In these trees, the MD always converged to a constant fraction of the size of the graph, and we found analytical expressions to this constant fraction.

In Chapter 5, we provided rigorous bounds on the SMD of dense Erdős-Rényi random graphs, and found that it is within a constant fraction of the MD, making Erdős-Rényi graphs the only known non-trivial graph family where adaptivity does not play an important role in source identification (with deterministic epidemic propagation).

We shifted our attention towards stochastic epidemics in Chapter 6, where we provided the first rigorous results in both adaptive and non-adaptive source identification in stochastic epidemic models. The analysis was done on the path graph, where we observed a double-logarithmic decrease in the number of required queries to identify the source. This result suggests that adaptivity has a very important role in stochastic epidemics.

In the final two chapters, we addressed the limitation that most source identification frameworks assume that the contact network is fully known to us. In Chapter 7, we focused on deterministic epidemic models, more specifically, we studied the robustness of the MD to changes in the network. We exhibited examples of graphs were even a single edge change can create an exponential increase in the MD, suggesting that in extreme cases, even a small uncertainty about the network structure can make source identification infeasible. However, we showed that such large increases are not possible in d-dimensional grid graphs, and in two dimensions, we even characterized the limiting distribution of the MD to a single random edge addition.

Finally, in Chapter 8, we suggested a new source identification framework, the SICTF, based on recent advances in contact tracing developed for the COVID-19 pandemic. We proposed two local search algorithms, which outperformed the state of the art source identification algorithms adapted to the SICTF, and we provided analytical bounds on the success probability of these algorithms.

9.1 Future Work

We discussed specific future research directions in Sections 2.3, 3.4, 4.1, 6.7, 7.1 and 8.6. In this section we give a high-level overview, and discuss a few additional future directions, that were not included in the previous chapters.

Although the S1/S2 source identification frameworks introduced in Chapter 1 are well-studied, many open problems remain. In the deterministic version, while the MD and the SMD of some classes of random graphs are have been studied (see Section 3.4), the classes of random graphs with spatial properties are either not fully understood (e.g., random geometric graphs), or have not even been addressed (e.g., geometric inhomogeneous random graphs). This is an important research direction because real contact networks often have spatial properties due to geographical effects. On the matter of robustness to edge additions, while Chapter 7 makes important advances in the case of grid graphs, the analysis on different (possibly random) graph families is an open research direction. Another direction would be to study a new, *coarse* variant of the MD, where nodes that are too close (say closer than distance d) do not need to be distinguished. The coarse MD would be a deterministic version of the idea of providing confidence sets around the source [65], instead of finding it exactly. We refer to the survey [153] for more open problems related to the MD. In case of the stochastic version of the S1/S2 frameworks, the main open problem is to provide rigorous results for graphs beyond the path graph, which we analyzed in Chapter 6.

This dissertation on the role of adaptivity in source identification was motivated by the limitations of the S1/S2 frameworks reviewed in Section 1.3, which were specific to the application of epidemics. Some of these limitations may not appear, or may appear differently in other applications of source identification. We believe that in light of the results presented in this dissertation, it would be timely to explore these potential other applications, or possibly find applications that have not been reported so far in the literature. The other well-explored application in the literature is rumor spreading, however, similarly to epidemics, we are not aware of actively pursued use cases (beyond proofs of principle) in this application either. It is possible that rumor source identification algorithms are used secretly by individuals or government agencies (e.g., to deanonymize Bitcoin transactions [89]), but such applications raise ethical concerns. Other applications such as identifying the source of train delays [168], or food-borne diseases [169] should be further explored.

In the application of epidemics, due to the limitations of the S1/S2 frameworks, we believe that the most important research direction is to find an analytically tractable source identification framework, which is close to a scenario that occurs in real epidemics. In Chapter 8, we attempted to propose such a framework (the SICTF), and we provided theoretical results about when the source identification task is feasible in this framework based on biological and demographic parameters. Such theoretical results are especially important, because we believe that governments currently do not have well-organized source identification programs, and that for them, the first important piece of information is whether the task is feasible in the first place. It would be important to find out whether this hypothesis is correct, and whether the SICTF is indeed appealing to professional epidemiologists, or if any further assumptions are missing, they should be included to further develop the SICTF.

Appendices Part VI

A Appendix for Chapter 2

In this Appendix, we explain the methods used in the simulation results of Chapter 2, and we state precise versions and give rigorous proofs of Theorems 2.2.2 and 2.2.3.

A.1 Methods

Data description

Settlement level daily COVID-19 infection data for Hungary

For the analysis presented in Figure 2.1, we used a dataset recording the daily number of newly infected cases in 3,118 Hungarian settlements. This data matches the officially reported total number of daily cases [7, 9], however, just as the official data, it suffers from some observational bias due to the limited capacity of testing in the country during certain periods of the pandemic. For the analysis presented in Figure 2.1 we considered all settlements, and obtained their population sizes from data shared by the Hungarian Statistical Office [2]. A version of this data aggregated on the county level is openly available [9].

Daily commuting network of Hungary

For the data-driven simulations of the Hungarian epidemic we use a microcensus collected and released by the Hungarian Statistical Office in 2016 [1]. The data contains the number of people commuting for work or school on a daily base between the 3,186 settlements in Hungary, with the districts of the capital considered as separate towns. In our analysis we concentrated only on settlements with populations larger than 1,000 inhabitants and kept commuting links with at least 25 daily commuters. From this data we constructed an undirected meta-population commuting network with 1,398 settlements as nodes (of which 97 were from the capital and its suburbs) and 8,322 commuting edges with weights computed as the average number of commuters between pairs of towns. The total population size of the network contained the 95% (9,285,286 individuals) of the Hungarian population. Despite the sparsity of the network (0.85% of the possible edges are present), 19% of individuals commute between settlements on a daily base.

Moran's I statistic

We compute the Moran's I statistic at time t as

$$I(t) = \frac{n\sum_{i,j} w_{ij}(y_i(t) - \bar{y}(t))(y_j(t) - \bar{y}(t))}{\sum_{i,j} w_{ij} \sum_i (y_i(t) - \bar{y}(t))^2}, \tag{A.1}$$

where n is the number of nodes, w_{ij} is the edge weight between the nodes i and j, $y_i(t)$ is the number of new infected cases at node i at time t and $\bar{y}(t) = (1/n) \cdot \sum_i y_i(t)$.

Generating Geometric Inhomogeneous Random Graphs

GIRG(τ, α) networks were generated by the following process: the location of n nodes are sampled uniformly at random from the square $[0,1]^2$, and each node u is assigned with a "fitness" value (w_u) sampled from a power-law distribution with exponent τ. Each pair of nodes are connected by an edge with a probability

$$P(u,v) = p \min\left\{\left(\frac{C w_u w_v}{n\|x_u - x_v\|^2}\right)^{\alpha}, 1\right\}, \tag{A.2}$$

which after only the largest connected component of the network is kept. To generate models with different parameters comparable to each other, we fix the number of edges to 5,000, by selecting the constant C and p accordingly, since these two parameters are responsible for the edge-density. For the exact implementation see [3]. When the fitness distribution w_u is set to be a power-law, node degrees also satisfy a power law. The abundance of long-range connections is tuned by α in (A.2): the smaller α, the more likely are long-range connections (b). The power-law exponent τ and the long-range parameter α tune the average graph distance in the network $\overline{\text{Dist}}(n)$, see [66, 42, 33, 68]:

$$\overline{\text{Dist}}(n) = \begin{cases} \Theta(\log\log n) & \text{when } \tau \in (2,3), \alpha > 1 \\ \Theta\big((\log n)^{\zeta}\big) & \text{when } \tau > 3, \alpha \in (1,2) \\ \Theta(\sqrt{n}) & \text{when } \tau > 3, \alpha > 2. \end{cases}$$

Comparing this to the average distance in the configuration model, where only the first two regimes are possible ($\Theta(\log\log n)$ when $\tau \in (2,3)$, $\Theta(\log n)$ when $\tau > 3$), and to distances in lattice models (where $\overline{\text{Dist}}_N$ is polynomial in n), we observe that the underlying geometry of GIRGs with the long-range connections play a role when $\tau > 3$, and the model interpolates between the small-world configuration model and the lattice.

Generating random networks from the Configuration Model

We generate a uniform sample from the set of graphs with power-law degree distribution with degree exponent τ by first generating a GIRG with given parameter τ (and $\alpha = 2.3$), and we swap the end-points of randomly selected pairs of edges [4] to remove all geometric and structural correlations from the structure, while conserving the degree of each node. We perform 10 × #number_of_edges swaps, which mixes the edges enough so that the resulting network becomes close to a uniform sample from the set of networks that have exactly the same degree sequence as the original GIRG network [102].

Core decomposition and seed selection

In metapopulation network models, we use k-shell decomposition to identify the largest k-core of the network [53, 5] and to select seeds in the central area. This algorithm computes the k-shell by recursively removing each node of the network that has degree less than k, until no more nodes can be removed. We take the largest k for which at least s nodes remain, and we select s nodes from this k-shell uniformly at random as our seed set in the central area. For the uniform seeding scenario, we select s nodes of the network uniformly at random. Finally, we infect $i_0 = 0.0005$ fraction of the agents in the total population, and we distribute these agents in the s settlements uniformly at random, irrespective of the size of the settlements.

In the theoretic computations and simulations, the s highest degree nodes are selected for the central area, and s uniformly random nodes are selected for the uniform seeding scenario. Node degrees and core-number of nodes in configuration network models are strongly correlated, allowing us to make this approximation.

SIR model on metapopulation networks

To make our simulations somewhat realistic, we set the hometown of each N_i agents to their initial settlement i. Each agent is assigned exactly one hometown, and the home assignments do not change for the rest of the simulation. We initialize the infection according to one of the seed selection scenarios and proceed with the simulation in each iteration t in three steps. In the diffusion step, each agent who is at its hometown i is selected to move to another town with probability p_m. The selected agents then chose a target town j with probability proportional to the weight w_{ij} of the link connecting town i to j, and move there. We set $p_m = 0.001$ in all simulations, which means that 0.1% of the total population moves in each iteration.

Agents that are not at their hometown simply move back to their home settlement. In the reaction step, each susceptible agent in town i becomes infected with probability $1-(1-\beta/N_i)^{I_i}$, where I_i is the number of infected agents in town i at iteration step t and β is the infection rate. In the final recovery step, each infected agent recovers with rate μ. For the exact implementation see [8].

The limiting function of the percolation pandemic size ratio

In Theorem 2.2.3, we identified the scaling of $f_G(p,s) = f_G(p_c + n^x, n^y) = \Theta(n^{\zeta}(x,y))$, where $\zeta(x,y)$ is a piecewise linear function. On each region A_1-A_6, ζ is given as follows:

$$\zeta(x,y) = \begin{cases} \zeta_1(x,y) = 1 + \left(\frac{1}{|\tau-3|} + \mathbb{1}_{\tau\in(3,4)}\right)x - y & \text{on } A_1 \\ \zeta_2(x,y) = 0 & \text{on } A_2 \\ \zeta_3(x,y) = \mathbb{1}_{\tau\in(3,4)}x + \left(1 - \frac{1}{\tau-1}\right)(1-y) & \text{on } A_3 \cup A_6 \\ \zeta_4(x,y) = -\frac{1}{|\tau-3|}x - \frac{1}{\tau-1}(1-y) & \text{on } A_4 \\ \zeta_5(x,y) = \frac{1}{(\tau-1)(\tau-2)}(1-y) & \text{on } A_5 \end{cases}$$

Finally, on region A_2, $f_G(p_c + n^x, n^y) = 1 - \Theta(n^{-\eta(x,y)})$ where

$$\eta(x,y) = \left(\frac{1}{|\tau-3|} + \mathbb{1}_{\tau\in(3,4)}\right)x - y.$$

A.2 Weak Switchover

In this section we focus on the Configuration model. We start by introducing new notation needed for stating the results and the proofs. See Table A.4 for a glossary of notations.

Definition A.2.1 (Configuration model and its percolation)**.** *Let us denote the Configuration model on n nodes and degree exponent τ by $\mathrm{CM}(n,\tau)$. Let G_n^p be the percolated $\mathrm{CM}(n,\tau)$ with edge-retention probability p on n vertices. Denote by n_c be the number of connected components of G_n^p, by $\mathcal{C}_i$ the i^{th} largest component of G_n^p, by $\mathcal{C}(u)$ the component in G_n^p which contains node u, and let $C_i = \mathbb{E}[|\mathcal{C}_i|]$. Let $p_{c,n,\tau}$ be the critical percolation parameter for $\mathrm{CM}(n,\tau)$ for the existence of a linear sized giant connected component. For edge-retention probability $p = p_n$ that may depend on n, we define $\theta_n = p_n - p_{c,n,\tau}$.*

Recall that $\mathbb{E}_p[\mathbf{Cl}(\mathcal{CI}_0(s))]$ is the expected cluster size of the seed set with the s highest degree nodes in the percolated graph with retention probability p, and $\mathbb{E}_p[\mathbf{Cl}(\mathcal{UI}_0(s))]$ is the same for the seed set with s uniformly chosen nodes. We are interested in the function $f_G(p,s)$, which is the ratio of these two expectations. In the next theorem, which is the precise version of Theorem 2.2.2, we prove existence of the weak switchover property, which was defined in Chapter 2 in terms of the function $f_G(p,s)$.

Theorem A.2.1. *The sequence of random graphs sampled from the Configuration model with exponent $\tau \in (2,4)$ and $n \to \infty$ exhibit weak switchover. Specifically, under the assumptions $1 \gg \theta_n \gg n^{-|\tau-3|/(\tau-1)}$, and $n \gg s_n \gg 1$,*

1. *if $\theta_n^{-\frac{1}{|\tau-3|}} \gg s_n$ or $s_n \gg n\theta^{\frac{\tau-1}{|\tau-3|}}$, then $f_{\mathrm{CM}(n,\tau)}(p_{c,n,\tau} + \theta_n, s_n) > 1$,*
2. *if $\theta_n^{-\frac{1}{|\tau-3|}} \ll s_n \ll n\theta_n^{\frac{\tau-1}{|\tau-3|}}$, then $f_{\mathrm{CM}(n,\tau)}(p_{c,n,\tau} + \theta_n, s_n) < 1$,*

with high probability as $n \to \infty$.

Theorem A.2.1 is a qualitative result that shows the existence of the weak switchover phenomenon. In Theorem A.2.2 we report our quantitative results on $f_{\mathrm{CM}(n,\tau)}(p,s)$, which will directly imply Theorem A.2.1. First we need some new definitions.

Definition A.2.2 (Phases of the parameter space)**.** *Given $\tau \in (2,4)$ and a sequence $(\theta_n)_{n\geq 1} = (p_n - p_{c,n,\tau})_{n\geq 1}$ with $(\theta_n > 0)$, and $(s_n)_{n\geq 1}$ with $(s_n > 0)$ for which the limits $x = \lim_{n\to\infty} \log_n(\theta_n)$ and $y = \lim_{n\to\infty} \log_n(s_n)$ both exist, let us partition the parameter space (θ_n, s_n) (equivalently, (x,y)) into six sets in the following way*

$$A_1 = \big\{(\theta_n, s_n) \mid s_n \ll \min(\theta_n^{-\frac{1}{|\tau-3|}}, n\theta_n^{\frac{\tau-1}{|\tau-3|}})\big\} \equiv \big\{(x,y) \mid y < \min\{-\tfrac{1}{|\tau-3|}x, 1 + \tfrac{\tau-1}{|\tau-3|}x\}\big\}$$

$$A_2 = \big\{(\theta_n, s_n) \mid \theta_n^{-\frac{1}{|\tau-3|}} \ll s_n \ll n\theta_n^{\frac{\tau-1}{|\tau-3|}}\big\} \equiv \big\{(x,y) \mid -\tfrac{1}{|\tau-3|}x < y < 1 + \tfrac{\tau-1}{|\tau-3|}x\big\}$$

$$A_3 = \big\{(\theta_n, s_n) \mid n\theta_n^{\frac{\tau-1}{|\tau-3|}} \ll s_n \ll \theta_n^{-\frac{1}{|\tau-3|}}\big\} \equiv \big\{(x,y) \mid 1 + \tfrac{\tau-1}{|\tau-3|}x \leq y \leq -\tfrac{1}{|\tau-3|}x\big\}$$

$$\begin{aligned} A_4 &= \big\{(\theta_n, s_n) \mid \max(\theta_n^{-\frac{1}{|\tau-3|}}, n\theta_n^{\frac{\tau-1}{|\tau-3|}}) \ll s_n \ll \min(n\theta_n^{\frac{\tau-2}{|\tau-3|}}, n\theta_n^{\frac{1}{|\tau-3|}})\big\} \\ &\equiv \big\{(x,y) \mid \max\{-\tfrac{1}{|\tau-3|}x, 1 + \tfrac{\tau-1}{|\tau-3|}x\} < y < \min\{1 + \tfrac{\tau-2}{|\tau-3|}x, 1 + \tfrac{1}{|\tau-3|}x\}\big\} \end{aligned}$$

$$A_5 = \big\{(\theta_n, s_n) \mid n\theta_n^{\frac{\tau-2}{|\tau-3|}} \ll s_n\big\} \equiv \big\{(x,y) \mid 1 + \tfrac{\tau-2}{|\tau-3|}x < y\big\}$$

$$A_6 = \big\{(\theta_n, s_n) \mid \max(\theta_n^{-\frac{1}{|\tau-3|}}, n\theta_n^{\frac{1}{|\tau-3|}}) \ll s_n \ll n\theta_n^{\frac{\tau-2}{|\tau-3|}})\big\} \equiv \big\{(x,y) \mid -\tfrac{1}{|\tau-3|}x < y < 1 + \tfrac{\tau-2}{|\tau-3|}x\big\}$$

See Figure A.1 for a visualization of the sets A_i. Intuitively, the union of A_1 and A_3 are the parameter ranges for which there is no uniformly selected seed in the giant. The union of A_1 and A_2 are the parameter ranges for which all of the high degree seeds are contained in the giant. The set A_4 is an intermediate regime, where there are high degree seeds outside the giant and uniform seeds inside the giant, and in A_5 the parameter s is so large that there are multiple uniformly selected seeds in medium sized components (in addition to the giant). We note that A_6 is an empty set for $\tau \in (3,4)$. For $\tau \in (2,3)$, the set A_6 contains the parameter ranges where the giant component is smaller than the contribution of small components with only a single uniformly selected seed.

For our quantitative results, we will have to condition on an event $\mathcal{E}_n$, which holds with high probability. This is a standard technique to rule out rare events that have too big of an impact on the expected value (see the analytic derivation for more details). In the next definition we extend the definition of $f_G(p,s)$ to incorporate this conditioning.

Definition A.2.3 (Percolation pandemic size ratio)**.** *On a graph G and two seeding sets $\mathcal{C}I_0(s), \mathcal{C}U_0(s)$ of size s, with edge-retention probability $p \in [0,1]$, let the pandemic size ratio function conditioned on an event $\mathcal{E}$ be*

$$f_G(p,s,\mathcal{E}) = \frac{\mathbb{E}_p[\mathbf{Cl}(\mathcal{CI}_0(s)) \mid \mathcal{E}]}{\mathbb{E}_p[\mathbf{Cl}(\mathcal{UI}_0(s)) \mid \mathcal{E}]}. \tag{A.3}$$

Now we are ready to state the precise version of Theorem 2.2.3 from Chapter 2.

Theorem A.2.2. *Let A_i defined in Definition A.2.2, for a sequence random graphs sampled from the Configuration model with exponent $\tau \in (2,4)$, and let $\mathcal{E}_n$ be the event that either $s_n \in A_2 \cup A_4 \cup A_5 \cup A_6$ or the event $\{\mathcal{UI}_0(s_n) \cap \mathcal{C}_1 = \emptyset, s_n \in A_1 \cup A_3\}$ hold. Then, if $p_{c,n,\tau}$ is the critical percolation parameter for* $\mathrm{CM}(n,\tau)$*, under assumptions $1 \gg \theta_n \gg n^{-(|\tau-3|)/(\tau-1)}$, and $n \gg s_n \gg 1$,*

$$f_{\mathrm{CM}(n,\tau)}(p_{c,n,\tau}+\theta_n, s_n, \mathcal{E}_n) = \begin{cases} \Theta(\theta_n^{\frac{1}{|\tau-3|}+\mathbb{1}_{\tau\in(3,4)}} n/s_n) & \textit{if} (s_n,\theta_n) \in A_1 \\ 1-\Theta(\theta_n^{-\frac{1}{|\tau-3|}-\mathbb{1}_{\tau\in(3,4)}} s_n/n) & \textit{if} (s_n,\theta_n) \in A_2 \\ \Theta(\theta_n^{\mathbb{1}_{\tau\in(3,4)}} (n/s_n)^{1-\frac{1}{\tau-1}}) & \textit{if} (s_n,\theta_n) \in A_3 \cup A_6 \\ \Theta(\theta_n^{-\frac{1}{|\tau-3|}} (n/s_n)^{-\frac{1}{\tau-1}}) & \textit{if} (s_n,\theta_n) \in A_4 \\ \Theta((n/s_n)^{\frac{1}{(\tau-1)(\tau-2)}}) & \textit{if} (s_n,\theta_n) \in A_5 \end{cases}, \tag{A.4}$$

and $\mathbb{P}(\mathcal{E}_n) \to 1$.

Since for A_1, A_3, A_4, A_5, A_6 we have $f_{\mathrm{CM}(n,\tau)}(p_n, s_n, \mathcal{E}_n) \to \infty$ and for A_2 we have $f_{\mathrm{CM}(n,\tau)}(p_n, s_n, \mathcal{E}_n) \to 1$, some of our results will be lost if we apply the same normalization to the limit of $f_{\mathrm{CM}(n,\tau)}(p_n, s_n, \mathcal{E}_n)$ for all A_i. For example if we normalize by applying the function $\log_n$, as we do in the definition of $\bar{f}_\tau$ in Definition A.2.4, the deviation of f_G below 1 in the region A_2 will disappear. To mitigate this issue, we propose a discontinuous normalization in addition to normalizing by $\log_n$.

Definition A.2.4. *Let $\theta = (\theta_n) = (p_n - p_{c,n,\tau})$, $s = (s_n)$ be a sequence of seed counts, $\mathcal{E} = (\mathcal{E}_n)$ be a sequence of events and for $z > 0$ let us define the normalisation*

$$\mathrm{dNorm}_n(z) = \begin{cases} \log_n(z) & \textit{if}\, z > 1 \\ -\log_n(1-z) - 1 & \textit{if}\, z < 1 \end{cases}. \tag{A.5}$$

Then, assuming the limits $x = \lim_{\to\infty} \log_n(\theta_n)$ and $y = \lim_{n\to\infty} \log_n(s_n)$ both exist, we define

$$\begin{aligned} \bar{f}_\tau(x,y,\mathcal{E}) &= \lim_{n\to\infty} \log_n(f_{\mathrm{CM}(n,\tau)}(p_{c,n,\tau}+\theta_n, s_n, \mathcal{E}_n)) \\ \tilde{f}_\tau(x,y,\mathcal{E}) &= \lim_{n\to\infty} \mathrm{dNorm}_n(f_{\mathrm{CM}(n,\tau)}(p_{c,n,\tau}+\theta_n, s_n, \mathcal{E}_n)). \end{aligned}$$

Note that the function $\bar{f}_\tau(x,y,\mathcal{E})$ defined above corresponds to the limit of $\log_n(\widetilde{f_G}(x,y))$ in Theorem 2.2.3 of Chapter 2. Now we are ready to apply the normalization and find the limiting curve. See Figure A.1 for a visualization in 3D.

Corollary A.2.1. *Let A_i and $\mathcal{E}$ be defined as in Theorem A.2.2, and $x, y, \tilde{f}_\tau(x,y,\mathcal{E})$ as given in*

Definition A.2.4. Then, under the assumptions $0 > x > -(|\tau-3|)/(\tau-1)$*, and* $1 > y > 0$,

$$\tilde{f}_\tau(x,y,\mathcal{E}) = \begin{cases} 1+\left(\frac{1}{|\tau-3|}+\mathbb{1}_{\tau\in(3,4)}\right)x-y & \textit{if}\,(x,y)\in A_1 \\ \left(\frac{1}{|\tau-3|}+\mathbb{1}_{\tau\in(3,4)}\right)x-y & \textit{if}\,(x,y)\in A_2 \\ \mathbb{1}_{\tau\in(3,4)}x+\left(1-\frac{1}{\tau-1}\right)(1-y) & \textit{if}\,(x,y)\in A_3\cup A_6 \\ -\frac{1}{|\tau-3|}x-\frac{1}{\tau-1}(1-y) & \textit{if}\,(x,y)\in A_4 \\ \frac{1}{(\tau-1)(\tau-2)}(1-y) & \textit{if}\,(x,y)\in A_5 \end{cases}, \tag{A.6}$$

and $\mathbb{P}(\mathcal{E}_n)\to 1$*. Moreover* $\bar{f}_\tau(x,y,\mathcal{E}) = \tilde{f}_\tau(x,y,\mathcal{E})$ *except if* $(x,y)\in A_2$*, when* $\bar{f}_\tau(x,y,\mathcal{E}) = 0$.

Remark A.2.1. *As shown in Figure A.1 (e) and (h), with the continuous normalizaton* $\bar{f}$*, there is a continuous transition between all of the regions except on the boundary of regions* A_1 *and* A_2 *and on the boundary of regions* A_3 *and* A_4*, where the transition is discontinuous.*

Remark A.2.2. *In our quantitative results, we conditioned on the event* $\mathcal{E}_n$*, which rules out the rare event that a small and uniform seed intersects with the giant component. When we perform simulations on finite size networks, if we use many independent samples in our sample average, it is possible that the rare event* $\mathcal{E}_n$ *occurs in certain parameter ranges, which can result in outliers that skew our sample average away from the theoretical prediction. Therefore, after we simulate independent estimates of* $\mathbf{Cl}(\mathcal{CI}_0(s))$ *and* $\mathbf{Cl}(\mathcal{UI}_0(s))$*, we perform an outlier removal technique, in which we remove datapoints that are larger than the mean plus* 2 *standard deviations of the dataset. As shown in Figure A.1 (j)-(m), in most cases the outlier removal does not remove more than* 1% *of the datapoints, except for the esimtates of* $\mathbf{Cl}(\mathcal{UI}_0(s))$*, for the parameter ranges close to the discontinuous phase transition (see Remark A.2.1), where in some cases* 15% *of the datapoints are removed. After the outlier detection, we compute the mean of the remaining datapoints for an estimate of* $\mathbb{E}_p[\mathbf{Cl}(\mathcal{CI}_0(s))\mid\mathcal{E}]$ *and* $\mathbb{E}_p[\mathbf{Cl}(\mathcal{UI}_0(s))\mid\mathcal{E}]$*, and compute their ratio to approximate* $f_G(p,s,\mathcal{E})$*. The resulting curves for* $\tau = 3.5$ *and* $\tau = 2.5$ *are shown in Figure A.1 (f) and (i).*

A.2.1 Previous Results on Random Networks with Power-Law Degree Distribution

Percolation cluster sizes in the near-critical regime has been extensively studied in the physics literature for various network models. For a-geometric networks with power-law degree distributions, the non-rigorous works [183, 59] predict *critical exponents*, some of it has been made rigorous for the configuration model [71, 72, 70, 236] for rank-1 inhomogeneous random graphs [31, 32, 237] and Erdős-Rényi graphs [73]. Based on these results, we summarize the cluster size distribution in the near-critical regime in Table A.1, and in a reparametrized form in Table A.1. To unify the notation we denote by θ_n the deviation of p_n from the critical point (denoted by $s^\star$ in the physics literature), and we denote the critical exponents as

$$\lambda = \begin{cases} \frac{2\tau-3}{\tau-2} & \text{if } 2 < \tau < 4 \\ \frac{5}{2} & \text{if } 4 < \tau \end{cases} \tag{A.7}$$

$$\sigma = \begin{cases} \frac{3-\tau}{\tau-2} & \text{if } 2 < \tau < 3 \\ \frac{\tau-3}{\tau-2} & \text{if } 3 < \tau < 4 \\ \frac{1}{2} & \text{if } 4 < \tau \end{cases} \tag{A.8}$$

$$\beta = \begin{cases} \frac{1}{3-\tau} & \text{if } 2 < \tau < 3 \\ \frac{1}{\tau-3} & \text{if } 3 < \tau < 4 \\ 1 & \text{if } 4 < \tau \end{cases} \tag{A.9}$$

$$p_c = \begin{cases} \frac{1}{\frac{\tau-2}{3-\tau}(d)_{min}^{\tau-2} n^{\frac{3-\tau}{\tau-1}} - 1} & \text{if } 2 < \tau < 3 \\ \frac{1}{\frac{\tau-2}{\tau-3}(d)_{min} - 1} & \text{if } \tau > 3 \end{cases} . \tag{A.10}$$

A.2.2 Proof of Theorem A.2.2

In the proof, first we are going to sample the graph and percolate the edges, which gives a random graph with a component structure described in Section A.2.1. Then, we are going to sample the seed sets, and we will understand which components the highest degree nodes and the uniform seed set are likely to "hit" (i.e., intersect).

The main difficulty of the proof is that in different parameter ranges, the highest degree nodes and the uniform seed set hit different types of clusters. We show that the highest degree nodes hit the components in decreasing order until a certain component size, which we call $C_{\min}^{(c)}$. For small s, $C_{\min}^{(c)}$ is exactly C_1, in which case the highest degree nodes are contained entirely in the giant, and for larger s, $C_{\min}^{(c)}$ is strictly smaller than C_1, in which case the highest degree nodes infect medium sized components in addition to the giant. We denote the contribution of these medium-sized components to the total size that the highest degree nodes infect by E_1 in the calculations below.

Similarly to the highest degree nodes, we denote the smallest component size for which all components of that size or larger are hit by the uniform seed set with high probability by $C_{\min}^{(u)}$.

In contrast with the highest degree nodes, for small s, the uniformly selected seeds hitting the giant component becomes a rare event (occurring with probability $q_{p,1} = o(1)$). To rule out this rare event, which would skew the expected value, we condition on $\mathcal{E}$, the complement of this rare event (a standard technique in the theory of random graphs with heterogenous degree distribution). Thus, the $C^{(u)}_{\min}$ for small s becomes undefined, and each uniformly selected seed hits a component with small expected size (denoted by E_3 below). As s increases, the uniform seed set starts hitting the giant to give $q_{p,1} = \Theta(1)$ and $C^{(u)}_{\min} = C_1$. Increasing s even further, similarly to the highest degree nodes, eventually we start having $C^{(u)}_{\min} < C_1$ and we denote by E_2 the contribution of these medium sized components to the total size of infected by the uniform seed set.

Since $C^{(c)}_{\min} \le C^{(u)}_{\min}$, the only way the uniform seed set can infect more nodes than the highest $C^{(c)}_{\min} = C^{(u)}_{\min} = C_1$, because in this case all highest degree nodes are contained in the giant ($E_1 = 0$), but the uniform seed set can still hit some small components ($E_3 > 0$), which implies that the uniform seed set has a small advantage.

This intuition is made formal in the proof below. See Table A.4 for a list of definitions used in the proof. In Claim A.2.1, in (A.11)-(A.15), we explicitly derive rows 3-7 and 12-18 of Table A.3. Rows 8-9 and 19-20 of Table A.3 contain the statements of Theorem A.2.2 and Corollary A.2.1. Entries of the rows 8-9 and 19-20 in Table A.3 can be computed by substituting in entries from the previous rows into (A.18), which we do in a case-by-case analysis after presenting the formal computations that support Claim A.2.1.

Notation. *In this section we drop the subscript n from sequences s, p, θ, and $\mathcal{E}$ and we use the simplified notation*

$$f_{n,\tau}(\theta, s) = f_{\mathrm{CM}(n,\tau)}(p_{c,n,\tau} + \theta_n, s_n)$$
$$f_{n,\tau}(\theta, s, \mathcal{E}) = f_{\mathrm{CM}(n,\tau)}(p_{c,n,\tau} + \theta_n, s_n, \mathcal{E}_n).$$

Claim A.2.1. *Under the assumptions $1 \gg \theta \gg n^{-(|\tau-3|)/(\tau-1)}$, $n \gg s \gg 1$ and definitions given in Table A.4, the following equations hold for $\tau \in (2,4)$*

$$q_{c,1} = \mathbb{P}(|\mathcal{CI}_0(s) \cap \mathcal{C}_1| > 0) = 1 - O(n^{-\log(n)}), \tag{A.11}$$

$$q_{p,1} = \mathbb{P}(|\mathcal{UI}_0(s) \cap \mathcal{C}_1| > 0) = \begin{cases} \Theta(s\theta^{\frac{1}{|\tau-3|}}) & \mathit{if}\, s \ll \theta^{-\frac{1}{|\tau-3|}}, \\ 1 - O(n^{-\log(n)}) & \mathit{if}\, s \gg \theta^{-\frac{1}{|\tau-3|}}, \end{cases} \tag{A.12}$$

$$E_1 = \mathbb{E}\left[\sum_{i=1}^{n_{c,p}} |\mathcal{C}_i| \mathbb{1}_{\{C_1 > |\mathcal{C}_i| > C^{(c)}_{\min}\}}\right] = \begin{cases} 0 & \mathit{if}\, s \ll n\theta^{\frac{\tau-1}{|\tau-3|}}, \\ \Theta(n^{1-\frac{1}{\tau-1}} s^{\frac{1}{\tau-1}}) & \mathit{if}\, s \gg n\theta^{\frac{\tau-1}{|\tau-3|}}, \end{cases} \tag{A.13}$$

$$E_2 = \mathbb{E}\left[\sum_{i=1}^{n_{c,p}} |\mathcal{C}_i| \mathbb{1}_{\{C_1 > |\mathcal{C}_i| > C^{(u)}_{\min}\}}\right] = \begin{cases} 0 & \mathit{if}\, s \ll n\theta^{\frac{\tau-2}{|\tau-3|}}, \\ \Theta(n^{\frac{\tau-3}{\tau-2}} s^{\frac{1}{\tau-2}}) & \mathit{if}\, s \gg n\theta^{\frac{\tau-2}{|\tau-3|}}. \end{cases} \tag{A.14}$$

(A.15)

When $\tau \in (2,3)$*, then*

$$E_3 = \mathbb{E}_{u \sim \mathcal{U}(V)} \left[|\mathcal{C}(u)| \mathbb{1}_{\{|\mathcal{C}(u)| < \min(C_{\min}^{(u)}, C_1)\}} \right] = \Theta(1), \tag{A.16}$$

while for $\tau \in (3,4)$,

$$E_3 = \mathbb{E}_{u \sim \mathcal{U}(V)} \left[|\mathcal{C}(u)| \mathbb{1}_{\{|\mathcal{C}(u)| < \min(C_{\min}^{(u)}, C_1)\}} \right] = \begin{cases} \Theta(\theta^{-1}) & \textit{if } s \ll n\theta^{\frac{\tau-2}{\tau-3}}, \\ \Theta\left(\left(\frac{n}{s}\right)^{\frac{\tau-3}{\tau-2}}\right) & \textit{if } s \gg n\theta^{\frac{\tau-2}{\tau-3}}. \end{cases} \tag{A.17}$$

Finally, for all cases, it holds that

$$f_{n,\tau}(\theta, s) = \frac{\mathbb{E}_{p_c+\theta}\left[\mathbf{Cl}(\mathcal{CI}_0(s))\right]}{\mathbb{E}_{p_c+\theta}\left[\mathbf{Cl}(\mathcal{UI}_0(s))\right]} = \frac{q_{c,1} C_1 + \Theta(E_1) + o(1/n)}{q_{p,1} C_1 + \Theta(E_2) + \Theta(sE_3)}. \tag{A.18}$$

In what follows, we derive each equation in this claim.

Proof of (A.11)**:** Using Table A.1, we estimate the probability from below by the probability that the largest degree vertex v_1 with degree $\Theta(n^{1/(\tau-1)})$ is not in the giant. Here, we use the fact that the number of edges in the giant component is $\Theta(n)$. Therefore, in an exploration process of the giant component, we need to match the $\Theta(C_1)$ many half-edges, and none of these half-edges can be matched to v_1, which means that the probability that v_1 avoids the giant is $(1 - d_1/n)^{\Theta(C_1)}$ with $C_1 = \Theta(\theta^{1/(3-\tau)})$. This yields that

$$q_{c,1} > 1 - \mathbb{P}(v_1 \notin \mathcal{C}_1) \approx 1 - \left(1 - \frac{d_1}{n}\right)^{\Theta(C_1)} \approx 1 - \Theta\left(\mathrm{e}^{-n^{\frac{1}{\tau-1}} \theta^{\frac{1}{|\tau-3|}}}\right) = 1 - O(n^{-\log(n)}) \tag{A.19}$$

because we assumed $\theta \gg n^{-(|\tau-3|)/(\tau-1)}$.

Proof of (A.12)**:** Using that $C_1 = \Theta(\theta^{1/(3-\tau)})$ from Table A.1, the probability that none of the uniformly selected seeds fall among the C_1 many vertices is

$$q_{p,1} = 1 - \left(1 - \frac{C_1}{n}\right)^s \approx 1 - (1 - \theta^{\frac{1}{|\tau-3|}})^s \approx 1 - \mathrm{e}^{-s\theta^{\frac{1}{|\tau-3|}}} \approx \begin{cases} \Theta(s\theta^{\frac{1}{|\tau-3|}}) & \text{if } s \ll \theta^{-\frac{1}{|\tau-3|}}, \\ 1 - O(n^{-\log(n)}) & \text{if } s \gg \theta^{-\frac{1}{|\tau-3|}}. \end{cases} \tag{A.20}$$

Proof of (A.13)**:** We start by counting the number of half-edges incident to $\mathcal{CI}_0(s) = \{v_1, \ldots, v_s\}$ as

$$H(\mathcal{CI}_0(s)) := \sum_{i=1}^{s} d_i = \sum_{i=1}^{s} \left(\frac{n}{i}\right)^{\frac{1}{\tau-1}} = n^{\frac{1}{\tau-1}} s^{1-\frac{1}{\tau-1}}. \tag{A.21}$$

Let us construct a (medium sized) component of given size K using an exploration process, by matching half-edges one-by-one in the component. We must match $\Theta(K)$ half-edges, so the chance that none of these half-edges are mathced to the half-edges attached to vertices in

$\mathcal{CI}_0(s)$ is

$$\begin{aligned}\mathbb{P}(\text{a half-edge is not matched with a half-edge attached to } \mathcal{CI}_0(s)) &= 1 - \frac{H(\mathcal{CI}_0(s))}{\Theta(n)} \\ &= 1 - \Theta\big((n/s)^{-(\tau-2)/(\tau-1)}\big),\end{aligned}$$

where the denominator is $\Theta(n)$ since during the whole procedure the available total number of half-edges is $\Theta(n)$. So, the probability that a component of size K is not containing any of the vertices in $\mathcal{CI}_0(s)$ is

$$\mathbb{P}(\mathcal{C}_u \cap \mathcal{CI}_0(s) = \varnothing | \mathcal{C}_u = K) = \exp\Big(-\Theta\big(K(\tfrac{n}{s})^{-(\tau-2)/(\tau-1)}\big)\Big). \tag{A.22}$$

Hence, components of size $K \gg (n/s)^{\frac{\tau-2}{\tau-1}}$ intersect with $\mathcal{CI}_0(s)$ with constant probability, whereas components of size $K \ll (n/s)^{\frac{\tau-2}{\tau-1}}$ do not. The threshold $(n/s)^{\frac{\tau-2}{\tau-1}}$ can either be larger than the size of the second largest component $C_2 = \theta^{-(\tau-2)/(|\tau-3|)}$, in which case the entire $\mathcal{CI}_0(s)$ is contained in the giant component, or $(n/s)^{\frac{\tau-2}{\tau-1}}$ is smaller than C_2, in which case $\mathcal{CI}_0(s)$ hits some medium components as well. Solving $(n/s)^{\frac{\tau-2}{\tau-1}} < \theta^{-\frac{\tau-2}{|\tau-3|}}$ for the latter case, we get

$$C^{(c)}_{\min} = \begin{cases} C_1 & \text{if } s \ll n\theta^{\frac{\tau-1}{|\tau-3|}}, \\ (n/s)^{\frac{\tau-2}{\tau-1}} & \text{if } s \gg n\theta^{\frac{\tau-1}{|\tau-3|}}. \end{cases} \tag{A.23}$$

This implies that $s \ll n\theta^{\frac{\tau-1}{|\tau-3|}}$ we have $E_1 = 0$ with high probability. Otherwise, by (A.22), we hit all components of size at least $C^{(c)}_{\min}$, and recalling that $n_{c,p}$ is the total number of percolated components, that is order n, we use by Table A.1 for the distribution of component sizes to calculate

$$\begin{aligned} E_1 = \mathbb{E}\left[\sum_{i=1}^{n_{c,p}} |\mathcal{C}_i| \mathbb{1}_{\{C_1 > |\mathcal{C}_i| > C^{(c)}_{\min}\}}\right] &\approx n_{c,p} \sum_{k=(n/s)^{\frac{\tau-2}{\tau-1}}}^{\theta^{-\frac{\tau-2}{|\tau-3|}}} k \cdot k^{-\frac{2|\tau-3|}{\tau-2}} \\ &\approx n \int_{(n/s)^{\frac{\tau-2}{\tau-1}}}^{\theta^{-\frac{\tau-2}{|\tau-3|}}} x^{1-\frac{2|\tau-3|}{\tau-2}}\, dx \approx n\left(\frac{n}{s}\right)^{\frac{\tau-2}{\tau-1}\left(2-\frac{2|\tau-3|}{\tau-2}\right)} = n^{\frac{\tau-2}{\tau-1}} s^{\frac{1}{\tau-1}}, \end{aligned} \tag{A.24}$$

because $1 - \frac{2|\tau-3|}{\tau-2} = -\big(1 + \frac{1}{\tau-2}\big) < -1$. We note that we used (and will use later) the simple result that $n_{c,p} = \Theta(n)$ because of the last row of Table A.1 substituted with constant k.

Proof of (A.14)**:** The expected number of uniformly chosen seeds in a cluster of size K is sK/n, and similarly to (A.22), the probability that s uniformly chosen seeds avoid a cluster of size K decays exponentially. Hence, we expect the uniform seed set get all clusters with $K \gg \frac{n}{s}$ and some of the clusters with size $K \ll \frac{n}{s}$, which implies

$$C_{\text{min}}^{(u)} = \begin{cases} \frac{n}{s} & \text{if } C_2 \gg \frac{n}{s}, \\ C_1 & \text{if } C_1 \gg \frac{n}{s} \gg C_2, \\ \text{undefined} & \text{if } \frac{n}{s} \gg C_1. \end{cases} \tag{A.25}$$

Then, if $s \ll n\theta^{\frac{\tau-2}{|\tau-3|}}$ we have $C_{\text{min}}^{(u)}$ equal C_1 or undefined, and therefore $E_2 = 0$. Otherwise, by Table A.1,

$$E_2 = \mathbb{E}\left[\sum_{i=1}^{n_{c,p}} |\mathcal{C}_i| \mathbb{1}_{\{C_1 > |\mathcal{C}_i| > C_{\text{min}}^{(u)}\}}\right] \approx n_{c,p} \sum_{k=n/s}^{\theta^{-\frac{\tau-2}{|\tau-3|}}} k \cdot k^{-\frac{2|\tau-3|}{\tau-2}} \approx n \int_{n/s}^{\theta^{-\frac{\tau-2}{|\tau-3|}}} x^{1-\frac{2|\tau-3|}{\tau-2}} dx \approx n\left(\frac{n}{s}\right)^{2-\frac{2|\tau-3|}{\tau-2}} = n^{\frac{\tau-3}{\tau-2}} s^{\frac{1}{\tau-2}} \tag{A.26}$$

because $1 - \frac{2|\tau-3|}{\tau-2} = -\left(1 + \frac{1}{\tau-2}\right) < -1$.

Proof of (A.15)**:** By Table A.1 and (A.25),

$$E_3 = \mathbb{E}_{u \sim \mathcal{U}(V)}\left[|\mathcal{C}(u)| \mathbb{1}_{\{|\mathcal{C}(u)| < \min(C_{\text{min}}^{(u)}, C_1)\}}\right] \approx \sum_{k=1}^{\min(C_{\text{min}}^{(u)}, C_2)} k \cdot k^{-\frac{\tau-1}{\tau-2}} \approx \int_1^{\min\left(\frac{n}{s}, \theta^{-\frac{\tau-2}{|\tau-3|}}\right)} x^{-\frac{1}{\tau-2}} dx. \tag{A.27}$$

There are three cases for what the integral in (A.27) could evaluate to. If $\tau \in (2,3)$, then $-\frac{1}{\tau-2} < -1$ and $E_3 = \Theta(1)$. In the other case, if $\tau \in (3,4)$, then $-\frac{1}{\tau-2} > -1$ and the integral in (A.27) evaluates to

$$E_3 = \min\left(\frac{n}{s}, \theta^{-\frac{\tau-2}{|\tau-3|}}\right)^{\frac{|\tau-3|}{\tau-2}}.$$

Therefore, if $s \ll n\theta^{\frac{\tau-2}{|\tau-3|}}$ we have $E_3 = \theta^{-1}$, otherwise, $E_3 = (n/s)^{\frac{|\tau-3|}{\tau-2}}$

Proof of (A.18)**:** We calculate the expected final size of the cluster of the uniform seed set first, i.e., the denominator in $f_{n,\tau}(\theta, s, \mathcal{E}_n)$. For the uniform seed set, following the definitions in Table A.4, since every cluster of size larger than $C_{\text{min}}^{(u)}$ is hit by the uniform seed set with constant probability (hidden in the Θ notation before E_2),

$$\mathbb{E}_p[\mathbf{Cl}(\mathcal{UI}_0(s))] = q_{p,1}C_1 + \Theta(E_2) + \mathbb{E}\left[\sum_{i=1}^{n_{c,p}} \mathcal{C}_i \mathbb{1}_{\{\mathcal{UI}_0(s) \cap \mathcal{C}_i \neq \emptyset\}} \mathbb{1}_{\{\mathcal{C}_i < C_{\text{min}}^{(u)}\}}\right]. \tag{A.28}$$

Denote the last term on the right hand side by T_3. Since there are at most s nodes in clusters that have size less than $C_{\text{min}}^{(u)}$, and assuming each of these s nodes hits a different cluster, we get the upper bound on the last term

$$T_3 < sE_3.$$

For the lower bound on T_3, first we argue that $\Theta(s)$ seeds fall into these small components. Indeed, note that since $C_1 + E_2 = o(n)$, we have that the probability of a uniformly random chosen seed being in a cluster of size less than $C^{(u)}_{\min}$ is strictly positive (tends to one, in fact). Moreover, since the expected number of uniformly chosen seeds in a cluster of size $K \ll C^{(u)}_{\min}$ is $\frac{sK}{n} \to 0$, the probability of a uniformly random chosen seed being the only seed in its cluster also tends to 1. Thus, we can ignore the seeds colliding or falling in larger components, and we can write

$$T_3 = \Omega(sE_3).$$

This establishes the bound on the denominator in (A.18).

We continue with the numerator and estimate $\mathbb{E}_p[\mathbf{Cl}(\mathcal{CI}_0(s))]$, i.e., the cluster size of the highest degree nodes. Following the definitions in Table A.4, we start with a lower bound that follows immediately from (A.11) and (A.13) and the fact that all components in E_1 will be infected with probability tending to (derived in (A.22)):

$$\mathbb{E}_p[\mathbf{Cl}(\mathcal{CI}_0(s))] > q_{c,1}C_1 + \Theta(E_1).$$

It is left to show an upper bound on $\mathbb{E}_p[\mathbf{Cl}(\mathcal{CI}_0(s))]$. Let us start with the case $s \ll n\theta^{\frac{\tau-1}{|\tau-3|}}$. In this case, using (A.22) and estimating each cluster-size trivially from above by n, we can write:

$$\begin{aligned}
\mathbb{E}_p[\mathbf{Cl}(\mathcal{CI}_0(s))]| &< C_1 + \mathbb{E}\left[\sum_{i=1}^{s} \mathbb{1}(v_i \notin C_1)\mathcal{C}(v)\right] \\
&< C_1 + \sum_{i=1}^{s}\left(1 - \frac{C_1}{n}\right)^{d_i} n \\
&< C_1 + \sum_{i=1}^{s} \exp\left(-\left(\frac{n}{i}\right)^{\frac{1}{\tau-1}} \theta^{\frac{1}{|\tau-3|}}\right) n \\
&< C_1 + \exp\left(2\log(n) - \left(\frac{n}{s}\right)^{\frac{1}{\tau-1}} \theta^{\frac{1}{|\tau-3|}}\right). \qquad \text{(A.29)}
\end{aligned}$$

Here we do a case distinction. Whenever $s \ll n\theta^{\frac{\tau-1}{|\tau-3|}}$, we have

$$2\log(n) - \left(\frac{n}{s}\right)^{\frac{1}{\tau-1}} \theta^{\frac{1}{|\tau-3|}} \to -\infty, \qquad \text{(A.30)}$$

and thus $\mathbb{E}_p[\mathbf{Cl}(\mathcal{CI}_0(s))] < C_1 + o(1/n)$, which is what we needed, because in this case $E_1 = 0$ in (A.14).

For the case $s \gg n\theta^{\frac{\tau-1}{|\tau-3|}}$ we use a coupling argument and monotonicity. Clearly, if we increase the edge-retention probability p to $p' > p$, the total size of infected clusters cannot decrease. So, let us consider percolation with p' satisfying $\theta' = (s/n)^{\frac{|\tau-3|}{\tau-1}} \gg \theta$, implying that $p' > p$. Repeating (A.29) with θ', we get

$$\mathbb{E}_p[\mathbf{Cl}(\mathcal{CI}_0(s))] < C_1' + o(1/n) = n\theta'^{\frac{1}{|\tau-3|}} + o(1/n) = n^{1-\frac{1}{\tau-1}} s^{\frac{1}{\tau-1}} + o(1/n),$$

since in this case (A.30) holds for the given choice of s and θ'. Hence, by the monotonicity property of the cluster sizes in variable p we arrive to

$$\mathbb{E}_p[\mathbf{Cl}(\mathcal{CI}_0(s))] < \mathbb{E}_{p'}[\mathbf{Cl}(\mathcal{CI}_0(s))] < n^{1-\frac{1}{\tau-1}} s^{\frac{1}{\tau-1}} + o(1/n) = \Theta(E_1) + o(1/n).$$

Completing the proof of Theorem A.2.2 and Corollary A.2.1: We will use (A.18) or a conditioned version of it to compute $f_{n,\tau}(\theta, s, \mathcal{E})$. We treat each region A_i in a case-by-case analysis to explain the final two rows of Table A.3. See Figure A.2 for an illustration of each case.

Case $A_1 = \{s \mid s \ll \min(\theta^{-\frac{1}{|\tau-3|}}, n\theta^{\frac{\tau-1}{|\tau-3|}})\}$**:** Recall the high probability event $\mathcal{E}_n$ that $\mathcal{UI}_0(s) \cap C_1 = \varnothing$ is required to hold on A_1. (Here, we assume $\mathcal{E}_n$ occurs and hence work with a conditioned version of (A.18)). Since $\mathcal{E}$ only concerns the uniform seeds, and since E_3 and E_2 counting contributions of clusters avoiding the giant component C_1, the only term that needs to be changed in (A.18) is $q_{p,1}$, which needs to be changed to 0. For the other terms, by (A.11)-(A.15), we have $q_{c,1} = 1 - O(n^{-\log(n)})$, $E_1 = E_2 = 0$ and $E_3 = \theta^{-\mathbb{1}_{\tau\in(3,4)}}$, so using (A.18) and the values from (A.11)–(A.15)

$$f_{n,\tau}(\theta, s, \mathcal{E}) = \frac{q_{c,1}C_1 + o(1/n)}{\Theta(sE_3)} = \Theta\left(\frac{n\theta^{\frac{1}{|\tau-3|}}}{s\theta^{-\mathbb{1}_{\tau\in(3,4)}}}\right) = \Theta\left(\theta^{\frac{1}{|\tau-3|}+\mathbb{1}_{\tau\in(3,4)}} n s^{-1}\right). \tag{A.31}$$

In the normalized form, we get the linear relation $\tilde{f}_\tau(x, y, \mathcal{E}) = 1 + \left(\frac{1}{|\tau-3|} + \mathbb{1}_{\tau\in(3,4)}\right)x - y$. By (A.12),

$$\mathbb{P}(\mathcal{E}) = 1 - q_{p,1} = 1 - \Theta(s\theta^{\frac{1}{|\tau-3|}}) \to 1$$

also holds.

Case $A_2 = \{s \mid \theta^{-\frac{1}{|\tau-3|}} \ll s \ll n\theta^{\frac{\tau-1}{|\tau-3|}}\}$**:** By (A.11)-(A.15), in this case, $q_{c,1} = 1 - O(n^{-\log(n)})$, $q_{p,1} = 1 - O(n^{-\log(n)})$, $E_1 = E_2 = 0$ and $E_3 = \theta^{-\mathbb{1}_{\tau\in(3,4)}}$, which means that

$$f_{n,\tau}(\theta, s) = \frac{(1 - O(n^{-\log(n)}))C_1 + o(1/n)}{(1 - O(n^{-\log(n)}))C_1 + \Theta(s\theta^{-\mathbb{1}_{\tau\in(3,4)}})} = 1 - \Theta\left(\frac{s\theta^{-\mathbb{1}_{\tau\in(3,4)}}}{C_1}\right) = 1 - \Theta\left(\frac{\theta^{-\frac{1}{|\tau-3|}-\mathbb{1}_{\tau\in(3,4)}} s}{n}\right). \tag{A.32}$$

In the normalized form, we get $\tilde{f}_\tau(x, y) = \left(\frac{1}{|\tau-3|} + \mathbb{1}_{\tau\in(3,4)}\right)x - y$. The event $\mathcal{E}$ must occur in this case by definition, hence the results directly apply to $\tilde{f}_\tau(x, y, \mathcal{E})$ and $f_{n,\tau}(\theta, s, \mathcal{E})$ as well.

We note that this is the only case in which the uniform seed set infects more nodes than the highest degree nodes. As opposed to the other cases where we only had asymptotic results for $f_{n,\tau}(\theta, s)$, in this case we compute that the main order of the ratio is 1, and even the asymptotics of the deviation from this main order. We can make such precise calculations only because both the numerator and the denominator of $f_{n,\tau}(\theta, s)$ are dominated by the expected size of the giant component, and these terms cancel each other. The deviation from 1 then comes from the contribution of small clusters that the uniform seed set can infect. Intuitively, in this case a "disassortative" choice of seeds helps the infection to spread more.

Case $A_3 = \{s \mid n\theta_n^{\frac{\tau-1}{|\tau-3|}} \ll s_n \ll \theta_n^{-\frac{1}{|\tau-3|}}\}$: In this case, it is possible that event $\mathcal{E}$ does not occur (with some probability tending to 0), and we have to work with a conditioned version of (A.18). Since $\mathcal{E}$ only concerns the uniform seed set, and since E_3 and E_2 are conditioned on an event that implies $\mathcal{E}$, the only term that needs to be changed in (A.18) is $q_{p,1}$, which needs to be changed to 0. For the other terms, by (A.11)-(A.15), we have $q_{c,1} = 1 - O(n^{-\log(n)})$, $E_1 = n^{1-\frac{1}{\tau-1}} s^{\frac{1}{\tau-1}}$, $E_2 = 0$ and $E_3 = \theta^{-\mathbb{1}_{\tau\in(3,4)}}$, which means that

$$f_{n,\tau}(\theta, s, \mathcal{E}) = \Theta\left(\frac{n^{1-\frac{1}{\tau-1}} s^{\frac{1}{\tau-1}}}{s\theta^{-\mathbb{1}_{\tau\in(3,4)}}}\right) = \Theta\left(\theta^{\mathbb{1}_{\tau\in(3,4)}} \left(\frac{n}{s}\right)^{1-\frac{1}{\tau-1}}\right). \tag{A.33}$$

In the normalized form, we get the linear relation $\tilde{f}_\tau(x, y, \mathcal{E}) = \mathbb{1}_{\tau\in(3,4)} x + \left(1 - \frac{1}{\tau-1}\right)(1 - y)$. By (A.12),

$$\mathbb{P}(\mathcal{E}) = 1 - q_{p,1} = 1 - \Theta(s\theta^{\frac{1}{|\tau-3|}}) \to 1$$

also holds.

Case $A_4 = \{s \mid \max(\theta_n^{-\frac{1}{|\tau-3|}}, n\theta_n^{\frac{\tau-1}{|\tau-3|}}) \ll s_n \ll \min(n\theta_n^{\frac{\tau-2}{|\tau-3|}}, n\theta_n^{\frac{1}{|\tau-3|}})\}$**:** By (A.11)-(A.15), in this case, $q_{c,1} = 1 - O(n^{-\log(n)})$, $q_{p,1} = 1 - O(n^{-\log(n)})$, $E_1 = n^{1-\frac{1}{\tau-1}} s^{\frac{1}{\tau-1}}$ $E_2 = 0$ and $E_3 = \theta^{-\mathbb{1}_{\tau\in(3,4)}}$, which means that

$$f_{n,\tau}(\theta, s) = \Theta\left(\frac{n^{1-\frac{1}{\tau-1}} s^{\frac{1}{\tau-1}}}{n\theta^{\frac{1}{|\tau-3|}} + s\theta^{-\mathbb{1}_{\tau\in(3,4)}}}\right) = \Theta\left(n^{-\frac{1}{\tau-1}} \theta^{-\frac{1}{|\tau-3|}} s^{\frac{1}{\tau-1}}\right) \tag{A.34}$$

because $n\theta^{\frac{1}{|\tau-3|}} \gg s\theta^{-\mathbb{1}_{\tau\in(3,4)}}$ holds due to $s \ll \min(n\theta_n^{\frac{\tau-2}{|\tau-3|}}, n\theta_n^{\frac{1}{|\tau-3|}})$ in the definition of A_4. In the normalized form, we get the linear relation $\tilde{f}_\tau(x, y) = -\frac{1}{|\tau-3|} x - \frac{1}{\tau-1}(1 - y)$. The event $\mathcal{E}$ must occur in this case by definition, hence the results directly apply to $\tilde{f}_\tau(x, y, \mathcal{E})$ and $f_{n,\tau}(\theta, s, \mathcal{E})$ as well.

Case $A_5 = \{s \mid n\theta^{\frac{\tau-2}{|\tau-3|}} \ll s\}$: By (A.11)-(A.15), in this case, $q_{c,1} = 1 - O(n^{-\log(n)})$, $q_{p,1} = 1 - O(n^{-\log(n)})$, $E_1 = n^{1-\frac{1}{\tau-1}} s^{\frac{1}{\tau-1}}$, $E_2 = n^{\frac{\tau-3}{\tau-2}} s^{\frac{1}{\tau-2}}$ and $sE_3 \le E_2$, which means that

$$f_{n,\tau}(\theta, s) = \Theta\left(\frac{n^{1-\frac{1}{\tau-1}} s^{\frac{1}{\tau-1}}}{n\theta^{\frac{1}{|\tau-3|}} + n^{\frac{\tau-3}{\tau-2}} s^{\frac{1}{\tau-2}}}\right) = \Theta\left(\left(\frac{n}{s}\right)^{\frac{1}{(\tau-1)(\tau-2)}}\right) \tag{A.35}$$

because $n\theta^{\frac{1}{|\tau-3|}} \ll n^{\frac{\tau-3}{\tau-2}} s^{\frac{1}{\tau-2}}$ holds due to $n\theta^{\frac{\tau-2}{|\tau-3|}} \ll s$ in the definition of A_5. In the normalized form, we get the linear relation $\tilde{f}_\tau(x, y) = \frac{1}{(\tau-1)(\tau-2)}(1 - y)$. The event $\mathcal{E}$ must occur in this case by definition, hence the results directly apply to $\tilde{f}_\tau(x, y, \mathcal{E})$ and $f_{n,\tau}(\theta, s, \mathcal{E})$ as well.

Case $A_6 = \{s \mid \max(\theta_n^{-\frac{1}{|\tau-3|}}, n\theta_n^{\frac{1}{|\tau-3|}}) \ll s_n \ll n\theta_n^{\frac{\tau-2}{|\tau-3|}})\}$: In this case (which occurs only for $\tau \in (2,3)$), by (A.11)-(A.15), we have $q_{c,1} = 1 - O(n^{-\log(n)})$, $q_{p,1} = 1 - O(n^{-\log(n)})$, $E_1 = n^{1-\frac{1}{\tau-1}} s^{\frac{1}{\tau-1}}$ $E_2 = 0$ and $E_3 = \Theta(1)$, which means that

$$f_{n,\tau}(\theta, s) = \Theta\left(\frac{n^{1-\frac{1}{\tau-1}} s^{\frac{1}{\tau-1}}}{n\theta^{\frac{1}{|\tau-3|}} + s}\right) = \Theta\left(\left(\frac{n}{s}\right)^{1-\frac{1}{\tau-1}}\right) \tag{A.36}$$

parameter region	slightly subcritical	critical window	slightly supercritical
$\mathrm{sign}(p_n - p_c)$	$-$	$-,0,+$	$+$
$\theta_n = \lvert p_n - p_c \rvert$	$1 \gg \theta_n \gg n^{-\sigma/(\lambda-1)}$	$n^{-\sigma/(\lambda-1)} \gg \theta_n$	$1 \gg \theta_n \gg n^{-\sigma/(\lambda-1)}$
C_1	$\theta_n^{-1/\sigma}$	$n^{1/(\lambda-1)}$	$n\theta_n^{\beta}$
C_2	$\theta_n^{-1/\sigma}$	$n^{1/(\lambda-1)}$	$\theta_n^{-1/\sigma}$
$\mathbb{P}_{i\sim\mathcal{U}(\{1,\ldots,n_{c,p}\})}(\mathcal{C}_i = k \mid u \neq 1)$	$k^{-\lambda}(\mathrm{e})^{-k/\theta_n^{-1/\sigma}}$		
$\mathbb{P}_{u\sim\mathcal{U}(V)}(\mathcal{C}(u) = k \mid \mathcal{C}(u) \neq \mathcal{C}_1)$	$k^{-(\lambda-1)}(\mathrm{e})^{-k/\theta_n^{-1/\sigma}}$		

parameter region	slightly subcritical	critical window	slightly supercritical
$\mathrm{sign}(p_n - p_c)$	$-$	$-,0,+$	$+$
$\theta_n = \lvert p_n - p_c \rvert$	$1 \gg \theta_n \gg n^{-\frac{\lvert\tau-3\rvert}{\tau-1}}$	$n^{-\frac{\lvert\tau-3\rvert}{\tau-1}} \gg \theta_n$	$1 \gg \theta_n \gg n^{-\frac{\lvert\tau-3\rvert}{\tau-1}}$
C_1	$\theta_n^{-\frac{\tau-2}{\lvert\tau-3\rvert}}$	$n^{\frac{\tau-2}{\tau-1}}$	$n\theta_n^{\frac{1}{\lvert\tau-3\rvert}}$
C_2	$\theta_n^{-\frac{\tau-2}{\lvert\tau-3\rvert}}$	$n^{\frac{\tau-2}{\tau-1}}$	$\theta_n^{-\frac{\tau-2}{\lvert\tau-3\rvert}}$
$\mathbb{P}_{i\sim\mathcal{U}(\{1,\ldots,n_{c,p}\})}(\mathcal{C}_i = k \mid u \neq 1)$	$k^{-\frac{2\tau-3}{\tau-2}}(\mathrm{e})^{-\frac{k}{C_1}}$		
$\mathbb{P}_{u\sim\mathcal{U}(V)}(\mathcal{C}(u) = k \mid \mathcal{C}(u) \neq \mathcal{C}_1)$	$k^{-\frac{\tau-1}{\tau-2}}(\mathrm{e})^{-\frac{k}{C_1}}$		

Table A.2: Table A.1 reparametrized with only the degree exponent τ.

because $n\theta^{\frac{1}{\lvert\tau-3\rvert}} \ll s$ holds due to the definition of A_6. In the normalized form, we get the linear relation $\tilde{f}_\tau(x,y) = \left(1 - \frac{1}{\tau-1}\right)(1-y)$. The event $\mathcal{E}$ must occur in this case by definition, hence the results directly apply to $\tilde{f}_\tau(x,y,\mathcal{E})$ and $f_{n,\tau}(\theta,s,\mathcal{E})$ as well.

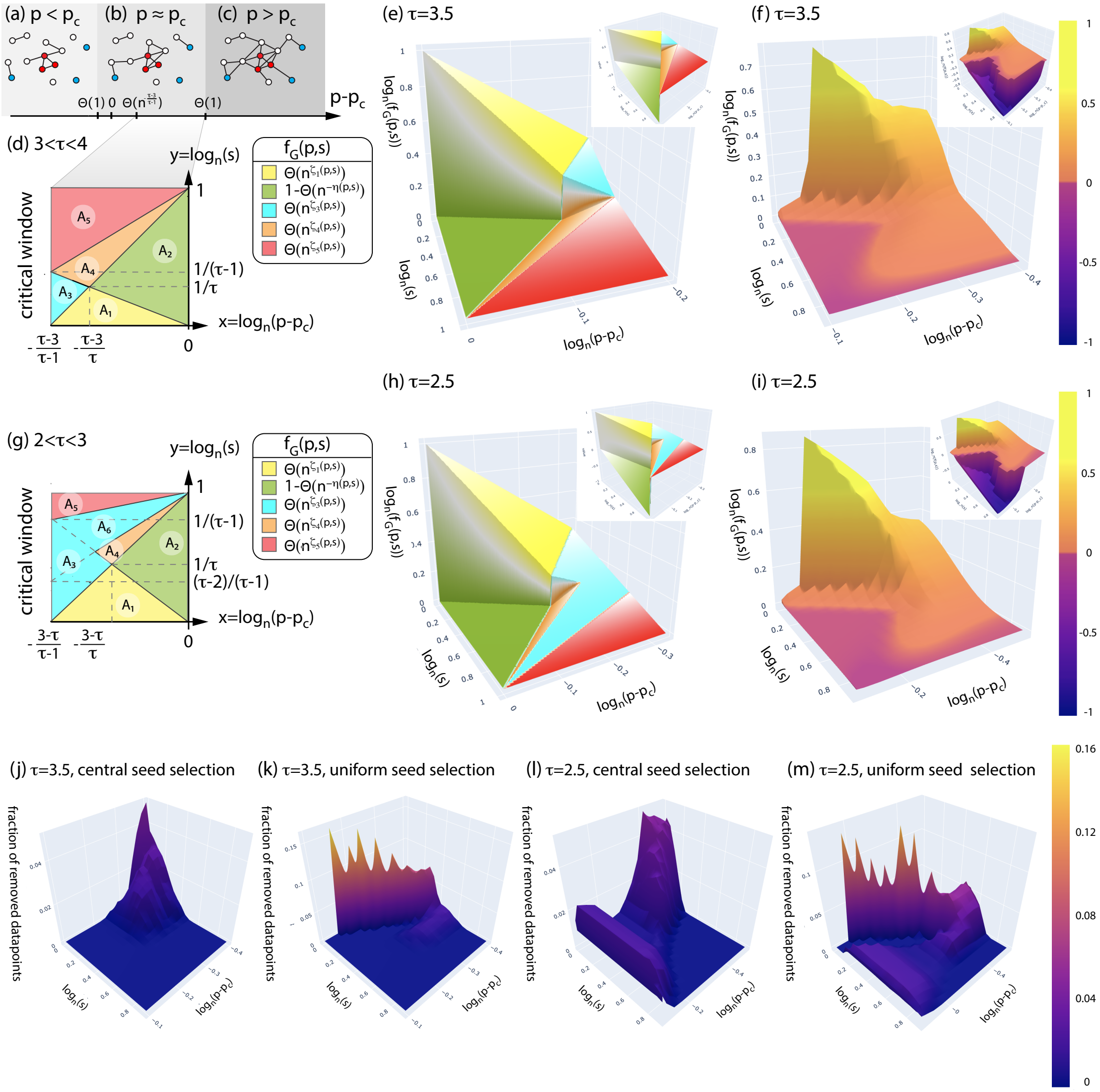
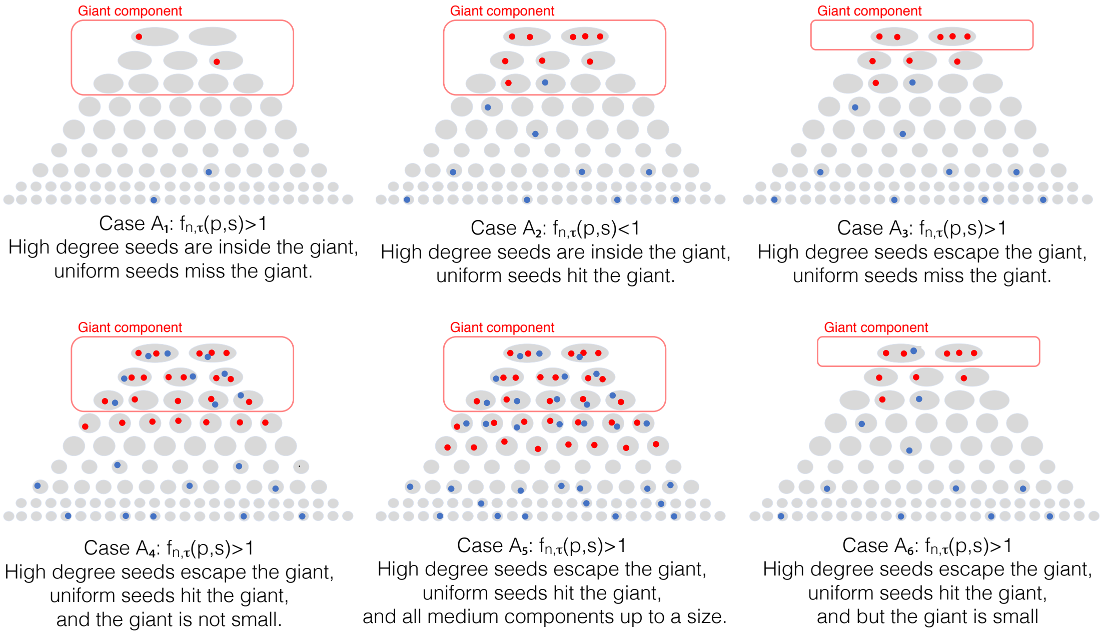


Figure A.1: Subfigures (a), (b) and (c) show the heuristic explanation of the switchover of the pandemic size ratio function f_G. Subfigures (d) and (g) show the phase diagram of the function f_G for $3 < \tau < 4$ and $2 < \tau < 3$, respectively, for values of p slightly above the percolation threshold and for various values of s. The asymptotics of f_G is different in the differenly colored parameter regions, which correspond to A_1-A_6 as given in Definition A.2.2. Subfigures (e) and (h) show the 3D plot of $\bar{f}_{3.5}$ and $\bar{f}_{2.5}$, the limit function $\log_n(f_G)$ for $\tau = 3.5$ and $\tau = 2.5$, respectively, as the number of nodes in G tends to infinity, and subfigures (f) and (i) show the corresponding simulation results on configuration model networks with $n = 10^7$ nodes. Each datapoint is an average of 1000 independent percolation instances on 10 independent random networks, after outlier removal is performed as explained in Remark A.2.2. The coloring on subfigures (e) and (h) follow the coloring on the phase diagram on subfigures (d) and (g), respectively. Since in the configuration model we only have weak switchover, the (green) part of the surface $\bar{f}_\tau$, which corresponds to $f_G < 1$, converges to 0. For a visualization of the precise deviation of $\bar{f}_\tau$ below 0, in the inset of subfigures (e) and (h) we plot the function $\tilde{f}_\tau$,and in the inset of subfigures (f) and (i) we plot the function $\mathrm{dNorm}(f_G)$. Subfigures (j)-(m) show the fraction of datapoints removed by the outlier detection when esimating $\mathbb{E}_p[\mathbf{Cl}(\mathcal{CI}_0(s)) \mid \mathcal{E}]$ (central seed selection) and $\mathbb{E}_p[\mathbf{Cl}(\mathcal{UI}_0(s)) \mid \mathcal{E}]$ (uniform seed selection) for $\tau = 3.5$ and $\tau = 2.5$.



Case A_1: $f_{n,\tau}(p,s)>1$
High degree seeds are inside the giant,
uniform seeds miss the giant.

Case A_2: $f_{n,\tau}(p,s)<1$
High degree seeds are inside the giant,
uniform seeds hit the giant.

Case A_3: $f_{n,\tau}(p,s)>1$
High degree seeds escape the giant,
uniform seeds miss the giant.

Case A_4: $f_{n,\tau}(p,s)>1$
High degree seeds escape the giant,
uniform seeds hit the giant,
and the giant is not small.

Case A_5: $f_{n,\tau}(p,s)>1$
High degree seeds escape the giant,
uniform seeds hit the giant,
and all medium components up to a size.

Case A_6: $f_{n,\tau}(p,s)>1$
High degree seeds escape the giant,
uniform seeds hit the giant,
and but the giant is small

Figure A.2: Illustration for the derivation of Claim A.2.2. Each subfigure shows a schematic of each of the 6 parameter regions defined by A_1 - A_6. The grey areas represent connected clusters in the percolated graph G^p, the red circles mark the s highest degree nodes and the blue circles mark s uniformly randomly chosen nodes.

τ	$\tau \in (3,4)$					
(s,θ)	A_1	A_2	A_3	A_4		A_5
$q_{c,1}$	$1-O(n^{-\log(n)})$					
$C_{\min}^{(c)}$	C_1		$(n/s)^{\frac{\tau-2}{\tau-1}}$			
E_1	0		$n^{1-\frac{1}{\tau-1}}s^{\frac{1}{\tau-1}}$			
$q_{p,1}$	$\Theta(s\theta^{\frac{1}{\lvert\tau-3\rvert}})$	$1-O(n^{-\log(n)})$	$\Theta(s\theta^{\frac{1}{\lvert\tau-3\rvert}})$	$1-O(n^{-\log(n)})$		
$C_{\min}^{(u)}$	$\varnothing$	C_1	$\varnothing$	C_1		n/s
E_2	0					$n^{\frac{\tau-3}{\tau-2}}s^{\frac{1}{\tau-2}}$
sE_3	$s\theta^{-1}$	$s\theta^{-1}$	$s\theta^{-1}$	$s\theta^{-1}$		$s(n/s)^{\frac{\tau-3}{\tau-2}}$
$f_{n,\tau}(\theta,s,\mathcal{E})$	$\theta^{\frac{\tau-2}{\tau-3}}n/s$	$1-\theta^{-\frac{\tau-2}{\tau-3}}s/n$	$\theta(n/s)^{1-\frac{1}{\tau-1}}$	$\theta^{-\frac{1}{\lvert\tau-3\rvert}}(n/s)^{-\frac{1}{\tau-1}}$		$(n/s)^{\frac{1}{(\tau-1)(\tau-2)}}$
$\tilde{f}_\tau(\tilde{x},\tilde{y},\mathcal{E})$	$1+\frac{\tau-2}{\tau-3}\tilde{x}-\tilde{y}$	$\frac{\tau-2}{\tau-3}\tilde{x}-\tilde{y}$	$\tilde{x}+\left(1-\frac{1}{\tau-1}\right)(1-\tilde{y})$	$-\frac{1}{\lvert\tau-3\rvert}\tilde{x}-\frac{1}{\tau-1}(1-\tilde{y})$		$\frac{1}{(\tau-1)(\tau-2)}(1-\tilde{y})$
τ	$\tau \in (2,3)$					
(s,θ)	A_1	A_2	A_3	A_6	A_4	A_5
$q_{c,1}$	$\Omega(1)$					
$C_{\min}^{(c)}$	C_1		$(n/s)^{\frac{\tau-2}{\tau-1}}$			
E_1	0		$n^{1-\frac{1}{\tau-1}}s^{\frac{1}{\tau-1}}$			
$q_{p,1}$	$\Theta(s\theta^{\frac{1}{\lvert\tau-3\rvert}})$	$1-O(n^{-\log(n)})$	$\Theta(s\theta^{\frac{1}{\lvert\tau-3\rvert}})$	$1-O(n^{-\log(n)})$		
$C_{\min}^{(u)}$	$\varnothing$	C_1	$\varnothing$	C_1	C_1	n/s
E_2	0					$n^{\frac{\tau-3}{\tau-2}}s^{\frac{1}{\tau-2}}$
sE_3	s	s	s		s	
$f_{n,\tau}(\theta,s,\mathcal{E})$	$\theta^{\frac{1}{3-\tau}}n/s$	$1-\theta^{-\frac{1}{3-\tau}}s/n$	$(n/s)^{1-\frac{1}{\tau-1}}$		$(n/s)^{-\frac{1}{\tau-1}}\theta^{-\frac{1}{3-\tau}}$	$(n/s)^{\frac{1}{(\tau-1)(\tau-2)}}$
$\tilde{f}_\tau(\tilde{x},\tilde{y},\mathcal{E})$	$1+\frac{1}{3-\tau}\tilde{x}-\tilde{y}$	$\frac{1}{3-\tau}\tilde{x}-\tilde{y}$	$\left(1-\frac{1}{\tau-1}\right)(1-\tilde{y})$		$-\frac{1}{3-\tau}\tilde{x}-\frac{1}{\tau-1}(1-\tilde{y})$	$\frac{1}{(\tau-1)(\tau-2)}(1-\tilde{y})$

Table A.3: Summary of the proof of Theorem A.2.2. See definitions for the notation in Table A.4. The colors of the columns A_1-A_6 are chosen to match Figure A.1. Dark grey signifies the leading term of the numerator, light grey signifies the leading term of the denominator of $f_{n,\tau}(p,s)$. For all rows (except for $q_{c,1}$ and $q_{p,1}$, where the O and Θ notation is made explicit) the values in the cells represent asymptotic values.

notation	definition/meaning
Network models	
$G = (V, E)$	graph with node set V and edge set E
$[d_i]_n$	expected degrees of the Configuration model
v_i	node with the i^{th} largest (expected) degree
τ	exponent of the power-law degree distribution
$\mathrm{CM}(n, \tau)$	Configuration model with size n and degree exponent τ
Bond percolation	
p	bond percolation retention probability
G^p	the bond percolated graph acquired by deleting each edge of G with probability $1-p$
$p_c, p_{c,n,\tau}$	critical point of bond percolation in general, and for the Configuration model $\mathrm{CM}(n,\tau)$
θ	$\lvert p - p_c\rvert$, deviation from critical point
$n_{c,p}$	number of connected components of G^p
$\mathcal{C}_i$	a set valued random variable that equals the i^{th} largest component of G^p
$\mathcal{C}(u)$	a set valued random variable that equals the component in G^p which contains node u
C_i	$\mathbb{E}[\lvert\mathcal{C}_i\rvert]$
Seed selection strategies	
s	the size of the seed set of an epidemic process
$\mathcal{CI}_0(s)$	central area seed set of size s, the s highest degree nodes
$\mathcal{UI}_0(s)$	uniform seed set of size s
$\mathbb{E}_p[\mathbf{Cl}(\mathcal{CI}_0(s))]$	expected cluster size of $\mathcal{CI}_0(s)$ in G^p
$\mathbb{E}_p[\mathbf{Cl}(\mathcal{UI}_0(s))]$	expected cluster size of $\mathcal{UI}_0(s)$ in G^p
$f_G(p,s)$	$\frac{\mathbb{E}_p[\mathbf{Cl}(\mathcal{CI}_0(s))]}{\mathbb{E}_p[\mathbf{Cl}(\mathcal{UI}_0(s))]}$, expected final infection size ratio
Variables defined for the Configuration model	
λ, σ, β	standard constants (depending on τ) defined in (A.7), (A.8) and (A.9)
$f_{n,\tau}(\theta, s)$	$f_{\mathrm{CM}(n,\tau)}(p,s)$, expected final infection size ratio in the Configuration model
dNorm_n	$\log_n(x)$ if $x > 1, -\log_n(1-x) - 1$ if $x < 1$
$\tilde{x}, \tilde{y}, \tilde{f}_\tau(\tilde{x}, \tilde{y})$	$\lim\limits_{n\to\infty} \log_n(\theta_n), \lim\limits_{n\to\infty} \log_n(s_n), \lim\limits_{n\to\infty} \mathrm{dNorm}_n(f_{n,\tau}(\theta, s))$
$\mathcal{E}$	the event that either $s \in A_2 \cup A_4 \cup A_5 \cup A_6$ or $s \in A_1 \cup A_3$ and $\mathcal{UI}_0(s) \cap \mathcal{C}_1 = \varnothing$ both hold
$\tilde{f}_\tau(\tilde{x}, \tilde{y}, \mathcal{E}), f_{n,\tau}(\theta, s, \mathcal{E})$	the same as $\tilde{f}_\tau(\tilde{x}, \tilde{y}), f_{n,\tau}(\theta, s)$ but conditioned on $\mathcal{E}$
Variables defined in the analytic derivation	
$q_{c,1}$	$\mathbb{P}(\lvert\mathcal{CI}_0(s) \cap \mathcal{C}_1\rvert > 0)$
$q_{p,1}$	$\mathbb{P}(\lvert\mathcal{UI}_0(s) \cap \mathcal{C}_1\rvert > 0)$
$C^{(c)}_{\min}$	$\min\{C_u \mid \mathbb{P}(\lvert\mathcal{CI}_0(s) \cap \mathcal{C}_u\rvert = \Theta(1))\}$
$C^{(u)}_{\min}$	$\min\{C_u \mid \mathbb{P}(\lvert\mathcal{UI}_0(s) \cap \mathcal{C}_u\rvert = \Theta(1))\}$
E_1	$\mathbb{E}\left[\sum_{i=1}^{n_{c,p}} \lvert\mathcal{C}_i\rvert \mathbb{1}_{\{C_1 > \lvert\mathcal{C}_i\rvert > C^{(c)}_{\min}\}}\right]$
E_2	$\mathbb{E}\left[\sum_{i=1}^{n_{c,p}} \lvert\mathcal{C}_i\rvert \mathbb{1}_{\{C_1 > \lvert\mathcal{C}_i\rvert > C^{(u)}_{\min}\}}\right]$
E_3	$\mathbb{E}_{u\sim\mathcal{U}(V)}\left[\lvert\mathcal{C}(u)\rvert \mathbb{1}_{\{\lvert\mathcal{C}(u)\rvert < \min(C^{(u)}_{\min}, C_1)\}}\right]$
A_1	$\{s \mid s \ll \min(\theta^{-\frac{1}{\lvert\tau-3\rvert}}, n\theta^{\frac{\tau-1}{\lvert\tau-3\rvert}})\}$
A_2	$\{s \mid \theta^{-\frac{1}{\lvert\tau-3\rvert}} \ll s \ll n\theta^{\frac{\tau-1}{\lvert\tau-3\rvert}}\}$
A_3	$\{s \mid n\theta_n^{\frac{\tau-1}{\lvert\tau-3\rvert}} \ll s_n \ll \theta_n^{-\frac{1}{\lvert\tau-3\rvert}}\} \cup \{s \mid \max(\theta_n^{-\frac{1}{\lvert\tau-3\rvert}}, n\theta_n^{\frac{\tau-2}{\lvert\tau-3\rvert}}) \ll s_n \ll n\theta_n^{\frac{1}{\lvert\tau-3\rvert}})\}$
A_4	$\{s \mid \max(\theta_n^{-\frac{1}{\lvert\tau-3\rvert}}, n\theta_n^{\frac{\tau-1}{\lvert\tau-3\rvert}}) \ll s_n \ll \min(n\theta_n^{\frac{\tau-2}{\lvert\tau-3\rvert}}, n\theta_n^{\frac{1}{\lvert\tau-3\rvert}})\}$
A_5	$\{s \mid n\theta^{\frac{\tau-2}{\lvert\tau-3\rvert}} \ll s\}$
A_6	$\{s \mid \max(\theta_n^{-\frac{1}{\lvert\tau-3\rvert}}, n\theta_n^{\frac{1}{\lvert\tau-3\rvert}}) \ll s_n \ll n\theta_n^{\frac{\tau-2}{\lvert\tau-3\rvert}})\}$

Table A.4: Definitions and glossary of notation

B Appendix for Chapter 5

B.1 Additional Proofs

B.1.1 Proof of Lemma 5.3.1

Let $T(n)$ denote the number of steps in which MAX-GAIN reduces the number of candidates from n to 1, and let C_N be the value (not depending on n) such that for all $n \geq C_N > 0$ the condition

$$T(n) \leq T(nq + f(n)) + 1 \text{ with } f(n) = o\left(\frac{n}{\log(n)}\right) \tag{B.1}$$

holds. Then we prove

$$T(n) < \log_{\frac{1}{q}}(n) + \log\log(n) + C_{q,f} + C_N, \tag{B.2}$$

where $C_{q,f}$ is a positive constant (it depends only on q and f but not n) computed implicitly at the end of the proof.

Proof by induction. Base case: if $n < C_{q,f} + C_N$ then $T(n) < C_{q,f} + C_N$ clearly holds as we can query each candidate. Induction step: Let now $n \geq C_{q,f} + C_N$ and we assume that for $M < n$ the induction hypothesis holds, that is

$$T(M) < \log_{\frac{1}{q}}(M) + \log\log(M) + C_{q,f} + C_N. \tag{B.3}$$

Then,

$$\begin{aligned} T(n) &\leq T(nq + f(n)) + 1 \\ &\overset{(B.3)}{\leq} \log_{\frac{1}{q}}(nq + f(n)) + \log(\log(nq + f(n))) + C_{q,f} + C_N + 1 \end{aligned} \tag{B.4}$$

For the induction hypothesis to hold we would like the last expression to be upper bounded by

$$\log_{\frac{1}{q}}(n) + \log(\log(n)) + C_{q,f} + C_N.$$

To compare these two quantities, we would like to transform $\log_{\frac{1}{q}}(nq + f(n))$. Using the fact that log is a concave function and by linearly approximating it at n,

$$\begin{aligned} \log_{\frac{1}{q}}(nq + f(n)) &= \log_{\frac{1}{q}}(n + \frac{f(n)}{q}) - 1 \\ &\leq \log_{\frac{1}{q}}(n) + \frac{f(n)}{q\log(\frac{1}{q})n} - 1 \end{aligned} \tag{B.5}$$

Plugging this into (B.4) we get

$$T(n) \leq \log_{\frac{1}{q}}(n) + \frac{f(n)}{q\log(\frac{1}{q})n} + \log(\log(nq + f(n))) + C_{q,f} + C_N \tag{B.6}$$

For the induction hypothesis to hold we need to show

$$\begin{aligned} \log_{\frac{1}{q}}(n) + \frac{f(n)}{q\log(\frac{1}{q})n} + \log(\log(nq + f(n))) + C_{q,f} + C_N \\ \leq \log_{\frac{1}{q}}(n) + \log(\log(n)) + C_{q,f} + C_N, \end{aligned} \tag{B.7}$$

which is equivalent to

$$\frac{f(n)}{q\log(\frac{1}{q})n} + \log(\log(nq + f(n))) \leq \log(\log(n)) \tag{B.8}$$

Again, by the concavity of $\log(\log(n))$ we can use a linear approximation

$$\log(\log(nq + f(n))) \leq \log(\log(n)) + \frac{n - (nq + f(n))}{n\log(n)} \tag{B.9}$$

So it is enough to show

$$\begin{aligned}
\frac{f(n)}{q\log(\frac{1}{q})n} &\leq \frac{n-(nq+f(n))}{n\log(n)} \\
\frac{f(n)}{n}\left(\frac{1}{\log(\frac{1}{q})q}+\frac{1}{log(n)}\right) &\leq \frac{1-q}{\log(n)} \\
\frac{f(n)\log(n)}{n} &\leq \left(\frac{1-q}{\log(\frac{1}{q})q}+\frac{1-q}{log(n)}\right)^{-1} \qquad \text{(B.10)}
\end{aligned}$$

Since the right hand side is bounded from below (for $n > 0$) and $f(n) = o\left(\frac{n}{\log(n)}\right)$, this last inequality must hold for $n \geq C_{q,f}$, for some constant $C_{q,f}$ (depending only on q and f but not n).

To conclude the proof, we showed that for all $n \in \mathbb{N}$

$$T(n) < \log_{\frac{1}{q}}(n) + \log\log(n) + C_{q,f} + C_N. \qquad \text{(B.11)}$$

This in particular implies

$$T(N) < \log_{\frac{1}{q}}(N) + \log\log(N) + C_{q,f} + C_N = (1+o(1))\log_{\frac{1}{q}}(N) \qquad \text{(B.12)}$$

for $C_N = o\left(\log_{\frac{1}{q}}(N)\right)$.

B.1.2 Proof of Lemma 5.4.1

The proof follows the proof of Lemma 2 (i) in [36] until the very last step, the evaluation of the multiplicative error term. There, the authors use $i = O(\log(n)/\log\log(n))$ and $\sqrt{\omega} \leq \log^2(N)\log\log(N)$ to get the asymptotic upper bound

$$\begin{aligned}
&\left(1+O\left(\frac{\delta}{N}\right)+O\left(\frac{1}{\sqrt{\omega}}\right)\right)\prod_{j=2}^{i}\left(1+O\left(\frac{\delta^j}{N}\right)+O\left(\frac{1}{\sqrt{\omega d^{j-1}}}\right)\right) \\
&\quad =\left(1+O\left(\frac{\delta^i}{N}\right)+O\left(\frac{1}{\sqrt{\omega}}\right)\right)\prod_{j=7}^{i-3}(1+O(\log^{-3}(N))) \\
&\quad =\left(1+O\left(\frac{\delta^i}{N}\right)+O\left(\frac{1}{\sqrt{\omega}}\right)\right)(1+O(\log^{-2}(N))) \\
&\quad =\left(1+O\left(\frac{\delta^i}{N}\right)+O\left(\frac{1}{\sqrt{\omega}}\right)\right)
\end{aligned}$$

However, the second upper bound on $\sqrt{\omega}$ is not necessary. Instead, we can write

$$\begin{aligned}
&\left(1+O\left(\frac{\delta}{N}\right)+O\left(\frac{1}{\sqrt{\omega}}\right)\right)\prod_{j=2}^{i}\left(1+O\left(\frac{\delta^j}{N}\right)+O\left(\frac{1}{\sqrt{\omega d^{j-1}}}\right)\right)\\
&\qquad=\left(1+O\left(\frac{\delta^i}{N}\right)+O\left(\frac{1}{\sqrt{\omega}}\right)\right)\prod_{j=5}^{i-2}\left(1+O\left(\frac{1}{\delta^2}\right)\right)\\
&\qquad=\left(1+O\left(\frac{\delta^i}{N}\right)+O\left(\frac{1}{\sqrt{\omega}}\right)\right)\left(1+O\left(\frac{1}{\delta}\right)\right)\\
&\qquad=\left(1+O\left(\frac{\delta^i}{N}\right)+O\left(\frac{1}{\sqrt{\omega}}\right)\right)
\end{aligned}$$

where the second to last inequality holds because $i < \delta$ and $(1+O(1/\delta^2))^{\delta} = (1+O(1/\delta))$, and the last inequality holds because $1/\delta = O(\delta^i/N)$.

The condition $\sqrt{\omega} \le \log^2(N)\log\log(N)$ is not used anywhere else in the proof of Lemma 2 (i) in [36], so we may remove this condition, which gives Lemma 5.4.1.

C Appendix for Chapter 6

C.1 Extending to Other Edge-Delay Distributions

Throughout Chapter 6, we assumed that the edge-delay distribution was Gaussian, however due to the Central Limit Theorem, it is natural to expect that our result generalizes to other distributions as well. However, there definitely are edge-delay distributions for which our result cannot generalize. Consider an edge-delay distribution $\mathcal{W}$ supported on two values: 1 and π. Since π is irrational, a single query node at one end of the path can determine the identity of the source with absolute certainty. Moreover, our results are not likely to generalize to heavy tailed $\mathcal{W}$ due to the lack of concentration in the answers.

We sketch how our proofs could be generalized to continuous sub-gaussian random variables. In the proofs of our main results, we exploit two types of properties of the edge-delay distribution; we are using the tight concentration of their sum in the non-adaptive upper bound and the adaptive upper and lower bounds, and we are using an anti-concentration result on their sum in the non-adaptive and adaptive lower bounds.

All of the concentration bounds are derived from Fact 6.3.2. This tail-bound result is easily extendable to sub-gaussian random variables (see Proposition 5.10 of [238]). The only difference in the results would be that σ would be replaced by the sub-gaussian norm

$$\|\mathcal{W}\|_{\psi_2} = \text{supp}_{p\geq 1} p^{-1/2} (\mathbb{E}|\mathcal{W}|^p)^{1/p}.$$

In the adaptive lower bound proof, when we make the anti-concentration arguments, our proof uses the density function of the Gaussian distribution. Therefore, we need that the density function of $\sum \mathcal{W}_i$ is pointwise close to the density function of the corresponding Gaussian distribution. Such statements are called local limit theorems for sums of independent random variables. In a sense, we are asking for much more than a tail-bound, but we can also be much looser than an exponential decay. In Lemma C.9.2 we need that the probability mass function of the posterior is bounded above by $\log(\mu_j)/(\sigma\sqrt{\mu_j})$. We prove this by writing the probability mass function (as a function of potential source v') explicitly using Bayes rule and the density of $\sum_{i=1}^{d'_l} \mathcal{W}_i$ and $\sum_{i=1}^{d'_r} \mathcal{W}_i$, where $d'_l = v' - l_j$ and $d'_r = r_j - v'$ are the distances between a node v' and the closest query nodes to the left and the right. We need these densities to be pointwise $o(\log(\mu_j)/(\sigma\sqrt{\mu_j}))$ close to the densities in the Gaussian case. Since in Claim C.9.1 we prove that $d'_l \in [1/2,2]a_{l_j}$ and $d'_r \in [1/2,2]a_{r_j}$, and by the definition $\mu_j = \min(a_{l_j}, a_{r_j})$, it is enough to show that the density of $\sum_{i=1}^{d'_l} \mathcal{W}_i$ is pointwise $o(\log(d'_l)/(\sigma\sqrt{d'_l}))$ close to the density in the Gaussian case (and we need the symmetric statement for d'_r). Such results are readily available for continuous distributions $\mathcal{W}$ with finite third moment (see Theorem 7.15 in [197]). We also point out, that similar results exist for discrete distributions $\mathcal{W}$ satisfying a certain lattice condition that can be used to rule out distributions like the one supported on 1 and π that we used as a counterexample in the beginning of the section (see Theorem 7.6 in [197]).

For the the anti-concentration result in the non-adaptive lower bound, we proved that the hypothesis testing problem cannot be solved between k neighboring nodes at distance d or more away from the query nodes. For this we upper bounded the union of the area under the probability density functions of the answers under each of the k hypotheses by another another function, which we could easily integrate. For general edge-delay distributions, again we aim to approximate the probability density functions of the answers by the the probability density function of Gaussian random variables. Since this time, instead of small l_∞ distance, we need small l_1 distance between the densities, a Berry-Esseen type theorem [30, 85] suffices instead of a local limit theorem.

We note that only the concentration arguments required the sub-gaussianity of the edge-delay distribution, the anti-concentration results held for a much more general class of distributions (finite third moment and continuity or lattice condition). We believe that with more advanced proof techniques the sub-gaussianity condition can also be relaxed.

C.2 The Difference Between S1 and S2

The only difference between two frameworks S1 and S2 defined in Section 1.2 is that the starting time of the epidemics is unknown in S1 and known in S2. We already mentioned that S2 is theoretically more appealing, and that there is little difference between the number of queries required in the two frameworks. The main consequence of the difference between the source identification algorithms in the two frameworks is that in S1, the answers that they can use are the relative differences between time measurements at different pairs of query nodes, whereas in S2 the answers they can use are the absolute differences between the (known) starting time of the epidemics and the time measurement at each query node. Since S1 is more restrictive than S2, our lower bounds on the number of required queries in S2 clearly also hold in S1. We comment on how the upper bounds can be extended in Remarks C.2.1 and C.2.2.

Additionally, we argue that while S2 has a simpler mathematical definition than S1, on the path network, proving lower bounds for S2 raises important challenges that would not have appeared in S1. Indeed, in the path network, the pair of query nodes that surround the source provide two independent answers about it in S2 (one from each direction between each query node and the source), but only one in S1 (because only the time difference between the measurements is meaningful). As a result, the analysis of the required number of queries is more challenging in S2 than in S1 because of the richer set of independent answers. Incorporating several independent measurements will be the main difficulty for the analysis of the number of queries needed to identify the source in more complex network models, such as bounded-degree trees. By focusing on S2 in the path network, our results therefore pave the way towards the analysis of more complex network models.

Remark C.2.1. *If the time of the first infection is not known (framework S1), we can model the answers by adding an unknown constant T_{start} to all of them. Then, Claim 6.4.1 (a) and (b) hold without modification and we can prove a version of (c) where the differences of the distances equal the differences of the answers rounded to the nearest integer (we just need a slightly tighter concentration result). Let us also define $q_{right} := q_{smallest} + d$, and a_{right} as the corresponding answer. Then, by Claim 6.4.1, if $\lfloor a_{smallest} - a_{left} \rceil = d$ then v^* is between $q_{smallest}$ and q_{right} and we can find v^* by computing*

$$\frac{\lfloor a_{smallest} - a_{right} \rceil + q_{smallest} + q_{right}}{2} = \frac{(v^* - q_{smallest}) - (q_{right} - v^*) + q_{smallest} + q_{right}}{2} = v^*$$

Otherwise, if $\lfloor a_{smallest} - a_{left} \rceil < d$ then v^* is between q_{left} and $q_{smallest}$, and v^* can be found analogously.

Remark C.2.2. *If the time of the first infection is not known (framework S1), and we model the answers by adding an unknown constant T_{start} to all of them, then a version of Lemma 6.5.1 shifted by T_{start} still holds. In this case, a slightly modified version of the algorithm finds the source. At each step the algorithm will query two nodes: one at l_i and r_i, with $l_0 = 1$ and $r_0 = n$. Then, we have a similar equation as* (6.9) *for the difference of the answers*

$$\mathrm{ans}_w(v^*, l_i) - \mathrm{ans}_w(v^*, r_i) \in (v^* - l_i) - (r_i - v^*) \pm 2C(\delta) \cdot \sigma \sqrt{d_i} \ln(1 + d_i),$$

where $d_i := r_i - l_i$, which means that we can keep track of a shrinking interval

$$\begin{cases} l_{i+1} = \max\left(l_i, \left\lceil \frac{\mathrm{ans}_w(v^*, l_i) - \mathrm{ans}_w(v^*, r_i) + l_i + r_i}{2} - C(\delta) \cdot \sigma \sqrt{d_i} \ln(1 + d_i) \right\rceil\right) \\ r_{i+1} = \min\left(r_i, \left\lfloor \frac{\mathrm{ans}_w(v^*, l_i) - \mathrm{ans}_w(v^*, r_i) + l_i + r_i}{2} + C(\delta) \cdot \sigma \sqrt{d_i} \ln(1 + d_i) \right\rfloor\right). \end{cases}$$

The rest of the proof can be written similarly to the case when the time of the first infection is known, and the only change in the final result is that we used twice as many queries to identify the source (which does not affect the asymptotic results).

C.3 Simulation Details for Figure 6.4

The simulation results were generated in the S1 source identification framework with the Python toolbox [220], which has been published in [224]. The underlying diffusion process was a Susceptible-Infected process (also called First Passage percolation) with uniform edge-delay distribution supported on the interval $[1-\sqrt{3}\sigma, 1+\sqrt{3}\sigma]$ (with mean 1 and standard deviation σ). Thereafter, a uniformly random query node was picked, and all further queries were selected by the Max-Gain algorithm as implemented in [220]. The algorithm was stopped when the candidate set reduced to a single node, which always had to be the source, since the Max-Gain algorithm always finds the correct source if enough queries are provided and the edge-delay distribution is bounded in some interval $[1-\epsilon, 1+\epsilon]$ for $\epsilon \in (0,1)$ (see Theorem 2 of [224]). The number of queries plotted in Figure 6.4 is the number of queries used by the Max-Gain until it was stopped, averaged over 192 simulations.

C.4 Proof of Claim 6.4.1

Claim 6.4.1. *For some $d = \Omega\left(\frac{1}{\sigma^2 \log(1/\delta)}\right)$, all of the following hold simultaneously with probability $\geq 1-\delta$:*

(a) among the query nodes located at or to the left of v^*, the closest one is the one with the smallest answer;

(b) among the query nodes located at or to the right of v^*, the closest one is the one with the smallest answer;

(c) the two closest query nodes to v^* on its left side and the closest query node on its right side all give a correct answer once rounded to the nearest integer.

Proof. In this proof, we will assume that

$$\sigma^2 \leq \frac{1}{16\ln(6/\delta)} \leq \frac{1}{2\ln(12/\delta)}. \tag{C.1}$$

If this is not the case, then $\sigma^2 = \Omega(1/\log(1/\delta))$, so we can simply query every node, which gives $d = 1 = \Omega\left(\frac{1}{\sigma^2 \log(1/\delta)}\right)$.

We need to choose d such that wherever v^* is identified, (a), (b), (c) simultaneously hold with probability $\geq 1-\delta$ over the choice of the weights $w(\cdot)$. Let us first study point (b) (point (a) is analogous). Let q be the closest query node at or to the right of v^*. Then (b) is true iff

- the sum of the weights of the edges between q and $q+d$ is positive;
- the sum of the weights of the edges between q and $q+2d$ is positive;

- …

- the sum of the weights of the edges between q and $1 + \lfloor \frac{n-1}{d} \rfloor d$ is positive.

More formally, (b) is true iff for all integers $i > 0$ such that $q + id \le n$, the sum of the weights of the edges between q and $q + id$ is positive.

A sufficient condition for this to hold is: for all positive integers x, the sum of the weights of the edges between nodes q and $q + x$ is positive. By Fact 6.3.1, each of these sums is distributed as a Gaussian $\mathcal{N}(x, x\sigma^2)$, so it is positive except with probability

$$\begin{aligned} \mathbb{P}_{X\sim\mathcal{N}(x,x\sigma^2)}[X < 0] &\le \mathbb{P}_{X\sim\mathcal{N}(x,x\sigma^2)}[X \notin x \pm x] \\ &\le \mathrm{e}^{-\frac{x^2}{2x\sigma^2}} && \text{(Fact 6.3.2)} \\ &= \mathrm{e}^{-\frac{x}{2\sigma^2}}. \end{aligned}$$

Therefore, by a union bound, (b) holds except with probability at most

$$\begin{aligned} \sum_{x=1}^{\infty} \left(\mathrm{e}^{-\frac{1}{2\sigma^2}}\right)^x &= \frac{\mathrm{e}^{-\frac{1}{2\sigma^2}}}{1 - \mathrm{e}^{-\frac{1}{2\sigma^2}}} \\ &< 3\mathrm{e}^{-\frac{1}{2\sigma^2}} && \text{(because } \sigma^2 \le 1 \Rightarrow \mathrm{e}^{-\frac{1}{\sigma^2}} < 2/3) \end{aligned}$$

which, assuming $\sigma^2 \le \frac{1}{2\ln(12/\delta)}$ (equation (C.1)), is at most $\delta/4$.

Finally, we study the probability that (c) holds. Let d_1, d_2, d_3 be the distances of those three query nodes to v^*. They are all at most $2d$ away from v^*. For $i = 1,2,3$, the corresponding answer is distributed as $X \sim \mathcal{N}(d_i, d_i\sigma^2)$, and is correct after rounding iff $X \in (d_i - 1/2, d_i + 1/2)$. Therefore, (c) holds except with probability

$$\begin{aligned} \sum_{i=1}^{3} \mathbb{P}_{X\sim\mathcal{N}(d_i,d_i\sigma^2)}[X \notin d_i \pm 1/2] &\le \sum_{i=1}^{3} \mathrm{e}^{-\frac{1}{8d_i\sigma^2}} && \text{(Fact 6.3.2)} \\ &\le 3\mathrm{e}^{-\frac{1}{16d\sigma^2}} \end{aligned}$$

which, assuming $d \le \frac{1}{16\sigma^2 \ln(6/\delta)}$, is at most $\delta/2$. Therefore, we set $d := \left\lfloor \frac{1}{16\sigma^2 \ln(6/\delta)} \right\rfloor$. By (C.1), $\frac{1}{16\sigma^2 \ln(6/\delta)} \ge 1$, so $d \ge (1/2) \cdot \frac{1}{16\sigma^2 \ln(6/\delta)} = \Omega\left(\frac{1}{\sigma^2 \log(1/\delta)}\right)$, as required.

Finally, by one more union bound, for our chosen value of d, all of (a), (b), (c) hold except with probability at most $\delta/4 + \delta/4 + \delta/2 = \delta$. □

C.5 Proof of Lemma 6.5.1

Lemma 6.5.1. *For any probability* $0 < \delta < 1/2$*, there is some constant* $C(\delta) = O\left(\sqrt{\log(1/\delta)}\right)$ *such that for any* n, σ *and any source* $v^* \in V$*, we have*

$$\mathbb{P}_w[\forall q \in V, \mathrm{ans}_w(v^*, q) \in |v^* - q| \pm C(\delta) \cdot \sigma \sqrt{|v^* - q|} \ln(1 + |v^* - q|)] \geq 1 - \delta. \tag{C.2}$$

Proof. We will use the quantity $|v^* - q|$ many times in this proof, so to simplify notation, let $d_q := |v^* - q|$. We will fix $C(\delta)$ later, but for the moment assume $C(\delta) \geq 2$ (this is clearly the case for the value we set it to in (C.4)).

First of all, for $q = v^*$, (C.2) holds trivially. For any other q, by Fact 6.3.1, $\mathrm{ans}_w(v^*, q)$ is distributed according to Gaussian $\mathcal{N}(d_q, d_q\sigma^2)$ (though not independently). Then, by Fact 6.3.2, for any $q \neq v^*$,

$$\begin{aligned}
\mathbb{P}[\mathrm{ans}_w(v^*, q) \notin d_q \pm C(\delta)\sigma\sqrt{d_q}\ln(1+d_q)] &\leq \mathrm{e}^{-\frac{(C(\delta)\sigma\sqrt{d_q}\ln(1+d_q))^2}{2d_q\sigma^2}} \\
&= \mathrm{e}^{-\frac{C(\delta)^2(\ln(1+d_q))^2}{2}} \\
&= \left(\frac{1}{1+d_q}\right)^{\frac{C(\delta)^2\ln(1+d_q)}{2}} \\
&= \left(\frac{1}{1+d_q}\right)^{\frac{(C(\delta)^2-1)\ln(1+d_q)+\ln(1+d_q)}{2}} \\
&= \left(\frac{1}{1+d_q}\right)^{\frac{(C(\delta)^2-1)\ln(1+d_q)}{2}} \left(\frac{1}{1+d_q}\right)^{\frac{\ln(1+d_q)}{2}} \\
&\leq \left(\frac{1}{1+d_q}\right)^{\frac{C(\delta)^2}{4}} \left(\frac{1}{1+d_q}\right)^{\frac{\ln(1+d_q)}{2}} \\
&\qquad (C(\delta) \geq 2 \text{ so } (C(\delta)^2-1)\ln(2) \geq C(\delta)^2/2) \\
&\leq \left(\frac{1}{2}\right)^{\frac{C(\delta)^2}{4}} \left(\frac{1}{1+d_q}\right)^{\frac{\ln(1+d_q)}{2}}.
\end{aligned}$$

By a union bound over all $q \neq v^*$, this implies that

$$\begin{aligned}
&\mathbb{P}[\exists\, q \in V, \mathrm{ans}_w(v^*, q) \notin d_q \pm C(\delta)\sigma\sqrt{d_q}\ln(1+d_q)] \\
&\quad\le 2\sum_{d=1}^{\infty}\left(\frac{1}{2}\right)^{\frac{C^2}{4}}\left(\frac{1}{1+d}\right)^{\frac{\ln(1+d)}{2}} \\
&\quad= 2\left(\frac{1}{2}\right)^{\frac{C^2}{4}}\sum_{d=1}^{\infty}\left(\frac{1}{1+d}\right)^{\frac{\ln(1+d)}{2}} \\
&\quad= 2\left(\frac{1}{2}\right)^{\frac{C^2}{4}}\left(\sum_{d=1}^{\lceil e^4+1\rceil-1}\left(\frac{1}{1+d}\right)^{\frac{\ln(1+d)}{2}} + \sum_{d=\lceil e^4+1\rceil}^{\infty}\left(\frac{1}{1+d}\right)^{\frac{\ln(1+d)}{2}}\right) \\
&\quad\le 2\left(\frac{1}{2}\right)^{\frac{C^2}{4}}\left(\lceil e^4+1\rceil - 1 + \sum_{d=\lceil e^4+1\rceil}^{\infty}\left(\frac{1}{1+d}\right)^{2}\right) \\
&\qquad\qquad (d \ge e^4 - 1 \text{ implies } \ln(1+d)/2 \ge 2) \\
&\quad\le 2\left(\frac{1}{2}\right)^{\frac{C^2}{4}}\left(\lceil e^4\rceil + \sum_{d=1}^{\infty}\frac{1}{d^2}\right) \\
&\quad= 2\left(\frac{1}{2}\right)^{\frac{C^2}{4}}\left(\lceil e^4\rceil + \frac{\pi^2}{6}\right) \\
&\quad\le 114\left(\frac{1}{2}\right)^{\frac{C^2}{4}} \qquad\qquad (C.3)
\end{aligned}$$

Setting

$$C(\delta) := \sqrt{4\log_2(114/\delta)} = O\left(\sqrt{\log(1/\delta)}\right), \qquad (C.4)$$

(C.3) becomes $\le \delta$, and we are done.

□

C.6 Proof of Claim 6.5.1

Claim 6.5.1. *Assume $d_0 \le n$, and $d_{i+1} \le C\cdot\sigma\sqrt{d_i}\ln(1+d_i)$ for some value $C > 0$. Then*

- *if $\sigma^2 \le 1/2$, there exists $k = O(\log\log_{1/\sigma} n)$ such that $d_k = \mathrm{poly}(C)$;*
- *if $\sigma^2 \ge 1/2$, there exists $k = O(\log\log n)$ such that $d_k = \sigma^2\cdot\mathrm{poly}(C, \log(1+\sigma^2))$.*

Proof. We track the value of d_i/σ^2 as i increases. First, as long as

$$d_i \ge \sigma^2(C\ln(1+d_i))^6 \qquad (C.5)$$

we have

$$\frac{d_{i+1}}{\sigma^2} \le \frac{C\sigma\sqrt{d_i}\ln(1+d_i)}{\sigma^2} = C\sqrt{\frac{d_i}{\sigma^2}}\ln(1+d_i) \overset{(C.5)}{\le} \left(\frac{d_i}{\sigma^2}\right)^{2/3}$$

(the last inequality can be deduced by dividing both sides by $\sqrt{\frac{d_i}{\sigma^2}}$ then raising both sides to the sixth power). Thus, by induction, as long as (C.5) holds, we have

$$\frac{d_i}{\sigma^2} \le \left(\frac{d_0}{\sigma^2}\right)^{(2/3)^i} \Rightarrow d_i \le \sigma^2\left(\frac{n}{\sigma^2}\right)^{(2/3)^i}.$$

Let $d_{\min}$ be the smallest value greater than $\max(2\sigma^2, 1)$ that we can assign to d_i such that (C.5) holds. Let k be the smallest integer for which $d_k \le d_{\min}$. Then we have

$$\sigma^2\left(\frac{n}{\sigma^2}\right)^{(2/3)^{k-1}} \ge d_{\min} \Leftrightarrow k \le 1 + \log_{3/2}\left(\frac{\log\left(\frac{n}{\sigma^2}\right)}{\log\left(\frac{d_{\min}}{\sigma^2}\right)}\right) \le 1 + \log_{3/2}\left(\frac{\log\left(\frac{n}{\sigma^2}\right)}{\log(\max(1/\sigma^2, 2))}\right).$$

- If $\sigma^2 \le 1/2$, then it is easy to verify that $d_{\min} = O((C\log C)^6) = \mathrm{poly}(C)$. Therefore, for $k = O(1 + \log(1 + \log_{1/\sigma} n))$, we have $d_k \le d_{\min} = \mathrm{poly}(C)$.
- If $\sigma^2 \ge 1/2$, then it is easy to verify that $d_{\min} = O(\sigma^2 C^6 \log(1+\sigma^2 C^6)^6) = \sigma^2\cdot\mathrm{poly}(C, \log(1+\sigma^2))$. Therefore, for $k = O(\log\log n)$, we have $d_k \le d_{\min} = \sigma^2 \cdot \mathrm{poly}(C, \log(1+\sigma^2))$. □

C.7 Proof of Lemma 6.5.2

Lemma 6.5.2. *For any probability* $\delta > 0$ *and any* $n \ge \Theta_\delta(\max(\sigma^2 \ln\sigma, 1))$, $\mathbb{P}_{v^*, w}[\mathrm{Typical}_\delta(v^*, w)] \ge 1 - \delta$.

Before proving Lemma 6.5.2, we first prove two claims.

Claim C.7.1. *For any probability* $\delta_1 > 0$, *there exists* $D_1(\delta_1) > 0$ *such that for any* n, σ *and any source* $v^* \in V$,

$$\mathbb{P}_w\left[\textit{for all } q \textit{ such that } d_q := |v^* - q| \ge D_1(\delta_1),\ \mathrm{ans}_w(v^*, q) \in d_q \pm \sigma\sqrt{d_q}\ln d_q\right] \ge 1 - \delta_1.$$

Proof. At first, let us consider only the case $q \ge v^*$. That is, consider node $q = v^* + d$ for some distance $d \ge \mathrm{e}^2$. Then $\mathrm{ans}_w(v^*, q) \sim \mathcal{N}(d, d\sigma^2)$, so

$$\begin{aligned}\mathbb{P}_w\left[\mathrm{ans}_w(v^*, q) \in d \pm \sigma\sqrt{d}\ln d\right] &\ge 1 - \mathrm{e}^{\frac{(\sigma\sqrt{d}\ln d)^2}{2d\sigma^2}} && \text{(from Fact 6.3.2)}\\ &= 1 - \mathrm{e}^{-\frac{(\ln d)^2}{2}}\\ &= 1 - \frac{1}{d^{(\ln d)/2}}.\end{aligned}$$

Now, for any integer $D_1 \geq \mathrm{e}$, by a union bound, this will hold for all $q \geq v^* + D_1$ with probability at least

$$1 - \sum_{d=D_1}^{\infty} \frac{1}{d^{(\ln d)/2}}.$$

Note that this sum converges, because $(\ln d)/2 > 1$ for large enough d. Thus the sequence of sums $(\sum_{d=k}^{\infty} 1/d^{(\ln d)/2})_{k\geq 3}$ converges to 0 and we can define

$$D_1(\delta_1) := \min\{k \geq 3 \mid \sum_{d=k}^{\infty} 1/d^{(\ln d)/2} \leq \delta_1/2\}.$$

Therefore, by going through the same reasoning for $q \leq v^*$ and taking a union bound, we get that

$$\mathrm{ans}_w(v^*, q) \in d_q \pm \sigma\sqrt{d_q}\ln d_q$$

will hold for all q at distance $d_q := |v^* - q| \geq D_1(\delta_1)$, except with probability at most $\delta_1/2 + \delta_1/2 = \delta_1$. □

Claim C.7.2. *For any probability $\delta_2 > 0$ and any $\epsilon \in (0,1)$, there exists $D_2(\delta_2,\epsilon,\sigma) > 0$ such that for any n,σ and any source $v^* \in V$,*

$$\mathbb{P}_w\left[\textit{for all } q \textit{ such that } d_q := |v^* - q| \geq D_2(\delta_2,\epsilon,\sigma), \mathrm{ans}_w(v^*,q) \in (1\pm\epsilon)d_q\right] \geq 1 - \delta_2,$$

and $D_2(\delta_2,\epsilon,\sigma) = O_{\delta_2,\epsilon}(\max(\sigma^2\log\sigma, 1))$.

Proof. At first, let us consider only the case $q \geq v^*$. That is, consider node $q = v^* + d$ for some distance d. Then $\mathrm{ans}_w(v^*,q) \sim \mathcal{N}(d, d\sigma^2)$, so

$$\begin{aligned}\mathbb{P}_w\left[\mathrm{ans}_w(v^*,q) \in (1\pm\epsilon)d\right] &\geq 1 - \mathrm{e}^{\frac{(\epsilon d)^2}{2d\sigma^2}} && \text{(from Fact 6.3.2)}\\ &= 1 - (\mathrm{e}^{-\frac{\epsilon^2}{2\sigma^2}})^d.\end{aligned}$$

Now, for any integer D_2, by a union bound, this will hold for all $q \geq v^* + D_2$ except with probability at most

$$\begin{aligned}\sum_{d=D_2}^{\infty}\left(\mathrm{e}^{-\frac{\epsilon^2}{2\sigma^2}}\right)^d &= \mathrm{e}^{-\frac{\epsilon^2}{2\sigma^2}D_2}\sum_{d=0}^{\infty}\left(\mathrm{e}^{-\frac{\epsilon^2}{2\sigma^2}}\right)^d\\ &= \frac{\mathrm{e}^{-\frac{\epsilon^2}{2\sigma^2}D_2}}{1-\mathrm{e}^{-\frac{\epsilon^2}{2\sigma^2}}}, \qquad \text{(C.6)}\end{aligned}$$

where the last step uses the fact that this is a geometric series.

- If $\frac{\epsilon^2}{2\sigma^2} \geq 1$, then (C.6) $\leq \frac{\mathrm{e}^{-D_2}}{1-1/\mathrm{e}}$, so if we set $D_2(\delta_2,\epsilon,\sigma) := \ln\left(\frac{2}{\delta_2(1-1/\mathrm{e})}\right)$, then (C.6) $\leq \delta_2/2$.

- If $\frac{\epsilon^2}{2\sigma^2} \le 1$, then we can use $\mathrm{e}^{-x} \le 1 - x/2$ on $[0,1]$ to obtain that

$$(C.6) \le \frac{2\mathrm{e}^{-\frac{\epsilon^2}{2\sigma^2} D_2}}{\frac{\epsilon^2}{2\sigma^2}},$$

so if we set $D_2(\delta_2, \epsilon, \sigma) := \frac{2\sigma^2}{\epsilon^2} \ln\left(\frac{4\sigma^2}{\epsilon^2 \delta_2}\right)$, then $(C.6) \le \delta_2/2$.

It is easy to check that both these values are $O_{\delta_2,\epsilon}(\max(\sigma^2 \ln \sigma, 1))$.

Finally, by going through the same reasoning for $q \le v^*$ and taking a union bound, we get that

$$\mathrm{ans}_w(v^*, q) \in (1 \pm \epsilon) d_q$$

will hold for all q at distance $d_q := |v^* - q| \ge D_2(\delta_2, \epsilon, \sigma)$, except with probability at most $\delta_2/2 + \delta_2/2 = \delta_2$. □

Definition C.7.1. *Let* $C(\delta) := 8/\delta$ *and* $D(\sigma, \delta) := \max(D_1(\delta/3), D_2(\delta/3, 1/4, \sigma), \sigma^2, \mathrm{e}^2)$.

Let us verify that this definition of $D(\sigma, \delta)$ satisfies the bounds claimed in equation (6.10). The lower bound of $\max(\sigma^2, \mathrm{e}^2)$ is trivial. The upper bound of $O_\delta(\max(\sigma^2 \log \sigma, 1))$ comes from the fact that $D_2(\delta_2, \epsilon, \sigma) = O_{\delta_2,\epsilon}(\max(\sigma^2 \log \sigma, 1))$.

We can now prove Lemma 6.5.2.

Proof of Lemma 6.5.2. Apply Claim C.7.1 with $\delta_1 := \delta/3$, and Claim C.7.2 with $\delta_2 := \delta/3$ and $\epsilon := 1/4$. Assume $n \ge 12/\delta$, and let $C' := 6/\delta$. Since v^* is uniformly distributed over $V = [n]$, we have

$$\mathbb{P}[\min(|v^* - 1|, |v^* - n|) < n/C'] \le \frac{2 + n/C'}{n} \le \frac{2}{n} + \frac{1}{C'} \le \delta/3.$$

By a union bound, the concentration bounds of both Claim C.7.1 and Claim C.7.2 as well as inequality $\min(|v^* - 1|, |v^* - n|) \ge n/C'$ will all hold with probability at least $1 - \delta/3 - \delta/3 - \delta/3 = 1 - \delta$.

Furthermore, by the concentration bound of Claim C.7.2, if $\min(|v^* - 1|, |v^* - n|) \ge n/C' \ge D_2(\delta/3, 1/4)$, then

$$\min(\mathrm{ans}_w(v^*, 1), \mathrm{ans}_w(v^*, n)) \ge (n/C')(1 - 1/4) = \frac{3n}{4C'} = \frac{n}{C(\delta)}. \tag{C.7}$$

Then for $n \ge \max(12/\delta, C' D_2(\delta/3, 1/4, \sigma)) = O_\delta(\max(\sigma^2 \ln \sigma, 1))$, with probability at least $1 - \delta$, we have

- $\min(\mathrm{ans}_w(v^*, 1), \mathrm{ans}_w(v^*, n)) \ge \frac{n}{C(\delta)}$ (from (C.7));
- for all q with $d_q := |v^* - q| \ge D(\sigma, \delta)$,

- $\mathrm{ans}_w(v^*, q) \in d_q \pm \sigma\sqrt{d_q}\ln d_q$ (from Claim C.7.1)
- $\mathrm{ans}_w(v^*, q) \in d_q(1 \pm 1/4)$ (from Claim C.7.1).

□

C.8 Proof of Fact 6.5.1

Fact 6.5.1. *For* $0 \leq j \leq j_{\mathrm{stop}}$, $\lambda_{j+1} < \lambda_j$.

Proof of Fact 6.5.1. Since $j \leq j_{\mathrm{stop}}$, by definition of j_{stop}, $\lambda_j \geq D$. Also, by equation (6.10), $D \geq \max(\sigma^2, \mathrm{e}^2)$. Therefore,

$$\begin{aligned}
\lambda_{j+1} &= \mathrm{reduce}_{n,\sigma}(\lambda_j) && \text{(Definition 6.5.7)}\\
&= \frac{\sigma\sqrt{\lambda_j}}{400\ln\lambda_j\log n} && \text{(Definition 6.5.6)}\\
&< \sigma\sqrt{\lambda_j} && (n \geq 3,\ \lambda_j \geq \mathrm{e}^2)\\
&\leq \lambda_j. && (\lambda_j \geq \sigma^2)
\end{aligned}$$

□

C.9 Proof of Lemma 6.5.3

Lemma 6.5.3. *If* $j < j_{\mathrm{stop}}$, *then* $A_j \Rightarrow B_j$.

The first step in proving Lemma 6.5.3 is to prove that only a small part of the information contained in K_j will actually influence the posterior of v^* given K_j: only the closest query nodes to the source q_{l_j}, q_{r_j} and the corresponding answers a_{l_j}, a_{r_j} will have an influence (they are introduced in Definition 6.5.5).

Definition C.9.1 ($E_{l,r,x,y}$)**.** *For any* l, r, x, y, *let* $E_{l,r,x,y}$ *be the event that* $v^* \in [l, r]$, $\mathrm{ans}_w(v^*, l) = x$, *and* $\mathrm{ans}_w(v^*, r) = y$.

Note that event $E_{l,r,x,y}$ depends purely on v^* and w, not on the actions of the algorithm.

Lemma C.9.1. *Recall that* R *is the internal randomness of the algorithm (see Definition 6.5.1). For any node* $v \in V$,

$$\mathbb{P}_{v^*,w,R}[v^* = v \mid K_j] = \mathbb{P}_{v^*,w}[v^* = v \mid E_{q_{l_j},q_{r_j},a_{l_j},a_{r_j}}].$$

Before proving this lemma, we need to show a simple property of independence and conditional probability.

Fact C.9.1. *Let X, Y be independent random variables. Let $E(X), F(X)$ be Boolean functions depending only on X, and let $G(F(X), Y)$ be a Boolean function depending only on $F(X)$ and Y. For simplicity, let us use $E(X)$ to denote the event $E(X) = 1$ (and similarly for F, G). Then we have*

$$\mathbb{P}[E(X) \mid F(X) \wedge G(F(X), Y)] = \mathbb{P}[E(X) \mid F(X)].$$

Proof. Intuitively, the reason this is true is that as G depends only on F (which is already provided in the conditioning) and Y (which is independent from X), adding G to the conditioning does not bring more information towards figuring out whether E will happen or not. Formally, let $G'(Y) = G(1, Y)$ (1 represents "true"). Then

$$\begin{aligned}
\mathbb{P}[E(X) \mid F(X) \wedge G(F(X), Y)] &= \mathbb{P}[E(X) \mid F(X) \wedge G'(Y)] \\
&= \frac{\mathbb{P}[E(X) \wedge F(X) \wedge G'(Y)]}{\mathbb{P}[F(X) \wedge G'(Y)]} \\
&= \frac{\mathbb{P}[E(X) \wedge F(X)] \cdot \mathbb{P}[G'(Y)]}{\mathbb{P}[F(X)] \cdot \mathbb{P}[G'(Y)]} && \text{(independence of } X \text{ and } Y\text{)} \\
&= \frac{\mathbb{P}[E(X) \wedge F(X)]}{\mathbb{P}[F(X)]} \\
&= \mathbb{P}[E(X) \mid F(X)]
\end{aligned}$$

□

We will notation $\mathbb{1}[\cdots]$ to denote the indicator Boolean function corresponding to some expression.

Proof of Lemma C.9.1. Fix any possible assignment k_j for K_j. We observe that event "$K_j = k_j$" is entirely determined by

(a) R (the internal randomness of the algorithm);

(b) the values of $\text{ans}_w(v^*, q_i)$ for each $i \le j$ (the answers to the first j queries);

(c) whether $v^* \in [q_{l_j}, q_{r_j}]$.

Conversely, (b) and (c) are entirely determined by K_j. In addition, (b) and (c) depend only on v^* and w, which are independent from R.

By the above, we can use Fact C.9.1, plugging in $E((v^*, w)) := \mathbb{1}[v^* = v]$, $F((v^*, w)) := \mathbb{1}[v^* \in$

$[q_{l_j}, q_{r_j}] \wedge \mathrm{ans}_w(v^*, q_i) = a_i \,\forall i \ge j]$ and $G(F((v^*, w)), R) := \mathbb{1}[K_j = k_j]$. This gives

$$\begin{aligned}
&\mathbb{P}_{v^*,w,R}[v^* = v \mid K_j = k_j] \\
&\quad = \mathbb{P}_{v^*,w,R}[v^* = v \mid \underbrace{v^* \in [q_{l_j}, q_{r_j}]}_{\text{(c)}} \wedge \underbrace{\mathrm{ans}_w(v^*, q_i) = a_i \,\forall i \ge j}_{\text{(b)}} \wedge K_j = k_j] \\
&\quad = \mathbb{P}_{v^*,w,R}[\underbrace{v^* = v}_{E} \mid \underbrace{v^* \in [q_{l_j}, q_{r_j}] \wedge \mathrm{ans}_w(v^*, q_i) = a_i \,\forall i \ge j}_{F} \wedge \underbrace{K_j = k_j}_{G}] \qquad \text{(C.8)} \\
&\quad = \mathbb{P}_{v^*,w}[\underbrace{v^* = v}_{E} \mid \underbrace{v^* \in [q_{l_j}, q_{r_j}] \wedge \mathrm{ans}_w(v^*, q_i) = a_i \,\forall i \ge j}_{F}]
\end{aligned}$$

Now, let us conceptually split w into two parts: w_{in}, which contains the weights of only the edges between nodes q_{l_j} and q_{r_j}, and w_{out}, which contains all the other weights. Since weights are distributed independently, (v^*, w_{in}) is independent from w_{out}.

Let $E((v^*, w_{\text{in}})) := \mathbb{1}[v^* = v]$, $F((v^*, w_{\text{in}})) := \mathbb{1}[v^* \in [q_{l_j}, q_{r_j}] \wedge \mathrm{ans}_w(v^*, q_{l_j}) = a_{l_j} \wedge \mathrm{ans}_w(v^*, q_{r_j}) = a_{r_j}]$, and $G(F((v^*, w_{\text{in}})), w_{\text{out}}) := \mathbb{1}[\mathrm{ans}_w(v^*, q_i) = a_i \,\forall i \ge j]$. It is easy to see why E and F depend only on v^* and w_{in}. For G, we note that all other answers can be deduced just from the fact that $v^* \in [q_{l_j}, q_{r_j}]$, the values of $\mathrm{ans}_w(v^*, q_{l_j})$ and $\mathrm{ans}_w(v^*, q_{r_j})$, and w_{out}. Therefore, G depends only on F and w_{out}. Thus we can again apply Fact C.9.1, to obtain

$$\begin{aligned}
&\mathbb{P}_{v^*,w}[v^* = v \mid v^* \in [q_{l_j}, q_{r_j}] \wedge \mathrm{ans}_w(v^*, q_i) = a_i \,\forall i \ge j] \\
&\quad = \mathbb{P}_{v^*,w}[\underbrace{v^* = v}_{E} \mid \underbrace{v^* \in [q_{l_j}, q_{r_j}] \wedge \mathrm{ans}_w(v^*, q_{l_j}) = a_{l_j} \wedge \mathrm{ans}_w(v^*, q_{r_j}) = a_{r_j}}_{F} \\
&\qquad\qquad \wedge \underbrace{\mathrm{ans}_w(v^*, q_i) = a_i \,\forall i \ge j}_{G}] \qquad \text{(C.9)} \\
&\quad = \mathbb{P}_{v^*,w}[\underbrace{v^* = v}_{E} \mid \underbrace{v^* \in [q_{l_j}, q_{r_j}] \wedge \mathrm{ans}_w(v^*, q_{l_j}) = a_{l_j} \wedge \mathrm{ans}_w(v^*, q_{r_j}) = a_{r_j}}_{F}] \\
&\quad = \mathbb{P}_{v^*,w}[v^* = v \mid E_{q_{l_j}, q_{r_j}, a_{l_j}, a_{r_j}}].
\end{aligned}$$

The result follows from combining (C.8) with (C.9). □

For the remainder of this section, to make the notation lighter, we will use the following shorthands.

$$l := q_{l_j} \quad r := q_{r_j} \quad a_l := a_{l_j} \quad a_r := a_{r_j} \quad \mu := \mu_j = \min(a_l, a_r)$$

Definition C.9.2 (μ'). *Let* $\mu' := \frac{\sigma\sqrt{\mu}}{400 \ln \mu}$.

Definition C.9.3 (I). *Let I be the interval of all sources v^* that are consistent with event T and E_{l,r,a_l,a_r}: more precisely,*

$$I := \{v \mid \mathbb{P}[T \wedge (v^* = v) \mid E_{l,r,a_l,a_r}] > 0\}.$$

Lemma C.9.2. *Assume $\mu, \mu' \geq D$. Then for any node $v \in I$,*

$$\mathbb{P}_{v^*,w}[v^* = v \mid E_{l,r,a_l,a_r}] \leq \frac{100 \ln \mu}{\sigma\sqrt{\mu}}.$$

Lemma C.9.2 has long and complicated proof, but its meaning is intuitive: it states is that when μ is large, the posteriors of v^* are not very concentrated at any point of segment I. The expression of this posterior is too complex to work with directly, so we will need the help of some facts and claims to prove what we want.

Before we prove Lemma C.9.2, we start with Fact C.9.2, which gives us a couple of useful inequalities, and Fact C.9.3, a simple calculus result that we will use in this section and the next.

Fact C.9.2. *Assume $\mu, \mu' \geq D$. Then*

$$\mu \geq 800^2 \cdot \max(\mu', D, \sigma^2, 1). \tag{C.10}$$

In particular, this implies

$$\begin{aligned} \mu &\geq (5/4)D && \text{(C.11)}\\ \sigma\sqrt{\mu} &\leq \mu/5 && \text{(C.12)}\\ \sigma\sqrt{\mu} &\leq \sigma\sqrt{(4/5)\mu}\ln((4/5)\mu) && \text{(C.13)}\\ (3/5)\mu &\geq 6000. && \text{(C.14)} \end{aligned}$$

Proof. Recall from equation (6.10) that $D \geq \sigma^2, \mathrm{e}^2$. Since $\mu \geq D \geq \mathrm{e}^2$, we have

$$400 \ln \mu \geq 400 \ln \mathrm{e}^2 = 800.$$

This implies

$$\sigma^2 \leq D \leq \mu' = \frac{\sigma\sqrt{\mu}}{400 \ln \mu} \leq \frac{\sigma\sqrt{\mu}}{800},$$

from which we get $\sigma \leq \sqrt{\mu}/800$ and $\mu' \leq \frac{\sigma\sqrt{\mu}}{800}$. Combining those, we get

$$\mu' \leq \frac{\sigma\sqrt{\mu}}{800} \leq \frac{\mu}{800^2},$$

which proves $\mu \geq 800^2\mu'$. The other three parts then follow directly from $\mu' \geq D \geq \sigma^2, \mathrm{e}^2$.

Among (C.11)–(C.14), all are trivial from (C.10), except for (C.13) which can be rewritten as $5/4 \leq \ln((4/5)\mu)$. This clearly holds for $\mu \geq 10^4$. □

Fact C.9.3. *Let $f(d) := \frac{\ln d}{\sqrt{d}}$ and $g(d) := \frac{\sqrt{d}}{\ln d}$. On $[\mathrm{e}^2, \infty)$, f is decreasing and g is increasing.*

Proof. The derivative of f is $\frac{2-\ln d}{2d\sqrt{d}}$. □

Proof of Lemma C.9.2. Since we have $\mu, \mu' \geq D$, Fact C.9.2 applies here.

Assume I is not empty (otherwise, the lemma holds vacuously). Let I' be I extended by $\sigma\sqrt{\mu}$ on both sides. Note that, by definition, the length of I' is at least $2\sigma\sqrt{\mu}$.

Claim C.9.1. *For any $v' \in I'$, let $d'_l := v' - l$ and $d'_r := r - v'$. Then*

$$d'_l \in [1/2, 2]\, a_l \tag{C.15}$$

$$d'_r \in [1/2, 2]\, a_r \tag{C.16}$$

$$|a_l - d'_l| \leq 3\sigma\sqrt{d'_l}\ln d'_l \tag{C.17}$$

$$|a_r - d'_r| \leq 3\sigma\sqrt{d'_r}\ln d'_r. \tag{C.18}$$

Proof. We will only prove (C.15) and (C.17); the proof of the other two is analogous.

Let us first take a look at the properties of source candidates in I. Take some candidate $v \in I$ and consider the (true) distance between v and the left side of the interval, $d_l := v - l$. First, let us show that $d_l \geq D$. Indeed, if $d_l < D$, then point $v - D$ would be strictly to the left of l. Thus we would have

$$\mathrm{ans}_w(v, v-D) > \mathrm{ans}_w(v, l) = a_l \geq \mu \overset{(C.11)}{\geq} (5/4)D,$$

which contradicts part (ii) in Definition 6.5.3 for $q := v - D$.

Now that we have $d_l \geq D$, we can apply part (ii) in Definition 6.5.3 again to obtain that

$$a_l \in [3/4, 5/4]\, d_l \;\Rightarrow\; d_l \in [4/5, 4/3]\, a_l \tag{C.19}$$

and

$$|a_l - d_l| \in \sigma\sqrt{d_l}\ln d_l. \tag{C.20}$$

Let us now extend these results to I'. Any point $v' \in I'$ is at most $\sigma\sqrt{\mu}$ away from some point $v \in I$. Therefore, if we continue using notation $d_l := v - l$, we have $d'_l \in d_l \pm \sigma\sqrt{\mu}$.

From (C.19), we get

$$d'_l \in d_l \pm \sigma\sqrt{\mu} \subseteq [4/5, 4/3]\, a_l \pm \sigma\sqrt{\mu}.$$

Besides, from (C.12), $\sigma\sqrt{\mu} \leq \mu/5 \leq a_l/5$, so we have

$$d'_l \in [4/5, 4/3]\, a_l \pm a_l/5 = [3/5, 23/15]\, a_l,$$

which proves (C.15).

In addition, we note that

$$d'_l \geq d_l - \sigma\sqrt{\mu} \geq d_l - a_l/5 \overset{\text{(C.19)}}{\geq} d_l - d_l/4 = (3/4)d_l. \tag{C.21}$$

Therefore,

$$\begin{aligned}
|a_l - d'_l| &\leq |a_l - d_l| + \sigma\sqrt{\mu} && \text{(definition of } I'\text{)}\\
&\leq \sigma\sqrt{d_l}\ln d_l + \sigma\sqrt{\mu} && \text{(from (C.20))}\\
&\leq \sigma\sqrt{d_l}\ln d_l + \sigma\sqrt{(4/5)\mu}\ln((4/5)\mu) && \text{(from (C.13))}\\
&\leq 2\sigma\sqrt{d_l}\ln d_l && (d_l \geq (4/5)a_l \geq (4/5)\mu)\\
&\leq 2\sigma\sqrt{4d'_l/3}\ln(4d'_l/3) && \text{(from (C.21))}\\
&\leq 3\sigma\sqrt{d'_l}\ln d'_l,
\end{aligned}$$

where the last step holds because $d'_l \geq (3/5)a_l \geq (3/5)\mu \overset{\text{(C.14)}}{\geq} 6000$, which is big enough. This proves (C.17). □

Note that Claim C.9.1 implies in particular that $I' \subseteq [l, r]$. Let us study the ratios of the posterior probabilities of $v^* = v'$ between different values of $v' \in I'$, conditioned on a_l, a_r (but regardless of whether T holds). We will use the shorthand

$$p(v') := \mathbb{P}[v^* = v' \mid v^* \in I' \wedge \text{ans}_w(v^*, l) = a_l \wedge \text{ans}_w(v^*, r) = a_r]. \tag{C.22}$$

Given that v^* is initially distributed uniformly, $p(v')$ is proportional to

$$\begin{aligned}
&\mathbb{P}[\text{ans}_w(v^*, l) = a_l \wedge \text{ans}_w(v^*, r) = a_r \mid v^* = v']\\
&\quad = \mathbb{P}[\text{ans}_w(v^*, l) = a_l \mid v^* = v']\mathbb{P}[\text{ans}_w(v^*, r) = a_r \mid v^* = v'],
\end{aligned} \tag{C.23}$$

where the independence comes from the fact that $\text{ans}_w(v^*, l)$ and $\text{ans}_w(v^*, r)$ depend on completely separate weights.

Because we assumed that all weights are independently distributed from $\mathcal{N}(1, \sigma^2)$, both factors in (C.23) follow a normal distribution. Those distributions are $\mathcal{N}(d'_l, \sigma^2 d'_l)$ and $\mathcal{N}(d'_r, \sigma^2 d'_r)$, where we continue notations $d'_l := v' - l$ and $d'_r := r - v'$. This means that $p(v')$ is proportional to

$$\frac{1}{\sqrt{2\pi\sigma^2 d'_l}} \mathrm{e}^{-\frac{(a_l - d'_l)^2}{2\sigma^2 d'_l}} \times \frac{1}{\sqrt{2\pi\sigma^2 d'_r}} \mathrm{e}^{-\frac{(a_r - d'_r)^2}{2\sigma^2 d'_r}}.$$

Note that in the above expression, a_l, a_r are fixed by the conditioning, while d'_l and d'_r depend on v'. From now on, we will denote them as $d'_l(v')$ and $d'_r(v')$ to make this clear.

Of course, the constant $\frac{1}{2\pi\sigma^2}$ does not matter. Besides, we know that $d'_l(v') \in [1/2, 2]a_l$ and $d'_r(v') \in [1/2, 2]a_r$ with a_l, a_r fixed, so the factor $\frac{1}{\sqrt{d'_l(v')d'_r(v')}}$ will vary only by a factor 4. Thus we can conclude that $p(v')$ is also proportional to

$$F(v') := \mathrm{e}^{-\frac{1}{2\sigma^2}\left(\frac{(a_l-d'_l(v'))^2}{d'_l(v')}+\frac{(a_r-d'_r(v'))^2}{d'_r(v')}\right)},$$

up to a factor 4 of error. More precisely, we know that there exists some $k>0$ such that for all $v' \in I'$,

$$p(v') \in [1,4]kF(v'). \tag{C.24}$$

Let

$$G(v') := \frac{1}{2\sigma^2}\left(\frac{(a_l-d'_l(v'))^2}{d'_l(v')}+\frac{(a_r-d'_r(v'))^2}{d'_r(v')}\right)$$

so that $F(v') = \mathrm{e}^{-G(v')}$, and let v'_{peak} be the value of $v' \in I'$ that maximizes the expression $F(v')$. We will show the existence of a relatively large interval $J \subseteq I'$ centered around v'_{peak} such that for all $v' \in J$, the value $F(v')$ is not much smaller than $F(v'_{\text{peak}})$.

Claim C.9.2. *There is some interval $J \subseteq I'$ of length at least $\frac{\sigma\sqrt{\mu}}{6\sqrt{2}\ln\mu}$, such that for all $v' \in J$,*

$$F(v') \geq \frac{F(v'_{\text{peak}})}{\mathrm{e}}.$$

Proof. The derivative of $F(v')$ is $G'(v')\mathrm{e}^{G(v')} = G'(v')F(v')$, where

$$G'(v') = \frac{1}{2\sigma^2}\left(-\frac{a_l-d'_l(v')}{d'_l(v')}-\left(\frac{a_l-d'_l(v')}{d'_l(v')}\right)^2+\frac{a_r-d'_r(v')}{d'_r(v')}+\left(\frac{a_r-d'_r(v')}{d'_r(v')}\right)^2\right).$$

First, note that by (C.15) and (C.16), we have

$$\left|\frac{a_l-d'_l(v')}{d'_l(v')}\right| \leq 1 \text{ and } \left|\frac{a_r-d'_r(v')}{d'_r(v')}\right| \leq 1.$$

Therefore,

$$|G'(v')| \leq \frac{1}{\sigma^2}\left(\left|\frac{a_l-d'_l(v')}{d'_l(v')}\right|+\left|\frac{a_r-d'_r(v')}{d'_r(v')}\right|\right).$$

And then we can use equations (C.17) and (C.18) to bound this further:

$$\begin{aligned}
|G'(v')| &\leq \frac{1}{\sigma^2}\left(\frac{3\sigma\sqrt{d'_l(v')}\ln d'_l(v')}{d'_l(v')} + \frac{3\sigma\sqrt{d'_r(v')}\ln d'_r(v')}{d'_r(v')}\right)\\
&\leq \frac{3}{\sigma}\left(\frac{\ln d'_l(v')}{\sqrt{d'_l(v')}} + \frac{\ln d'_r(v')}{\sqrt{d'_r(v')}}\right)\\
&\leq \frac{3}{\sigma}\left(\frac{\ln(\mu/2)}{\sqrt{\mu/2}} + \frac{\ln(\mu/2)}{\sqrt{\mu/2}}\right) && (d'_l \geq a_l/2 \geq \mu/2 \geq \mathrm{e}^2 \text{ and Fact C.9.3; same for } d'_r)\\
&\leq \frac{6\sqrt{2}\ln\mu}{\sigma\sqrt{\mu}}.
\end{aligned}$$

We now have

$$|F'(v')| \leq \frac{6\sqrt{2}\ln\mu}{\sigma\sqrt{\mu}}F(v'),$$

for all $v' \in I'$, which from $F(v'_{\text{peak}}) > 0$ and Grönwall's Lemma for ordinary differential inequalities can be seen to imply that for all $v' \in I'$,

$$F(v') \geq \mathrm{e}^{-\left|v'-v'_{\text{peak}}\right|\frac{6\sqrt{2}\ln\mu}{\sigma\sqrt{\mu}}}F(v'_{\text{peak}}).$$

This means that for any $v' \in I'$ within distance at most $\frac{\sigma\sqrt{\mu}}{6\sqrt{2}\ln\mu}$ of v'_{peak},

$$F(v') \geq \mathrm{e}^{-1}F(v'_{\text{peak}}) \geq \frac{F(v'_{\text{peak}})}{\mathrm{e}}.$$

We can then set

$$J := I' \cap \left(v'_{\text{peak}} \pm \frac{\sigma\sqrt{\mu}}{6\sqrt{2}\ln\mu}\right).$$

Defined this way, J will clearly have length at least $\frac{\sigma\sqrt{\mu}}{6\sqrt{2}\ln\mu}$. Indeed v'_{peak} is in I', and the length of I' is at least $2\sigma\sqrt{\mu} \geq 2 \times \frac{\sigma\sqrt{\mu}}{6\sqrt{2}\ln\mu}$. □

Now, recall from (C.24) that $p(v') \in [1,4]kF(v')$. Thus for any $v' \in I'$,

$$p(v') \leq 4kF(v') \leq 4kF(v'_{\text{peak}}). \tag{C.25}$$

But in particular, by Claim C.9.2, for any $v' \in J$,

$$p(v') \geq kF(v') \geq \frac{kF(v'_{\text{peak}})}{\mathrm{e}}. \tag{C.26}$$

Besides, being a probability distribution, $p(v')$ must sum up to 1, so we have

$$1 = \sum_{v' \in I'} p(v') \geq \sum_{v' \in J} p(v') \overset{(C.26)}{\geq} |J| \frac{kF(v'_{\text{peak}})}{\mathrm{e}},$$

thus $kF(v'_{\text{peak}}) \leq \frac{\mathrm{e}}{|J|}$. Therefore, for any $v' \in I'$,

$$p(v') \overset{(C.25)}{\leq} 4kF(v'_{\text{peak}}) \leq \frac{4\mathrm{e}}{|J|} \leq \frac{24\mathrm{e}\sqrt{2}\ln\mu}{\sigma\sqrt{\mu}} \leq \frac{100\ln\mu}{\sigma\sqrt{\mu}}. \tag{C.27}$$

We are finally ready to prove the lemma. For $v \in I$, we can observe that

$$\begin{aligned}
&\mathbb{P}[v^* = v \mid E_{l,r,a_l,a_r}] \\
&\quad = \mathbb{P}[v^* = v \mid v^* \in [l,r] \wedge \text{ans}_w(v^*, l) = a_l \wedge \text{ans}_w(v^*, r) = a_r] \\
&\quad \leq \mathbb{P}[v^* = v \mid v^* \in I' \wedge \text{ans}_w(v^*, l) = a_l \wedge \text{ans}_w(v^*, r) = a_r] && \text{(strengthen the condition)} \\
&\quad = p(v) && \text{(defined in (C.22))} \\
&\quad \leq \frac{100\ln\mu}{\sigma\sqrt{\mu}}. && (v \in I \subset I' \text{ and (C.27)})
\end{aligned}$$

□

With Lemma C.9.2 in hand, we are finally ready to prove Lemma 6.5.3.

Proof of Lemma 6.5.3. First, we show that $j < j_{\text{stop}}$ and A_j imply that $\mu, \mu' \geq D$. Indeed, $j < j_{\text{stop}}$ implies $\lambda_j, \lambda_{j+1} \geq D$. If in addition A_j holds, then $\mu = \mu_j \geq \lambda_j \geq D$, and

$$\mu' = \frac{\sigma\sqrt{\mu}}{400\ln\mu} \geq \frac{\sigma\sqrt{\lambda_j}}{400\ln\lambda_j} > \frac{\sigma\sqrt{\lambda_j}}{400\ln\lambda_j \log n} = \lambda_{j+1} \geq D.$$

Therefore, the assumptions of Lemma C.9.2 hold.

We need to prove that for any $v \in V$,

$$\mathbb{P}[T \wedge (v^* = v) \mid K_j] \leq \frac{1}{\left(\frac{8}{3}\lambda_{j+1} + 1\right)\log n}.$$

If $v \notin I$, then by definition of I, $\mathbb{P}[T \wedge (v^* = v) \mid K_j] = 0$, so the inequality holds trivially. On the other hand, if $v \in I$,

$$\begin{aligned}
&\mathbb{P}[T \wedge (v^* = v) \mid K_j] \\
&\qquad \le \mathbb{P}[v^* = v \mid K_j] \\
&\qquad = \mathbb{P}[v^* = v \mid E_{l,r,a_l,a_r}] && \text{(by Lemma C.9.1)} \\
&\qquad \le \frac{100 \ln \mu}{\sigma \sqrt{\mu}} && \text{(by Lemma C.9.2)} \\
&\qquad \le \frac{100 \ln \lambda_j}{\sigma \sqrt{\lambda_j}} \\
&\qquad = \frac{1}{4\lambda_{j+1} \log n} \\
&\qquad \le \frac{1}{\left(\frac{8}{3}\lambda_{j+1} + 1\right) \log n}. && (\lambda_{j+1} \ge D \ge \mathrm{e}^2)
\end{aligned}$$

□

C.10 Proof of Lemma 6.5.6

Lemma 6.5.6. *For* $n \ge \Theta_p(\max(\sigma^3, 1))$*, we have*

$$j_{\text{stop}} + 1 = \begin{cases} \Omega_p(1 + \log(1 + \log_{1/\sigma} n)) \text{ } \mathit{if}\, \sigma^2 \le 1/2 \\ \Omega_p(\log\log n) \text{ } \mathit{if}\, \sigma^2 \ge 1/2. \end{cases}$$

This proof is very similar in spirit to the proof of Claim 6.5.1 (C.6).

Proof of Lemma 6.5.6. We track the value of λ_j/σ^2 as j increases. First, as long as

$$\frac{\lambda_j}{(\ln \lambda_j)^6} \ge \sigma^2 (400 \log n)^6, \tag{C.28}$$

we have

$$\frac{\lambda_{j+1}}{\sigma^2} = \frac{\text{reduce}_{n,\sigma}(\lambda_{j+1})}{\sigma^2} = \frac{\sigma\sqrt{\lambda_j}}{400\sigma^2 \ln \lambda_j \log n} = \sqrt{\frac{\lambda_j}{\sigma^2}} \overset{\text{(C.28)}}{\ge} \left(\frac{\lambda_j}{\sigma^2}\right)^{1/3}.$$

Let $\lambda_{\min}$ be the smallest possible value for λ_j at least as large as D such that (C.28) holds. Then, by induction, as long as $\lambda_j \ge \lambda_{\min}$, we have

$$\frac{\lambda_j}{\sigma^2} \ge \left(\frac{\lambda_0}{\sigma^2}\right)^{1/3^j} \Rightarrow \lambda_j \ge \sigma^2 \left(\frac{n}{C\sigma^2}\right)^{1/3^j}. \tag{C.29}$$

Recall that $j_{\min}$ is the smallest integer $j \ge 0$ such that $\lambda_j < D$. Let j^* be the smallest integer

$j \geq 0$ such that $\lambda_j < \lambda_{\min}$. Then $j^* \leq j_{\min}$, and applying (C.29) to j^* we get

$$\sigma^2 \left(\frac{n}{C\sigma^2}\right)^{1/3^{j^*}} < D \Rightarrow j_{\min} \geq j^* > \log_3 \left(\frac{\log\left(\frac{n}{C\sigma^2}\right)}{\log\left(\frac{\lambda_{\min}}{\sigma^2}\right)}\right). \tag{C.30}$$

Since $\lambda_0 = n/C$, for $n \geq CD = O_p(\sigma^2 \ln \sigma)$, we have $j_{\min} \geq 1$. We separate into cases to obtain more lower bounds of $j_{\min}$. First, if $\sigma^2 \leq 1/2$, it is easy to verify that $\lambda_{\min} = \max(D, (\log n)^{O(1)}) = (\log n)^{O_p(1)}$. Therefore, by (C.30), there exists $n_1 = O_p(1)$ such that for $n \geq n_1$, we have

$$j_{\min} \geq \Omega_p(\log(1 + \log_{1/\sigma} n)).$$

Second, if $\sigma^2 \geq 1/2$, it is easy to verify that $\lambda_{\min} = \max(D, \sigma^2(\log n)^{O(1)}) = \sigma^2(\log n)^{O_p(1)}$. Therefore, by (C.30), there exists $n_2 = O_p(\sigma^3)$ such that for $n \geq n_2$, we have

$$j_{\min} \geq \Omega_p(\log\log n).$$

In summary, there exists $n_3 = \max(n_1, n_2) = O_p(\max(\sigma^3, 1))$ such that for $n \geq n_3$,

$$j_{\min} = \begin{cases} \Omega_p(1 + \log(1 + \log_{1/\sigma} n)) \text{ if } \sigma^2 \leq 1/2 \\ \Omega_p(\log\log n) \text{ if } \sigma^2 \geq 1/2, \end{cases}$$

which is exactly the bounds we want for $j_{\text{stop}} + 1$. We can then conclude by observing that

$$j_{\text{stop}} + 1 = \min\left(j_{\min}, 1 + \left\lfloor \frac{p \log n}{2} \right\rfloor\right)$$

and that $1 + \left\lfloor \frac{p \log n}{2} \right\rfloor$ is larger than the claimed lower bounds for n large enough (which is covered in the $\Theta_p(\max(\sigma^3, 1))$ lower bound on n in the statement of the lemma). □

D Appendix for Chapter 7

The goal of this appendix is to state and prove Lemmas D.3.1-D.3.3, which we will do in Section D.3. Before introducing these lemmas, we prove some claims that will be useful later. We start by elementary claims in Section D.1, then in Section D.2 we prove more involved results that characterize the normal and special regions of G'.

D.1 Elementary Claims

The following claim shows that resolving sets must have nodes on the boundaries of the grid, which helps us reduce the number of subsets that we must prove are non-resolving.

Claim D.1.1. *If R is any resolving set of G' (the grid with extra edge EF) satisfying Assumption 7.4.2 and 7.4.3, there must be two vertices X and Y in R which satisfy following two properties:*

1. *They are on opposite boundaries of G'*
2. *If one of them is a corner, the other one must be an adjacent corner.*

Proof. Consider vertices $A = (1,2)$ and $B = (2,1)$. It is easy to see that only vertices on boundaries PQ and PS except corner P will be able to distinguish A and B as none of E and F is on the boundaries. So we need at least one vertex on the union of the boundaries PQ and PS, excluding P, in the resolving set. A similar argument holds for the other 4 corners, hence we can deduce the two required conditions. □

Claim D.1.2. *For all $A, B \in V$ if* $\text{Gain}(A,B)$ *is positive, it will have same parity as* Gain *and* Gain$'$ *as defined in Definition 7.4.2.*

Proof. Note that $\text{Gain}(A,B) > 0$ indicates that A uses e to reach B. For this to happen, we must have $A \in R_F$ and $B \in R_E$ (or the other way around), in which case $d_{G'}(A,B) = AE + 1 + FB$. This gives

$$\begin{aligned}\text{Gain}(A,B) &= AB-(AE+1+FB)\\ &= |x_A-x_B|+|y_A-y_B|-(|x_A-x_E|+|y_A-y_E|+1+|x_F-x_B|+|y_F-y_B|),\end{aligned}$$

which has same parity as

$$(x_A-x_B)+(y_A-y_B)-((x_A-x_E)+(y_A-y_E)+1+(x_F-x_B)+(y_F-y_B))=(y_E-y_F)-(x_E-x_F)-1,$$

which has the same parity as Gain $=(y_F-y_E)+(x_E-x_F)-1$ and Gain$'=(y_F-y_E)-(x_E-x_F)-1$. □

Remark D.1.1. *Consider a vertex X and its 4 neighbouring vertices X_1, X_2, X_3, X_4. A single vertex in the graph cannot distinguish all of these 4 vertices.*

Proof. Suppose that vertex A distinguishes all 4 vertices. By triangular inequality, AX_i can only take 3 distinct values, namely $AX-1$, AX and $AX+1$. Hence, by pigeon hole principle, at least 2 vertices will have same distance to A. □

D.2 Exact Characterization of Normal and Special Regions

For better readability, we repeat the two assumptions introduced in Chapter 7.

Assumption 7.4.2. *We assume that*

1. $x_F \le x_E$
2. $y_E \le y_F$
3. $x_E - x_F \le y_F - y_E$.

Assumption 7.4.3. *We assume that*

1. $x_F \ne x_E$
2. Gain$' \ge 2$
3. *none of E and F lie on the boundary of the grid.*

We prove that in two dimensions, under Assumptions 7.4.2 and 7.4.3, the normal region takes a fairly regular shape. As shown in Figure D.1, we only have two cases based on the parity of Gain$'$. This is in sharp contrast with higher dimensions, where the normal region can take very different shapes (see Figure 7.4).

The following quantities will be useful to describe the shape of the normal region.

Definition D.2.1 (α, β)**.** *Let*

$$\alpha = \frac{1 + y_F + y_E + x_F - x_E}{2}$$

and

$$\beta = \frac{1 + y_F + y_E + x_E - x_F}{2}.$$

Remark D.2.1. *We make the following observations about α and β under Assumptions 7.4.2 and 7.4.3.*

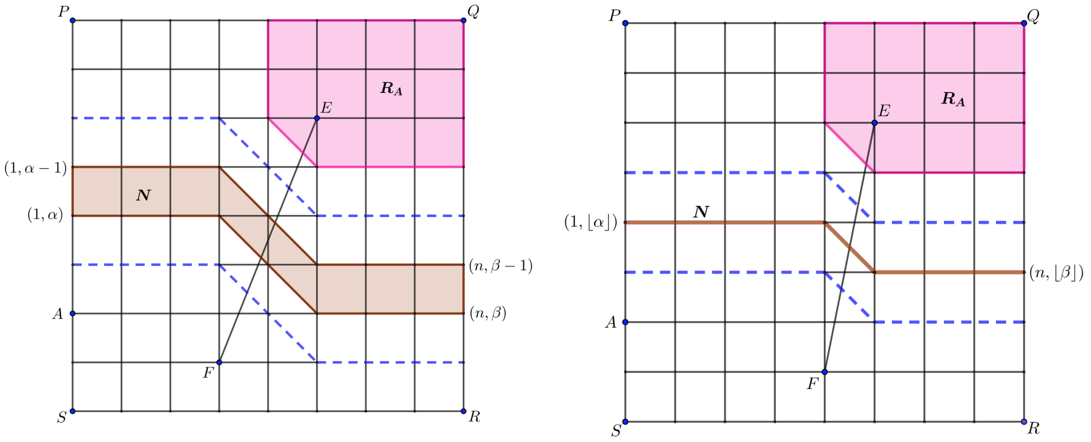
1. *Note that $\beta - \alpha = x_E - x_F$. Assumptions 7.4.2 and 7.4.3 imply $x_E > x_F$, and since x_E, x_F are both integers, we know that $\beta - \alpha \geq 1$. Consequently, $\lfloor\beta\rfloor > \lfloor\alpha\rfloor$ holds.*

2. *Using Assumptions 7.4.2 and 7.4.3, we find the following useful equalities and inequalities:*

$$\alpha = y_F - \frac{\text{Gain}}{2} = y_E + \frac{\text{Gain}' + 2}{2} \geq y_E + 2, \tag{D.1}$$

and

$$\beta = y_E + \frac{\text{Gain} + 2}{2} = y_F - \frac{\text{Gain}'}{2} \leq y_F + 1. \tag{D.2}$$



(a) When Gain' is even, the height of N is 2.

(b) When Gain$'$ is odd, the height of N is 1.

Figure D.1: Illustration for Claims D.2.1 and D.2.2. Brown region(including the boundary), which is just a set of line segments in the case when Gain$'$ is odd, is the normal region of the grid. Pink region(including the boundary) indicates the special region of a point A belonging to R_E which lies on boundary PS of G'.

Next, we express precisely the normal region of the grid.

Claim D.2.1. *Under Assumptions 7.4.2 and 7.4.3,*

$$\begin{aligned} N = &\{(x,y) \mid x < x_F, \alpha - 1 \leq y \leq \alpha\} \cup \\ &\{(x,y) \mid x_F \leq x \leq x_E, x_F - \alpha \leq x - y \leq x_F - \alpha + 1\} \cup \\ &\{(x,y) \mid x > x_E, \beta - 1 \leq y \leq \beta\} \end{aligned} \tag{D.3}$$

Geometrically, the normal region will be the union of three strips of "height" 1 or 2: two horizontal strips with y-coordinates around α and β respectively, and a third strip at 45 degree angle joining the two horizontal strips (see Figure D.1). By "height" here we mean the number of vertices corresponding to each x coordinate. The height of the strips depends on whether α and β are integers or not (and thus on the parity of Gain'): when Gain' is even, α and β are integers and the height of the strip is 2; when Gain' is odd, α and β are odd integers divided by two, and the height of the strip is 1.

The union of the three strips forms a single continuous strip, which separates the grid into two connected components along the y-axis. The y-coordinates of the strip lie completely between the y-coordinates of nodes E and F, which means that for every x-coordinate, the normal vertices are sandwiched between non-normal vertices along the y-axis. Here we rely heavily on the inequality $\text{Gain}' \geq 2$ in Assumption 7.4.3; for $\text{Gain}' = 0$ the normal region can touch the PQ and RS boundaries of the grid.

Proof of Claim D.2.1. Let $A = (x_A, y_A)$ be a vertex in N. By Claim 7.2.1, being a normal vertex is equivalent to

$$|AE - AF| = ||x_A - x_F| + |y_A - y_F| - |x_A - x_E| - |y_A - y_E|| \leq 1. \tag{D.4}$$

First we show that we cannot have $y_A < y_E$. If $y_A < y_E$, equation (D.4) reduces to

$$-1 \leq (y_F - y_E) + |x_A - x_F| - |x_A - x_E| \leq 1. \tag{D.5}$$

Next, by the triangular inequality we have

$$-(x_E - x_F) \leq |x_A - x_F| - |x_A - x_E| \leq (x_E - x_F),$$

and therefore by Definition 7.4.2,

$$\text{Gain}' + 1 \leq (y_F - y_E) + |x_A - x_F| - |x_A - x_E| \tag{D.6}$$

Due to the assumption $\text{Gain}' \geq 2$, equations (D.5) and (D.6) contradict each other. Hence, we cannot have $y_A < y_E$. Similarly it can be shown that we cannot have $y_A > y_F$. In short, for A to be a normal point, we must have $y_E \leq y_A \leq y_F$ (i.e., the y_A coordinate must be between E and F). This reduces equation (D.4) to

$$-1 \leq |x_A - x_F| + y_F - y_A - |x_A - x_E| - y_A + y_E \leq 1 \tag{D.7}$$

Now we are going to have three cases depending on whether $x_A < x_F$, $x_A > x_E$ or $x_F \leq x_A \leq x_E$. When $x_A < x_F < x_E$, equation (D.7) reduces to

$$\alpha - 1 = \frac{1 + y_F + y_E + x_F - x_E}{2} - 1 \leq y_A \leq \frac{1 + y_F + y_E + x_F - x_E}{2} = \alpha,$$

where α is given in Definition D.2.1. This gives the first line of equation (D.3). Similarly, it can be verified that for the other two possibilities $x_F \le x_A \le x_E$ and $x_E < x_A$, we get the remaining two lines. □

Now that N is explicitly written in terms of the coordinates of the nodes, we can leverage the partitioning in Claim 7.2.1 to do the same for R_E and R_F. However, we find it more instructive to express R_E and R_F implicitly using N, instead of explicit equations similar to equation (D.3).

Remark D.2.2. *Under Assumptions 7.4.2 and 7.4.3,*

$$R_F = \{(x,y) \notin N \mid \exists k \in \mathbb{N} \textit{ with } (x, y+k) \in N\}, \tag{D.8}$$

and

$$R_E = \{(x,y) \notin N \mid \exists k \in \mathbb{N} \textit{ with } (x, y-k) \in N\}. \tag{D.9}$$

Proof. By Claim D.2.1, the normal region splits V into two connected components, one containing E, which we denote by V_F, and one containing F, which we denote by V_E. Now we show that $V_E = R_E$ and $V_F = R_F$. By Claim 7.2.1 it is clear that we cannot have two neighboring vertices A, B with $A \in R_E$ and $B \in R_F$. Indeed the equations $AE - AF > 1$, $BE - BF < -1$, $|AE - BE| \le 1$ and $|AF - BF| \le 1$ cannot hold at the same time. By Claim 7.2.1, the vertices $V \setminus N$ are partitioned into R_E and R_F, and since we cannot have two neighboring vertices split between R_E and R_F, each connected component V_E and V_F must be contained entirely in R_E or R_F. We also know that $E \in R_F$ and $F \in R_E$, which implies that the only way to assign the vertices of $V \setminus N$ into R_E and R_F is to have $R_E = V_E$ and $R_F = V_F$. □

In the next claim, we characterize the special regions of the nodes on the boundary PS of the grid. This will be useful in the subsequent results as we will be mainly dealing with nodes on the boundaries.

Claim D.2.2. *Let $A = (1,k)$ be a point on boundary PS, and* $\mathrm{Gain}_{\max}(A)$ *given in Remark 7.2.4. Then, under Assumptions 7.4.2 and 7.4.3,*

1. *if A belongs to R_E, i.e., $k > \alpha$, with α given in Definition D.2.1,*

$$\begin{aligned} R_A = &\{(x,y) \mid x_E \le x, y \le y_E\} \cup \\ &\left\{(x,y) \;\middle|\; x_E \le x, 0 \le y - y_E < \frac{\mathrm{Gain}_{\max}(A)}{2}\right\} \cup \\ &\left\{(x,y) \;\middle|\; y \le y_E, 0 \le x_E - x < \frac{\mathrm{Gain}_{\max}(A)}{2}\right\} \cup \\ &\left\{(x,y) \;\middle|\; x \le x_E, y_E \le y, (x_E - x) + (y - y_E) < \frac{\mathrm{Gain}_{\max}(A)}{2}\right\}, \end{aligned} \tag{D.10}$$

and

$$\text{Gain}_{\max}(A) = \begin{cases} \text{Gain} & \textit{for } k \geq y_F \\ \text{Gain} - 2(y_F - k) & \textit{for } \alpha < k < y_F \end{cases} \tag{D.11}$$

2. *if A belongs to R_F i.e.* $k < \alpha - 1$ *and* $\text{Gain}' \geq 2$*,*

$$\begin{aligned} R_A = &\{(x,y) \mid x_F \leq x, y_F \leq y\} \cup \\ &\left\{(x,y) \;\middle|\; x_F \leq x, 0 \leq y_F - y < \frac{\text{Gain}_{\max}(A)}{2}\right\} \cup \\ &\left\{(x,y) \;\middle|\; y_F \leq y, 0 \leq x_F - x < \frac{\text{Gain}_{\max}(A)}{2}\right\} \cup \\ &\left\{(x,y) \;\middle|\; x \leq x_F, y \leq y_F, (x_F - x) + (y_F - y) < \frac{\text{Gain}_{\max}(A)}{2}\right\}, \end{aligned} \tag{D.12}$$

and

$$\text{Gain}_{\max}(A) = \begin{cases} \text{Gain}' & \textit{for } k \leq y_E \\ \text{Gain}' - 2(k - y_E) & \textit{for } y_E < k < \alpha - 1. \end{cases} \tag{D.13}$$

By symmetry, the vertices $A = (n, k)$ on boundary QR have a similar expression for their special region, however, we do not include this in this dissertation in the interest of space.

We will only cover the $A \in R_F$ case; the other case is analogous. We are interested in the nodes $T \in R_A$, i.e., nodes that use edge e to reach A. By Remark 7.2.4, vertex E gets the maximum benefit from the extra edge, hence we expect R_A to be a neighbourhood "centered" at E. However, R_A cannot be a ball centered at E, because the directions are not equivalent. For instance, if T is in the rectangle formed by E and Q, then we can go to node E for "free", without sacrificing any of the gain we get by using the extra edge. This is because the shortest path from T to A in G passed through E anyways, so $\text{Gain}(A, T) = \text{Gain}_{\max}(A)$ (i.e., T also gets maximum benefit). Hence, all nodes in this rectangle will be in R_A. For a different example, if T is in the rectangle formed by E and P, then going along the y axis towards E is "free", but going along the x axis towards is a "detour", hence there may be a threshold for x_T below which the shortest path does not use the extra edge. We will make this intuition rigorous below.

Proof. First, we check that equation (D.11) agrees with the definition of $\text{Gain}_{\max}$. By Remark 7.2.4, we have

$$\text{Gain}_{\max}(A) = AE - (1 + AF), \tag{D.14}$$

which for $k \geq y_F$ implies

$$\text{Gain}_{\max}(A) = (x_E - 1) + (k - y_E) - (1 + (x_F - 1) + (k - y_F)) = y_F - y_E + x_E - x_F - 1 = \text{Gain}$$

because of Definition 7.2.3, and for $\alpha < k < y_F$ implies

$$\text{Gain}_{\max}(A) = (x_E - 1) + (k - y_E) - (1 + (x_F - 1) + (y_F - k)) = \text{Gain} - 2(y_F - k).$$

Next, we need to find nodes T such that

$$TE + 1 + FA < AT. \tag{D.15}$$

Combining equations (D.14) and (D.15) we get that

$$\begin{gathered} TE + AE - \text{Gain}_{\max}(A) < AT \\ TE_x + TE_y + AE_x + AE_y - \text{Gain}_{\max}(A) < AT_x + AT_y, \end{gathered} \tag{D.16}$$

where TE_x and TE_y denote the distance along the x and y axes, respectively (e.g., $TE_x = |x_T - x_E|$).

There are five cases depending on where T could be:

Case 1: T is in the rectangle formed by nodes E and Q

In this case there exists a shortest path in grid from T to A which passes through E, and T will certainly use the edge e to reach A. Hence, this rectangle belongs to R_A, which accounts for the first line of in equation (D.10).

Case 2: T is in the rectangle formed by E and P

In this case $AE_x = AT_x + TE_x$ and $AE_y = AT_y - TE_y$, which reduces equation (D.16) to

$$TE_x < \frac{\text{Gain}_{\max}(A)}{2}.$$

This accounts for the second set in the equation (D.10).

Case 3: T is in the rectangle formed by E and R, and has y-coordinate less than k

In this case $AE_x = AT_x - TE_x$ and $AE_y = TE_y + AT_y$, which reduces equation (D.16) to

$$TE_y < \frac{\text{Gain}_{\max}(A)}{2}.$$

This accounts for the third set in the (D.10).

Case 4: T is in the rectangle formed by E and S, and has y-coordinate less than k

In this case $AE_x = AT_x + TE_x$ and $AE_y = TE_y + AT_y$, which reduces equation (D.16) to

$$TE_x + TE_y < \frac{\text{Gain}_{\max}(A)}{2}.$$

This accounts for the fourth set in the (D.10).

Case 5: T has y-coordinate greater than or equal to k

In this case $TE_y = AE_y + AT_y$, which reduces equation (D.16) to

$$2AE_y - \text{Gain}_{\max}(A) < AT_x - (AE_x + TE_x) \leq 0, \tag{D.17}$$

where the last inequality follows from the triangle inequality. However, using equation (D.11), for $k \geq y_F$ we have

$$\begin{aligned} 2AE_y - \text{Gain}_{\max}(A) &= 2(k - y_E) - \text{Gain} \\ &\geq (y_F - y_E) - (x_E - x_F) + 1 \\ &= \text{Gain}' + 2 > 0, \end{aligned} \tag{D.18}$$

and for $k < y_F$ we have

$$\begin{aligned} 2AE_y - \text{Gain}_{\max}(A) &= 2(k - y_E) - \text{Gain} + 2(y_F - k) \\ &= (y_F - y_E) - (x_E - x_F) + 1 \\ &= \text{Gain}' + 2 > 0, \end{aligned} \tag{D.19}$$

which contradicts equation (D.17). Therefore Case 5 is impossible.

Since the five cases cover the entire node set, the necessary and sufficient conditions for $T \in R_A$ are characterized, and this completes the proof. □

D.3 Technical Lemmas for the Proof of Theorem 7.4.3

Lemma D.3.1. *Under Assumptions 7.4.2 and 7.4.3, the metric dimension of G' is at least 3.*

Proof. Suppose that there exist two points X and Y that distinguish all points in the grid. By Claim D.1.1, they have to be on opposite boundaries. Next, their maximum gains cannot exceed 1. Indeed, suppose for contradiction that $\text{Gain}_{\max}(X) > 1$, and that $X \in R_F$. Then, by Remark 7.2.4 we have $XF - (1 + XE) > 1$, and thus the four neighboring vertices of F will all have distance $\min(XF \pm 1, XE + 2) = XE + 2$ to X. By Remark D.1.1, the four neighboring vertices of F cannot be distinguished by a single vertex Y, which contradicts our assumption that $\{X, Y\}$ is a resolving set, and hence we must have $\text{Gain}_{\max}(X) \leq 1$. By a symmetric argument, we also have $\text{Gain}_{\max}(Y) \leq 1$

We have two cases depending on the parity of Gain'.

Case 1: Gain' is even.

By Claim D.1.2, we know that $\text{Gain}_{\max}(X)$ is also even, and since $\text{Gain}_{\max}(X) \leq 1$, it must equal to 0, which in turn implies that X is a normal vertex. By a symmetric argument, Y must be normal vertex too. Moreover, recall that X and Y must lie on opposite boundaries. Therefore, because of Claim D.2.1, X is either $(1, \alpha - 1)$ or $(1, \alpha)$ and Y is either $(n, \beta - 1)$ or (n, β), as these are the only normal vertices on the boundaries of G'. As X and Y are normal vertices, edge e has no effect on the distances from any vertex of G' to X and Y, which therefore remain the same as in the original grid G. But we know that the only resolving sets of the grid G that have cardinality 2 are two adjacent corners of G, which disqualifies X and Y from being a resolving set of G and thus G'.

Case 2: Gain$'$ is odd.

Recall, that $\text{Gain}_{\max}(X) \leq 1$, $\text{Gain}_{\max}(Y) \leq 1$, and both X and Y must lie on the boundary of G'. Let us first rule out the possibility of X or Y being on the top/bottom boundaries PQ and RS. More specifically, we will show that there is no point $X = (k,1)$ with $\text{Gain}_{\max}(X) \leq 1$. If $X = (k,1)$, then $X \in R_F$ and

$$\begin{aligned} \text{Gain}_{\max}(X) &= XF - XE - 1 \\ &= |k - x_F| + y_F - 1 - |k - x_E| - (y_E - 1) \\ &= (|k - x_F| - |k - x_E|) + y_F - y_E - 1 \\ &\geq -(x_E - x_F) + y_F - y_E - 1 \\ &= \text{Gain}', \end{aligned} \tag{D.20}$$

where the inequality follows from the triangle inequality. Now, Assumption 7.4.3 states that $\text{Gain}' \geq 2$, and thus no point $X = (k,1)$ can have $\text{Gain}_{\max}(X) \leq 1$.

Now we consider the case when X and Y lie on PS and QR, respectively. We will check which vertices $X = (1,k)$ and $Y = (n,k)$ have $\text{Gain}_{\max} \leq 1$. The $\text{Gain}_{\max}$ of vertices $X = (1,k)$ is expressed in equation (D.11). Since $\text{Gain} > \text{Gain}' \geq 2$, only the $\text{Gain} - 2(y_F - k)$ term can equal 1. The term $\text{Gain} - 2(y_F - k)$ is an increasing linear function of k that takes the value 1 for only a single value of k, namely $k = \alpha + 1/2 = \lfloor \alpha \rfloor + 1$. Similarly, in (D.13), the only value that satisfies $\text{Gain}' - 2(k - y_E) = 1$ is $k = \lfloor \alpha \rfloor - 1$. Consequently, the only vertices on PS that have $\text{Gain}_{\max}(X) = 1$ are $X_1 = (1, \lfloor \alpha \rfloor - 1)$ and $X_3 = (1, \lfloor \alpha \rfloor + 1)$. Similarly, the only vertices on QR that have $\text{Gain}_{\max}(X) = 1$ are $Y_1 = (n, \lfloor \beta \rfloor - 1)$ and $Y_3 = (n, \lfloor \beta \rfloor + 1)$. The only vertices on PS and QR that have $\text{Gain}_{\max}(X) = 0$ are the normal vertices $X_2 = (1, \lfloor \alpha \rfloor)$ and $Y_2 = (n, \lfloor \beta \rfloor)$. Hence, we have

$$X \in \{X_1 = (1, \lfloor \alpha \rfloor - 1), X_2 = (1, \lfloor \alpha \rfloor), X_3 = (1, \lfloor \alpha \rfloor + 1)\},$$

and,

$$Y \in \{Y_1 = (n, \lfloor \beta \rfloor - 1), Y_2 = (n, \lfloor \beta \rfloor), Y_3 = (n, \lfloor \beta \rfloor + 1)\}.$$

To finish the proof, we are going to rule out the remaining nine resolving sets that can be formed by X_1, X_2, X_3 and Y_1, Y_2, Y_3. Since $\text{Gain}_{\max}(X_1) = \text{Gain}_{\max}(X_3) = 1$, the expression for the special regions of X_1 and X_3 in Claim D.2.2 simplifies to

$$R_{X_1} = \{(x,y) \mid x_F \leq x, y_F \leq y\} \tag{D.21}$$

and

$$R_{X_3} = \{(x,y) \mid x_E \leq x, y \leq y_E\}. \tag{D.22}$$

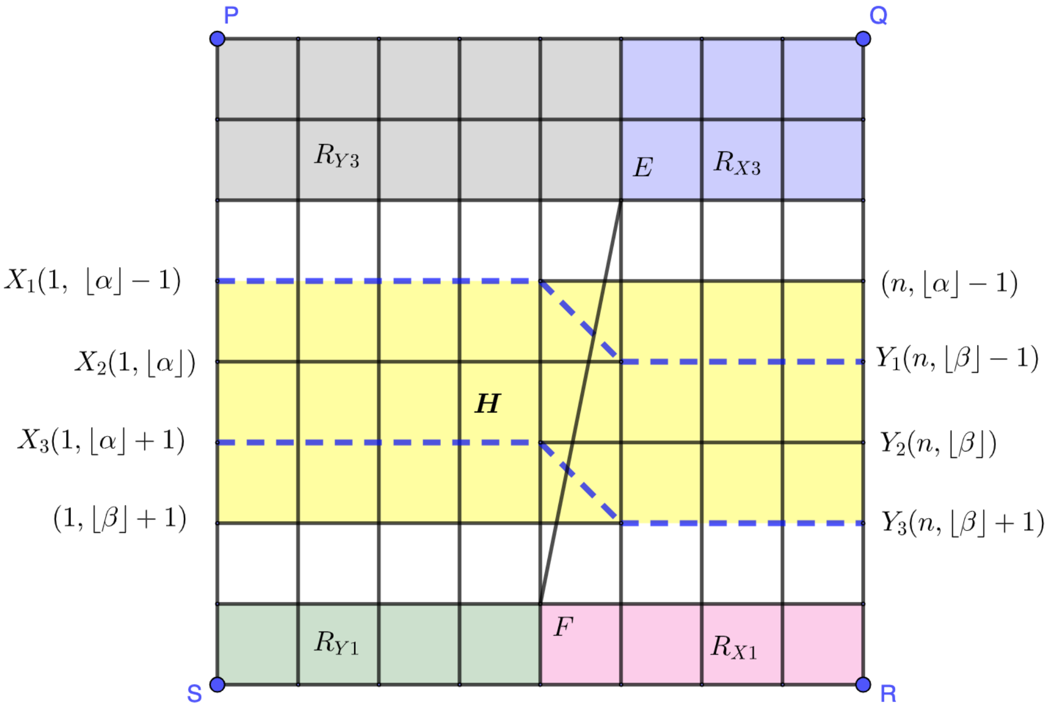

P
Q
S
R
R_{Y3}
E
R_{X3}
$X_1(1,\ \lfloor\alpha\rfloor-1)$
$(n, \lfloor\alpha\rfloor-1)$
$X_2(1, \lfloor\alpha\rfloor)$
$Y_1(n, \lfloor\beta\rfloor-1)$
H
$X_3(1, \lfloor\alpha\rfloor+1)$
$Y_2(n, \lfloor\beta\rfloor)$
$(1, \lfloor\beta\rfloor+1)$
$Y_3(n, \lfloor\beta\rfloor+1)$
R_{Y1}
F
R_{X1}

Figure D.2: Figure for the proof of Lemma D.3.1 for the case when Gain′ is odd. Note that sub-grid H does not intersect with any R_{Y_i} or any R_{X_i} for $i = 1,3$.

By a symmetric argument,

$$R_{Y_1} = \{(x, y) \mid x \le x_F, y_F \le y\} \tag{D.23}$$

and

$$R_{Y_3} = \{(x, y) \mid x \le x_E, y \le y_E\}. \tag{D.24}$$

Consider the rectangular sub-grid H formed by points X_1, $(n, \lfloor\alpha\rfloor-1)$, Y_3, $(1, \lfloor\beta\rfloor+1)$ (see Figure D.2). The sub-grid H cannot intersect the special regions of X_1, X_3, Y_1 and Y_3 because (i) by equations (D.21) and (D.23), the special regions of X_1 and Y_1 have y-coordinate at least y_F, (ii) by equations (D.22) and (D.24), the special regions of X_3 and Y_3 have y-coordinate at most y_E, and (iii) the sub-grid H has y-coordinates more than y_E and less than y_F. The statement (iii) follows by equation (D.1) and the inequality $\text{Gain}' \ge 2$ from Assumption 7.4.3, since the lowest y-coordinate value of a vertex in H is

$$\lfloor\alpha\rfloor - 1 = \left\lfloor y_E + \frac{\text{Gain}'+2}{2} \right\rfloor - 1 \ge y_E + \frac{\text{Gain}'+1}{2} - 1 > y_E,$$

and by equation (D.2) and the inequality $\text{Gain}' \ge 3$ (which follows from the assumption that Gain′ is odd in addition to Assumption 7.4.3), since the highest y-coordinate value of a vertex in H is

$$\lfloor\beta\rfloor + 1 = \left\lfloor y_F - \frac{\text{Gain}'}{2} \right\rfloor - 1 \leq y_F - \frac{\text{Gain}'}{2} - 1 < y_F. \tag{D.25}$$

Consequently, the distances in graph G' between any point in H and X or Y are same as in G. By Remark D.2.1, we also have $\lfloor\alpha\rfloor < \lfloor\beta\rfloor$, which implies that X and Y cannot be adjacent corners of H, and they cannot resolve the sub-grid H.

Since we ruled out every pair of vertices X, Y for being a resolving set, the proof is concluded. □

Lemma D.3.2. *Under Assumptions 7.4.2 and 7.4.3, if* Gain$'$ *is odd, the set* $\{X = (1, \lfloor\beta\rfloor), Y = (n, \lfloor\beta\rfloor), Q = (n, 1)\}$ *is a resolving set in* G'.

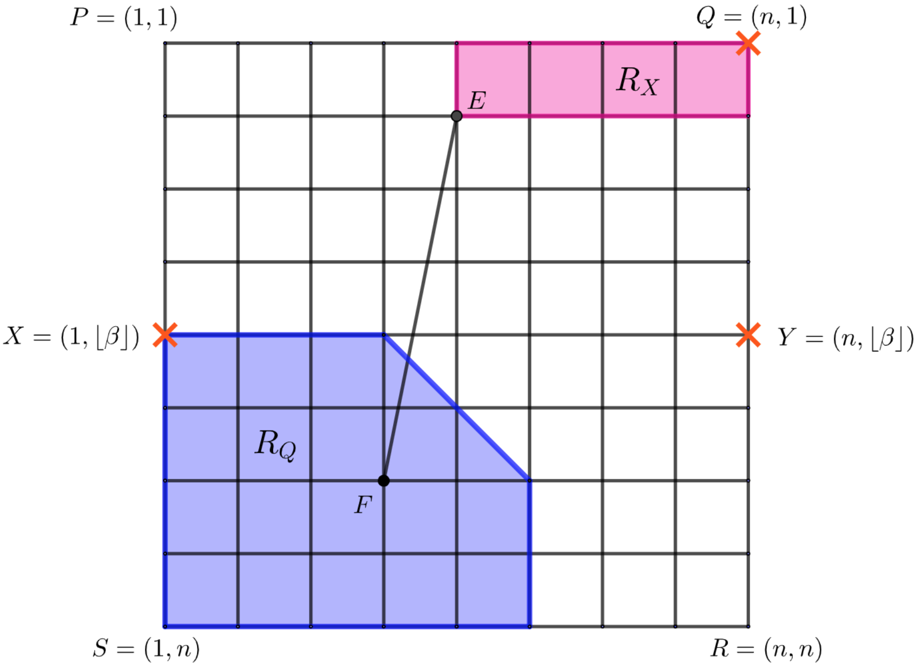

$P = (1,1)$
$Q = (n,1)$
R_X
E
$X = (1, \lfloor\beta\rfloor)$
$Y = (n, \lfloor\beta\rfloor)$
R_Q
F
$S = (1,n)$
$R = (n,n)$

Figure D.3: This is the illustration for the proof of Lemma D.3.2. Points X, Y, Q marked with red cross form a resolving set. Y is a normal point. Blue and pink regions(boundaries included) are special regions of Q and X, respectively.

Proof. By Claim D.2.1, Y is a normal vertex. The only normal vertex on boundary PS is vertex $(1, \lfloor\alpha\rfloor)$, and since by Remark D.2.1 we have $\lfloor\beta\rfloor > \lfloor\alpha\rfloor$, X cannot be a normal vertex. By Claim D.2.2 we have $X \in R_E$, and by Remark 7.2.3, vertex X has non-empty special region $R_X \subseteq R_F$ (see the pink region in Figure D.3).

Suppose for contradiction that there exist two distinct points A and B, which are not distinguished by the three points X, Y, Q in G'. We separate three cases depending on the position of A and B:

Case 1: One of A and B is in R_X, and the other is in N_X

Without loss of generality, we assume $A \in R_X$ and $B \in N_X$.

Since Y is a normal point and it does not distinguish $A = (x_A, y_A)$ and $B = (x_B, y_B)$, we have $AY = BY$, which can be expanded as

$$n - x_A + |y_A - \lfloor\beta\rfloor| = n - x_B + |y_B - \lfloor\beta\rfloor|,$$

whence

$$|y_A - \lfloor\beta\rfloor| - |y_B - \lfloor\beta\rfloor| = x_A - x_B. \tag{D.26}$$

By the assumption that A and B are not distinguished by X, we have that $d_{G'}(A, X) = d_{G'}(B, X)$. Since $A \in R_X$ and $B \in N_X$, this yields that

$$AX - \text{Gain}(A, X) = BX. \tag{D.27}$$

Therefore,

$$\begin{aligned}\text{Gain}(A, X) &= AX - BX \\ &= (x_A - 1) + |y_A - \lfloor\beta\rfloor| - (x_B - 1) - |y_B - \lfloor\beta\rfloor| \\ &= 2(x_A - x_B),\end{aligned} \tag{D.28}$$

where the last line follows form equation (D.26). Equation (D.28) implies that Gain(A, X) must be even. By Claim D.1.2, if Gain(A, X) is even then Gain$'$ must be even too, which contradicts our assumption that Gain$'$ is odd.

Case 2: $A, B \in N_X$

In this case, the distances between X, Y and A, B are the same in graph G' as in G, which implies that A, B are not distinguished by X nor Y in G. The only pairs of vertices that are not distinguished by X, Y in the grid G are vertices that are symmetric to the horizontal line passing through X and Y. Therefore A, B must be such a pair. By a similar parity based argument as in Case 1, if one of A and B is in R_Q and the other is not, then they are distinguished by either Y or Q. Indeed, substituting Q instead of X into equations (D.27) and (D.26), we get

$$\begin{aligned}\text{Gain}(A, Q) &\overset{\text{(D.27)}}{=} AQ - BQ \\ &= n - x_A + y_A - 1 - (n - x_B) - (y_B - 1) \\ &\overset{\text{(D.26)}}{=} |y_B - \lfloor\beta\rfloor| - |y_A - \lfloor\beta\rfloor| + y_A - y_B \\ &\equiv 0 \pmod 2.\end{aligned} \tag{D.29}$$

Then, Gain$'$ should also be even by Claim D.1.2, contradicting our assumption that Gain$'$ is odd.

We are left with the cases $A, B \in N_Q$ and $A, B \in R_Q$. Notice that since we showed $\lfloor\beta\rfloor + 1 < y_F$ for odd Gain$'$ in equation (D.25), and since by equation (D.2) we have $\lfloor\beta\rfloor = \lfloor y_E + (\text{Gain}+2)/2\rfloor > 1$, neither F nor Q are on the horizontal line through X and Y. Hence, any pair of nodes A, B that are symmetric to the XY line are distinguished by both Q and F in graph G. We immediately see that if $A, B \in N_Q$, the pair A, B is also by Q in G'. If $A, B \in R_Q$, by Claim 7.2.3 together with $Q \in R_F$, and since F distinguishes A, B in G, we have

$$d_{G'}(Q, A) = QE + 1 + FA \neq QE + 1 + FB = d_{G'}(Q, B).$$

Hence, in every sub-case of Case 2 we showed that A, B must be distinguished by at least one of Q, X and Y in G'.

Case 3: $A, B \in R_X$:

By Claim D.2.2, we have $Q \in R_X$. The anti-transitivity property of special regions (Remark 7.2.2) implies that if $A \in R_X$ and $Q \in R_X$, then $A \notin R_Q$ and therefore $A \in N_Q$. Similarly, we have $B \in N_Q$, and we can deduce that the distances between Q and A, B are the same in graph G' as in graph G. Moreover, since Y is a normal vertex, the distances between Y and A, B are the same in G' as in G too.

Remark 7.2.3 together with $X \in R_E$ implies that we have $R_X \subseteq R_F$, and Remark D.2.2 and Claim D.2.1 together imply that every vertex in R_F has y-coordinate at most $\beta - 1 < \lfloor\beta\rfloor$. Hence, both A and B are contained in the rectangular sub-grid with corners $QYXP$. Since Q and Y are adjacent corners of the sub-grid $QYXP$, they must resolve the entire sub-grid $QYXP$ in graph G, including vertices A and B. Since distances from Y and Q to A, B are the same in graph G' as in G, vertices Q and Y must distinguish A and B in G' as well.

Thus, every vertex pair A, B is distinguished by some vertex in the set $\{X, Y, Q\}$, and the proof is concluded. □

Lemma D.3.3. *Under Assumptions 7.4.2 and 7.4.3, if* Gain$'$ *is even and* $x_E - x_F < \frac{\text{Gain}'}{2} + 2$, *the set* $\{X = (1, \beta - 1), Y = (n, \beta - 1), Z = (1, \alpha - 1)\}$ *is a resolving set in* G'.

Proof of Lemma D.3.3. See Figure D.4 for an illustration. Note that Y and Z are normal points, and that $X \in R_E$. First we calculate $\text{Gain}_{\max}(X)$. By equation (D.11), since $\beta - 1 \leq y_F$ because of equation (D.2),

$$\begin{aligned} \text{Gain}_{\max}(X) &= \text{Gain} - 2(y_F - \beta + 1) \\ &= 2(x_E - x_F) - 2. \end{aligned} \tag{D.30}$$

Now we show that R_X completely lies inside the rectangle $PQTZ$. Indeed, according to Claim

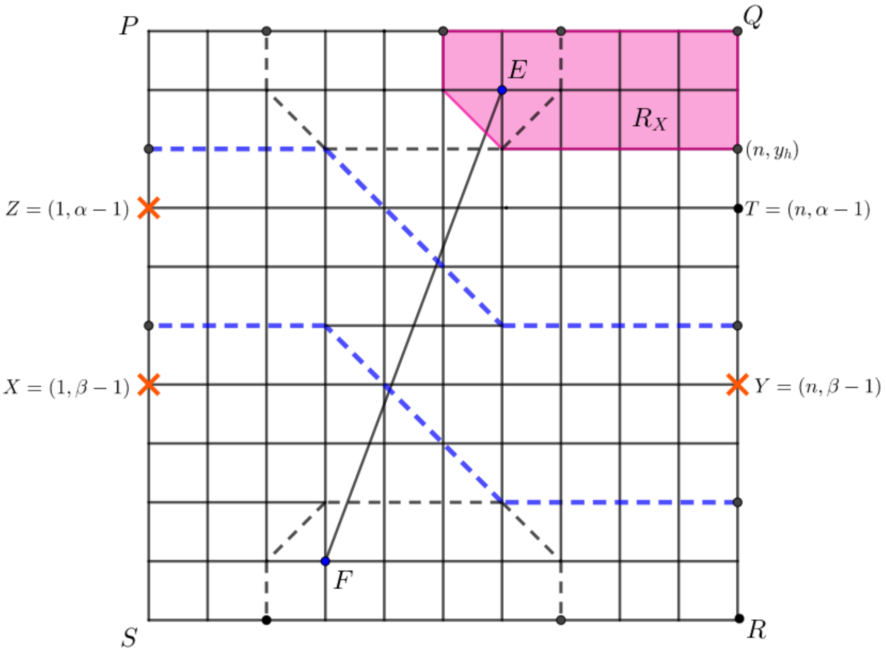


Figure D.4: Illustration for the proof of Lemma D.3.3. Points X, Y, Z marked with red crosses form a resolving set. Y and Z are normal points, the pink region(including the boundary) is R_X.

D.2.2, the largest y-coordinate of a point in special region of X will be

$$\begin{aligned} y_{\max} &= y_E + \frac{\text{Gain}_{\max}(X)}{2} - 1 \\ &\overset{(D.30)}{=} y_E + \frac{2(x_E - x_F) - 2}{2} - 1 \\ &= \frac{1 + y_F + y_E + x_F - x_E}{2} - 1 - \frac{(y_F - y_E) - (x_E - x_F) - 1}{2} + (x_E - x_F) - 2 \\ &= \alpha - 1 - \frac{\text{Gain}'}{2} + (x_E - x_F) - 2 \\ &< \alpha - 1, \end{aligned} \tag{D.31}$$

where the inequality follows by the assumption $x_E - x_F < \text{Gain}'/2 + 2$. Hence, since we also have $\alpha < \beta$ by Remark D.2.1, all points in the special region of X will have y-coordinate less than that of Z. Alternatively, denoting vertex $(n, \alpha - 1)$ by T, we have that R_X is contained in the rectangle $PQTZ$.

Let us suppose for contradiction that there exist two distinct points $A = (x_A, y_A)$ and $B = (x_B, y_B)$ which are not distinguished by X, Y, Z. We distinguish three cases based on the positions of A and B:

Case 1: $A, B \in N_X$

In this case all distances between X, Y, Z and A, B are the same in graph G' as in graph G. It is easy to see that to be equidistant from X and Y, vertices A and B must be symmetric to the horizontal line through X and Y, in which case Z can distinguish A and B.

Case 2: $A, B \in R_X$

In this case, we show that A and B cannot be equidistant from both Y and Z. Both A and B lie inside of R_X, and thus the region $PQTZ$. Now we show that Y and Z resolve $PQTZ$ in G, which implies that they resolve $PQYZ$ in G' because they are normal vertices. Our argument will be similar to the standard argument that shows that two adjacent corners resolve the grid. To be equidistant from Z, both of them should lie on a diagonal line parallel to PR, or equivalently,

$$x_A - y_A = x_B - y_B. \tag{D.32}$$

To be equidistant from Y, they should lie on a diagonal line parallel to QS, or equivalently,

$$x_A + y_A = x_B + y_B. \tag{D.33}$$

However, equations (D.32) and (D.33) cannot hold simultaneously for $A \neq B$.

Case 3: One of A and B is in R_X, and the other is in N_X

Without loss of generality, we assume that $A \in R_X$ and $B \in N_X$. Since R_X lies inside of $PQTZ$, we know that the y-coordinate of A is less than that of Z and Y, i.e., $y_A < \alpha - 1 < \beta - 1$. Since we have shown in Case 2 that Y and Z resolve $PQTZ$ in G', B cannot lie in the region $PQTZ$. There are two other possibilities for where B could lie:

1. Let us assume that B lies in the region $ZTYX$. Since Y does not distinguish A and B, we have $AY = BY$, which implies that $x_A + y_A = x_B + y_B$ as in equation (D.33). Similarly, since Z does not distinguish A and B, we have $AZ = BZ$, which implies that $x_A + (\alpha - 1 - y_A) = x_B + y_B - (\alpha - 1)$. Subtracting the second equation from the first gives $y_A = \alpha - 1$ which contradicts equation (D.31).

2. Let us assume that B lies in the region $XYRS$, or equivalently, $y_B \geq \beta - 1$. Since we have assumed A and B to be equidistant from X and Z, we have $AZ = BZ$ and $BX = d_{G'}(A, X) = AX - \text{Gain}(A, X)$. Writing these equations in terms of the variables x_A, y_A, x_B and y_B gives

$$x_A - 1 + (\alpha - 1 - y_A) = x_B - 1 + y_B - (\alpha - 1), \tag{D.34}$$

and

$$x_A - 1 + (\beta - 1 - y_A) - \text{Gain}(A, X) = x_B - 1 + y_B - (\beta - 1). \tag{D.35}$$

Subtracting equation (D.35) from equation (D.34) yields

$$\text{Gain}(A, X) = 2(\beta - \alpha) = 2(x_E - x_F). \tag{D.36}$$

Equations (D.36) and (D.30) together contradict the fact that $\mathrm{Gain}(A, X) \le \mathrm{Gain}_{\max}(X)$.

We considered all cases and the proof is concluded. □

E Appendix for Chapter 8

E.1 Additional Proofs

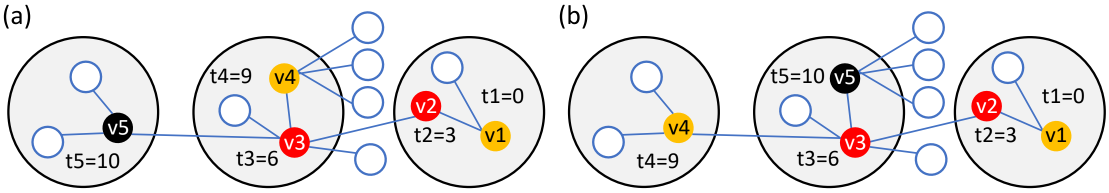


Figure E.1: Illustration for Lemma 8.4.1 using the same coloring as Figure 8.2 (a). (a) An example for an epidemic where among the nodes of the transmission path (v_1, v_2, v_3, v_5), the middle household contains no symptomatic node (only the asymptomatic node v_3), but the LS+ algorithm still succeeds. Indeed, at iteration 0 we set $s_{c,0} = v_5$, after which we find that v_3 is asymptomatic, and next that v_2 is asymptomatic and v_4 is symptomatic, with a lower symptom onset time then v_5. Hence, in iteration 1 we set $s_{c,1} = v_4$, and we find that v_3, v_2 are asymptomatic and v_1 is is symptomatic, with a lower symptom onset time then v_4. Finally, in iteration 2 we set $s_{c,2} = v_1$, and we find $s'_c = v_1 = s_{c,2}$, which implies that the algorithm stops, and returns the correct source v_1. (b) An example for an epidemic where the LS+ algorithm would fail if we would update the candidate before the test queue becomes empty. Similarly to subfigure (a), in iteration 0 of the algorithm first learns about asymptomatic node v_3 and next about asymptomatic node v_2 and symptomatic node v_4. If the algorithm updates the candidate to v_4 and continues further, instead of scheduling the tests of the household members of v_2, then it is not hard to check that v_4 will be the final estimate and the algorithm fails. However, if the algorithm waits until the test queue becomes empty and tests the household members of v_2, then v_1 becomes the next candidate and the algorithm succeeds.

E.1.1 Proof of Lemma 8.4.1

We start by restating the lemma here for convenience.

Lemma 8.4.1. *In the RB tree network, the LS algorithm succeeds if and only if all nodes on the transmission path are symptomatic, and the LS+ algorithm succeeds if among the nodes of the transmission path, there exists a symptomatic node in each household, and the source is symptomatic.*

Proof. Throughout the proof we assume that there is no limitation on the number available tests. We can make this assumption because in the SICTF there is only a daily limit on the number tests, there is no limitation on the number of days, and neither the LS nor the LS+ algorithms proceed in an iteration until the test queue becomes empty, which implies that all nodes that enter the test queue get eventually tested.

Suppose that the LS algorithm succeeds. Then the list of candidate nodes s_c at different iterations forms a path that consists entirely of symptomatic nodes between the source and the first hospitalized node. In tree networks, the transmission path is the only path between the source and the first hospitalized node, which yields the "only if" part of the statement on the LS algorithm.

Next, suppose that all nodes on the transmission path are symptomatic. Then, we claim that the candidate node $s_{c,i}$ computed in the i^{th} iteration of the LS algorithm is v_{l-i}, the i^{th} node of the reverse transmission path. Our claim is definitely true for $i = 0$, because $s_{c,0}$ is initialized to be the first hospitalized node v_l. Then, the proof proceeds by induction. By the induction hypothesis, in the i^{th} step, $s_{c,i} = v_{l-i}$, and since we are on a tree, the symptom onset time of $v_{l-(i+1)}$ (which is revealed because all nodes on the transmission path are symptomatic by assumption) is the only symptom onset time among the neighbors of $s_{c,i}$ that have a lower symptom onset time than $s_{c,i}$ itself. Therefore $s'_c = v_{l-(i+1)}$, and $s_{c,i+1}$ is updated to be $v_{l-(i+1)}$ in the beginning of the next iteration, which proves that the induction hypothesis holds until the source is reached.

Finally, suppose that among the nodes of the transmission path, there exists a symptomatic node in each household, and the source is symptomatic. Let us denote by w_i the i^{th} symptomatic node of the *reverse* transmission path. Then, we claim that the candidate list $s_{c,i}$ computed in the i^{th} iteration of the LS+ algorithm equals w_i. Similarly to the case of the LS algorithm, the $i = 0$ case holds by definition, and we proceed by induction. Suppose that $s_{c,i} = w_i$. It will also be useful to define the index of w_i on the *forward* transmission path (without skipping asymptomatic nodes). Let j be this index, for which therefore $w_i = v_j$. Now we distinguish 3 cases: (i) $v_{j-1} = w_{i+1}$ is symptomatic, (ii) v_{j-1} is asymptomatic and $v_{j-2} = w_{i+1}$ is symptomatic, and (iii) v_{j-1} and v_{j-2} are asymptomatic and $v_{j-3} = w_{i+1}$ is symptomatic. We claim that there are no more cases, and that in all three cases w_{i+1} is tested in the i^{th} iteration of the LS+ algorithm. Case (i) is immediate because all neighbors of $s_{c,i}$ are tested. Case (ii) is only possible if either $v_{j-1} \in H(s_{c,i})$ or $v_{j-2} \in H(v_{j-1})$, otherwise v_{j-1} would be a lone asymptomatic node in a household, which contradicts the assumption that there is a symptomatic node in each household. Since all the contacts of asymptomatic nodes in $H(s_{c,i})$ (see Figure 8.3 (d)) and all nodes in the household of asymptomatic nodes are tested in the LS+ algorithm (see Figure 8.3 (e)), v_{j-2} must be tested too. Finally, case (iii) is possible only if $v_{j-1} \in H(s_{c,i})$ and $v_{j-3} \in H(v_{j-2})$ both hold, otherwise v_{j-1} or v_{j-2} would be a lone asymptomatic node in a household. Similarly to the previous case, v_{j-3} must be tested (see Figure 8.3 (f)). There are no more cases because, by Remark 8.4.2, on the RB tree a transmission path can only have two nodes in each household, and we assumed that there exists a symptomatic node in each household among the nodes of the transmission path.

After we proved that w_{i+1} is tested in the i^{th} iteration of the LS+ algorithm, we must still show that it will be the next candidate $s_{c,i+1}$ for the induction hypothesis to hold. This is true because once the symptom onset time of w_{i+1} is revealed, none of its neighbors are scheduled for testing, and therefore all tested nodes have w_{i+1} on their path to the source, which means that w_{i+1} must have the lowest revealed symptom onset time, and therefore that it will be the next candidate $s_{c,i+1}$. □

E.1.2 Proof of Theorem 8.4.1

We are going to need prove a few intermediate results before proving Theorem 8.4.1. A first step is to count all the possible paths from the source with a given length.

Definition E.1.1. *Let $G(s)$ be the RB tree with parameters (d_c, d_h), and let s be the source. A Red-Blue (RB) path of length n is any path of nodes in $(s = v_0, v_1, ...v_n)$ such that $(v_i, v_{i+1}) \in E'$ for $0 \le i < n$. Let $\mathcal{C}_n$ be the set of RB paths of length n.*

Lemma E.1.1. *In the RB tree with parameters (d_c, d_h), $|C_0| = 1$, while for $n \ge 1$,*

$$|\mathcal{C}_n| = \lambda_1 \left(\frac{d_c - 1 + D}{2}\right)^n + \lambda_2 \left(\frac{d_c - 1 - D}{2}\right)^n \tag{E.1}$$

where

$$D = \sqrt{(d_c - 1)^2 + 4 d_c d_h} \tag{E.2}$$

$$\lambda_1 = \frac{(d_c + 1 + D)(2d_h + d_c - 1 + D)}{2D(d_c - 1 + D)} \tag{E.3}$$

$$\lambda_2 = \frac{(D - d_c - 1)(2d_h + d_c - 1 - D)}{2D(d_c - 1 - D)}. \tag{E.4}$$

Proof. Let us keep track of the number of RB paths of length n depending on the color of the last node in the path. Let r_n and b_n be the numbers of RB paths of length n such that the last node is red and blue, respectively. A RB path of length 0 consists only of the source, which implies that $r_0 = 1$ and $b_0 = 0$. The source has d_c red and d_h blue neighbours, which implies that $r_1 = d_c$ and $b_1 = d_h$.

Suppose that P is an RB path of length $n \ge 2$. If the last node of P is red, then the node before the last node can be both blue or red. Red nodes other than the source have $d_c - 1$ red children, while blue nodes have d_c red children, yielding

$$r_n = (d_c - 1)r_{n-1} + d_c b_{n-1}, \text{ for } n \geq 2. \tag{E.5}$$

If the last node of P is blue, then the node before has to be red. Since every red node, including the source, has d_h blue children, we have

$$b_n = d_h r_{n-1}, \text{ for } n \geq 1. \tag{E.6}$$

By substituting equation (E.6) into equation (E.5), we obtain the recurrence

$$r_n = (d_c - 1)r_{n-1} + d_c d_h r_{n-2}, \text{ for } n \geq 2. \tag{E.7}$$

We solve this recurrence equation by calculating the characteristic equation

$$t^2 - (d_c - 1)t - d_c d_h = 0, \tag{E.8}$$

whose roots are

$$t_1 = \frac{d_c - 1 + \sqrt{(d_c-1)^2 + 4d_c d_h}}{2} = \frac{d_c - 1 + D}{2} \tag{E.9}$$

$$t_2 = \frac{d_c - 1 - \sqrt{(d_c-1)^2 + 4d_c d_h}}{2} = \frac{d_c - 1 - D}{2} \tag{E.10}$$

yielding the the general solution

$$r_n = c_1 t_1^n + c_2 t_2^n, \tag{E.11}$$

where c_1, c_2 are given by the initial conditions for $n = 0, 1$

$$c_1 + c_2 = r_0 = 1 \tag{E.12}$$

$$c_1 t_1 + c_2 t_2 = r_1 = d_c, \tag{E.13}$$

which are

$$c_1 = \frac{1}{2} + \frac{d_c + 1}{2\sqrt{(d_c-1)^2 + 4d_c d_h}} = \frac{1}{2} + \frac{d_c + 1}{2D} \tag{E.14}$$

$$c_2 = \frac{1}{2} - \frac{d_c + 1}{2\sqrt{(d_c-1)^2 + 4d_c d_h}} = \frac{1}{2} - \frac{d_c + 1}{2D}. \tag{E.15}$$

From equations (E.5) and (E.6) we conclude that for $n \geq 1$,

$$b_n = d_h(c_1 t_1^{n-1} + c_2 t_2^{n-1}) \tag{E.16}$$

and therefore

$$|C_n| = r_n + b_n = \lambda_1 t_1^n + \lambda_2 t_2^n, \tag{E.17}$$

where

$$\lambda_1 = c_1\left(1+\frac{d_h}{t_1}\right) \tag{E.18}$$

$$\lambda_2 = c_2\left(1+\frac{d_h}{t_2}\right). \tag{E.19}$$

Inserting the values for t_1, t_2, c_1, c_2 we obtain the desired result. □

Since LS+ improves on LS by making use of the household structure of the network, we need further information about the household structure of the transmission paths. Recall that by Remark 8.4.2, households on transmission paths on an RB tree were characterized either by a single red node (that is followed by a red node), or a pair of consecutive red and blue nodes. The following definition and lemma refine our previous result on counting the number of RB paths by taking the household structure into account.

Definition E.1.2. *Let $P = \{s = v_0, v_1, \ldots, v_n = h\}$ be a RB path of length n. We say that a node v on the path P is in a P-single-household if no other node from P is in the same household as v. Otherwise, we say v is in a P-multi-household. Given a path P, let $M_s : \mathcal{C}_n \to \{0,1\}$ be the indicator function that the source is in a P-multi-household. Similarly, let $M_l : \mathcal{C}_n \to \{0,1\}$ be the indicator function that the last node of path P is in a P-multi-household. Finally, for $0 \le k \le n+1$ and $\alpha, \beta \in \{0,1\}$, let*

$$\begin{aligned} C_{n,k,\alpha,\beta} = \{P \in \mathcal{C}_n : (&\textit{there are exactly } k \textit{ nodes in } P - \textit{single-households}) \\ &\wedge (M_s(P) = \alpha) \wedge (M_l(P) = \beta)\}. \end{aligned} \tag{E.20}$$

The set $C_{n,k,\alpha,\beta}$ depends on 4 parameters, but only some combinations of these parameters make it non-empty. The following definition will be useful in this regard.

Condition E.1.1. *Let $\alpha, \beta \in \{0,1\}$ and $n \ge 2$. We say $k \in \mathbb{N}$ satisfies Condition E.1.1 if and only if k and n have different parity and $n+1-2(\alpha+\beta) \ge k \ge 2-(\alpha+\beta)$.*

Lemma E.1.2. *It holds that $|C_{0,1,0,0}| = 1$, $|C_{1,0,1,1}| = d_h$ and $|C_{1,2,0,0}| = d_c$. Let $\alpha, \beta \in \{0,1\}$, let $n \ge 2$ and let $k \in \mathbb{N}$ satisfy Condition E.1.1. Then*

$$|\mathcal{C}_{n,k,\alpha,\beta}| = \binom{\frac{n+k-3}{2}}{k-2+\alpha+\beta} d_h^{\frac{n-k+1}{2}} d_c^{\frac{n-k+3}{2}-\beta-\alpha} (d_c-1)^{k+\alpha+\beta-2}. \tag{E.21}$$

In all other cases $|C_{n,k,\alpha,\beta}| = 0$.

Proof. Since there are $n+1$ nodes on path P, with k in P-single households and thus $n+1-k$ of them in P-multi-households, we must have

$$k + \frac{n+1-k}{2} = \frac{n+k+1}{2}$$

households along path P in total. Clearly, the numbers n and k cannot be of the same parity for any RB path P, which is thus assumed for the rest of the proof (this assumption is also part of Condition E.1.1).

If $n = 0$, then the source is also the first hospitalized node, and it is in a P-single-household, which implies that $|C_{0,1,0,0}| = 1$. If $n = 1$, then there are two cases: either the source is in the same P-multi-household with the first hospitalized node, or both of them are in P-single-households. The former case is possible via d_h edges from the source, which gives $|C_{1,0,1,1}| = d_h$, while the latter case is possible via d_c edges, and gives $|C_{1,0,1,1}| = d_c$. Since these are the only possible RB paths of length $n \leq 1$, we must have $|C_{0,k,\alpha,\beta}| = |C_{1,k,\alpha,\beta}| = 0$ for any other choice of parameters k, α and β.

Let us assume that $n \geq 2$. Then, the source and the first hospitalized node are not in the same household. Let us denote the household of the source by H_s and the household of the first hospitalized node by H_h. Note that $(1-\alpha)$ and $(1-\beta)$ are the indicators of H_s and H_h being P-single-households, and therefore $k \geq (1-\alpha)+(1-\beta)$. If this inequality (which is also part of Condition E.1.1) does not hold, then clearly $|\mathcal{C}_{n,k,\alpha,\beta}| = 0$. Similarly, the number of P-multi-households is $\frac{n-k+1}{2}$ and we must have $\frac{n-k+1}{2} \geq \alpha+\beta$ for $|\mathcal{C}_{n,k,\alpha,\beta}| > 0$, which implies the inequality $n+1-2\alpha-2\beta \geq k$. Therefore $\mathcal{C}_{n,k,\alpha,\beta}$ is empty if Condition E.1.1 does not hold. For the rest of the proof we assume that Condition E.1.1 does hold.

There are $\frac{n+k-3}{2}$ households along path P, excluding H_s and H_h. Among them, there are $k-(1-\alpha)-(1-\beta)$ P-single-households, which can be chosen in $\binom{\frac{n+k-3}{2}}{k-2+\alpha+\beta}$ ways. Once we know the color of each node along the path, the number of RB paths can be computed by multiplying the numbers of children with the appropriate color of each node. P-single-households have no blue nodes, and P-multi-households have exactly one, which implies that there are $\frac{n-k+1}{2}$ blue nodes. Since blue nodes are preceded by red nodes that have d_h blue children, they give the multiplicative factor $d_h^{\frac{n-k+1}{2}}$. Blue nodes, except from the first hospitalized node (if it is blue), have d_c red children. So far we have accounted for all of the nodes in P-multi-households and none of the nodes in P-single-households. If the source is in a P-single-household, then we must count its red children, whose number is d_c. This implies that there exist $\frac{n-k+1}{2}-\beta+(1-\alpha)$ nodes with d_c red children. Finally, each P-single-household, except H_s and/or H_h in case they are P-single households, has d_c-1 red children. There are $k-(1-\alpha)-(1-\beta)$ such P-single-households, which gives the final term in equation (E.21). □

The sets $\mathcal{C}_{n,k,\alpha,\beta}$ define equivalence classes on the transmission paths based on their household structure. In the next lemma we show that once we know which equivalence class we are in, it is possible compute the success probability of the LS+ algorithm.

Lemma E.1.3. *Let P be the transmission path in the* DDE$_{\mathrm{NR}}$ *epidemic model with parameters (p_i, p_a, p_h) on the RB tree with parameters (d_c, d_h), and let p be as computed in* (8.5)*. Then, it holds that*

$$\mathbb{P}(LS+\ \ succeeds|P \in \mathcal{C}_{0,1,0,0}) = 1$$

and

$$\mathbb{P}(LS+\ succeeds|P \in \mathcal{C}_{1,0,1,1}) = \mathbb{P}(LS+\ succeeds|\mathcal{C}_{1,2,0,0}) = 1-p.$$

Let $\alpha, \beta \in \{0,1\}$, let $n \geq 2$ and let $k \in \mathbb{N}$ satisfy Condition E.1.1. Then, it holds that

$$\mathbb{P}(LS+\ succeeds|P \in \mathcal{C}_{n,k,\alpha,\beta}) \geq (1-p)^{\frac{n+k-1}{2}}(1+p)^{\frac{n-k+1}{2}-\alpha-\beta}. \tag{E.22}$$

In all other cases $\mathbb{P}(LS+\ succeeds|P \in \mathcal{C}_{n,k,\alpha,\beta})$ is not defined.

Proof. If $n = 0$, then $k = 1$ and $\alpha = \beta = 0$. In that case, the source is the first hospitalized node and LS+ always succeeds. If $n = 1$, then the first hospitalized node is in the neighbourhood of the source, and LS+ succeeds if and only if the source is symptomatic, which happens with probability $1-p$.

Let us assume that $n \geq 2$ and that k satisfies Condition E.1.1 (otherwise $|\mathcal{C}_{n,k,\alpha,\beta}| = 0$ and $\mathbb{P}(LS+\ \text{succeeds}|P \in \mathcal{C}_{n,k,\alpha,\beta})$ is not defined). By Lemma 8.4.1 the LS+ algorithm succeeds in the DDE$_{\text{NR}}$ model on the RB tree if, among the nodes of the transmission path, there exists a symptomatic node in each household, and the source is symptomatic, which means that we can prove a lower bound on the success probability of LS+. Let us assume that the source is indeed symptomatic. Since the first hospitalized node is symptomatic by definition, the households of the source and of the first hospitalized node cannot make the LS+ algorithm fail. Let us denote these two households by H_s and H_h, respectively. Also, let M and S be the sets of all P-multi- and P-single-households, respectively, excluding H_s and H_h. Then, LS+ succeeds if all nodes in the households of S are symptomatic, and if at least one node in the households of M is symptomatic, which has probability $1-p$ and $1-p^2$ for each type of household, respectively, by equation (8.5). These observations yield that

$$\begin{aligned}\mathbb{P}(LS+\ \text{succeeds}|P \in \mathcal{C}_{n,k,\alpha,\beta}) &\geq \mathbb{P}(\text{source is sym})(1-p)^{|S|}(1-p^2)^{|M|}\\ &= (1-p)(1-p)^{k-2+\alpha+\beta}(1-p^2)^{\frac{n-k+1}{2}-\alpha-\beta}\\ &= (1-p)^{k-1+\alpha+\beta}(1-p^2)^{\frac{n-k+1}{2}-\alpha-\beta}.\end{aligned} \tag{E.23}$$

□

Finally, we are ready to state and prove Theorem 8.4.1 on the success probability of LS+, which we restate here for convenience.

Theorem 8.4.1. *Let p be as in* (8.5) *and let $\mathcal{S}(n,\alpha,\beta)$ be the set of k values that satisfy Condition E.1.1. Then, for the* DDE$_{\text{NR}}$ *epidemic model with parameters (p_i, p_a, p_h) on the RB tree* *with parameters* (d_c, d_h) *we have*

$$\mathbb{P}(LS+\ \textit{succeeds}) \geq \mathbb{P}(d(s,h)=0) + (1-p)\mathbb{P}(d(s,h)=1) +$$
$$\sum_{n=2}^{\infty} \sum_{\substack{\alpha,\beta\in\{0,1\}\\ k\in\mathcal{S}(n,\alpha,\beta)}} \binom{\frac{n+k-3}{2}}{k-2+\alpha+\beta} \frac{(d_h(1-p))^{\frac{n+k-1}{2}}(d_c(1+p))^{\frac{n-k+1}{2}-\alpha-\beta} d_c(d_c-1)^{k+\alpha+\beta-2}}{\lambda_1\left(\frac{d_c-1+D}{2}\right)^n + \lambda_2\left(\frac{d_c-1-D}{2}\right)^n} \mathbb{P}(d(s,h)=n), \tag{E.24}$$

where D, λ_1 *and* λ_2 *are terms depending on parameters* d_c *and* d_h *and are computed explicitly in Lemma E.1.1.*

Proof. Let us extend the domain of $\mathbb{P}(LS+\ \text{succeeds}|P \in C_{n,k,\alpha,\beta})$ by function g defined as $g : \mathbb{N}\times\mathbb{N}\times\{0,1\}\times\{0,1\} \to [0,1]$ such that

$$g(n,k,\alpha,\beta) = \begin{cases} \mathbb{P}(LS+\ \text{succeeds}|P \in C_{n,k,\alpha,\beta}) & \text{if } k \in \mathcal{S}(n,\alpha,\beta) \\ 0 & \text{if } k \notin \mathcal{S}(n,\alpha,\beta). \end{cases} \tag{E.25}$$

Unlike $\mathbb{P}(LS+\ \text{succeeds}|P \in C_{n,k,\alpha,\beta})$, g is defined for every 4-tuple of parameters $(n,k,\alpha,\beta) \in \mathbb{N}\times\mathbb{N}\times\{0,1\}\times\{0,1\}$. By the law of total probability we expand the success probability by conditioning on the path P being of length n as

$$\begin{aligned} \mathbb{P}(LS+\ \text{succeeds}) &= \sum_{n=0}^{\infty}\sum_{k=0}^{\infty}\sum_{\alpha,\beta\in\{0,1\}} g(n,k,\alpha,\beta)\mathbb{P}(P \in C_{n,k,\alpha,\beta}) \\ &= \sum_{n=0}^{\infty}\sum_{k=0}^{\infty}\sum_{\alpha,\beta\in\{0,1\}} g(n,k,\alpha,\beta)\mathbb{P}(P \in C_{n,k,\alpha,\beta}|P \in C_n)\mathbb{P}(d(s,h)=n). \end{aligned} \tag{E.26}$$

Next, we exchange the sums over α, β and k. This allows us to sum over only those k values that satisfy Condition E.1.1, which implies that $\mathbb{P}(LS+\ \text{succeeds}|P \in C_{n,k,\alpha,\beta})$ is well-defined. As in Lemma E.1.2, we need to treat the $n=0$ and $n=1$ cases separately. Continuing equation (E.26), we arrive to

$$\begin{aligned} \mathbb{P}(LS+\ \text{succeeds}) =& \mathbb{P}(d(s,h)=0) + (1-p)\mathbb{P}(d(s,h)=1) + \\ & \sum_{n=2}^{\infty}\sum_{\alpha,\beta\in\{0,1\}}\sum_{k\in\mathcal{S}(n,\alpha,\beta)} \mathbb{P}(LS+\ \text{succeeds}|P \in C_{n,k,\alpha,\beta})\frac{|C_{n,k,\alpha,\beta}|}{|C_n|}\mathbb{P}(d(s,h)=n) \end{aligned} \tag{E.27}$$

Substituting in the results from Lemmas E.1.1, E.1.2 and E.1.3 into equation (E.27) gives the desired result. □

E.1.3 Proof of Theorem 8.4.2

We start by restating Theorem 8.4.2 for convenience.

Theorem 8.4.2. *In the (d_r, d)-RET with parameters p_i, p_a, p_h, let $a_{t,l}$ be as in Definition 8.4.5. Then*

$$a_{t,0} = 1 \tag{E.28}$$

$$a_{t,l} = d_r p_i \sum_{m=l-1}^{t-1} \binom{m}{l-1} (1-p_i)^{m-l+1} d^{l-1} p_i^{l-1}, \textit{for } t \geq l \geq 1 \tag{E.29}$$

$$a_{t,l} = 0, \textit{for } l > t. \tag{E.30}$$

Proof. Similarly to [93, 167], the proof relies on generating functions. We start by addressing the boundary cases. For all $t \geq 0$, it holds that $A_{t,0} = 1$, and therefore $a_{t,0} = 1$. Similarly, for all l, t such that $l > t$, it holds that $A_{t,l} = 0$, and therefore $a_{t,l} = 0$. Suppose that $t \geq l = 1$. During day $t-1$, on the first level, there are $A_{t-1,1}$ infected (internal) nodes and $d_r - A_{t-1,1}$ (external) nodes that may be infected with probability p_i during day t. Thus,

$$A_{t,1} = A_{t-1,1} + \mathrm{Bin}(d_r - A_{t-1,1}; p_i). \tag{E.31}$$

Taking the expectation of both sides in equation (E.31) yields

$$a_{t,1} = a_{t-1,1}(1-p_i) + d_r p_i, \text{ for } t \geq 1. \tag{E.32}$$

By subtracting the appropriate recurrence equations for $a_{t,1}$ and $a_{t-1,1}$ for $t \geq 2$ we obtain the homogeneous recurrence equation

$$a_{t,1} - a_{t-1,1}(2-p_i) + (1-p_i)a_{t-2,1} = 0, \text{ for } t \geq 2 \tag{E.33}$$

and boundary conditions $a_{0,1} = 0$ and $a_{1,1} = d_r p_i$. We solve for $a_{t,1}$ using the same methods as in the proof of Lemma E.1.1 and obtain

$$a_{t,1} = d_r \left(1 - (1-p_i)^t\right), \text{ for } t \geq 0. \tag{E.34}$$

Next, let us consider the general case $t \geq l > 1$. On day $t-1$, there are $A_{t-1,l-1}$ nodes on level $l-1$. Since, each node on level $l-1$ has d children, there are $dA_{t-1,l-1}$ nodes on level l that have an infectious parent on level $l-1$. However, $A_{t-1,l}$ of them are already infected. Therefore $dA_{t-1,l-1} - A_{t-1,l}$ nodes of level l may be infected on day t, each with probability p_i, which implies

$$A_{t,l} = A_{t-1,l} + \mathrm{Bin}(dA_{t-1,l-1} - A_{t-1;l}, p_i), \text{ for } t \geq l \geq 2. \tag{E.35}$$

Taking the expectation of both sides in equation (E.35) yields

$$\begin{aligned} a_{t,l} &= a_{t-1,l} + (d a_{t-1,l-1} - a_{t-1,l}) p_i \\ &= a_{t-1,l}(1-p_i) + d p_i a_{t-1,l-1}, \text{ for } t \geq l \geq 2. \end{aligned} \tag{E.36}$$

For convenience, let us introduce $\lambda = 1 - p_i$ and $\mu = d p_i$, and also let

$$f(x,y) = \sum_{t=1}^{\infty}\sum_{l=1}^{\infty} a_{t,l} x^t y^l = \sum_{t=1}^{\infty}\sum_{l=1}^{t} a_{t,l} x^t y^l \tag{E.37}$$

be the generating function for $a_{t,l}$ with $t, l \geq 1$. By multiplying (E.36) by $x^t y^l$ and summing it over $t, l \geq 2$ we obtain

$$\begin{aligned} \sum_{t=2}^{\infty}\sum_{l=2}^{t} a_{t,l} x^t y^l &= \lambda \sum_{t=2}^{\infty}\sum_{l=2}^{t} a_{t-1,l} x^t y^l + \mu \sum_{t=2}^{\infty}\sum_{l=2}^{t} a_{t-1,l-1} x^t y^l \\ &= \lambda x \sum_{t=1}^{\infty}\sum_{l=2}^{t} a_{t,l} x^t y^l + \mu x y \sum_{t=1}^{\infty}\sum_{l=1}^{t} a_{t,l} x^t y^l. \end{aligned} \tag{E.38}$$

Since $a_{1,l} = 0$ for $l \geq 2$,

$$\sum_{t=1}^{\infty}\sum_{l=2}^{t} a_{t,l} x^t y^l = \sum_{t=2}^{\infty}\sum_{l=2}^{t} a_{t,l} x^t y^l, \tag{E.39}$$

and by inserting (E.39) into (E.38), we obtain

$$(1-\lambda x) \sum_{t=1}^{\infty}\sum_{l=2}^{t} a_{t,l} x^t y^l = \mu x y \sum_{t=1}^{\infty}\sum_{l=1}^{t} a_{t,l} x^t y^l \overset{\text{(E.37)}}{=} \mu x y f(x,y). \tag{E.40}$$

Now, we can also decompose the sum (E.39) using geometric series as

$$\begin{aligned} \sum_{t=1}^{\infty}\sum_{l=2}^{t} a_{t,l} x^t y^l &= \sum_{t=1}^{\infty}\sum_{l=1}^{t} a_{t,l} x^t y^l - \sum_{t=1}^{\infty} a_{t,1} x^t y \\ &\overset{\text{(E.34)}}{=} f(x,y) - d_r y \sum_{t=1}^{\infty} (1-\lambda^t) x^t \\ &= f(x,y) - d_r x y \left(\frac{1}{1-x} - \frac{\lambda}{1-\lambda x} \right). \end{aligned} \tag{E.41}$$

By plugging (E.41) into (E.40), we obtain the expression

$$f(x,y) = d_r (1-\lambda) x y \frac{1}{1-x} \frac{1}{1-\lambda x - \mu x y}. \tag{E.42}$$

Then, we expand the fractions in (E.42) into a power series and we next apply the binomial theorem, we arrive to

$$\begin{aligned} f(x,y) &= d_r(1-\lambda)xy\sum_{n=0}^{\infty}x^n\sum_{m=0}^{\infty}x^m(\lambda+\mu y)^m \\ &= d_r(1-\lambda)xy\sum_{n=0}^{\infty}x^n\sum_{m=0}^{\infty}x^m\sum_{j=0}^{m}\binom{m}{j}\lambda^{m-j}(\mu y)^j \\ &= d_r(1-\lambda)\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\sum_{j=0}^{m}\binom{m}{j}\lambda^{m-j}\mu^j x^{1+n+m}y^{j+1}. \end{aligned} \tag{E.43}$$

Let $t=1+n+m$ and $l=j+1$. In order to obtain an expression for $a_{t,l}$, we must change the variables in the sums of equation (E.43) from (n,m,k) to (t,m,l). Changing the inner sum from variable j to l is simple. Changing the variables in the two outer sums is more challenging because t,n and m depend on each other in a nontrivial way. More precisely, since $m,n\geq 0$ we have $t\geq 1$ and also $m\leq t-1$, which means that we have to set the lower limit of t and the upper limit of m accordingly. As for the remaining limits, variable t can be arbitrary large, and m can take any integer value starting from 0 independently of t, which yields the expression

$$f(x,y) = d_r(1-\lambda)\sum_{t=1}^{\infty}\sum_{m=0}^{t-1}\sum_{l=1}^{m+1}\binom{m}{l-1}\lambda^{m-l+1}\mu^{l-1}x^t y^l. \tag{E.44}$$

For the values of l with $l\geq m+1$, the binomial coefficient $\binom{m}{l-1}$ is 0, which implies that we can increase the upper limit of the inner sum from $m+1$ to t in equation (E.44). Then,

$$\begin{aligned} f(x,y) &= d_r(1-\lambda)\sum_{t=1}^{\infty}\sum_{m=0}^{t-1}\sum_{l=1}^{t}\binom{m}{l-1}\lambda^{m-l+1}\mu^{l-1}x^t y^l \\ &= \sum_{t=1}^{\infty}\sum_{l=1}^{t}d_r(1-\lambda)\sum_{m=0}^{t-1}\binom{m}{l-1}\lambda^{m-l+1}\mu^{l-1}x^t y^l. \end{aligned} \tag{E.45}$$

Finally we can read off the value of $a_{t,l}$ from equation (E.45) as

$$a_{t,l} = d_r(1-\lambda)\sum_{m=0}^{t-1}\binom{m}{l-1}\mu^{l-1}\lambda^{m-l+1} = d_r p_i\sum_{m=0}^{t-1}\binom{m}{l-1}(dp_i)^{l-1}(1-p_i)^{m-l+1}. \tag{E.46}$$

□

E.1.4 Proof of Corollary 8.4.1

We start by restating Corollary 8.4.1 for convenience.

Corollary 8.4.1. *In the RET(p_i,d_r,d), let a_t be the expectation of* (8.12)*, as in Definition 8.4.5. For $t\geq 0$,*

$$a_t = 1+d_r\frac{(1-p_i+dp_i)^t-1}{d-1}. \tag{E.47}$$

Proof. By using linearity of expectation, equation (8.12) and Theorem 8.4.2 we obtain:

$$\begin{aligned} a_t &= \sum_{l=0}^{+\infty} a_{t,l} \\ &= 1 + \sum_{l=1}^{+\infty} a_{t,l} \\ &= 1 + d_r p_i \sum_{l=1}^{t} \sum_{m=l-1}^{t-1} \binom{m}{l-1} (1-p_i)^{m-l+1} d^{l-1} p_i^{l-1} \end{aligned} \tag{E.48}$$

Before we use binomial theorem, we need to swap the sums. Boundaries from (E.48) are equivalent to $t-1 \geq m \geq l-1 \geq 0$, so we can rewrite this as 2 conditions: $m+1 \geq l \geq 1$ and $t \geq m \geq 0$.

$$\begin{aligned} a_{t,l} &= 1 + d_r p_i \sum_{m=0}^{t-1} \sum_{l=1}^{m+1} \binom{m}{l-1} (1-p_i)^{m-l+1} d^{l-1} p_i^{l-1} \\ &= 1 + d_r p_i \sum_{m=0}^{t-1} \sum_{l=0}^{m} \binom{m}{l} (1-p_i)^{m-l} d^{l} p_i^{l} \end{aligned} \tag{E.49}$$

Finally, by applying the binomial theorem and summing the geometric series, we obtain the desired equation:

$$\begin{aligned} a_{t,l} &= 1 + d_r p_i \sum_{m=0}^{t-1} (1 - p_i + d p_i)^m \\ &= 1 + d_r \frac{(1-p_i+dp_i)^t - 1}{d-1}. \end{aligned} \tag{E.50}$$

□

E.1.5 Proof or Lemma 8.4.3

We restate Lemma 8.4.3 here for convenience.

Lemma 8.4.3. *Let us consider the stopped DET model with parameters* $(c_{t,l}), p_a, p_h$*, and let* h *denote the first hospitalized node. Then*

$$\mathbb{P}(d(s,h) = l) = \sum_{t=0}^{+\infty} \frac{c_{t,l} - c_{t-1,l}}{c_t - c_{t-1}} (1-(1-p_a)p_h)^{c_{t-1}} \left(1 - (1-(1-p_a)p_h)^{c_t - c_{t-1}}\right). \tag{E.51}$$

Proof. Recall that a node added at day t is uniformly distributed among the $c_t - c_{t-1} > 0$ nodes added that day, and that the number of nodes added to level l is $c_{t,l} - c_{t-1,l}$ on day t. If we condition on the time of the first hospitalized case, denoted by TI_h, then

$$\begin{aligned}\mathbb{P}(d(s,h)=l) &= \sum_{t=0}^{+\infty} \mathbb{P}(d(s,h)=l|TI_h=t)\mathbb{P}(TI_h=t)\\ &= \sum_{t=0}^{+\infty} \frac{c_{t,l}-c_{t-1,l}}{c_t-c_{t-1}} \mathbb{P}(\text{node is not hosp})^{c_{t-1}}(1-\mathbb{P}(\text{node is not hosp })^{c_t-c_{t-1}})\\ &= \sum_{t=0}^{+\infty} \frac{c_{t,l}-c_{t-1,l}}{c_t-c_{t-1}} (1-(1-p_a)p_h)^{c_{t-1}}\left(1-(1-(1-p_a)p_h)^{c_t-c_{t-1}}\right). \end{aligned} \tag{E.52}$$

□

E.2 Dynamic Message Passing for the DDE model

In this section, we explain how we derived and implemented the DMP equations for the DDE+HNM model. We start by reviewing the previous work on the DMP equations for the SIR model in Appendix E.2.1, and then we proceed to our derivations in Appendix E.2.2. In Appendix E.2.3, we explain how we find candidate (node,time) pairs for the DMP equations, and in Appendix E.2.4 we conclude by combining Appendices E.2.2 and E.2.3 into a source-identification algorithm.

E.2.1 DMP Equations for the SIR Model

The DMP equations were first derived by [161] for the SIR model in the context of source identification. Their goal is to compute the marginal probabilities that node i is in a given state at time t (denoted by $P^i_S(t), P^i_I(t)$ and $P^i_R(t)$ for the susceptible, infected and recovered states, respectively), given initial conditions $P^i_S(t_0), P^i_I(t_0)$ and $P^i_R(t_0)$ at some initial time t_0. To solve this problem in tree networks, we may consider a dynamic programming approach, where we delete a node i, we compute the marginal probabilities of $P^j_S(t-1)$ for all neighbors j of i in the remaining subtrees, and use this information to compute $P^i_S(t)$ (as the marginals are independent in each of the subtrees conditioned on the state of i). The DMP equations make the dynamic programming intuition explicit. Originally, the DMP equations were developed for static networks, but since the generalization to time-varying networks is straightforward, and has already been foreshadowed in a similar heuristic algorithm [130], we include it in this preliminary section. For time-varying networks, we define $N_i(t)$ as the set of neighbors of node i in the time-window $[t, t+1)$.

To formalize the dynamic programming approach, [161] introduces some new notation. Let λ be the probability that an infectious node infects a susceptible neighbor, and let μ be the probability that an infectious node recovers. Let D_i be the auxiliary dynamics, where node i receives infection signals, but ignores them, and thus remains in the S state at all times. Let $P^{j\to i}_S(t)$ be the probability that node j is in the state S at time t in the dynamics D_i, and let $\theta^{k\to i}(t)$ be the probability that the infection signal has not been passed from node k to node i

up to time t in the dynamics D_i. Finally, let $\phi^{k\to i}(t)$ be the probability that the infection signal has not been passed from node k to node i up to time t, and that node k is in the state I at time t, in the dynamics D_i. With these definitions, the dynamic programming approach is formalized by the following equations for $t \geq t_0$:

$$P_S^{i\to j}(t+1) = P_S^i(t_0) \prod_{k\in N_i(t)\setminus j} \theta^{k\to i}(t+1), \tag{E.53}$$

$$\theta^{k\to i}(t+1) - \theta^{k\to i}(t) = -\lambda\phi^{k\to i}(t), \tag{E.54}$$

$$\phi^{k\to i}(t) = (1-\lambda)(1-\mu)\phi^{k\to i}(t-1) + \left(P_S^{k\to i}(t-1) - P_S^{k\to i}(t)\right). \tag{E.55}$$

The marginal probabilities that node i is in a given state at time t are then given by

$$P_S^i(t+1) = P_S^i(t_0) \prod_{k\in N_i(t)} \theta^{k\to i}(t+1)\,, \tag{E.56}$$

$$P_R^i(t+1) = P_R^i(t) + \mu P_I^i(t)\,, \tag{E.57}$$

$$P_I^i(t+1) = 1 - P_S^i(t+1) - P_R^i(t+1)\,. \tag{E.58}$$

These equations are only exact on trees, but they can also be applied to networks with cycles as a heuristic approach. The heuristic gives good approximations to the true marginals if the network is at least locally tree-like [138].

E.2.2 DMP Equations for the DDE+HNM Model

There are several differences between the SIR model on locally tree-like networks and the DDE+HNM model (see Figure 8.2 (a)). First, the DDE model has additional compartments (exposed nodes, asymptomatic nodes), which motivates the introduction of several new variables. Let $\lambda_{(a)}$ (resp., $\lambda_{(s)}$) be the probability that an asymptomatic (resp., symptomatic) node infects a susceptible node. Let $\phi^{k\to i}(t)^{(a)}$ (resp., $\phi^{k\to i}(t)^{(s)}$) be the probability that the infection signal has not been passed from node k to node i up to time t, and that node k is asymptomatic (resp., symptomatic) infectious at time t, in the dynamics D_i.

The second important difference is that in the DDE model, the transition times between different compartments are deterministic instead of following a geometric distribution as in the standard SIR model. While deterministic transition times sound simpler at first, it turns out that they make the DMP equations more complex, because the Markovian property that each marginal probability depends only on the previous timestep is lost if the transition times are larger than 1. Recall that the times for the transitions $E \to I$ and $I \to R$ (with their default values) are $T_E = 3$ and $T_I = 14$.

Let us incorporate these two differences into equations (E.53)–(E.55) to derive the DMP equations for the DDE model. Equation (E.59) is essentially a copy of (E.53). Equation (E.60) follows equation (E.54), but we incorporate the two different variants of infected (asymptomatic and symptomatic) patients with their respective infection probabilities $\lambda_{(a)}$ and $\lambda_{(s)}$. Equation (E.61) is a new equation, which is necessary because recovery times are no longer geometric random variables; instead we need to check the probabilities of infection $T_E + T_I$ timesteps earlier than the current time t. Finally, equation (E.62) (resp., (E.63)) is the asymptomatic (resp., symptomatic) version of equation (E.55), while also incorporating the deterministic time for the transition $E \rightarrow I$. For $t \geq t_0$, this yields equations

$$P_S^{i\to j}(t+1) = P_S^i(t_0) \prod_{k\in N_i(t)\setminus j} \theta^{k\to i}(t+1) = P_S^i(t_0)\frac{P_S^i(t+1)}{\theta^{j\to i}(t+1)}, \tag{E.59}$$

$$\theta^{k\to i}(t+1) - \theta^{k\to i}(t) = -\lambda_{(a)}\phi_{(a)}^{k\to i}(t) - \lambda_{(s)}\phi_{(s)}^{k\to i}(t), \tag{E.60}$$

$$P_R^{k\to i}(t) = P_S^{k\to i}(t - T_E - T_I - 1) - P_S^{k\to i}(t - T_E - T_I) \tag{E.61}$$

$$\begin{aligned}\phi_{(a)}^{k\to i}(t) = {} & (1-\lambda_{(a)})(1-P_R^{k\to i}(t))\phi_{(a)}^{k\to i}(t-1) \\ & + p_a[P_S^{k\to i}(t-T_E-1) - P_S^{k\to i}(t-T_E)].\end{aligned} \tag{E.62}$$

$$\begin{aligned}\phi_{(s)}^{k\to i}(t) = {} & (1-\lambda_{(s)})(1-P_R^{k\to i}(t))\phi_{(s)}^{k\to i}(t-1) \\ & + (1-p_a)[P_S^{k\to i}(t-T_E-1) - P_S^{k\to i}(t-T_E)].\end{aligned} \tag{E.63}$$

We note that for early values of t, equations (E.61)–(E.63) depend on $P_S^{k\to i}$ before t_0, which we initialize to be 1 (all nodes are susceptible before the first node develops the infection). The marginal probability that node i is susceptible at time t is still computed by equation (E.56) as before. Equations (E.57)–(E.58) do not apply anymore; we explain it in Appendix E.2.4 how to take into account observations for nodes in the infectious compartments.

The third difference between the the SIR model on locally tree-like networks and the DDE+HNM model is that the HNM model contains many short cycles inside the households. Short cycles can cause unwanted feedback loops in the DMP equations where, loosely speaking, nodes are treated as if they could reinfect themselves. We solve this issue by modifying the underlying graph to be locally tree-like (only for the computation of the DMP equations). Specifically, we introduce a new central household-node for each household, and we replace the cliques inside the households by a star graph centered at this new household-node node. Introducing such a central household-node does of course alter epidemic process, in particular it makes household infections less independent and slower (all household infections need to pass through an extra node). To mitigate this issue, we assume that central household-nodes have $T_E = 1$ and that they are infected with probability 1 by any node in the same household. We tested the validity of the resulting DMP equations against simulations of the epidemic progressions and we found the results to be quite accurate, in particular, more accurate than the version without the introduction of these central household-nodes.

Note that we derived the DMP equations for the DDE+HNM model, however, since (i) the compartments are the same, (ii) the equations support temporal networks, and (iii) we have separate infection probabilities $\lambda_{(a)}$ and $\lambda_{(s)}$ for asymptomatic and symptomatic nodes, our equations can also be applied to the DCS+TU model after a discretizing (rounding) the time observations.

Finally, we touch upon the computational complexity of computing the DMP equations. In principle, we need to update $O(dN)$ equations (for each edge) over $t_{\max}$ timesteps, where $t_{\max}$ is the maximum time during which the marginals can still change, which can be as large as $O(N)$. However, since we are only interested in computing the likelihood of the 5 earliest observations, $t_{\max}$ is typically quite low. Moreover, since we assume to be in an early stage of the epidemic, most of the equations remain unchanged. For better computational scalability, we only compute $P_S^i(t)$ and $\theta^{k\to i}(t)$ for nodes k, i that have $P_I^{k\to i}(t) > 0.01$, i.e., we only update nodes that are at least somewhat likely to have received the infection. Otherwise, we set $P_I^{k\to i}(t) = P_I^{k\to i}(t-1)$, $\theta^{k\to i}(t) = \theta^{k\to i}(t-1)$, and in the implementation we can perform these assignments implicitly using appropriate data structures. With these adjustments, the time-complexity of the algorithm becomes independent of N, but remains dependent on the network parameters, the epidemic parameters and the number of sensors in a non-trivial way.

E.2.3 Feasible Source-time Pairs for Source Identification

In this section we explain how we implemented the feasible source identification algorithm, which was suggested as a preprocessing step for a method very similar to the DMP equations by [130]. Let us define the directed graph G_2 on (node,infecton_time) pairs (we use “nodes” for the nodes of the original graph G and “pairs” for the nodes of G_2), and draw an edge between two pairs $(v_1, t_1) \to (v_2, t_2)$ if v_1 and v_2 are in contact at t_2, and t_2 is in the interval $[t_1 + T_E, t_1 + T_E + T_I]$. Observe that in the DDE model there is an edge $(v_1, t_1) \to (v_2, t_2)$ if and only if v_1 becoming infected at time t_1 can infect v_2 at time t_2. The definition of G_2 is applicable to the DCS model as well after discretization (rounding), however, since the infection times are not deterministic anymore, not all possible infections $(v_1, t_1) \to (v_2, t_2)$ have a corresponding edge in G_2.

Then, we perform a breadth-first search backwards on the directed edges of G_2, starting from each pair $(v_i, t_i - T_E - T_P)$, where v_i is a symptomatic sensor node, and t_i is the symptom onset time of v_i (for the DCS model, we start from integer times in the $t_i - T_E - T_P \pm (\sigma_E + \sigma_P)$ interval to account for the randomness of the transition times). To limit the time complexity of the algorithm, we only consider the k_1 earliest observations, which means that we start k_1 breadth-first searches. With this construction, each pair (v, t) discovered by a breadth-first search started from $(v_i, t_i - T_E - T_P)$ could have caused the infection in v_i; we say that (v, t) is an explanation for observation i. We perform the breath-first searches until we find k_2 pairs that explain all of the k_1 earliest observations. See the pseudocode in Algorithm E.2.1.

Claim E.2.1. *In the DDE model, Algorithm E.2.1 with $\sigma_E = \sigma_P = 0$ finds the k_2 feasible explanations with the latest starting time of the k_1 earliest symptomatic nodes.*

Proof. By construction, a source node v that becomes infectious at time t can cause an observation (v_i, t_i) if and only if there is a directed path from (v, t) to $(v_i, t_i - T_E - T_P)$. Therefore, the breadth-first search algorithm finds all of the closest feasible sources in time. □

E.2.4 Source Identification via Feasible Source Identification and DMP

In this section we explain how to combine Algorithm E.2.1 with the DMP equations derived in Appendix E.2.2. See the pseudocode in Algorithm E.2.2.

We start by computing the DMP equations (E.59)-(E.63) and (E.53) for the k_2 tuples of node and time pairs that can explain the first k_1 symptomatic observations returned by Algorithm E.2.1. Next, our goal is to use these DMP equations to compute the likelihood of each of the k_2 tuples using the k_1 observations. Similarly to [161], we make the assumption that the first k_1 observations are independent, and we can compute the likelihood by multiplying their respective marginals together. For symptomatic observed nodes v, we know the time of symptom onset, which we denote by $S(v)$. Then, the marginal probability of v developing symptoms exactly at time t can be computed by taking the difference of $P_S^v(S(v) - T_P - T_E - 1)$ and $P_S^v(S(v) - T_P - T_E)$ and multiplying the difference by $(1 - p_a)$. In Algorithm E.2.2 we drop the multiplicative factor $(1 - p_a)$ because it is present for all of the tuples, and it does not change the final order of their scores. For asymptomatic (resp., negative) observations, we only know that at the time of testing, denoted by $A(v)$ (resp., $NE(v)$), at least a time interval of length T_E has passed (resp., T_E has not passed) since the time of infection. Therefore, dropping the p_a factor similarly to the symptomatic case, we compute the marginal of asymptomatic observations as $1 - P_S^v(A(v) - T_E)$, and we compute the marginal of negative observations as $P_S^v(NE(v) - T_E)$. Finally, the contributions of the observations are multiplied together for each of the k_2 tuples returned by Algorithm E.2.1, and the scores approximating the likelihoods are returned.

Algorithm E.2.1: Feasible source identification (reverse dissemination [130])

Input:

- The mean exposed time T_E the mean pre-infectious time T_P, the mean infectious time T_I,

  the std of the exposed time σ_E and the std of the pre-infectious time σ_P
- $F(v)_{min}$ and $F(v)_{max}$ returns the minimum and maximum times when v could have been exposed based on all of its (possibly asymptomatic or negative) test results
- $S(v)$ returns the time of symptom onset for a node v tested positive symtomatic.
- $N(v, [t_{min}, t_{max}])$ returns the set of neighbors of node v in the interval $[t_{min}, t_{max}]$
- A lower estimate of the time the source became infectious t_{min}
- Integers k_1, k_2

Output: A list of at most k_2 tuples of node and time pairs that can explain the first k_1 symptomatic nodes

```
$l \leftarrow \{\}$;   // if the list $l[t]$ contains the tuple $(v, w)$, then the infection started at $w$ at time $t$ can explain $v$
$D \leftarrow \{\}$; // if the list $D[w, t]$ contains the node $v$, then the infection started at $w$ at time $t$ can explain $v$
$doneList \leftarrow []$;
for $v \in SortIncreasingByValues(S)[0 : k_1]$ do
    $t'_{min} \leftarrow S(v) - (T_E + T_P) - (\sigma_E + \sigma_P)$;
    $t'_{max} \leftarrow S(v) - (T_E + T_P) + (\sigma_E + \sigma_P)$;
    for $t' \leftarrow t'_{min}$ to $t'_{max}$ do
        $Append((v, v), l[t'])$;
        $Append(v, D[v, t'])$;
        if $Length(D[v, t']) = k_1$ then
            $Append((v, t'), doneList)$
$t \leftarrow SortIncreasingByValues(S)[k_1 - 1]$;
$stopCondition \leftarrow False$ ;
while not $stopCondition$ and $t > t_{min}$ do
    for $v, w \in l[t]$ do
        for $u \in N(w, [t, t - 1])$ do
            $t'_{min} \leftarrow \max(F(u)_{min}, t - T_E - T_I)$;
            $t'_{max} \leftarrow \min(F(u)_{max}, t - T_E)$;
            for $t' \leftarrow t'_{max}$ to $t'_{min}$ do
                $Append((v, u), l[t'])$;
                $Append(v, D[(u, t')])$;
                if $Length(D[u, t']) = k_1$ then
                    $Append((u, t'), doneList)$
    $doneList \leftarrow SortBySecondElement(doneList)$;
    if $Length(doneList \geq k2)$ and $t - T_E \leq doneList[k2][1]$ then
        $stopCondition \leftarrow True$;
    else
        $t \leftarrow t - 1$;
return $doneList$
```

Algorithm E.2.2: Source identification via DMP

Input:

- The mean exposed time T_E, the mean pre-infectious time T_P, the mean infectious time T_I
- $S(v)$ returns the time of symptom onset for a node v tested positive symtomatic.
- $A(v)$ and $NE(v)$ return the time of asymptomatic and negative test results, respectively

Output: A dictionary L of k_2 elements, which contains a score for each (v, t) pair that explains the first k_1 observations. Higher scores signify higher confidence of being the source.

```
$L \leftarrow \{\}$;
$doneList \leftarrow$ Algorihtm E.2.1$(k_1, k_2)$;
for $v, t_0 \in doneList$ do
    $P_S \leftarrow$ eq. (E.53) based on DMP eq. (E.59)-(E.63) with $P_S^v(t_0) = 0$, and $P_S^w(t_0) = 1$ for all $w \neq v$;
    $L[v, t_0] \leftarrow 1$;
    for $w \in S$ do
        $L[v, t_0] \leftarrow L[v, t_0] \cdot (P_S^v(S(v) - T_P - T_E - 1) - P_S^v(S(v) - T_P - T_E))$;
    if $w \in A$ then
        $L[v, t_0] \leftarrow L[v, t_0] \cdot (1 - P_S^v(A(v) - T_E))$;
    for $w \in NE$ do
        $L[v, t_0] \leftarrow L[v, t_0] \cdot P_S^v(NE(v) - T_E)$;
return $L$
```

Bibliography

[1] (2016). Hungarian microcensus. https://www.ksh.hu/mikrocenzus2016/?lang=en. (referenced on pages: 16, 17, and 203)

[2] (2020). Hungarian statistical office: Settlement and population registry. http://www.ksh.hu/apps/hntr.egyeb?p_lang=EN&p_sablon=LETOLTES [date of access: 01/04/2020]. (referenced on page: 203)

[3] (2020). Implementation of geometric inhomogeneous random graphs. https://github.com/joostjor/random-graphs/blob/master/girg.py. (referenced on page: 204)

[4] (2020). Networkx: Connected double edge swap function. https://networkx.org/documentation/networkx-1.9/reference/generated/networkx.algorithms.swap.connected_double_edge_swap.html [date of access: 13/10/2020]. (referenced on page: 205)

[5] (2020). Networkx: K-shell decomposition. https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.core.k_shell.html [date of access: 13/10/2020]. (referenced on page: 205)

[6] (2021). Central hungary, wikipedia. https://en.wikipedia.org/wiki/Central_Hungary [date of access: 17/05/2021]. (referenced on page: 18)

[7] (2021a). Daily new covid-19 infections in hungary. https://koronavirus.gov.hu [date of access: 17/05/2021]. (referenced on page: 203)

[8] (2021). Implementation of epidemic modelling with metapopulation and percolation models. https://github.com/dczifra/epidemic_seeding. (referenced on pages: 28, 205)

[9] (2021b). Wikipedia: Covid-19 corona virus pandemic in hungary. https://hu.wikipedia.org/wiki/Covid19-koronav{í}rus-j{á}rv{á}ny_Magyarorsz{á}gon [date of access: 17/05/2021]. (referenced on page: 203)

[10] (2022). Semantic Scholar API. https://www.semanticscholar.org/product/api. (referenced on page: 13)

[11] Abraham, I., Chechik, S., and Krinninger, S. (2017). Fully dynamic all-pairs shortest paths with worst-case update-time revisited. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 440–452. SIAM. (referenced on page: 140)

[12] Aldous, D. et al. (1991). Asymptotic fringe distributions for general families of random trees. *The Annals of Applied Probability*, 1(2):228–266. (referenced on pages: 44, 48, and 49)

[13] Aleta, A., Martin-Corral, D., y Piontti, A. P., Ajelli, M., Litvinova, M., Chinazzi, M., Dean, N. E., Halloran, M. E., Longini Jr, I. M., Merler, S., et al. (2020). Modelling the impact of testing, contact tracing and household quarantine on second waves of covid-19. *Nature Human Behaviour*, 4(9):964–971. (referenced on page: 16)

[14] Altarelli, F., Braunstein, A., Dall'Asta, L., Ingrosso, A., and Zecchina, R. (2014a). The patient-zero problem with noisy observations. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(10):P10016. (referenced on pages: 6, 166)

[15] Altarelli, F., Braunstein, A., Dall'Asta, L., Lage-Castellanos, A., and Zecchina, R. (2014b). Bayesian inference of epidemics on networks via belief propagation. *Phys. Rev. Lett.*, 112(11):118701. (referenced on page: 6)

[16] Apolloni, A., Poletto, C., Ramasco, J. J., Jensen, P., and Colizza, V. (2014). Metapopulation epidemic models with heterogeneous mixing and travel behaviour. *Theoretical Biology and Medical Modelling*, 11(1):1–26. (referenced on page: 15)

[17] Arratia, R. and DeSalvo, S. (2013). On the singularity of random bernoulli matrices—novel integer partitions and lower bound expansions. *Annals of Combinatorics*, 17(2):251–274. (referenced on page: 67)

[18] Arratia, R., Goldstein, L., Gordon, L., et al. (1989). Two moments suffice for poisson approximations: the chen-stein method. *The Annals of Probability*, 17(1):9–25. (referenced on page: 67)

[19] Babai, L., Erdős, P., and Selkow, S. M. (1980). Random graph isomorphism. *SIaM Journal on computing*, 9(3):628–635. (referenced on page: 34)

[20] Bailey, R. F. and Cameron, P. J. (2011). Base size, metric dimension and other invariants of groups and graphs. *Bulletin of the London Mathematical Society*, 43(2):209–242. (referenced on page: 34)

[21] Bajardi, P., Poletto, C., Ramasco, J. J., Tizzoni, M., Colizza, V., and Vespignani, A. (2011). Human mobility networks, travel restrictions, and the global spread of 2009 h1n1 pandemic. *PloS one*, 6(1):e16591. (referenced on page: 16)

[22] Ball, F., Sirl, D., and Trapman, P. (2009). Threshold behaviour and final outcome of an epidemic on a random network with household structure. *Advances in Applied Probability*, 41(3):765–796. (referenced on page: 174)

[23] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512. (referenced on pages: 4, 43, 47, and 135)

[24] Bartha, Z., Komjáthy, J., and Raes, J. (2021). Sharp bound on the threshold metric dimension of trees. *arXiv preprint arXiv:2111.08813*. (referenced on page: 35)

[25] Barthélemy, M., Godreche, C., and Luck, J.-M. (2010). Fluctuation effects in metapopulation models: percolation and pandemic threshold. *Journal of theoretical biology*, 267(4):554–564. (referenced on page: 23)

[26] Beerliova, Z., Eberhard, F., Erlebach, T., Hall, A., Hoffmann, M., Mihal'ák, M., and Ram, L. S. (2006). Network discovery and verification. *IEEE Journal on selected areas in communications*, 24(12):2168–2181. (referenced on page: 34)

[27] Beidas, R. S., Buttenheim, A. M., Feuerstein-Simon, R., Kilaru, A. S., Asch, D. A., Volpp, K. G. M., Lawman, H. G., and Cannuscio, C. C. (2020). Optimizing and implementing contact tracing through behavioral economics. *Nejm Catalyst Innovations in Care Delivery*. (referenced on page: 166)

[28] Benjamini, I. and Schramm, O. (2011). Recurrence of distributional limits of finite planar graphs. In *Selected Works of Oded Schramm*, pages 533–545. Springer. (referenced on page: 174)

[29] Bensmail, J., Mazauric, D., Mc Inerney, F., Nisse, N., and Pérennes, S. (2018). Sequential metric dimension. In *International Workshop on Approximation and Online Algorithms*, pages 36–50. Springer. (referenced on page: 36)

[30] Berry, A. C. (1941). The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136. (referenced on page: 228)

[31] Bhamidi, S., van der Hofstad, R., and van Leeuwaarden, J. S. (2010). Scaling limits for critical inhomogeneous random graphs with finite third moments. *Electronic Journal of Probability*, 15:1682–1702. (referenced on pages: 27, 210)

[32] Bhamidi, S., van der Hofstad, R., Van Leeuwaarden, J. S., et al. (2012). Novel scaling limits for critical inhomogeneous random graphs. *Annals of Probability*, 40(6):2299–2361. (referenced on pages: 27, 210)

[33] Biskup, M. et al. (2004). On the scaling of the chemical distance in long-range percolation models. *Annals of Probability*, 32(4):2938–2977. (referenced on page: 204)

[34] Blum, M. G. B., François, O., and Janson, S. (2006). The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *Ann. Appl. Probab.*, 16(4):2195–2214. (referenced on page: 42)

[35] Bojja Venkatakrishnan, S., Fanti, G., and Viswanath, P. (2017). Dandelion: Redesigning the bitcoin network for anonymity. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(1):1–34. (referenced on page: 8)

[36] Bollobás, B., Mitsche, D., and Prałat, P. (2013). Metric dimension for random graphs. *The Electronic Journal of Combinatorics*, 20(4):P1. (referenced on pages: 9, 10, 14, 34, 38, 39, 65, 66, 68, 71, 78, 79, 80, 81, 85, 99, 225, and 226)

[37] Bollobás, B., Riordan, O., Spencer, J., and Tusnády, G. (2001). The degree sequence of a scale-free random graph process. *Random Structures Algorithms*, 18(3):279–290. (referenced on page: 43)

[38] Bradshaw, W. J., Alley, E. C., Huggins, J. H., Lloyd, A. L., and Esvelt, K. M. (2021). Bidirectional contact tracing could dramatically improve covid-19 control. *Nature communications*, 12(1):1–9. (referenced on pages: 165, 166)

[39] Braithwaite, I., Callender, T., Bullock, M., and Aldridge, R. W. (2020). Automated and partly automated contact tracing: a systematic review to inform the control of covid-19. *The Lancet Digital Health*. (referenced on page: 165)

[40] Brandt, A., Diemunsch, J., Erbes, C., LeGrand, J., and Moffatt, C. (2017). A robber locating strategy for trees. *Discrete Applied Mathematics*, 232:99–106. (referenced on page: 36)

[41] Brauner, J. M., Mindermann, S., Sharma, M., Johnston, D., Salvatier, J., Gavenčiak, T., Stephenson, A. B., Leech, G., Altman, G., Mikulik, V., et al. (2021). Inferring the effectiveness of government interventions against covid-19. *Science*, 371(6531). (referenced on page: 15)

[42] Bringmann, K., Keusch, R., and Lengler, J. (2016). Average distance in a general class of scale-free networks with underlying geometry. *arXiv preprint arXiv:1602.05712*, -1. (referenced on page: 204)

[43] Bringmann, K., Keusch, R., and Lengler, J. (2019). Geometric inhomogeneous random graphs. *Theoretical Computer Science*, 760:35–54. (referenced on page: 20)

[44] Brockmann, D. and Helbing, D. (2013). The hidden geometry of complex, network-driven contagion phenomena. *science*, 342(6164):1337–1342. (referenced on pages: 6, 13)

[45] Bubeck, S., Devroye, L., and Lugosi, G. (2017). Finding adam in random growing trees. *Random Structures & Algorithms*, 50(2):158–172. (referenced on page: 9)

[46] Burton, J. K., Bayne, G., Evans, C., Garbe, F., Gorman, D., Honhold, N., McCormick, D., Othieno, R., Stevenson, J. E., Swietlik, S., et al. (2020). Evolution and effects of covid-19 outbreaks in care homes: a population analysis in 189 care homes in one geographical region of the uk. *The Lancet Healthy Longevity*, 1(1):e21–e31. (referenced on page: 16)

[47] Cáceres, J., Garijo, D., Puertas, M. L., and Seara, C. (2010). On the determining number and the metric dimension of graphs. *the electronic journal of combinatorics*, pages R63–R63. (referenced on page: 34)

[48] Cáceres, J., Hernando, C., Mora, M., Pelayo, I. M., Puertas, M. L., Seara, C., and Wood, D. R. (2007a). On the metric dimension of cartesian products of graphs. *SIAM Journal on Discrete Mathematics*, 21(2):423–441. (referenced on page: 34)

[49] Cáceres, J., Hernando, C., Mora, M., Pelayo, I. M., Puertas, M. L., Seara, C., and Wood, D. R. (2007b). On the metric dimension of cartesian products of graphs. *SIAM Journal on Discrete Mathematics*, 21(2):423–441. (referenced on pages: 34, 148)

[50] Canonne, C. L. and Gur, T. (2018). An adaptivity hierarchy theorem for property testing. *Computational Complexity*, 27(4):671–716. (referenced on page: 132)

[51] Cărbune, V. (2014). Active learning for source localization. Master's thesis, ETH-Zürich. (referenced on page: 9)

[52] Carinci, F. (2020). Covid-19: preparedness, decentralisation, and the hunt for patient zero. (referenced on page: 3)

[53] Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y., and Shir, E. (2007). A model of internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences*, 104(27):11150–11154. (referenced on page: 205)

[54] Celis, L. E., Pavetic, F., Spinelli, B., and Thiran, P. (2015). Budgeted sensor placement for source localization on trees. *Electronic Notes in Discrete Mathematics*, 50:65–70. (referenced on page: 34)

[55] Chai, Y., Wang, Y., and Zhu, L. (2021). Information sources estimation in time-varying networks. *IEEE Transactions on Information Forensics and Security*, 16:2621–2636. (referenced on page: 187)

[56] Chartrand, G., Eroh, L., Johnson, M. A., and Oellermann, O. R. (2000). Resolvability in graphs and the metric dimension of a graph. *Discrete Applied Mathematics*, 105(1-3):99–113. (referenced on pages: 34, 35)

[57] Chen, X. and Wang, C. (2014). Approximability of the minimum weighted doubly resolving set problem. In *International Computing and Combinatorics Conference*, pages 357–368. Springer. (referenced on page: 34)

[58] Chiu, S.-E. and Javidi, T. (2016). Sequential measurement-dependent noisy search. In *2016 IEEE Information Theory Workshop (ITW)*, pages 221–225. IEEE. (referenced on pages: 132, 133)

[59] Cohen, R., Ben-Avraham, D., and Havlin, S. (2002). Percolation critical exponents in scale-free networks. *Physical Review E*, 66(3):036113. (referenced on pages: 25, 27, 29, and 210)

[60] Colizza, V., Pastor-Satorras, R., and Vespignani, A. (2007). Reaction–diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics*, 3(4):276–282. (referenced on pages: 5, 17)

[61] Colizza, V. and Vespignani, A. (2007). Invasion threshold in heterogeneous metapopulation networks. *Physical review letters*, 99(14):148701. (referenced on page: 15)

[62] Colizza, V. and Vespignani, A. (2008). Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations. *Journal of theoretical biology*, 251(3):450–467. (referenced on page: 23)

[63] Crepey, P., Alvarez, F. P., and Barthélemy, M. (2006). Epidemic variability in complex networks. *Physical Review E*, 73(4):046131. (referenced on page: 15)

[64] Danon, L., House, T., and Keeling, M. J. (2009). The role of routine versus random movements on the spread of disease in great britain. *Epidemics*, 1(4):250–258. (referenced on page: 5)

[65] Dawkins, Q., Li, T., and Xu, H. (2021). Diffusion source identification on networks with statistical confidence. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2500–2509. PMLR. (referenced on pages: 9, 198)

[66] Deijfen, M., van der Hofstad, R., and Hooghiemstra, G. (2013). Scale-free percolation. In *Annales de l'IHP Probabilités et statistiques*, volume 49, pages 817–838. (referenced on page: 204)

[67] Demetrescu, C. and Italiano, G. F. (2004). A new approach to dynamic all pairs shortest paths. *Journal of the ACM (JACM)*, 51(6):968–992. (referenced on page: 140)

[68] Deprez, P., Hazra, R. S., and Wüthrich, M. V. (2015). Inhomogeneous long-range percolation for real-life network modeling. *Risks*, 3(1):1–23. (referenced on page: 204)

[69] Devroye, L. (2002/03). Limit laws for sums of functions of subtrees of random binary search trees. *SIAM J. Comput.*, 32(1):152–171. (referenced on page: 42)

[70] Dhara, S., van der Hofstad, R., van Leeuwaarden, J., and Sen, S. (2017). Critical window for the configuration model: finite third moment degrees. *Electron. J. Probab*, 22(16):1–33. (referenced on pages: 27, 210)

[71] Dhara, S., van der Hofstad, R., and van Leeuwaarden, J. S. (2021). Critical percolation on scale-free random graphs: new universality class for the configuration model. *Communications in Mathematical Physics*, -1:1–49. (referenced on pages: 27, 210)

[72] Dhara, S., van der Hofstad, R., van Leeuwaarden, J. S., and Sen, S. (2016). Heavy-tailed configuration models at criticality. *arXiv preprint arXiv:1612.00650*, -1. (referenced on pages: 27, 210)

[73] Ding, J., Kim, J. H., Lubetzky, E., and Peres, Y. (2011). Anatomy of a young giant component in the random graph. *Random Structures & Algorithms*, 39(2):139–178. (referenced on pages: 27, 210)

[74] Dong, W., Zhang, W., and Tan, C. W. (2013). Rooting out the rumor culprit from suspects. In *2013 IEEE International Symposium on Information Theory*. IEEE. (referenced on page: 6)

[75] Drmota, M. (2009). *Random trees: an interplay between combinatorics and probability*. Springer Science & Business Media. (referenced on pages: 45, 167)

[76] Dudek, A., English, S., Frieze, A., MacRury, C., and Prałat, P. (2022). Localization game for random graphs. *Discrete Applied Mathematics*, 309:202–214. (referenced on pages: 36, 66)

[77] Dudek, A., Frieze, A., and Pegden, W. (2019). A note on the localization number of random graphs: diameter two case. *Discrete Applied Mathematics*, 254:107–112. (referenced on page: 66)

[78] Dye, C., Cheng, R. C., Dagpunar, J. S., and Williams, B. G. (2020). The scale and dynamics of covid-19 epidemics across europe. *Royal Society open science*, 7(11):201726. (referenced on page: 15)

[79] Eames, K., Bansal, S., Frost, S., and Riley, S. (2015). Six challenges in measuring contact networks for use in modelling. *Epidemics*, 10:72–77. (referenced on page: 11)

[80] Emamjomeh-Zadeh, E., Kempe, D., and Singhal, V. (2016). Deterministic and probabilistic binary search in graphs. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 519–532. ACM. (referenced on pages: 66, 108, and 134)

[81] Endo, A. et al. (2020). Implication of backward contact tracing in the presence of overdispersed transmission in covid-19 outbreaks. *Wellcome open research*, 5. (referenced on page: 165)

[82] Erdős, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae (Debrecen)*. (referenced on page: 135)

[83] Eroh, L., Feit, P., Kang, C. X., and Yi, E. (2015). The effect of vertex or edge deletion on the metric dimension of graphs. *Journal of Combinatorics*, 6(4):433–444. (referenced on pages: 35, 139, and 146)

[84] Eroh, L., Kang, C. X., and Yi, E. (2017). A comparison between the metric dimension and zero forcing number of trees and unicyclic graphs. *Acta Mathematica Sinica, English Series*, 33(6):731–747. (referenced on page: 35)

[85] Esseen, C.-G. et al. (1945). Fourier analysis of distribution functions. a mathematical study of the laplace-gaussian law. *Acta Mathematica*, 77:1–125. (referenced on page: 228)

[86] Estrada-Moreno, A., Rodriguez-Velazquez, J. A., and Gonzalez Yero, I. (2013). The k-metric dimension of a graph. *Applied Mathematics & Information Sciences*. (referenced on page: 35)

[87] Fan, L., Li, B., Liu, D., Dai, H., and Ru, Y. (2020). Identifying propagation source in temporal networks based on label propagation. In *International Conference of Pioneering Computer Scientists, Engineers and Educators*, pages 72–88. Springer. (referenced on page: 187)

[88] Fanti, G., Kairouz, P., Oh, S., and Viswanath, P. (2015). Spy vs. spy: Rumor source obfuscation. In *ACM SIGMETRICS Performance Evaluation Review*, volume 43, pages 271–284. ACM. (referenced on page: 8)

[89] Fanti, G. and Viswanath, P. (2017). Deanonymization in the bitcoin p2p network. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1364–1373. (referenced on page: 199)

[90] Farajtabar, M., Rodriguez, M. G., Zamani, M., Du, N., Zha, H., and Song, L. (2015). Back to the past: Source identification in diffusion networks from partially observed cascades. In *Artificial Intelligence and Statistics*, pages 232–240. PMLR. (referenced on page: 9)

[91] Fauver, J. R., Petrone, M. E., Hodcroft, E. B., Shioda, K., Ehrlich, H. Y., Watts, A. G., Vogels, C. B., Brito, A. F., Alpert, T., Muyombwe, A., et al. (2020). Coast-to-coast spread of sars-cov-2 during the early epidemic in the united states. *Cell*, 181(5):990–996. (referenced on page: 17)

[92] Feige, U., Raghavan, P., Peleg, D., and Upfal, E. (1994). Computing with noisy information. *SIAM Journal on Computing*, 23(5):1001–1018. (referenced on page: 134)

[93] Feng, Y. and Mahmoud, H. (2018). Profile of random exponential binary trees. *Methodology and Computing in Applied Probability*, 20(2):575–587. (referenced on pages: 167, 179, 185, and 273)

[94] Fernholz, D. and Ramachandran, V. (2003). The giant k-core of a random graph with a specified degree sequence. (referenced on page: 24)

[95] Frieze, A., Martin, R., Moncel, J., Ruszinkó, M., and Smyth, C. (2007). Codes identifying sets of vertices in random networks. *Discrete Mathematics*, 307(9-10):1094–1107. (referenced on pages: 35, 71)

[96] Fuchs, M. (2008). Subtree sizes in recursive trees and binary search trees: Berry-Esseen bounds and Poisson approximations. *Combin. Probab. Comput.*, 17(5):661–680. (referenced on page: 42)

[97] Funk, S., Gilad, E., Watkins, C., and Jansen, V. A. (2009). The spread of awareness and its impact on epidemic outbreaks. *Proceedings of the National Academy of Sciences*, 106(16):6872–6877. (referenced on page: 15)

[98] Garijo, D., González, A., and Márquez, A. (2014). The difference between the metric dimension and the determining number of a graph. *Applied Mathematics and Computation*, 249:487–501. (referenced on page: 34)

[99] Gautreau, A., Barrat, A., and Barthélemy, M. (2007). Arrival time statistics in global disease spread. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(09):L09001. (referenced on page: 15)

[100] Geneson, J., Kaustav, S., and Labelle, A. (2020). Extremal results for graphs of bounded metric dimension. *arXiv preprint arXiv:2008.13302*. (referenced on page: 149)

[101] Geneson, J. and Yi, E. (2020). Broadcast dimension of graphs. *arXiv preprint arXiv:2005.07311*. (referenced on page: 35)

[102] Gkantsidis, C., Mihail, M., and Zegura, E. W. (2003). The markov chain simulation method for generating connected power law random graphs. In *ALENEX*, pages 16–25. (referenced on page: 205)

[103] Glebov, R., Liebenau, A., and Szabó, T. (2015). On the concentration of the domination number of the random graph. *SIAM Journal on discrete mathematics*, 29(3):1186–1206. (referenced on page: 77)

[104] Gomez-Rodriguez, M., Leskovec, J., and Krause, A. (2012). Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):1–37. (referenced on page: 11)

[105] Gozzi, N., Tizzoni, M., Chinazzi, M., Ferres, L., Vespignani, A., and Perra, N. (2021). Estimating the effect of social inequalities on the mitigation of covid-19 across communities in santiago de chile. *Nature communications*, 12(1):1–9. (referenced on pages: 5, 29)

[106] Hall, M., Woolhouse, M., and Rambaut, A. (2015). Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set. *PLoS computational biology*, 11(12):e1004613. (referenced on page: 5)

[107] Harary, F. and Melter, R. A. (1976). On the metric dimension of a graph. *Ars Combin*, 2(191-195):1. (referenced on page: 33)

[108] Haslegrave, J., Johnson, R. A., and Koch, S. (2018). Locating a robber with multiple probes. *Discrete Mathematics*, 341(1):184–193. (referenced on page: 36)

[109] Hauptmann, M., Schmied, R., and Viehmann, C. (2012). Approximation complexity of metric dimension problem. *Journal of Discrete Algorithms*, 14:214–222. (referenced on page: 34)

[110] Hernando, C., Mora, M., Slater, P. J., and Wood, D. R. (2008). Fault-tolerant metric dimension of graphs. *Convexity in discrete structures*, 5:81–85. (referenced on pages: 35, 166)

[111] Hofstad, R. v. d. (2017). *Random graphs and complex networks Volume 1*. Cambridge Series in Statistical and Probabilistic Mathematics. (referenced on page: 43)

[112] Hofstad, R. v. d. (2020+). *Random graphs and complex networks Volume 2*. Cambridge Series in Statistical and Probabilistic Mathematics. (referenced on page: 53)

[113] Hofstad, R. v. d. and Litvak, N. (2014). Degree-degree dependencies in random graphs with heavy-tailed degrees. *Internet mathematics*, 10(3-4):287–334. (referenced on page: 27)

[114] Holmager, T. L., Lynge, E., Kann, C. E., and St-Martin, G. (2020). Geography of covid-19 in denmark. *Scandinavian Journal of Public Health*, -1:1403494820975607. (referenced on page: 29)

[115] Holmgren, C. and Janson, S. (2015). Limit laws for functions of fringe trees for binary search trees and random recursive trees. *Electron. J. Probab.*, 20:51 pp. (referenced on page: 45)

[116] Holmgren, C. and Janson, S. (2017). Fringe trees, Crump-Mode-Jagers branching processes and m-ary search trees. *Probab. Surveys*, 14:53–154. (referenced on pages: 42, 43, 44, 50, and 57)

[117] Hoorn, P. v. d. and Litvak, N. (2015). Degree-degree dependencies in directed networks with heavy-tailed degrees. *Internet mathematics*, 11(2):155–179. (referenced on page: 27)

[118] Hu, Y.-C., Perrig, A., and Johnson, D. B. (2006). Wormhole attacks in wireless networks. *IEEE journal on selected areas in communications*, 24(2):370–380. (referenced on page: 140)

[119] Hu, Z.-L., Shen, Z., Tang, C.-B., Xie, B.-B., and Lu, J.-F. (2018). Localization of diffusion sources in complex networks with sparse observations. *Physics Letters A*, 382(14):931–937. (referenced on page: 8)

[120] Huang, Q. (2017). Source locating of spreading dynamics in temporal networks. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 723–727. (referenced on page: 187)

[121] Inc., W. R. (2019). Mathematica online, Version 12.0. Champaign, IL, 2019. (referenced on page: 47)

[122] Ingraham, N. E. and Ingbar, D. H. (2021). The omicron variant of sars-cov-2: Understanding the known and living with unknowns. *Clinical and Translational Medicine*, 11(12):e685. (referenced on page: 3)

[123] Jagers, P. et al. (1975). *Branching processes with biological applications*. Wiley. (referenced on page: 50)

[124] Jagers, P. and Nerman, O. (1984). The growth and composition of branching populations. *Advances in applied probability*, 16(2):221–259. (referenced on page: 169)

[125] Janson, S. (1998). New versions of suen's correlation inequality. *Random Struct. Algorithms*, 13:467–483. (referenced on pages: 68, 69)

[126] Janson, S. (2013). Asymptotic normality of fringe subtrees and additive functionals in conditioned Galton-Watson trees. *Random Struct. Algorithms*, 48:57–101. (referenced on pages: 44, 48)

[127] Janson, S. and Luczak, M. J. (2007). A simple solution to the k-core problem. *Random Structures & Algorithms*, 30(1-2):50–62. (referenced on page: 24)

[128] Ji, F., Tay, W. P., and Varshney, L. R. (2017). An algorithmic framework for estimating rumor sources with different start times. *IEEE Trans. Signal Process.*, 65(10):2517–2530. (referenced on page: 6)

[129] Jiang, J. (2017). *Rumor source identification in complex networks*. PhD thesis, Deakin University. (referenced on page: 6)

[130] Jiang, J., Wen, S., Yu, S., Xiang, Y., and Zhou, W. (2016). Rumor source identification in social networks with time-varying topology. *IEEE Transactions on Dependable and Secure Computing*, 15(1):166–179. (referenced on pages: 167, 187, 188, 277, 280, and 282)

[131] Jiang, J., Wen, S., Yu, S., Xiang, Y., and Zhou, W. (2017). Identifying propagation sources in networks: State-of-the-art and comparative studies. *IEEE Communications Surveys & Tutorials*, 19(1):465–481. (referenced on page: 6)

[132] Jog, V. and Loh, P.-L. (2016). Analysis of centrality in sublinear preferential attachment trees via the Crump-Mode-Jagers branching process. *IEEE Transactions on Network Science and Engineering*, 4(1):1–12. (referenced on page: 43)

[133] Kahn, J., Komlós, J., and Szemerédi, E. (1995). On the probability that a random±1-matrix is singular. *Journal of the American Mathematical Society*, 8(1):223–240. (referenced on page: 67)

[134] Kandeel, M., Mohamed, M. E. M., Abd El-Lateef, H. M., Venugopala, K. N., and El-Beltagi, H. S. (2021). Omicron variant genome evolution and phylogenetics. *Journal of Medical Virology*. (referenced on page: 3)

[135] Kang, D., Choi, H., Kim, J.-H., and Choi, J. (2020). Spatial epidemic dynamics of the covid-19 outbreak in china. *International Journal of Infectious Diseases*, 94:96–102. (referenced on page: 17)

[136] Karp, R. M. and Kleinberg, R. (2007). Noisy binary search and its applications. In *Proceedings SODA*, pages 881–890. (referenced on page: 134)

[137] Karpovsky, M. G., Chakrabarty, K., and Levitin, L. B. (1998). On a new class of codes for identifying vertices in graphs. *IEEE Transactions on Information Theory*, 44(2):599–611. (referenced on page: 71)

[138] Karrer, B. and Newman, M. E. J. (2010). Message passing approach for general epidemic models. *Phys. Rev. E*, 82:016101. (referenced on page: 278)

[139] Karsai, M., Koltai, J., Vásárhelyi, O., and Röst, G. (2020). Hungary in mask/maszk in hungary. *Corvinus Journal of Sociology and Social Policy*, 1(2). (referenced on page: 17)

[140] Khim, J. and Loh, P.-L. (2016). Confidence sets for the source of a diffusion in regular trees. *IEEE Transactions on Network Science and Engineering*, 4(1):27–40. (referenced on page: 9)

[141] Khuller, S., Raghavachari, B., and Rosenfeld, A. (1996). Landmarks in graphs. *Discrete Applied Mathematics*, 70(3):217–229. (referenced on pages: 33, 34, and 148)

[142] Kim, Y., Kumbhat, M., Nagy, Z. L., Patkós, B., Pokrovskiy, A., and Vizer, M. (2015). Identifying codes and searching with balls in graphs. *Discrete Applied Mathematics*, 193:39–47. (referenced on pages: 38, 65, and 66)

[143] Kitsak, M., Gallos, L., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H., and Makse, H. (2010). Identification of influential spreaders in complex networks. *Nature Physics*, 6. (referenced on pages: 5, 15)

[144] Knuth, D. E. (1997). *The art of computer programming. Vol. 1.* Addison-Wesley, Reading, MA. Fundamental algorithms, Third edition [of MR0286317]. (referenced on page: 42)

[145] Knuth, D. E. (1998). *The art of computer programming. Vol. 3.* Addison-Wesley, Reading, MA. Sorting and searching, Second edition [of MR0445948]. (referenced on page: 42)

[146] Kojaku, S., Hébert-Dufresne, L., Mones, E., Lehmann, S., and Ahn, Y.-Y. (2021). The effectiveness of backward contact tracing in networks. *Nature Physics*, pages 1–7. (referenced on page: 165)

[147] Komjáthy, J. and Ódor, G. (2021). Metric dimension of critical galton–watson trees and linear preferential attachment trees. *European Journal of Combinatorics*, 95:103317. (referenced on pages: 9, 13, 14, and 41)

[148] Komlós, J. (1967). On determinant of (0, 1) matrices. *Studia Science Mathematics Hungarica*, 2:7–21. (referenced on page: 67)

[149] Kraemer, M. U., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D. M., Du Plessis, L., Faria, N. R., Li, R., Hanage, W. P., et al. (2020). The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497. (referenced on page: 16)

[150] Kretzschmar, M. E., Rozhnova, G., Bootsma, M. C., van Boven, M., van de Wijgert, J. H., and Bonten, M. J. (2020). Impact of delays on effectiveness of contact tracing strategies for covid-19: a modelling study. *The Lancet Public Health*, 5(8):e452–e459. (referenced on pages: 165, 166)

[151] Kumar, A., Borkar, V. S., and Karamchandani, N. (2017). Temporally agnostic rumor-source detection. *IEEE Trans. Signal Inf. Process. Netw.*, 3(2):316–329. (referenced on page: 8)

[152] Kupferschmidt, K. (2021). *Where did 'weird'Omicron come from?* American Association for the Advancement of Science. (referenced on page: 3)

[153] Kuziak, D. and Yero, I. G. (2021). Metric dimension related parameters in graphs: A survey on combinatorial, computational and applied results. *arXiv preprint arXiv:2107.04877.* (referenced on pages: 34, 198)

[154] Lalitha, A., Ronquillo, N., and Javidi, T. (2017). Measurement dependent noisy search: The gaussian case. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3090–3094. IEEE. (referenced on pages: 132, 133, and 134)

[155] Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR. (referenced on page: 13)

[156] Lecomte, V., Ódor, G., and Thiran, P. (2022). The power of adaptivity in source identification with time queries on the path. *in press at Theoretical Computer Science.* (referenced on pages: 11, 13, 14, 38, 108, and 111)

[157] Lee, J., Choi, B. Y., and Jung, E. (2018). Metapopulation model using commuting flow for national spread of the 2009 h1n1 influenza virus in the republic of korea. *Journal of theoretical biology*, 454:320–329. (referenced on page: 5)

[158] Li, X., Wang, X., Zhao, C., Zhang, X., and Yi, D. (2019). Locating the source of diffusion in complex networks via gaussian-based localization and deduction. *Applied Sciences*, 9(18):3758. (referenced on page: 8)

[159] Lichev, L., Mitsche, D., and Pralat, P. (2021). Localization game for random geometric graphs. *arXiv preprint arXiv:2102.10352.* (referenced on pages: 9, 34, and 38)

[160] Litvak, N. and Hofstad, R. v. d. (2013). Uncovering disassortativity in large scale-free networks. *Physical Review E*, 87(2):022801. (referenced on page: 27)

[161] Lokhov, A. Y., Mézard, M., Ohta, H., and Zdeborová, L. (2014). Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Physical Review E*, 90(1):012801. (referenced on pages: 6, 167, 187, 188, 277, and 281)

[162] Lorch, L., Kremer, H., Trouleau, W., Tsirtsis, S., Szanto, A., Schölkopf, B., and Gomez-Rodriguez, M. (2020). Quantifying the effects of contact tracing, testing, and containment measures in the presence of infection hotspots. *arXiv preprint arXiv:2004.07641.* (referenced on pages: 12, 166, 172, 173, 174, and 176)

[163] Louni, A., Santhanakrishnan, A., and Subbalakshmi, K. (2015). Identification of source of rumors in social networks with incomplete information. *arXiv preprint arXiv:1509.00557.* (referenced on page: 166)

[164] Louni, A. and Subbalakshmi, K. P. (2014). A two-stage algorithm to estimate the source of information diffusion in social media networks. In *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE. (referenced on page: 10)

[165] Łuczak, T. (1991). Size and connectivity of the k-core of a random graph. *Discrete Mathematics*, 91(1):61–68. (referenced on page: 24)

[166] Luo, W., Tay, W. P., and Leng, M. (2014). How to identify an infection source with limited observations. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):586–597. (referenced on page: 6)

[167] Mahmoud, H. (2021). Profile of random exponential recursive trees. *Methodology and Computing in Applied Probability*, pages 1–17. (referenced on pages: 167, 171, and 273)

[168] Manitz, J., Harbering, J., Schmidt, M., Kneib, T., and Schöbel, A. (2014a). Network-based source detection: From infectious disease spreading to train delay propagation. In *29th International Workshop on Statistical Modelling*, volume 1, pages 201–205. (referenced on pages: 3, 199)

[169] Manitz, J., Kneib, T., Schlather, M., Helbing, D., and Brockmann, D. (2014b). Origin detection during food-borne disease outbreaks-a case study of the 2011 ehec/hus outbreak in germany. *PLoS currents*, 6. (referenced on pages: 3, 199)

[170] Manuel, P., Rajan, B., Rajasingh, I., and Monica, M. C. (2006). Landmarks in torus networks. *Journal of Discrete Mathematical Sciences and Cryptography*, 9(2):263–271. (referenced on page: 140)

[171] Mase, S. (1992). Approximations to the birthday problem with unequal occurrence probabilities and their application to the surname problem in japan. *Annals of the Institute of Statistical Mathematics*, 44(3):479–499. (referenced on page: 67)

[172] Mashkaria, S. (2020). Verification of a conjecture regarding metric dimension of a grid augmented with one edge. https://zenodo.org/record/3999323. (referenced on page: 163)

[173] Mashkaria, S., Ódor, G., and Thiran, P. (2020). On the robustness of the metric dimension to adding a single edge. *arXiv preprint arXiv:2010.11023*. (referenced on pages: 11, 13, 14, and 139)

[174] Melter, R. A. and Tomescu, I. (1984). Metric bases in digital geometry. *Computer Vision, Graphics, and Image Processing*, 25(1):113–121. (referenced on page: 156)

[175] Memish, Z. A., Aljerian, N., and Ebrahim, S. H. (2021). Tale of three seeding patterns of sars-cov-2 in saudi arabia. *The Lancet Infectious Diseases*, 21(1):26–27. (referenced on page: 15)

[176] Merow, C. and Urban, M. C. (2020). Seasonality and uncertainty in global covid-19 growth rates. *Proceedings of the National Academy of Sciences*, 117(44):27456–27464. (referenced on page: 16)

[177] Mitsche, D. and Rué, J. (2015). On the limiting distribution of the metric dimension for random forests. *European Journal of Combinatorics*, 49:68–89. (referenced on pages: 9, 34, 43, and 44)

[178] Mol, L., Murphy, M. J., and Oellermann, O. R. (2020). The threshold dimension of a graph. *arXiv preprint arXiv:2001.09168*. (referenced on page: 35)

[179] Mossel, E. and Ross, N. (2017). Shotgun assembly of labeled graphs. *IEEE Transactions on Network Science and Engineering*, 6(2):145–157. (referenced on page: 11)

[180] Müller, S. A., Balmer, M., Charlton, W., Ewert, R., Neumann, A., Rakow, C., Schlenther, T., and Nagel, K. (2021). Predicting the effects of covid-19 related interventions in urban settings by combining activity-based modelling, agent-based simulation, and mobile phone data. *PloS one*, 16(10):e0259037. (referenced on page: 4)

[181] Nerman, O. (1981). On the convergence of supercritical general (CMJ) branching processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(3):365–395. (referenced on page: 50)

[182] Newman, M. (2018). *Networks*. Oxford university press. (referenced on pages: 4, 20)

[183] Newman, M. E., Strogatz, S. H., and Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 64(2):026118. (referenced on pages: 25, 27, 29, and 210)

[184] Nowak, R. D. (2011). The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12):7893–7906. (referenced on page: 66)

[185] Nowakowski, R. and Winkler, P. (1983). Vertex-to-vertex pursuit in a graph. *Discrete Mathematics*, 43(2-3):235–239. (referenced on page: 36)

[186] Ódor, G. (2020). dynamic_source_localization_toolbox. https://github.com/odorgergo/dynamic_source_localization_toolbox. (referenced on page: 40)

[187] Ódor, G., Czifra, D., Komjáthy, J., Lovász, L., and Karsai, M. (2021a). Switchover phenomenon induced by epidemic seeding on geometric networks. *Proceedings of the National Academy of Sciences*, 118(41). (referenced on pages: 5, 13, 14, 15, and 25)

[188] Ódor, G. and Thiran, P. (2021). Sequential metric dimension for random graphs. *Journal of Applied Probability*, 58(4):909–951. (referenced on pages: 10, 13, 14, 39, 65, and 111)

[189] Ódor, G., Vuckovic, J., Ndoye, M.-A. S., and Thiran, P. (2021b). Source detection via contact tracing in the presence of asymptomatic patients. *arXiv preprint arXiv:2112.14530*. (referenced on pages: 12, 13, 14, and 165)

[190] Onak, K. and Parys, P. (2006). Generalization of binary search: Searching in trees and forest-like partial orders. In *Proceedings FOCS'06*, pages 379–388. IEEE. (referenced on page: 134)

[191] Paluch, R., Gajewski, Ł. G., Hołyst, J. A., and Szymanski, B. K. (2020a). Optimizing sensors placement in complex networks for localization of hidden signal source: A review. *Future Generation Computer Systems*, 112:1070–1092. (referenced on pages: 8, 9)

[192] Paluch, R., Lu, X., Suchecki, K., Szymański, B. K., and Hołyst, J. A. (2018). Fast and accurate detection of spread source in large complex networks. *Scientific reports*, 8(1):1–10. (referenced on pages: 8, 170)

[193] Paluch, R., Suchecki, K., and Hołyst, J. A. (2020b). Locating the source of interacting signal in complex networks. *arXiv preprint arXiv:2012.02039*. (referenced on page: 8)

[194] Park, Y. J., Choe, Y. J., Park, O., Park, S. Y., Kim, Y.-M., Kim, J., Kweon, S., Woo, Y., Gwack, J., Kim, S. S., et al. (2020). Contact tracing during coronavirus disease outbreak, south korea, 2020. *Emerging infectious diseases*, 26(10):2465. (referenced on page: 165)

[195] Pastor-Satorras, R., Castellano, C., Van Mieghem, P., and Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925. (referenced on pages: 4, 15)

[196] Penrose, M. et al. (2003). *Random geometric graphs*, volume 5. Oxford university press. (referenced on pages: 38, 135)

[197] Petrov, V. V. (2012). *Sums of independent random variables*, volume 82. Springer Science & Business Media. (referenced on page: 228)

[198] Pinto, P. C., Thiran, P., and Vetterli, M. (2012). Locating the source of diffusion in large-scale networks. *Phys. Rev. Lett.*, 109:068702. (referenced on pages: 3, 4, 6, 8, 14, 120, 134, and 166)

[199] Prakash, B. A., Vreeken, J., and Faloutsos, C. (2012). Spotting culprits in epidemics: How many and which ones? In *2012 IEEE 12th International Conference on Data Mining*, pages 11–20. (referenced on page: 6)

[200] Quilliot, A. (1985). A short note about pursuit games played on a graph with a given genus. *Journal of combinatorial theory, Series B*, 38(1):89–92. (referenced on page: 36)

[201] Rácz, M. Z. and Richey, J. (2020). Rumor source detection with multiple observations under adaptive diffusions. *IEEE Transactions on Network Science and Engineering*, 8(1):2–12. (referenced on page: 9)

[202] Raj, F. S. and George, A. (2017). On the metric dimension of HDN 3 and PHDN 3. In *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, pages 1333–1336. (referenced on page: 34)

[203] Reicher, S. and Drury, J. (2021). Pandemic fatigue? how adherence to covid-19 regulations has been misrepresented and why it matters. *bmj*, 372. (referenced on page: 16)

[204] Rényi, A. (1961). On a problem of information theory. *MTA Matematikai Kutato Intezet Kozlemeny B*, 6:505–516. (referenced on page: 133)

[205] Röst, G., Bartha, F. A., Bogya, N., Boldog, P., Dénes, A., Ferenci, T., Horváth, K. J., Juhász, A., Nagy, C., Tekeli, T., et al. (2020). Early phase of the covid-19 outbreak in hungary and post-lockdown scenarios. *Viruses*, 12(7):708. (referenced on page: 17)

[206] Russo, L., Anastassopoulou, C., Tsakris, A., Bifulco, G. N., Campana, E. F., Toraldo, G., and Siettos, C. (2020). Tracing day-zero and forecasting the covid-19 outbreak in lombardy, italy: A compartmental modelling and numerical optimization approach. *Plos one*, 15(10):e0240649. (referenced on page: 3)

[207] Salathe, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., Campbell, E. M., Cattuto, C., Khandelwal, S., Mabry, P. L., et al. (2012). Digital epidemiology. (referenced on page: 11)

[208] Seager, S. (2012). Locating a robber on a graph. *Discrete Mathematics*, 312(22):3265–3269. (referenced on page: 36)

[209] Seager, S. (2014). Locating a backtracking robber on a tree. *Theoretical Computer Science*, 539:28–37. (referenced on page: 36)

[210] Seager, S. M. (2013). A sequential locating game on graphs. *Ars Comb*, 110. (referenced on page: 36)

[211] Sebő, A. and Tannier, E. (2004). On metric generators of graphs. *Mathematics of Operations Research*, 29(2):383–393. (referenced on page: 149)

[212] Seo, E., Mohapatra, P., and Abdelzaher, T. (2012). Identifying rumors and their sources in social networks. In *Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR III*. SPIE. (referenced on page: 6)

[213] Settles, B. (2009). Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences. (referenced on page: 132)

[214] Shah, D. and Zaman, T. (2010). Detecting sources of computer viruses in networks: Theory and experiment. In *Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '10, pages 203–214, New York, NY, USA. ACM. (referenced on page: 6)

[215] Shah, D. and Zaman, T. (2011). Rumors in a network: Who's the culprit? *IEEE Transactions on information theory*, 57(8):5163–5181. (referenced on pages: 3, 6, and 14)

[216] Shao, Z., Wu, P., Zhu, E., and Chen, L. (2019). On metric dimension in some hex derived networks. *Sensors*, 19(1):94. (referenced on page: 33)

[217] Shen, Z., Cao, S., Wang, W.-X., Di, Z., and Stanley, H. E. (2016). Locating the source of diffusion in complex networks by time-reversal backward spreading. *Physical Review E*, 93(3):032301. (referenced on page: 8)

[218] Slater, P. J. (1975). Leaves of trees. *Congr. Numer*, 14(549-559):37. (referenced on pages: 9, 33, and 45)

[219] Sooryanarayana, B., Kunikullaya, S., and Swamy, N. (2016). k-metric dimension of a graph. *International Journal of Mathematical Combinatorics*, 4. (referenced on page: 35)

[220] Spinelli, B. (2018a). Code for the paper: Back to the source: An online approach forsensor placement and source localization. https://github.com/bmspinelli/back_to_the_source. Accessed on 2021.12.21. (referenced on page: 230)

[221] Spinelli, B. (2018b). *Localizing the Source of an Epidemic Using Few Observations*. PhD thesis, EPFL. (referenced on pages: 4, 6, and 9)

[222] Spinelli, B., Celis, E., and Thiran, P. (2019). A general framework for sensor placement in source localization. *IEEE Transactions on Network Science and Engineering*, 6:86–102. (referenced on pages: 10, 108)

[223] Spinelli, B., Celis, L. E., and Thiran, P. (2016). Observer placement for source localization: The effect of budgets and transmission variance. *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 743–751. (referenced on page: 108)

[224] Spinelli, B., Celis, L. E., and Thiran, P. (2017a). Back to the source: An online approach for sensor placement and source localization. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1151–1160. (referenced on pages: 10, 135, 167, 188, 189, and 230)

[225] Spinelli, B., Celis, L. E., and Thiran, P. (2017b). The effect of transmission variance on observer placement for source-localization. *Applied network science*, 2(1):20. (referenced on page: 8)

[226] Spinelli, B., Celis, L. E., and Thiran, P. (2018). How many sensors to localize the source? the double metric dimension of random networks. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1036–1043. IEEE. (referenced on pages: 34, 111)

[227] Stone, L., Olinky, R., and Huppert, A. (2007). Seasonal dynamics of recurrent epidemics. *Nature*, 446(7135):533–536. (referenced on page: 15)

[228] Suen, W. S. (1990). A correlation inequality and a Poisson limit theorem for nonoverlapping balanced subgraphs of a random graph. *Random Structures & Algorithms*, 1(2):231–242. (referenced on pages: 68, 69)

[229] Szocska, M., Pollner, P., Schiszler, I., Joo, T., Palicz, T., McKee, M., Asztalos, A., Bencze, L., Kapronczay, M., Petrecz, P., et al. (2021). Countrywide population movement monitoring using mobile devices generated (big) data during the covid-19 crisis. *Scientific reports*, 11(1):1–9. (referenced on page: 16)

[230] Tang, W., Ji, F., and Tay, W. P. (2018). Estimating infection sources in networks using partial timestamps. *IEEE Transactions on Information Forensics and Security*, 13(12):3035–3049. (referenced on page: 8)

[231] Tao, T. and Vu, V. (2006). On random±1 matrices: singularity and determinant. *Random Structures & Algorithms*, 28(1):1–23. (referenced on page: 67)

[232] Tillquist, R. C., Frongillo, R. M., and Lladser, M. E. (2021a). Getting the lay of the land in discrete space: A survey of metric dimension and its applications. *arXiv preprint arXiv:2104.07201*. (referenced on page: 34)

[233] Tillquist, R. C., Frongillo, R. M., and Lladser, M. E. (2021b). Truncated metric dimension for finite graphs. *arXiv preprint arXiv:2106.14314*. (referenced on page: 35)

[234] Troncoso, C., Payer, M., Hubaux, J.-P., Salathé, M., Larus, J., Bugnion, E., Lueks, W., Stadler, T., Pyrgelis, A., Antonioli, D., et al. (2020). Decentralized privacy-preserving proximity tracing. *arXiv preprint arXiv:2005.12273*. (referenced on pages: 165, 191)

[235] Tsitsiklis, J., Xu, K., and Xu, Z. (2018). Private sequential learning. In *Conference On Learning Theory*, pages 721–727. (referenced on page: 135)

[236] van der Hofstad, R., Janson, S., and Luczak, M. (2019). Component structure of the configuration model: barely supercritical case. *Random Structures & Algorithms*, 55(1):3–55. (referenced on pages: 27, 210)

[237] van der Hofstad, R., Kliem, S., and van Leeuwaarden, J. S. (2018). Cluster tails for critical power-law inhomogeneous random graphs. *Journal of statistical physics*, 171(1):38–95. (referenced on pages: 27, 210)

[238] Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press. (referenced on page: 227)

[239] Volz, E. M. and Frost, S. D. (2013). Inferring the source of transmission with phylogenetic data. *PLoS computational biology*, 9(12):e1003397. (referenced on page: 5)

[240] Wang, Z., Dong, W., Zhang, W., and Tan, C. W. (2014). Rumor source detection with multiple observations. In *The 2014 ACM international conference on Measurement and modeling of computer systems - SIGMETRICS '14*, New York, New York, USA. ACM Press. (referenced on page: 9)

[241] White, E. R. and Hébert-Dufresne, L. (2020). State-level variation of initial covid-19 dynamics in the united states. *PloS one*, 15(10):e0240648. (referenced on page: 15)

[242] Wormald, N. C. et al. (1999). Models of random regular graphs. *London Mathematical Society Lecture Note Series*, pages 239–298. (referenced on page: 174)

[243] Xie, Y., Sekar, V., Maltz, D. A., Reiter, M. K., and Zhang, H. (2005). Worm origin identification using random moonwalks. In *2005 IEEE Symposium on Security and Privacy (S&P'05)*, pages 242–256. IEEE. (referenced on page: 3)

[244] Xu, J., Xu, K., and Yang, D. (2021). Optimal query complexity for private sequential learning against eavesdropping. In *International Conference on Artificial Intelligence and Statistics*, pages 2296–2304. PMLR. (referenced on page: 135)

[245] Xu, K. (2018). Query complexity of bayesian private learning. In *Advances in Neural Information Processing Systems*, pages 2431–2440. (referenced on page: 135)

[246] Xu, S., Teng, C., Zhou, Y., Peng, J., Zhang, Y., and Zhang, Z.-K. (2019). Identifying the diffusion source in complex networks with limited observers. *Physica A: Statistical Mechanics and its Applications*, 527:121267. (referenced on page: 8)

[247] Yartseva, L., Simoes, J. E., and Grossglauser, M. (2016). Assembling a network out of ambiguous patches. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 883–890. IEEE. (referenced on page: 11)

[248] Zejnilovic, S., Gomes, J., and Sinopoli, B. (2013). Network observability and localization of the source of diffusion based on a subset of nodes. In *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, pages 847–852. IEEE. (referenced on pages: 9, 33, and 111)

[249] Zejnilović, S., Gomes, J., and Sinopoli, B. (2015). Sequential observer selection for source localization. In *Signal and Information Processing (GlobalSIP), 2015 IEEE Global Conference on*, pages 1220–1224. IEEE. (referenced on pages: 10, 188)

[250] Zejnilović, S., Gomes, J., and Sinopoli, B. (2017). Sequential source localization on graphs: A case study of cholera outbreak. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1010–1014. IEEE. (referenced on pages: 8, 10)

[251] Zejnilović, S., Mitsche, D., Gomes, J., and Sinopoli, B. (2016). Extending the metric dimension to graphs with missing edges. *Theoretical Computer Science*, 609:384–394. (referenced on pages: 11, 35)

[252] Zejnilovic, S., Xavier, J., Gomes, J., and Sinopoli, B. (2015). Selecting observers for source localization via error exponents. In *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE. (referenced on page: 8)

[253] Zhang, X., Zhang, Y., Lv, T., and Yin, Y. (2016). Identification of efficient observers for locating spreading source in complex networks. *Physica A: Statistical Mechanics and its Applications*, 442:100–109. (referenced on page: 8)

[254] Zhang, Z., Xu, W., Wu, W., and Du, D.-Z. (2017). A novel approach for detecting multiple rumor sources in networks with partial observations. *J. Comb. Optim.*, 33(1):132–146. (referenced on page: 34)

[255] Zhou, L. and Hero, A. O. (2021). Resolution limits for the noisy non-adaptive 20 questions problem. *IEEE Transactions on Information Theory*, 67(4):2055–2073. (referenced on page: 133)

[256] Zhu, K., Chen, Z., and Ying, L. (2016). Locating the contagion source in networks with partial timestamps. *Data Mining and Knowledge Discovery*, 30(5):1217–1248. (referenced on page: 8)

[257] Zhu, K. and Ying, L. (2014). A robust information source estimator with sparse observations. In *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*. IEEE. (referenced on page: 6)

[258] Zhu, K. and Ying, L. (2016). Information source detection in the SIR model: A sample-path-based approach. *IEEE ACM Trans. Netw.*, 24(1):408–421. (referenced on page: 6)

Gergely Ódor

Ph.D. student in Computer Science at EPFL
Email: gergely.odor@epfl.ch
Website: https://www.gergelyodor.com/

Education

Ph.D. in Computer Science, *École polytechnique fédérale de Lausanne*, Lausanne, Switzerland

2017 – present I Thesis topic: The role of adaptivity in source location I Advisor: Prof. Patrick Thiran

M.S. in Mathematics, *Central European University*, Budapest, Hungary

2016 – 2017 I Thesis: Global information loss and criticality in resistance matrices I Advisor: Prof. Bálint Virág

B.S. in Mathematics with Computer Science, *Massachusetts Institute of Technology*, Cambridge, MA

2012 – 2016 I Cumulative GPA**:** 4.9 (out of 5.0)

Preprints

G. Ódor, J. Vuckovic, M.S. Ndoye and P. Thiran
"Source Detection via Contact Tracing in the Presence of Asymptomatic Patients"
arXiv preprint arXiv: 2112.14530 (2021)

Y. Meirovitch, A. Matveev, H. Saribekyan, D. Budden, D. Rolnick, G. Odor, S. Knowles-Barley, T.R. Jones, H. Pfister, J.W. Lichtman, N. Shavit,
"A Multi-Pass Approach to Large-Scale Connectomics,"
arXiv preprint arXiv:1612.02120 (2016)

Peer-reviewed Journal Publications

V. Lecomte, G. Ódor, and P. Thiran
"The power of adaptivity in source identification with time queries on the path"
Theoretical Computer Science, Volume 911, pp. 92 – 123; (2022)

S. Mashkaria, G. Ódor, and P. Thiran
"On the robustness of the metric dimension of grid graphs to adding a single edge"
Discrete Applied Mathematics, in press (2022)

G. Ódor, D. Czifra, J. Komjáthy, L. Lovász, and M. Karsai,
"Switchover phenomenon induced by epidemic seeding on geometric networks"
Proceedings of the National Academy of Sciences, 118(41); (2021)

G. Ódor, and P. Thiran,
"Sequential metric dimension for random graphs"
Journal of Applied Probability, Volume 58, Issue 4, December pp. 909 – 951; (2021)

J. Komjáthy, and G. Ódor,
"Metric dimension of critical Galton–Watson trees and linear preferential attachment trees"
European Journal of Combinatorics, 95, p.103317; (2021)

A. Matveev, Y. Meirovitch, H. Saribekyan, W. Jakubiuk, T. Kaler, G.O., D. Budden, A. Zlateski, N.Shavit,
"A multicore path to connectomics-on-demand"
Proceedings of the 22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming;
(2017) (Best Paper nominee).

G. [Géza] Ódor, R. Dickman, G. [Gergely] Ódor,
"Griffiths phases and localization in hierarchical modular networks"
Sci. Rep. 5, 14451 (Nature Publishing Group); (2015).

H. Schulz, G. [Géza] Ódor, G. [Gergely] Ódor, M. F. Nagy,
"Simulation of 1+1 dimensional surface growth and lattices gases using GPUs"
Comp. Phys. Comm. 182; (2011)

Peer-reviewed Conference Publications

G. Odor, Y.-H. Li, A. Yurtsever, Y.-P. Hsieh, Q. Tran-Dinh, M. El Halabi and V. Cevher,
"Frank-Wolfe works for non-Lipschitz continuous gradient objectives: scalable Poisson phase retrieval,"
ICASSP (2016)

Presentations

2021 November: Budapest Semesters in Mathematics Colloquium (Budapest)
Title: Switchover phenomenon induced by epidemic seeding on geometric networks (invited speaker)

2021 July: Franco-Dutch meeting "Bézout-Eurandom" (IHP Paris)
Title: Switchover phenomenon induced by epidemic seeding on geometric networks

2021 July: Rátz László Conference of Mathematics Teachers (online)
Title: Random trees and epidemic spreading (invited speaker)

2020 February: Budapest University of Technology and Economics Stochastic Seminar (TU Budapest)
Title: Sequential metric dimension for random graphs (invited speaker)

2019 July: 19th International Conference on Random Structures and Algorithms (ETH Zurich)
Title: Sequential metric dimension for random graphs (contributed talk)

2019 March: YEP XV "Information Diffusion on Random Networks" (TU Eindhoven)
Title: Source localization with adaptive sensor selection in random graphs (contributed talk)

2018 April: Wiki Workshop at The Web Conference (WWW2018 Lyon)
Title: How did Wikipedia become navigable (poster)

2014, 2015, 2018 December: Statistical Physics Holiday Seminar, Eötvös Lóránd University (ELTE)

Reseach internships

Alfréd Rényi Institute of Mathematics (Hungarian Academy of Science) – Budapest, Hungary

Temporary research position in the group of Prof. Bálint Virág 06.2017 – 08.2017

Computational Connectomics Group, MIT CSAIL – Cambridge, MA

Undergrad Researcher under the direction of Prof. Nir Shavit 09.2014 – 05.2016

Laboratory for Information and Inference Systems, EPFL – Lausanne, Switzerland

Research Intern under the direction of Prof. Volkan Cevher 06.2015 – 08.2015

Bear Lab, MIT – Cambridge, MA

Undergrad Researcher under the direction of Profs Mark Bear and Arnold Heynen 02.2013 – 05.2014

Teaching activities

Teaching Assistantship:

- Dynamical system theory for engineers (EPFL, 09.2018 – 01.2019, 09.2019 –01.2020)
- Probabilities and statistics (EPFL, 02.2019 – 06.2019)
- Theory of Computation (EPFL, 02.2018 – 06.2018)
- Matrix Computations with Applications (CEU, 02.2017 – 06.2017)

Tutoring:

- Mathematics and English for disadvantaged children (Menetszél Association 02.2020 – 06.2021)
- Introductory mathematics classes for MIT students (MIT Math Learning Center, 09.2014 – 05.2015)
- Advanced computer science classes for MIT students (HKN Tutoring, 02.2015. – 05.2015)

Supervision of students/junior researchers

- Jana Vuckovic (summer@EPFL intern 2021)
- Miguel-Angel Sanchez Ndoye (EPFL Student Assistant Spring 2021, Summer 2021)
- Stanislas Jouven (EPFL BA semester project, Spring 2019, Fall 2019)
- Victor Lecomte (summer@EPFL intern, 2019)
- Satvik Mashkaria (summer@EPFL intern, 2019)
- Nicolas D'Argenlieu (EPFL BA semester project, Spring 2019)
- Constantin Isabela (EPFL MS semester project, Fall 2018)
- Farzad Pourkamali (summer@EPFL intern, 2018)
- Shivani Angappan and Kejia Wang (MIT PRIMES Circle, Spring 2016)

Outreach activities

- Tutored online Hungarian disadvantaged students in 5th and 7th grade in Mathematics and English (Menetszél Association from 2020-2021)
- Tutored at an after-school program for Boston-area public high school students that offers a mathematical enrichment curriculum and an introductory research experience to talented students with disadvantaged backgrounds. (MIT PRIMES Circle tutor in Spring 2016)

Awards

- The paper “A multicore path to connectomics-on-demand” was nominated for Best Paper Award at PPoPP17
- International Mathematical Olympiad, Mar del Plata, 2012 – Honorable Mention
- International Olympiad in Informatics, Hungarian Qualifiers 2012 – 5th place
- W. L. Putnam Math. Comp. 2012, 2013, 2014, and 2015 – Top 12% each year (top 7% in 2013)

Languages

Fluent in English and Hungarian, intermediate in French