This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2023.3337281

Recovering static and time-varying communities using persistent edges

Konstantin Avrachenkov, Maximilien Dreveton, and Lasse Leskelä

Abstract—This article focuses on spectral methods for recovering communities in temporal networks. In the case of fixed communities, spectral clustering on the simple time-aggregated graph (i.e., the weighted graph formed by the sum of the interactions over all temporal snapshots) does not always produce satisfying results. To utilise information carried by temporal correlations, we propose to employ different weights on freshly appearing and persistent edges. We show that spectral clustering on such weighted graphs can be explained as a relaxation of the maximum likelihood estimator of an extension of the degree-corrected stochastic block model with Markov interactions. We also study the setting of evolving communities, for which we use the prediction at time $t-1$ as an oracle for inferring the community labels at time t. We demonstrate the accuracy of the proposed methods on synthetic and real data sets.

Index Terms—graph clustering, temporal networks, spectral methods, stochastic block model

1 Introduction

Complex networks are commonly used to describe and analyze interactions between entities. A natural problem arising in complex network analysis consists of identifying meaningful structures within the network. Community recovery, i.e., partitioning the set of nodes of a network into *communities* based on some common properties of the vertices, is now a well-established area [1]. In many situations, interactions between node pairs vary over time, and classical graph-based models are replaced by temporal network models [2]. Temporal networks naturally arise in various situations, such as communication or face-to-face interactions between individuals. In brain networks, interactions between neurons are also time-varying. Finally, ecological networks such as food webs experience seasonal changes. The longitudinal dimension of data raises new challenges to traditional clustering algorithms. Moreover, the community structure of temporal networks can either be static or time-varying, and different clustering algorithms need to be used in each case. For example, community structures in a brain network are likely to be static, while the communities in a social network can vary.

If communities are static, then each additional snapshot makes clustering easier. Furthermore, when snapshots are temporally uncorrelated, simple time-aggregation of data does not lose any information and recovers communities optimally [3]. Nonetheless, for time-correlated models, this simple temporal aggregation of the data might lose important features such as temporal patterns. As such, [4] describes information-theoretic recovery criteria for Markov edge evolution dynamics, while [5], [6] study a spectral algorithm based on the squared adjacency matrix.

Previous research on community detection in temporal networks has also focused on evolving communities, for example by generalizing belief-propagation methods [7], developing variational EM algorithms [8] and introducing new spectral methods [9], [10], [11] or modularity-based methods [12], [13]. We refer to [14], [15] for an overview. All of the aforementioned works focus on evolving communities for which the interactions between nodes are resampled at every time step. One can then treat each layer independently by applying static community detection and smoothing the community predictions. A more challenging situation is when the edges at a given time step also depend on the edges at the previous time step. It is argued in [16] that link persistence makes the inference of communities harder.

In this work, we first study the case of static community memberships. We introduce a temporal extension of the degree-corrected stochastic block model [17], [18], in which the community structure is fixed and the interactions between node pairs follow a Markov evolution which only depends on the community structure and the degree correction parameters. We show that maximum likelihood estimation reduces to the maximisation of a regularized modularity of a time-aggregated graph, in the limit of a large number of snapshots and sparse interactions. This weighted graph is not given by a simple sum of the adjacency matrices over all snapshots. Instead, we show that the maximum likelihood estimator distinguishes, at each snapshot, the newly formed edges from the persistent ones. Indeed, the information contained in these two types of temporal edges is different. Thus, we propose a time-aggregation corresponding to the sum of all snapshots of the signal coming from the newly formed and persistent edges, but with different weights. This leads to a weighted graph that better takes into account the temporal nature of the signal than the vanilla time-aggregation. A continuous relaxation of this maximum

- *K. Avrachenkov is with Inria Sophia Antipolis, 2004 Route des Lucioles, 06902 Valbonne, France. E-mail: k.avrachenkov@inria.fr.*
- *M. Dreveton is with EPFL, School of Computer and Communication Sciences, CH-1015 Lausanne, Switzerland. E-mail: maximilien.dreveton@epfl.ch.*
- *L. Leskelä is with Aalto University, Department of Mathematics and Systems Analysis, Otakaari 1, 02150 Espoo, Finland. E-mail: lasse.leskela@aalto.fi.*

Manuscript received XXX; revised XXX.

likelihood estimator then leads to a normalized spectral clustering algorithm. We finally extend this study to higher-order temporal correlations.

We also study evolving communities with temporal correlation in the edge formation. In this case, we propose to recover the community at a given time step by using the prediction of the previous time step as a noisy oracle. This naturally leads to a noisy semi-supervised inference problem [15], which we tackle using spectral methods.

The article is structured as follows. Section 2 studies the static communities, while Section 3 focuses on evolving communities. Numerical experiments are presented in Section 4 and the conclusions and future research directions are given in Section 5.

This article is an extension of [19]. Compared to the conference article, the section on evolving communities is entirely new. Moreover, the spectral algorithm for fixed communities is modified (in order not to require any hyper-parameters besides the number of communities). The extension to higher-order Markov chains is also new. Finally, we present experiments on a wider collection of data sets.

2 Static communities

This section studies the recovery of static communities in a temporal network. We introduce in Section 2.1 a temporal stochastic block model where dynamic interaction patterns between node pairs follow a Markov chain. We render this model more versatile by adding degree correction parameters, describing the tendency of each node to start new interactions. We show that in some cases the likelihood of this model is approximated by a regularized modularity of a time-aggregated graph. This time-aggregated graph involves a trade-off between new edges and persistent edges. A continuous relaxation reduces the regularized modularity maximisation to a normalized spectral clustering.

2.1 Degree-corrected temporal network model

A degree-corrected temporal stochastic block model with N nodes, K blocks and T snapshots is a probability distribution

$$\mathbb{P}(A \,|\, \sigma, F, \theta) \;=\; \prod_{1\le i<j\le N} F^{\theta_i\theta_j}_{\sigma_i\sigma_j}(A^1_{ij},\dots,A^T_{ij}) \qquad (2.1)$$

of a symmetric adjacency tensor $A \in \{0,1\}^{N\times N\times T}$ with zero diagonal entries, defined by $A^t_{ij} = 1$ if nodes i and j interact at time t, and $A^t_{ij} = 0$ otherwise. Moreover, $\sigma = (\sigma_1,\dots,\sigma_N)$ is a community assignment with $\sigma_i \in \{1,\dots,K\}$ indicating the community of node i, $F = \left(F^{\eta\theta}_{k\ell}\right)$ is a collection of probability distributions on $\{0,1\}^T$ indexed by community labels $k,\ell \in [K] = \{1,\dots,K\}$ and numbers $\eta,\theta \ge 0$ (representing the degree correction parameters), and $\theta = (\theta_1,\dots,\theta_N)$ is a vector of node-specific degree correction parameters $\theta_i \ge 0$.

In the following, we will restrict ourselves to homogeneous models with Markov edge dynamics, in which the nodes' static community memberships are sampled independently and uniformly at random from $[K]$, and for $x \in \{0,1\}^T$ we have

$$F^{\theta_i\theta_j}_{\sigma_i\sigma_j}(x) \;=\; \begin{cases} \mu^{\theta_i\theta_j}_{x_1}\prod_{t=2}^T P^{\theta_i\theta_j}_{x_{t-1},x_t} & \text{if } \sigma_i=\sigma_j,\\ \nu^{\theta_i\theta_j}_{x_1}\prod_{t=2}^T Q^{\theta_i\theta_j}_{x_{t-1},x_t} & \text{otherwise,}\end{cases} \qquad (2.2)$$

with initial distributions

$$\begin{aligned}\mu^{\theta_i\theta_j} &= \begin{pmatrix}1-\theta_i\theta_j\mu_1\\ \theta_i\theta_j\mu_1\end{pmatrix},\\ \nu^{\theta_i\theta_j} &= \begin{pmatrix}1-\theta_i\theta_j\nu_1\\ \theta_i\theta_j\nu_1\end{pmatrix},\end{aligned} \qquad (2.3)$$

and transition probability matrices

$$\begin{aligned}P^{\theta_i\theta_j} &= \begin{pmatrix}1-\theta_i\theta_jP_{01} & \theta_i\theta_jP_{01}\\ 1-P_{11} & P_{11}\end{pmatrix},\\ Q^{\theta_i\theta_j} &= \begin{pmatrix}1-\theta_i\theta_jQ_{01} & \theta_i\theta_jQ_{01}\\ 1-Q_{11} & Q_{11}\end{pmatrix}.\end{aligned} \qquad (2.4)$$

The parameters θ_i account for the fact that some nodes might be more inclined than others to start new connections, similarly to the degree-corrected block model of [18]. More precisely, $\theta_i\theta_j\mu_1$ (resp., $\theta_i\theta_j\nu_1$) is the probability of an interaction between nodes i and j belonging to the same (resp., different) cluster at the first snapshots. The first snapshot is therefore an instance of the degree-corrected block model of [18]. Moreover, the interaction between i and j is a Markov chain whose transition matrix is $P^{\theta_i\theta_j}$ (resp., $Q^{\theta_i\theta_j}$) if the two nodes belong to the same (resp., different) cluster. To obtain tractable analytical expressions for the model likelihood, we do not add degree correction parameters in front of P_{11} and Q_{11}. This corresponds to assuming that popularity does not affect the duration of interactions. Moreover, we assume that $\max_{i,j}\{\theta_i\theta_j\delta\} \le 1$, where $\delta = \max\{\mu_1,\nu_1,P_{01},Q_{01}\}$. Finally, we normalise the degree correction parameters so that $\frac{1}{|C_k|}\sum_{i\in C_k}\theta_i = 1$ for all i, where $C_k = \{i : \sigma(i) = k\}$ indicates community k.

2.2 Maximum likelihood estimator

The following proposition presents a concise formulation of the model likelihood, with proof presented in Appendix A.1.

Proposition 2.1. *A maximum likelihood estimator for the Markov block model defined by (2.1)–(2.2) is any community assignment $\sigma \in [K]^N$ that maximises*

$$\sum_{\substack{i,j\\ \sigma_i=\sigma_j}}\left\{A^1_{ij}\left(\rho_1^{\theta_i\theta_j}-\rho_0^{\theta_i\theta_j}\right)+\rho_0^{\theta_i\theta_j}+\left(A^1_{ij}-A^T_{ij}\right)\ell_{10}^{\theta_i\theta_j}\right\}$$
$$+\sum_{\substack{i,j\\ \sigma_i=\sigma_j}}\sum_{t=2}^{T}\left\{\left(\ell_{01}^{\theta_i\theta_j}+\ell_{10}^{\theta_i\theta_j}\right)\left(A^t_{ij}-A^{t-1}_{ij}A^t_{ij}\right)\right.$$
$$\left.+\ell_{11}^{\theta_i\theta_j}A^{t-1}_{ij}A^t_{ij}-\log\frac{Q_{00}^{\theta_i\theta_j}}{P_{00}^{\theta_i\theta_j}}\right\}$$

where $\rho_a^{\theta_i\theta_j} = \log\frac{\mu_a^{\theta_i\theta_j}}{\nu_a^{\theta_i\theta_j}}$ and $\ell_{ab}^{\theta_i\theta_j} = \log\frac{P_{ab}^{\theta_i\theta_j}}{Q_{ab}^{\theta_i\theta_j}} - \log\frac{P_{00}^{\theta_i\theta_j}}{Q_{00}^{\theta_i\theta_j}}$.

The MLE derived in Proposition 2.1 is more complex than summing all snapshots independently. In particular, the terms $A^{t-1}_{ij}A^t_{ij}$ account for *persistent edges* over two consecutive snapshots. Denote by $A^t_{\text{pers}} = A^{t-1}\odot A^t$ the entrywise product of adjacency matrices A^{t-1} and A^t. Then A^t_{pers} is the adjacency matrix of the graph containing the persistent edges between $t-1$ and t, and $A^t_{\text{new}} = A^t - A^t_{\text{pers}}$ corresponds to the graph containing the *edges freshly appearing* at time t.

Assuming that the number of snapshots T is large, we can ignore the boundary terms, and the MLE expressed in Proposition 2.1 reduces to maximising

$$\sum_{t=2}^{T} \sum_{\substack{i,j: \\ \sigma_i=\sigma_j}} \left\{ \left(\ell_{01}^{\theta_i\theta_j} + \ell_{10}^{\theta_i\theta_j}\right)\left(A_{ij}^t - A_{ij}^{t-1}A_{ij}^t\right) + \ell_{11}^{\theta_i\theta_j} A_{ij}^{t-1}A_{ij}^t - \log \frac{Q_{00}^{\theta_i\theta_j}}{P_{00}^{\theta_i\theta_j}} \right\}.$$

By utilising (2.3)–(2.4), we can further simplify it to express this as a modularity. Recall that given a weighted graph W, a partition σ, and a resolution parameter γ, the regularized modularity is defined as [20], [21]

$$\mathcal{M}(W,\sigma,\gamma) \;=\; \sum_{i,j} \delta(\sigma_i,\sigma_j)\left(W_{ij} - \gamma\frac{d_i d_j}{2m}\right),$$

where $d_i = \sum_j W_{ij}$ and $m = \sum_i d_i$. Hence, suppose that $P^{\theta_i\theta_j}$ and $Q^{\theta_i\theta_j}$ are nondegenerate, and $\mu^{\theta_i\theta_j}$ (resp. $\nu^{\theta_i\theta_j}$) is the stationary distribution of $P^{\theta_i\theta_j}$ (resp. $Q^{\theta_i\theta_j}$). In a sparse setting, P_{01} and Q_{01} are small, and after a Taylor expansion (see Section A.2 for detailed derivations) the previous expression is approximately equal to $\mathcal{M}(W,\sigma,\gamma)$, where W is defined by

$$W \;=\; \sum_{t=2}^{T}\left(\alpha A_{\text{new}}^t + \beta A_{\text{pers}}^t\right), \tag{2.5}$$

with

$$\begin{aligned} \alpha &= \log\frac{P_{01}}{Q_{01}} + \log\frac{1-P_{11}}{1-Q_{11}}, \\ \beta &= \log\frac{P_{11}}{Q_{11}}, \end{aligned} \tag{2.6}$$

$$\gamma \;=\; (P_{01}-Q_{01})\,\frac{\alpha(\mu_1+(K-1)\nu_1)+(\beta-\alpha)(\mu_1 P_{11}+(K-1)\nu_1 Q_{11})}{K}.$$

2.3 Comparison with previous work on static communities

The correspondence between maximum likelihood estimation and modularity maximisation is known in static block models [22]. Analogously to the single-layer case, the modularity of a temporal network, with possibly time-dependent community structure, was previously defined by Mucha et al. [12] and Pamfil et al. [13] as

$$\sum_{t=1}^{T}\mathcal{M}(A^t,\sigma^t,\gamma_t) + \sum_{t=1}^{T}\sum_{s\neq t}\sum_{i}\omega_i^{st}\delta\left(\sigma_i^s,\sigma_i^t\right) \tag{2.7}$$

where γ_t is the resolution parameter for layer t, σ_i^t is the community membership of node i at time step t, and w_i^{st} denotes a coupling between time instants s and t. For a static community structure, the second term in (2.7) is irrelevant. When the resolution is constant over time, the relevant term in (2.7) can be written as

$$\sum_{t=1}^{T}\mathcal{M}(A^t,\sigma,\gamma) \;=\; \mathcal{M}(A^{\text{agg}},\sigma,\gamma),$$

where $A^{\text{agg}} = \sum_{t=1}^{T} A^t$ is the weighted adjacency matrix of the time-aggregated data. In contrast, the matrix W in (2.5) involves a trade-off between new edges and persistent edges. We notice that $W \propto A^{\text{agg}}$ only if $\alpha = \beta$. This is the case when freshly appearing and persistent edges are equally important, and only in this case using the simple time-aggregated adjacency matrix provides optimal inference.

2.4 Temporal spectral clustering combining new and persistent edges

Following our analysis in Section 2.2, the community prediction is given by

$$\hat{\sigma} \;=\; \arg\max_{\sigma\in[K]^N}\mathcal{M}(W,\sigma,\gamma),$$

where W is defined in Equation (2.5) and γ is a proper resolution parameter. This optimisation problem is NP-complete in general [23] but can be approximately solved by continuous relaxation. We can choose the relaxation so that the optimisation problem reduces to normalized spectral clustering on the weighted graph W (we refer to [24] and to Section A.2 for detailed computations). We note that in order to compute the normalized Laplacian of W, we should assume $\alpha,\beta\geq 0$, which we observe in all our experiments.

Since α and β are likely to be unknown in practical situations, we propose to estimate them as follows. For $a,b\in\{0,1\}$, the empirical probability of observing a transition $a\to b$ in the interaction pattern between nodes i and j is

$$\frac{n_{ab}(i,j)}{n_a(i,j)},$$

where $n_{ab}(i,j) = \sum_{t=2}^{T} 1(A_{ij}^{t-1}=a,\, A_{ij}^t = b)$ is the observed number of transitions $a\to b$ between i and j, and $n_a(i,j) = \sum_b n_{ab}(i,j)$. Thus, if $\hat{\sigma}$ is an estimator of the community labelling, then

$$\begin{aligned} \widehat{P}_{ab} &= \frac{1}{|\{i,j\colon \hat{\sigma}_i=\hat{\sigma}_j\}|}\sum_{i,j\colon \hat{\sigma}_i=\hat{\sigma}_j}\frac{n_{ab}(i,j)}{n_a(i,j)}, \\ \widehat{Q}_{ab} &= \frac{1}{|\{i,j\colon \hat{\sigma}_i\neq\hat{\sigma}_j\}|}\sum_{i,j\colon \hat{\sigma}_i\neq\hat{\sigma}_j}\frac{n_{ab}(i,j)}{n_a(i,j)}, \end{aligned} \tag{2.8}$$

are estimators of P and Q from which one can compute estimates of α and β. These observations motivate the following algorithm. We start with $\alpha=\beta=1$, and we let $\hat{\sigma}$ be the output of spectral clustering on $W=\sum_{t=2}^{T}\alpha A_{\text{new}}^t+\beta A_{\text{pers}}^t$. Using $\hat{\sigma}$ we compute estimates $\widehat{P}$ and $\widehat{Q}$, which lead to a new value of α and β. We iterate the procedure until some convergence criterion is met. This leads to Algorithm 1.

2.5 Extension to higher-order Markov chains

Interactions with a longer memory can be modelledusing higher-order Markov chains [25], [26]. We demonstrate below how Algorithm 1 can be extended to a long-memory context. Suppose the interactions are Markov chains of order M. Namely, the intra- and inter-block interaction connectivity functions, defined previously by (2.2), are now given by

$$F_{\sigma_i\sigma_j}(x) \;=\; \begin{cases} \mu(x_{1:M})\displaystyle\prod_{t=M+1}^{T} P\left(x_t \mid s_t\right) & \text{if } \sigma_i=\sigma_j, \\ \nu(x_{1:M})\displaystyle\prod_{t=M+1}^{T} Q\left(x_t \mid s_t\right) & \text{otherwise,} \end{cases}$$

Algorithm 1: Spectral clustering for temporal networks with Markov edge dynamics and static node labelling.

Input: Adjacency matrices $A^1, \dots, A^T$, number of clusters K, stopping parameter ϵ
Output: Predicted membership matrix $\hat{\sigma} \in [K]^N$
Process:
Let $\alpha = \beta = 1$ and *stopping* = False;
while *stopping* = *False* **do**
- Let $W = \sum_{t=2}^{T} \alpha A^t_{\text{new}} + \beta A^t_{\text{pers}}$ where $A^t_{\text{new}} = A^t - A^{t-1} \odot A^t$ and $A^t_{\text{pers}} = A^{t-1} \odot A^t$;
- $\hat{\sigma} \leftarrow$ SPECTRAL CLUSTERING (W, K);
- Let $\widehat{P}$ and $\widehat{Q}$ as defined by (2.8);
- Let $\tilde{\alpha} = \log \frac{\widehat{P}_{01}}{\widehat{Q}_{01}} + \log \frac{\widehat{P}_{10}}{\widehat{Q}_{10}}$ and $\tilde{\beta} = \log \frac{\widehat{P}_{11}}{\widehat{Q}_{11}}$.

if $\|(\alpha, \beta) - (\tilde{\alpha}, \tilde{\beta})\|_2 < \epsilon$ **then**
  stopping = True;
else
  $\alpha = \tilde{\alpha}$ and $\beta = \tilde{\beta}$.
Return: $\hat{\sigma}$

Algorithm 2: Spectral clustering for temporal networks with higher-order Markov edge dynamics.

Input: Adjacency matrices $A^1, \dots, A^T$, number of clusters K, Markov order M, stopping parameter ϵ
Output: Predicted membership matrix $\hat{\sigma} \in [K]^N$
Process:
Let $\alpha = (\alpha_{sb})_{s \in \{0,1\}^M, b \in \{0,1\}}$ and initialise $\alpha_{sb} = 1$ for all s, b. Let *stopping* = False;
while *stopping* = *False* **do**
- $W_{ij} := \sum_{t=2}^{T} \sum_{s,b} \alpha_{sb} 1\left(A_{ij}^{(t-M):(t-1)} = s, A^t_{ij} = b\right)$;
- $\hat{\sigma} \leftarrow$ SPECTRAL CLUSTERING (W, K);
- Let $\widehat{P}$ and $\widehat{Q}$ as defined by (2.9);
- Let $\tilde{\alpha}_{sb} := \log \frac{\widehat{P}}{\widehat{Q}}(b \,|\, s)$ for all $s \in \{0,1\}^M$ and $b \in \{0,1\}$.

if $\|\alpha - \tilde{\alpha}\|_2 < \epsilon$ **then**
  stopping = True;
else
  $\alpha = \tilde{\alpha}$.
Return: $\hat{\sigma}$

where $s_t = (x_{t-M}, \cdots, x_{t-1})$. Following a similar reasoning, we propose to apply spectral clustering on the weighted matrix W defined by

$$W_{ij} = \sum_{t=M+1}^{T} \sum_{\substack{s \in \{0,1\}^M \\ b \in \{0,1\}}} \alpha_{sb} \, 1\left(A_{ij}^{(t-M):(t-1)} = s, A^t_{ij} = b\right),$$

with $\alpha_{sb} = \log \frac{P(b \,|\, s)}{Q(b \,|\, s)}$ and $A_{ij}^{(t-M):(t-1)} = (A_{ij}^{t-M}, \cdots, A_{ij}^{t-1})$. Moreover, an estimator of P is given by

$$\begin{aligned} \widehat{P}(b \,|\, s) &= \frac{1}{|\{i,j \colon \hat{\sigma}_i = \hat{\sigma}_j\}|} \sum_{\substack{i,j \\ \hat{\sigma}_i = \hat{\sigma}_j}} \frac{n_{sb}(i,j)}{n_s(i,j)}, \\ \widehat{Q}(b \,|\, s) &= \frac{1}{|\{i,j \colon \hat{\sigma}_i \neq \hat{\sigma}_j\}|} \sum_{\substack{i,j \\ \hat{\sigma}_i \neq \hat{\sigma}_j}} \frac{n_{sb}(i,j)}{n_s(i,j)}, \end{aligned} \tag{2.9}$$

where $n_{sb}(i,j) = \sum_{t=M}^{T} 1\left(A_{ij}^{(t-M):(t-1)} = s, A^t_{ij} = b\right)$ and $n_s(i,j) = \sum_{b \in \{0,1\}} n_{sb}(i,j)$. This can be formalised as Algorithm 2.

3 Evolving communities

In this section, we consider a population of N nodes partitioned into K time-evolving communities. At time t, we denote by $\sigma^t_i \in [K]$ the community membership of node i and by $A^t_{ij} \in \{0,1\}$ the observed interaction between nodes i and j. We investigate methods of recovering the community structure, denoted by $(\sigma^1, \cdots, \sigma^T)$ where $\sigma^t \in [K]^N$, from an observed adjacency tensor $A = \left(A^t_{ij}\right)$.

3.1 Model description

Similarly to several articles on dynamic SBM [7], [8], [16], we first assume that each node community labels $\sigma_i^{1:T} \in [K]^T$ is a Markov chain of length T with initial probability α and transition probability matrix π. Hence,

$$\mathbb{P}\left(\sigma^{1:T}\right) = \prod_{i=1}^{N} \alpha\left(\sigma_i^1\right) \prod_{t=2}^{T} \pi\left(\sigma_i^{t-1}, \sigma_i^t\right). \tag{3.1}$$

For simplicity, we will assume that the initial labels and the transitions are uniform, that is

$$\alpha = \frac{1}{K} 1_K \quad \text{and} \quad \pi = \eta I_K + \frac{1-\eta}{K} 1_K 1_K^T.$$

In other words, a node keeps its label with probability $\eta \in [0,1]$ and chooses a label uniformly at random with probability $1 - \eta$.

We then assume that the interaction between two nodes i and j is a Markov chain depending only on the community labelling[1]. Hence,

$$\mathbb{P}(A \,|\, \sigma) = \prod_{1 \leq i < j \leq N} \mathbb{P}\left(A^1_{ij} \,|\, \sigma^1_i, \sigma^1_j\right) \prod_{t=2}^{T} \mathbb{P}\left(A^t_{ij} \,|\, A^{t-1}_{ij}, \sigma^t_i, \sigma^t_j\right).$$

We consider a homogeneous model in which the initial distribution is given by

$$\mathbb{P}\left(A^1_{ij} \,|\, \sigma^1_i, \sigma^1_j\right) = \begin{cases} \mu(A^1_{ij}), & \text{if } \sigma^1_i = \sigma^1_j, \\ \nu(A^1_{ij}), & \text{otherwise,} \end{cases}$$

and the transition probabilities are

$$\mathbb{P}\left(A^t_{ij} = b \,|\, A^{t-1}_{ij} = a, \sigma^t_i, \sigma^t_j\right) = \begin{cases} P_{ab} & \text{if } \sigma^t_i = \sigma^t_j, \\ Q_{ab} & \text{otherwise.} \end{cases}$$

1. For simplicity, in this section we do not add any degree correction parameters.

3.2 Community inference

The inference of the temporal community structure is complicated because interactions observed in a distant past are less informative for inferring the current community membership of a node. To mitigate the bias caused by outdated information, we propose a two-stage online algorithm:

1) at time $t = 1$, we use a static community detection algorithm to output $\hat{\sigma}^1$, a prediction of the initial node labels σ^1 from the observation of the first snapshot A^1;
2) at time $t > 1$, we will use the observation of the first t snapshots $A^1, \dots, A^t$ as well as the previous predictions $\hat{\sigma}^1, \cdots \hat{\sigma}^{t-1}$. This will be treated as a semi-supervised learning problem, where the prediction $\hat{\sigma}^{t-1}$ is seen as a noisy oracle for the true node labelling σ^t.

From the Markov structure, the prediction at time $t > 1$ reduces to predicting σ^t using only the network at time $t-1$ and t and the previous prediction $\hat{\sigma}^{t-1}$. This can be interpreted as a noisy semi-supervised problem, where the previous prediction $\hat{\sigma}^{t-1}$ plays the role of an oracle for the node labels at time t. This oracle is biased, as it bears two kinds of potential mistakes. Firstly, $\hat{\sigma}^{t-1}$ is not necessarily exactly equal to the perfect community labelling σ^{t-1}. Furthermore, since the node labels vary through time, σ^{t-1} does not precisely correspond to σ^t. Assume that the network data A and community labels σ come from the model described in Section 3.1. We denote the oracle bias at time t by

$$\rho_t \;=\; \mathbb{P}\left(\sigma_i^t \neq \hat{\sigma}_i^{t-1}\right), \tag{3.2}$$

and we will make a simplifying assumption that it does not depend on node i.

The MAP estimator for the online learning problem is

$$\begin{aligned}\hat{\sigma}^t \;&=\; \underset{z\in[K]^N}{\arg\max}\, \mathbb{P}\left(z \,|\, A^t, A^{t-1}, \hat{\sigma}^{t-1}\right)\\ &=\; \underset{z\in[K]^N}{\arg\max}\, \mathbb{P}\left(A^t \,|\, A^{t-1}, z, \hat{\sigma}^{t-1}\right)\mathbb{P}\left(z \,|\, A^{t-1}, \hat{\sigma}^{t-1}\right)\end{aligned}$$

by Bayes' rule.

Since $\mathbb{P}\left(A^t \,|\, A^{t-1}, z, \hat{\sigma}^{t-1}\right) = \mathbb{P}\left(A^t \,|\, A^{t-1}, z\right)$, then by proceeding similarly to the proof of Proposition 2.1, the log-likelihood term $\log \mathbb{P}\left(A^t \,|\, A^{t-1}, z\right)$ can be rewritten as

$$\frac{1}{2}\sum_{\substack{i,j\\ z_i=z_j}}\left\{\ell_{01}^{\theta_i\theta_j}\left(A_{ij}^t - A_{ij}^{t-1}A_{ij}^t\right) + \ell_{10}^{\theta_i\theta_j}\left(A_{ij}^{t-1} - A_{ij}^{t-1}A_{ij}^t\right) + \ell_{11}^{\theta_i\theta_j}A_{ij}^{t-1}A_{ij}^t - \log\frac{Q_{00}^{\theta_i\theta_j}}{P_{00}^{\theta_i\theta_j}}\right\}.$$

The oracle information is equal to

$$\begin{aligned}\mathbb{P}\left(z \,|\, \hat{\sigma}^{t-1}\right) \;&=\; \prod_{i=1}^N \frac{\mathbb{P}\left(\hat{\sigma}_i^{t-1} \,|\, z_i\right)}{\mathbb{P}\left(\hat{\sigma}_i^{t-1}\right)}\mathbb{P}\left(z_i\right)\\ &=\; \left(\frac{\rho_t}{1-\rho_t}\right)^{\left|\{i\in[N]\colon z_i\neq\hat{\sigma}_i^{t-1}\}\right|}(1-\rho_t)^N K^{-N},\end{aligned}$$

where we used the uniformity of the node labels and where ρ_t is defined in (3.2).

The MAP estimator at time t is therefore any labelling σ that maximises

$$\sum_{\substack{i,j\\ z_i=z_j}}\left\{\ell_{01}^{\theta_i\theta_j}\left(A_{ij}^t - A_{ij}^{t-1}A_{ij}^t\right) + \ell_{10}^{\theta_i\theta_j}\left(A_{ij}^{t-1} - A_{ij}^{t-1}A_{ij}^t\right) + \ell_{11}^{\theta_i\theta_j}A_{ij}^{t-1}A_{ij}^t - \log\frac{Q_{00}^{\theta_i\theta_j}}{P_{00}^{\theta_i\theta_j}}\right\} + 2\lambda_t\sum_{i=1}^N 1\left(z_i = s_i\right),$$

where $\ell_{ab}^{\theta_i\theta_j} = \log\frac{P_{ab}^{\theta_i\theta_j}}{P_{ab}^{\theta_i\theta_j}} - \log\frac{P_{00}^{\theta_i\theta_j}}{P_{00}^{\theta_i\theta_j}}$ and $\lambda_t = \log\frac{1-\rho_t}{\rho_t}$, and where $s_i = \sigma_i^{t-1}$.

3.3 Continuous relaxation

Denote by $A_{\text{pers}}^t = A^{t-1} \odot A^t$ the *persistent edges*, by $A_{\text{new}} = A^t - A_{\text{pers}}^t$ the *freshly formed* edges, and by $A_{\text{old}} = A^{t-1} - A_{\text{pers}}$ the *disappearing edges* between time $t-1$ and t. Denote by $\Sigma \in \{0,1\}^{N\times K}$ (resp., by S) the one-hot encoding of $\sigma \in [K]^N$ (resp., of s). In other words, we have $\Sigma_{ik} = 1(\sigma_i = k)$ and $S_{ij} = 1(s_i = k)$. We observe that $\sum_i 1(\sigma_i = s_i) = \Sigma^T S$. Thus, using a Taylor expansion as in Section 2.1, we can approximate the MAP estimator at time t as the maximisation over $\Sigma \in \{0,1\}^{N\times K}$ of

$$\operatorname{Tr}\left(\Sigma^T\left(W^t - \tau 1_N 1_N^T\right)\Sigma + 2\lambda\Sigma^T S\right), \tag{3.3}$$

where $W^t = \alpha_{01}A_{\text{new}}^t + \alpha_{10}A_{\text{old}}^t + \alpha_{11}A_{\text{pers}}^t$ with $\alpha_{ab} = \log\frac{P_{ab}}{Q_{ab}}$ and $\tau = P_{01} - Q_{01}$ is a resolution parameter, and Σ verifies the constraints $\sum_{k\in[K]}\Sigma_{ik} = 1$ for all $i \in [N]$.

We perform a continuous relaxation (mirroring what is commonly done for spectral methods [24]) and solve instead

$$\underset{\substack{X\in\mathbb{R}^{N\times K}\\ \sum_{i=1}^N \kappa_i X_{ik}^2 = \sum_{i=1}^N \kappa_i}}{\arg\max}\quad \operatorname{Tr}\left(X^T W_\tau X + 2\lambda X^T S\right), \tag{3.4}$$

where $\kappa = (\kappa_1, \dots, \kappa_N)$ is a vector of positive entries. In what follows, we choose to constrain X to verify $\left(X^T D X\right)_{kk} = 1$ by taking $\kappa_i = \frac{d_i}{2m}$[2]. This leads to the optimisation problem

$$\underset{\substack{X\in\mathbb{R}^{N\times K}\\ \forall k\in[K]\colon (X^T D X)_{kk}=1}}{\arg\max}\quad \operatorname{Tr}\left(X^T W_\tau X + 2\lambda X^T S\right). \tag{3.5}$$

In order to solve (3.5), let us introduce the eigenvalue decomposition of $-D^{-1/2}W_\tau D^{-1/2}$, written as

$$-D^{-1/2}W_\tau D^{-1/2} \;=\; Q\Delta Q^T, \tag{3.6}$$

where $\Delta = \operatorname{diag}(\delta_1, \dots, \delta_N)$ with $\delta_1 \leq \cdots \leq \delta_N$ and $Q^T Q = I_N$.

The following Proposition, whose proof is deferred to Appendix B, gives an expression for the solution of (3.5).

2. Another possible relaxation is to constraint X such that $X^T D X = I_K$, so that $D^{-1/2}X$ belongs to a Stiefel manifold. The minimisation problem obtained with this (stronger) constraint is similar to an *unbalanced orthogonal Procrustes problem*, for which necessary or sufficient conditions for the local and/or global minimiser exist, but no closed-form expression of the solution is available [27], [28].

Proposition 3.1. *Let* $B = \lambda Q^T D^{-1/2} S$ *where* Q *is defined in* (3.6)*. The solution* $\hat{X} = (\hat{X}_{\cdot 1}, \cdots, \hat{X}_{\cdot K})$ *of the optimisation problem* (3.5) *satisfies for all* $k \in [K]$

$$(-W_\tau - \gamma_k^* D)\, \hat{X}_{\cdot k} \;=\; \lambda S_{\cdot k}, \tag{3.7}$$

where γ_k^* *is the smallest solution of*

$$\sum_{i=1}^{N} \left(\frac{B_{ik}}{\delta_i - \gamma} \right)^2 - 1 \;=\; 0. \tag{3.8}$$

The expression of the solution of (3.5) given in Proposition 3.1 leads to Algorithm 3.

Algorithm 3: Online clustering of time-varying communities.

Input: Observed graph sequence $A^{1:T} = (A^1, \ldots, A^T)$; number of communities K; static graph clustering algorithm `algo`; parameters $\alpha_{01}, \alpha_{10}, \alpha_{11}$ and $\lambda_1, \ldots, \lambda_T$

Output: Node labelling $\hat{\sigma}^{1:T}$

Initialise: Compute $\hat{\sigma}^1 \leftarrow \texttt{algo}\,(A^1)$.

for $t = 2, \ldots, T$ **do**

- Compute $W = \alpha_{01} A^t_{\text{new}} + \alpha_{10} A^t_{\text{old}} + \alpha_{11} A^t_{\text{pers}}$.
- For all $k \in [K]$, let γ_k^* be the smallest solution of Equation (3.8).
- Compute $\hat{X}_{\cdot k}$ as the solution of Equation (3.7).
- Let $\hat{\sigma}_i^t = \text{K-means}(\hat{X}_{i\cdot})$ for all $i \in [N]$.

3.4 Comparison with previous work on evolving communities

It is often natural to assume that the pairwise interactions evolve over time, but the temporal changes in connectivity patterns occur gradually. Therefore, many works in the literature propose dynamic extensions of the stochastic block model, where both community memberships and connections can vary through time.

To model a smooth evolution of communities across time, the majority of works describe changes in the memberships via a Markov evolution. Moreover, previous works almost always suppose that the interaction at time t depends only on the block structure at time t. For example, Yang et al. [29] assume that the community membership of each node forms a Markov chain independent of the memberships of other nodes. Xu and Hero [30] extend this model and allow for both the community memberships and connectivity parameters to vary. This renders the model unidentifiable. This issue was raised by Matias and Miele [8], who established identifiability conditions for models with both community memberships and connectivity parameters varying over time. Pensky and Zhang [31] assume that at most s nodes can switch their memberships between two consecutive time instants, and that the connection probabilities vary smoothly over time.

All of the aforementioned works suppose that the interactions at a given snapshot t are conditionally independent of all past snapshots given the current community memberships. Namely, it is assumed that the temporal interactions are resampled at each time instant (i.e., the interaction A^t_{ij} between nodes i and j at time j depends only on the community labels σ_i^t and σ_j^t at time t, and not on the previous interactions A^{t-1}_{ij} between i and j). Barucca et al. [16] propose a model similar to ours but assume that the edge persistence is the same for intra- and inter-community interactions. More precisely, using the notations of Section 3.1, [16] supposes that

$$\mathbb{P}\left(A^t_{ij} = b \,|\, A^{t-1}_{ij} = a, \sigma_i^t, \sigma_j^t\right) \;=\; \begin{cases} P_{ab}, & \text{if } \sigma_i^t = \sigma_j^t, \\ Q_{ab}, & \text{otherwise,} \end{cases}$$

where the transition matrices P and Q can be written as

$$\begin{aligned} P &= \xi 1_2^T 1_2 + (1-\xi)\tilde{P}, \\ Q &= \xi 1_2^T 1_2 + (1-\xi)\tilde{Q}, \end{aligned} \tag{3.9}$$

and where $\tilde{P}$ and $\tilde{Q}$ are given transition matrices and $\xi \in [0,1]$ is an edge persistence parameter. Such a decomposition implicitly supposes that the Markov chain is positively autocorrelated[3] ($\text{Cov}(X_0, X_1) \geq 0$) for a stationary Markov chain (X_t)). While positive autocorrelation can be a reasonable assumption for social networks, it might not be suitable for other situations (for example, in biological networks spiking phenomena might occur).

To generalise the model even further, in addition to time-varying interactions and community memberships, the Markov transition matrices describing the persistence of edges may also evolve over time. An instance of such a model was discussed by Xu [32], by assuming a Gaussian evolution for log-transformed transition matrices, and experimenting with an extended Kalman filter for inferring community memberships and model parameters. Further exploration of such more general Markov models appears an interesting direction for future research.

4 Numerical Experiments

4.1 Static communities

We compare the performances of temporal spectral clustering data sets with static communities. In this section, *improved time-aggregation* refers to Algorithm 1, while *simple time-aggregation* refers to spectral clustering on the time-aggregated graph $\sum_t A^t$.

In our experiments, we set $\epsilon = 0.1$ and a maximum number of iterations = 10. We observed empirically that Algorithm 1 frequently required less than 5 iterations to converge. The only cases when Algorithm 1 did not stop within 10 iterations were the cases when the time-aggregated graphs were too sparse (this may happen when the time horizon T is not enough long).

4.1.1 Synthetic data sets

We plot in Figure 1 the averaged adjusted rand-index (ARI) obtained on 25 realisations of stochastic block models with

3. This comes from the following Lemma. A 2-by-2 binary stochastic matrix $P \neq I_2$ admits a representation $P = \xi I_2 + (1-\xi)\begin{pmatrix} \pi_0 & \pi_1 \\ \pi_1 & \pi_0 \end{pmatrix}$ for some $\xi \in [0,1]$ and some probability distribution $\pi = (\pi_0, \pi_1)$ iff both eigenvalues of P are non-negative iff $\text{Cov}(X_0, X_1) \geq 0$ for a stationary Markov chain $(X_0, X_1, \cdots)$ on $\{0,1\}$ with transition matrix P.

Markov edge dynamics for various degree-correction parameters θ. More precisely:

- $\theta_i = 1$ for all $i \in [N]$ in Figure 1a;
- in Figure 1b, the θ_i's are generated θ_i according to $|\mathcal{N}(0,\sigma^2)| + 1 - \sigma\sqrt{2/\pi}$ where $|\mathcal{N}(0,\sigma^2)|$ denotes the absolute value of a normal random variable with mean 0 and variance σ^2, where $\sigma = 0.25$;
- in Figure 1c, the θ_i's are generated from a Pareto distribution with density function $f(x) = \frac{am^a}{x^{a+1}} 1(x \geq m)$ with $a = 2.5$ and $m = 2/3$.

Note that the sampling of the θ_i's enforces $\mathbb{E}\theta_i = 1$ in all settings.

While the *simple time aggregation* works well, it is striking to notice that Algorithm 1 provides better results. For example, when the degree-corrected parameters are generated according to a Gaussian distribution, Figure 1b shows that the *simple time-aggregation* requires around 100 snapshots to get a perfect ARI, while Algorithm 1 only requires 30 snapshots.

4.1.2 Real world data

We investigate data sets of face-to-face contacts between individuals, for which ground-truth communities are known. The data sets are available on SocioPatterns website[4]. Three data sets were collected in a high school [33], [34], one in a primary school [35] and two in a work environment [36], [37]. For each data set, nodes correspond to individuals (students or workers), time-varying network edges correspond to close proximity interactions, measured by wearable sensors able to sense proximity, and communities correspond to school classes or work departments. The data set dimensions are given in Table 1.

Data set name	N	T	K
High school 2011	118	5391	3
High school 2012	180	11273	5
High school 2013	327	7375	9
Workplace 2013	88	7027	4
Workplace 2015	161	17508	5
Primary school	232	3100	10

TABLE 1
Dimensions of six data sets from SocioPattern library.

We plot in Figure 2 the accuracy obtained by Algorithm 1 on each of these data sets, compared to the spectral clustering done on the time aggregated graph $\sum_{t=1}^{T} A^t$. We observe a clear improvement in the accuracy of most data sets. This highlights the importance of weighting differently the persistent edges and the freshly appearing edges. Additionally, we would like to emphasise that Algorithm 1 estimates the values of α and β using the network data. Only two data sets (workplace 2015 and primary school) show no improvements. Nonetheless, we demonstrate in Figure 3 that we can improve the accuracy of these two data sets by taking into account higher-order correlations.

4. http://www.sociopatterns.org/

4.2 Evolving communities

In this section, we evaluate the performance of Algorithm 3 on synthetic data sets.

We compare in Figure 4 the averaged accuracy obtained by Algorithm 3 with Algorithm 1 (spectral clustering with persistent edges) and an algorithm performing spectral clustering on each snapshot individually. In particular, we observe that when $\eta = 1$ (i.e., static community structure), Algorithm 1 is extremely efficient, as expected. Since it takes into account all previous snapshots, it in particular outperforms Algorithm 3. On the contrary, when $\eta \neq 1$, the lagging problem arises, and Algorithm 1 ends up with a very poor accuracy after a few snapshots. On the contrary, Algorithm 3 keeps a very high accuracy over all snapshots.

In Figure 4, we choose λ_t to be constant with respect to time and equal to 0.5, while Figure 5 explores other constant possible values of λ. We observe that when λ is in the range $[0.1, 1]$, the algorithm provides good accuracy. On the other hand, when λ becomes too large, Algorithm 3 gives too much importance to the oracle, and the accuracy becomes worse. In practice, the choice of the parameters λ_t could be optimised from the data, e.g. based on η or on the transition matrices P and Q. Moreover, it would be intuitive to increase λ_t with t, as the confidence in the oracle is higher when more temporal data is available. We leave these research ideas for future work.

Finally, we compare in Figure 6 the performance of Algorithm 3 with the VEM algorithm of [8][5]. The VEM algorithm of [8] is tailored for models with time-varying communities but where the interactions are resampled at every time step. Thus, we observe that its performances are excellent when the links are resampled at every time step (Figure 6(a)), but the performance decreases when the time-persistency increases (Figures 6(b) and 6(c)). In contrast, Algorithm 3 performs better than VEM when the correlation in the edge dynamics is strong.

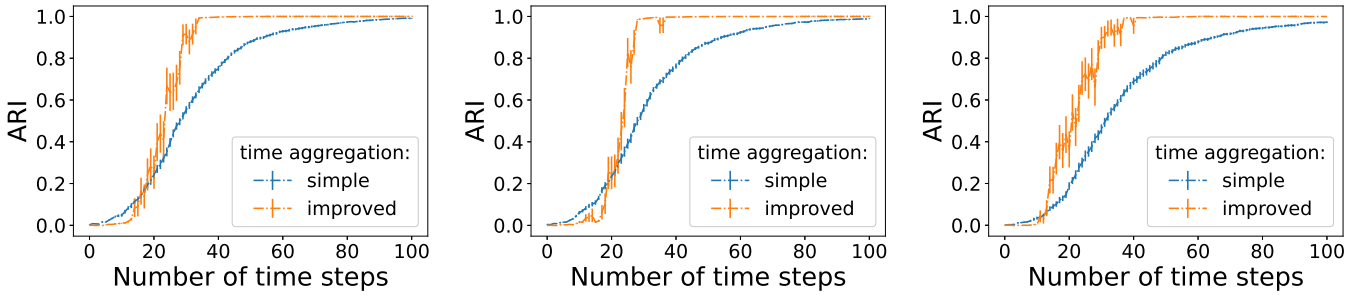
5 Conclusion

In this article, we derived spectral algorithms for clustering temporal networks, both in the case of static and evolving communities. For static communities, the Markov chain assumption for the edge formation implies a difference in the information carried by persistent and newly formed edges. This difference is lost by performing a simple time aggregation that indifferently sums all the interactions. Instead, we proposed to weigh differently the persistent and the newly formed edges. In the evolving community setting, we highlighted the possibility of using a past prediction as a semi-supervised oracle for the prediction at a current snapshot. Overall, the derived spectral algorithms for clustering temporal networks provide a valuable tool for inferring community structure in complex systems, and the proposed weighing methods for persistent and newly formed edges offer a refined approach to analyse dynamic networks.

One direction for future research is online/adaptive tuning of hyperparameters such as estimates of the bias of the noisy semi-supervised oracle.

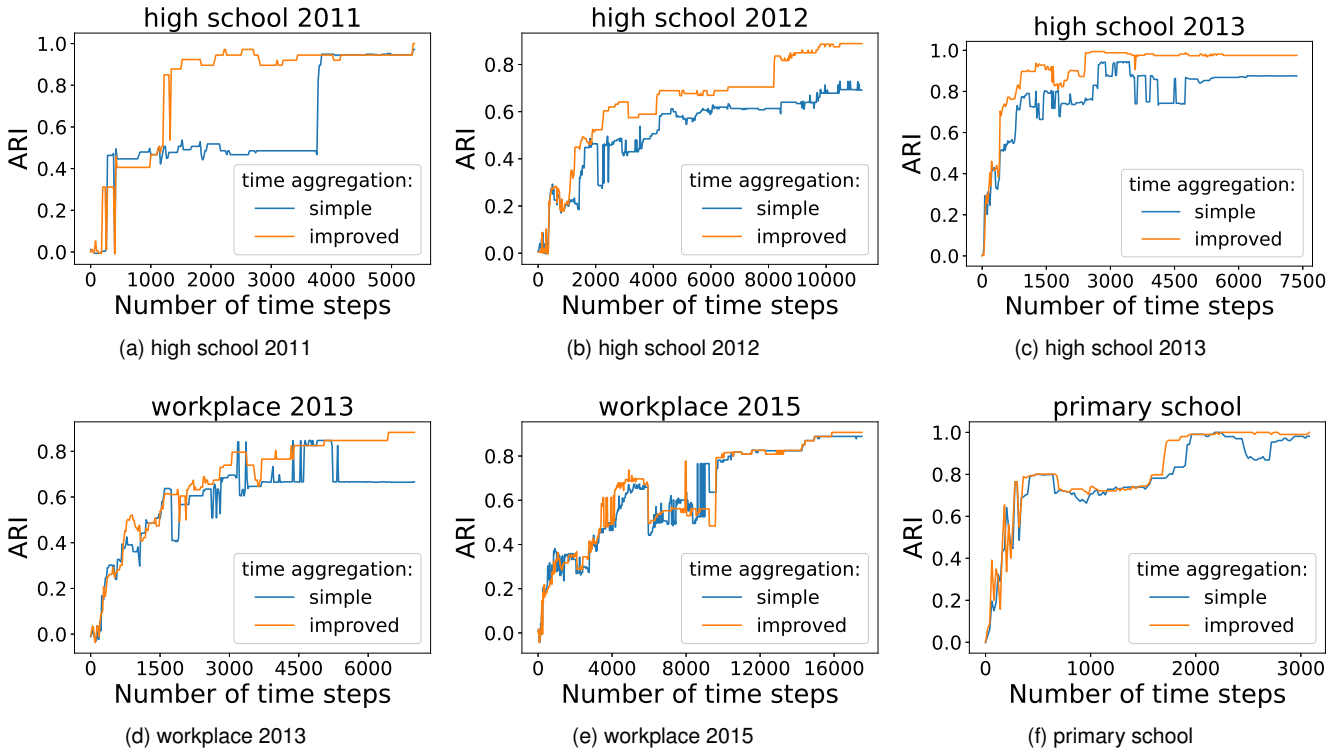
5. To implement this VEM algorithm, we used the R package *dynsbm* provided by the authors of [8] and at http://lbbe.univ-lyon1.fr/dynsbm.

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2023.3337281



(a) No degree correction (b) Gaussian (c) Pareto

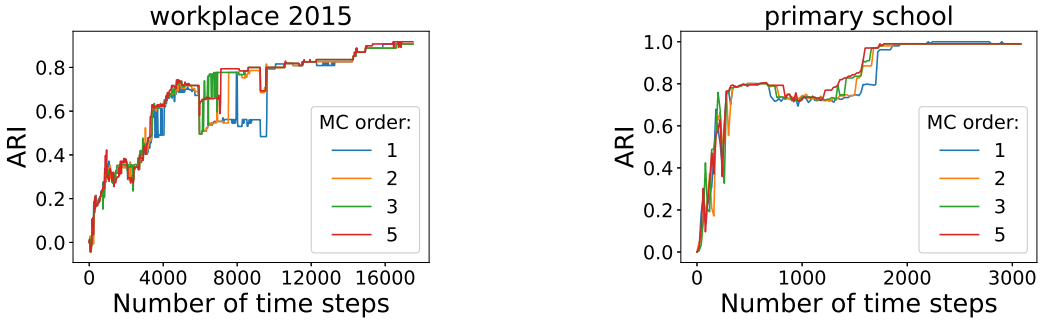
Fig. 1. Accuracy of Algorithm 1 with $\alpha = 1$ and different β, on a degree-corrected temporal SBM with 300 nodes and $K = 3$ blocks (with a uniform prior), and a stationary Markov edge evolution $\mu_1 = 0.02$, $\nu_1 = 0.01$, $P_{11} = 0.7$ and $Q_{11} = 0.2$, for a different generation of the degree correction parameters θ. The results are averaged over 25 synthetic graphs, and error bars show the standard error.



(a) high school 2011 (b) high school 2012 (c) high school 2013

(d) workplace 2013 (e) workplace 2015 (f) primary school

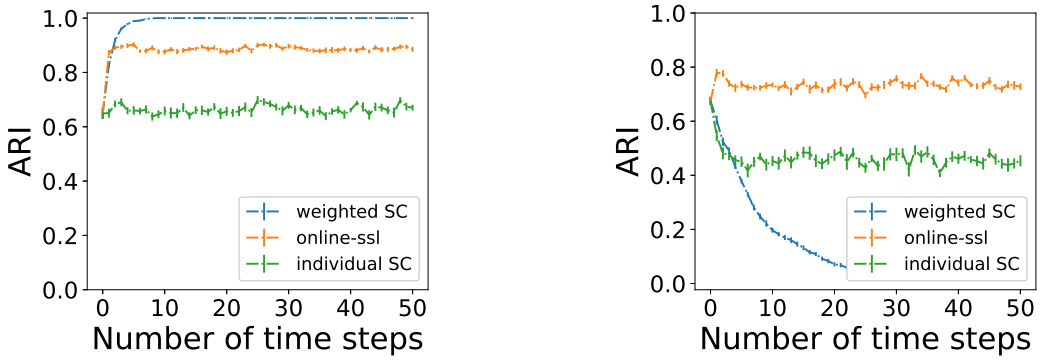
Fig. 2. Comparison of accuracy obtained on the SocioPatterns data sets by using spectral clustering with simple time-aggregation versus improved time-aggregation (Algorithm 1).



(a) workplace 2015 (b) primary school

Fig. 3. Comparison of the ARI obtained on *workplace 2015* and *primary school* data sets by using temporal spectral clustering with higher order correlations (Algorithm 2).

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2023.3337281



(a) $\eta = 1$ (b) $\eta = 0.85$

Fig. 4. ARI obtained by various clustering algorithms on time-varying Markov Block Models with 300 nodes and $K = 2$ blocks (with a uniform prior), and a stationary Markov edge evolution $\mu_1 = 0.1$, $\nu_1 = 0.05$, $P_{11} = 0.6$ and $Q_{11} = 0.2$. The results are averaged over 25 synthetic graphs, and error bars show the standard error. We ran Algorithm 3 (*online-ssl*) with parameters $\alpha_{01} = 1, \alpha_{10} = 0$ and $\alpha_{11} = 2$), and we compare with *weighted SC* (Algorithm 1 with $\alpha = 1, \beta = 2$) and an algorithm performing *spectral clustering* on each snapshot individually (*individual SC*).

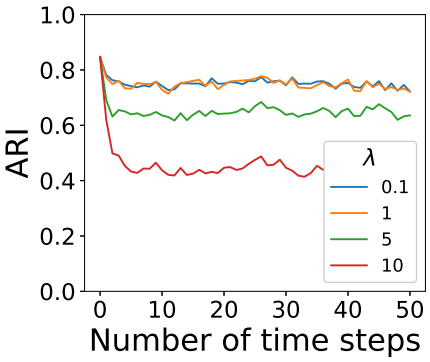


Fig. 5. ARI of Algorithm 3 with $\alpha_{01} = 1, \alpha_{10} = 0$ and $\alpha_{11} = 2$ for various choices of λ. Simulations are performed on time-varying Markov Block Models with $n = 200$, $K = 2$, $\mu_1 = 0.2$, $\nu_1 = 0.1$, $P_{11} = 0.7$, $Q_{11} = 0.3$ and $\eta = 0.8$. The results are averaged over 25 synthetic graphs, and error bars show the standard error.

Several approximations and relaxations were made in deriving a weighted similarity matrix used for spectral clustering and for tuning its weights. Analysing the accuracy of these approximations, and proving that the proposed algorithms terminate in a finite time remain topics for future research.

ACKNOWLEDGMENTS

This work has been done within the project of Inria - Nokia Bell Labs "Distributed Learning and Control for Network Analysis" and was partly supported by COSTNET EU Cost Action CA15109. We thank anonymous reviewers for helping to improve the presentation of the article.

APPENDIX A
PROOFS FOR SECTION 2

A.1 MLE computation

Proof of Proposition 2.1. By the temporal Markov property, the log-likelihood of the model can be written as $\log \mathbb{P}(A \mid \sigma, \theta) = \log \mathbb{P}(A^1 \mid \sigma, \theta) + \sum_{t=2}^{T} \log \mathbb{P}(A^t \mid A^{t-1}, \sigma, \theta)$. By denoting $\rho_a^{\theta_i\theta_j} = \log \frac{\mu_a^{\theta_i\theta_j}}{\nu_a^{\theta_i\theta_j}}$, we find that

$$\begin{aligned}
&\log \mathbb{P}(A^1 \mid \sigma, \theta) \\
&= \frac{1}{2} \sum_{i,j} \sum_a \delta(A_{ij}^1, a) \Big(\delta(\sigma_i, \sigma_j) \rho_a^{\theta_i\theta_j} + \log \nu_a^{\theta_i\theta_j} \Big) \\
&= \frac{1}{2} \sum_{i,j} \delta(\sigma_i, \sigma_j) \sum_a \delta(A_{ij}^1, a) \rho_a^{\theta_i\theta_j} + c_1(A),
\end{aligned}$$

where $c_1(A) = \frac{1}{2} \sum_{i,j} \sum_a \delta(A_{ij}^1, a) \log \nu_a^{\theta_i\theta_j}$ does not depend on the community structure. Similarly, by denoting $R_{ab}^{\theta_i\theta_j} = \log \frac{P_{ab}^{\theta_i\theta_j}}{Q_{ab}^{\theta_i\theta_j}}$, we find that

$$\begin{aligned}
&\log \mathbb{P}(A^t \mid A^{t-1}, \sigma, \theta) \\
&= \frac{1}{2} \sum_{i,j} \sum_{a,b} \delta(A_{ij}^{t-1}, a) \delta(A_{ij}^t, b) \Big(\delta(\sigma_i, \sigma_j) R_{ab}^{\theta_i\theta_j} + \log Q_{ab}^{\theta_i\theta_j} \Big) \\
&= \frac{1}{2} \sum_{i,j} \delta(\sigma_i, \sigma_j) \sum_{a,b} \delta(A_{ij}^{t-1}, a) \delta(A_{ij}^t, b) R_{ab}^{\theta_i\theta_j} + c_t(A),
\end{aligned}$$

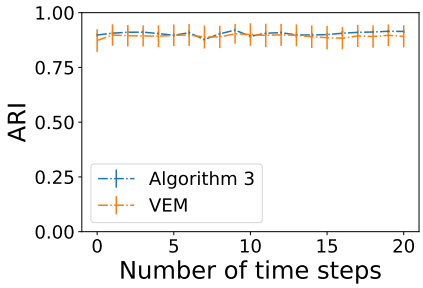
where $c_t(A) = \frac{1}{2} \sum_{i,j} \sum_{a,b} \delta(A_{ij}^{t-1}, a) \delta(A_{ij}^t, b) \log Q_{ab}^{\theta_i\theta_j}$ does not depend on the community structure. Simple computations show that

$$\sum_a \delta(A_{ij}^1, a) \rho_a^{\theta_i\theta_j} = A_{ij}^1 (\rho_1^{\theta_i\theta_j} - \rho_0^{\theta_i\theta_j}) + \rho_0^{\theta_i\theta_j}$$

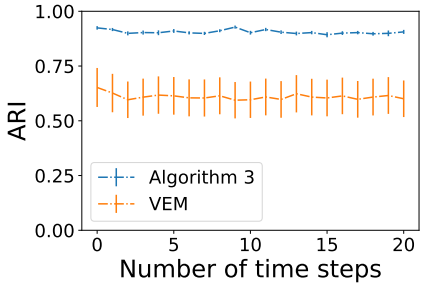
and

$$\begin{aligned}
&\sum_{a,b} \delta(A_{ij}^{t-1}, a) \delta(A_{ij}^t, b) R_{ab}^{\theta_i\theta_j} \\
&= R_{00}^{\theta_i\theta_j} + A_{ij}^{t-1} \left(R_{10}^{\theta_i\theta_j} - R_{00}^{\theta_i\theta_j} \right) + A_{ij}^t \left(R_{01}^{\theta_i\theta_j} - R_{00}^{\theta_i\theta_j} \right) \\
&\qquad + A_{ij}^{t-1} A_{ij}^t \left(R_{11}^{\theta_i\theta_j} - R_{01}^{\theta_i\theta_j} - R_{10}^{\theta_i\theta_j} + R_{00}^{\theta_i\theta_j} \right) \\
&= R_{00}^{\theta_i\theta_j} + A_{ij}^{t-1} \ell_{10}^{\theta_i\theta_j} + A_{ij}^t \ell_{01}^{\theta_i\theta_j} \\
&\qquad + A_{ij}^{t-1} A_{ij}^t \left(\ell_{11}^{\theta_i\theta_j} - \ell_{01}^{\theta_i\theta_j} - \ell_{10}^{\theta_i\theta_j} \right).
\end{aligned}$$

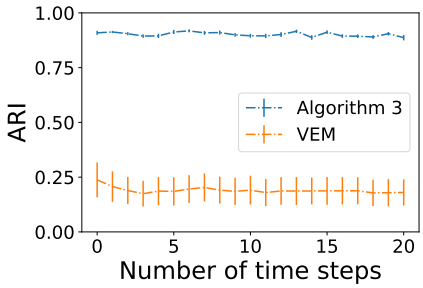
This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2023.3337281



(a) $P_{11} = 0.2$



(b) $P_{11} = 0.5$



(c) $P_{11} = 0.7$

Fig. 6. Comparison of ARIs obtained by Algorithm 3 and the VEM algorithm of [8] on time-varying Markov Block Models with 250 nodes and $K = 2$ blocks (with a uniform prior), $\eta = 0.75$, and a stationary Markov edge evolution $\mu_1 = 0.2$, $\nu_1 = 0.1$, $Q_{11} = 0.1$ and for various P_{11}. Figure 6(a) is a setting where links are re-sampled at each time step, while Figures 6(b) and 6(c) models medium and high intra-community link persistency. The results are averaged over 25 synthetic graphs, and error bars show the standard error.

By collecting the above observations, we now find that $\log \mathbb{P}(A \,|\, \sigma, \theta)$ equals

$$\frac{1}{2} \sum_{\substack{i,j \\ \sigma_i = \sigma_j}} \left\{ A^1_{ij} (\rho_1^{\theta_i\theta_j} - \rho_0^{\theta_i\theta_j}) + \rho_0^{\theta_i\theta_j} + (A^1_{ij} - A^T_{ij}) \ell_{10}^{\theta_i\theta_j} \right\}$$
$$+ \frac{1}{2} \sum_{\substack{i,j \\ \sigma_i = \sigma_j}} \sum_{t=2}^{T} \left\{ (\ell_{01}^{\theta_i\theta_j} + \ell_{10}^{\theta_i\theta_j}) \left(A^t_{ij} - A^{t-1}_{ij} A^t_{ij} \right) \right.$$
$$\left. + \ell_{11}^{\theta_i\theta_j} A^{t-1}_{ij} A^t_{ij} - \log \frac{Q_{00}^{\theta_i\theta_j}}{P_{00}^{\theta_i\theta_j}} \right\} + c(A),$$

where $c(A) = \sum_t c_t(A)$ does not depend on σ. Hence the claim follows. □

A.2 Sparse MLE and modularity

Recall the structural assumptions (2.3)–(2.4) about the degree correction parameters. Because $P_{01}, Q_{01} = o(1)$, a first-order Taylor expansion yields

$$\log \frac{1 - \theta_i\theta_j Q_{01}}{1 - \theta_i\theta_j P_{01}} \;=\; \theta_i\theta_j \left(P_{01} - Q_{01}\right) + o\left(P_{01} + Q_{01}\right),$$

as well as $\ell_{01}^{\theta_i\theta_j} \approx \log \frac{P_{01}}{Q_{01}}$, $\ell_{10}^{\theta_i\theta_j} \approx \log \frac{1-P_{11}}{1-Q_{11}}$ and $\ell_{11}^{\theta_i\theta_j} \approx \log \frac{P_{11}}{Q_{11}}$. Using these approximations in the MLE expression leads to the maximisation of

$$\sum_{t=2}^{T} \sum_{i,j\,:\,\sigma_i = \sigma_j} \left(\tilde{a}^t_{ij} - \theta_i\theta_j \left(P_{01} - Q_{01}\right) \right), \qquad \text{(A.1)}$$

where $\tilde{a}^t_{ij} = \alpha \left(A^t_{\text{new}}\right)_{ij} + \beta \left(A^t_{\text{pers}}\right)_{ij}$. Since μ and ν are stationary distributions,

$$\mathbb{E}\left(A^t_{\text{new}}\right)_{ij} \;=\; \begin{cases} \theta_i\theta_j\mu_1(1 - P_{11}) & \text{if } \sigma_i = \sigma_j, \\ \theta_i\theta_j\nu_1(1 - Q_{11}) & \text{otherwise,} \end{cases}$$
$$\mathbb{E}\left(A^t_{\text{pers}}\right)_{ij} \;=\; \begin{cases} \theta_i\theta_j\mu_1 P_{11} & \text{if } \sigma_i = \sigma_j, \\ \theta_i\theta_j\nu_1 Q_{11} & \text{otherwise.} \end{cases}$$

Therefore, using $W_{ij} = \sum_{t=2}^{T} \tilde{a}_{ij}$, we have

$$\mathbb{E}W_{ij} \;=\; \begin{cases} (T-1)\theta_i\theta_j\mu_1 \left(\alpha(1 - P_{11}) + \beta P_{11}\right) & \text{if } \sigma_i = \sigma_j \\ (T-1)\theta_i\theta_j\nu_1 \left(\alpha(1 - Q_{11}) + \beta Q_{11}\right) & \text{otherwise.} \end{cases}$$

Since the community labelling is sampled uniformly at random, using the normalization for the θ_i's, we obtain $\bar{d}_i = (T-1)\theta_i N \frac{\mu_1(\alpha(1-P_{11})+\beta P_{11})+(K-1)\nu_1(\alpha(1-Q_{11})+\beta Q_{11})}{K}$, together with $\bar{m} = \frac{(T-1)N^2}{2} \frac{\mu_1(\alpha(1-P_{11})+\beta P_{11})+(K-1)\nu_1(\alpha(1-Q_{11})+\beta Q_{11})}{K}$. Hence, we observe that

$$\theta_i\theta_j(P_{01} - Q_{01}) = \gamma \frac{\bar{d}_i \bar{d}_j}{2\bar{m}},$$

where $\gamma = (P_{01} - Q_{01})(T-1) \frac{\mu_1(\alpha(1-P_{11})+\beta P_{11})+(K-1)\nu_1(\alpha(1-Q_{11})+\beta Q_{11})}{K}$. We end the proof by employing equation (A.1). □

A.3 Modularity and spectral clustering

The regularized modularity of a partition $\sigma \in [K]^N$ of the graph A is defined as

$$\mathcal{M}\left(A, \sigma, \gamma\right) \;=\; \sum_{i,j} \delta\left(\sigma_i, \sigma_j\right) \left(A_{ij} - \gamma \frac{d_i d_j}{2m} \right),$$

where $d = A1_n$ is a vector of degrees and γ is a resolution parameter. The above equation can be rewritten as

$$\mathcal{M}\left(A, \sigma, \gamma\right) \;=\; \operatorname{Tr} \tilde{Z}^T \left(A - \gamma \frac{dd^T}{2m} \right) \tilde{Z},$$

where $\tilde{Z} \in \{0,1\}^{N \times K}$ is the membership matrix associated with the vector σ, that is $\tilde{Z}_{ik} = 1$ for $k = \sigma_i$, and $\tilde{Z}_{ik} = 0$ otherwise. As maximising the modularity over $\sigma \in [K]^N$ is in general NP-complete [23], it is convenient to perform a continuous relaxation. Following [24], we transform the problem into

$$\hat{X} \;=\; \underset{\substack{X \in \mathbb{R}^{N \times K} \\ X^T D X = I_K}}{\arg\max} \operatorname{Tr} X^T \left(A - \gamma \frac{dd^T}{2m} \right) X. \qquad \text{(A.2)}$$

The predicted membership matrix $\hat{Z}$ is then recovered by performing an approximated solution to the following k-means problem (see [38])

$$\left(\hat{Z}, \hat{Y}\right) \;=\; \underset{Z\in\mathcal{Z}_{N,K}, Y\in\mathbb{R}^{K\times K}}{\arg\min} \left\| ZY - \hat{X} \right\|_F . \tag{A.3}$$

The Lagrangian associated to the optimisation problem (A.2) is

$$\operatorname{Tr} X^T \left(A - \gamma \frac{dd^T}{2m} \right) X - \operatorname{Tr}\left(\Lambda^T \left(X^T X - I_K \right) \right),$$

where $\Lambda \in \mathbb{R}^{K\times K}$ is a symmetric matrix of Lagrangian multipliers. Up to a change of basis, we can assume that Λ is diagonal. The stationarity condition for the solution of (A.2) becomes

$$\left(A - \gamma \frac{dd^T}{2m} \right) X \;=\; DX\Lambda \quad \text{and} \quad X^T D X \;=\; I_K,$$

which is a generalized eigenvalue problem: the columns of X are the generalized eigenvectors, and the diagonal elements of Λ are the eigenvalues. In particular, since the constant vector 1_n satisfies $(A - \gamma\frac{dd^T}{2m})1_n = (1-\gamma)D1_n$, we conclude that the eigenvalues should be larger than $1-\gamma$ for the partition to be meaningful.

Multiplying the first equation by 1_n^T leads to $(1-\gamma)d^T X = d^T X \Lambda$, and therefore $d^T X = 0$ (using the previous remark on Λ). The system then simplifies in

$$AX \;=\; DX\Lambda \quad \text{and} \quad X^T D X \;=\; I_K.$$

Defining a re-scaled vector $U = D^{-1/2}X$ shows that U satisfies $D^{-1/2}AD^{-1/2}U = U\Lambda$ and $U^TU = I_K$. Thus, the columns of U are eigenvectors of $D^{-1/2}AD^{-1/2}$ associated with the K largest eigenvalues (or equivalently, the eigenvectors of $\mathcal{L} = I_N - D^{-1/2}AD^{-1/2}$ associated to the K smallest eigenvalues).

Appendix B
Proofs for Section 3

Proof of Proposition 3.1. By letting $\Gamma \in \mathbb{R}^{K\times K}$ be the diagonal matrix whose diagonal entries Γ_{kk} are the Lagrange multipliers associated with the constraints $(X^TDX)_{kk} = 1$, the Lagrangian of the relaxed optimisation problem (3.5) is

$$\operatorname{Tr}\left(X^T W_\tau X + 2\lambda S^T X \right) + \operatorname{Tr}\left(\left(X^T D X - I_k \right) \Gamma \right).$$

This leads to the *constrained* linear system

$$\begin{cases} W_\tau X + DX\Gamma &= -\lambda S, \\ (X^TDX)_{kk} &= 1, \end{cases} \tag{B.1}$$

whose unknowns are Γ and X.

Firstly, we note that if (Γ^1, X^1) and (Γ^2, X^2) are solutions of the system (B.1), then (see Lemma B.1 for the computations)

$$\begin{aligned} \mathcal{C}\left(X^1\right) - \mathcal{C}\left(X^2\right) \;&=\; \operatorname{Tr}\left[(\Gamma^1 - \Gamma^2)(X^1 - X^2)^T D (X^1 - X^2) \right] \\ &= \sum_{k\in[K]} \left(\Gamma^2_{kk} - \Gamma^1_{kk}\right) \left\| X^1_{\cdot k} - X^2_{\cdot k} \right\|_D \end{aligned}$$

where $\mathcal{C}(x) = \operatorname{Tr}\left(X^T W_\tau X + 2\lambda S^T X\right)$ is the cost function maximised in (3.5) and $\|x\|_D = x^T D x$. Hence, among the solution pairs (Γ, X) of the system (B.1), the solution of the minimisation problem (3.5) is the vector X associated with the diagonal matrix Γ whose entries are the smallest.

Secondly, the eigenvalue decomposition of $-D^{-1/2}W_\tau D^{-1/2}$ reads as

$$-D^{-1/2}W_\tau D^{-1/2} \;=\; Q\Delta Q^T,$$

where $\Delta = \operatorname{diag}(\delta_1, \ldots, \delta_N)$ with $\delta_1 \le \cdots \le \delta_N$ and $Q^TQ = I_N$. Therefore, after the change of variables $U = Q^T D^{1/2} X$ and $B = \lambda Q^T D^{-1/2} S$, the system (B.1) is transformed to

$$\begin{cases} \Delta U &= U\Gamma + B, \\ (U^TU)_{kk} &= 1 \quad \forall k \in [K]. \end{cases}$$

Thus, the solution $\hat{X} = (\hat{X}_{\cdot 1}, \cdots, \hat{X}_{\cdot K})$ of the optimisation problem (3.5) satisfies

$$\left(-W_\tau - \gamma^*_k D\right) \hat{X}_{\cdot k} \;=\; \lambda S_{\cdot k},$$

where for all $k \in [K]$, γ^*_k is the smallest solution of the *explicit secular equation* [39]

$$\sum_{i=1}^{N} \left(\frac{B_{ik}}{\delta_i - \gamma} \right)^2 - 1 \;=\; 0.$$

□

Lemma B.1. *If (Γ^1, X^1) and (Γ^2, X^2) are solutions of the system (B.1), then*

$$\mathcal{C}\left(X^1\right) - \mathcal{C}\left(X^2\right) \;=\; \sum_{k\in[K]} \left(\Gamma^2_{kk} - \Gamma^1_{kk}\right) \left\| X^1_{\cdot k} - X^2_{\cdot k} \right\|_D ,$$

where $\mathcal{C}(X) = \operatorname{Tr}\left(X^T W_\tau X + 2\lambda S^T X\right)$ and $\|x\|_D = x^T D x$.

Proof. Because (Γ^1, X^1) and (Γ^2, X^2) are solutions of (B.1), it holds that

$$\begin{aligned} -W_\tau X^1 - DX^1\Gamma^1 &= \lambda S, \\ -W_\tau X^2 - DX^2\Gamma^2 &= \lambda S. \end{aligned} \tag{B.2}$$

Using the definition of $\mathcal{C}(X)$ and the fact that (Γ^1, X^1) is solution of (B.1), we notice that

$$\mathcal{C}(X^1) \;=\; \operatorname{Tr}\left(-(X^1)^T D X^1 \Gamma^1 + \lambda (X^1)^T S \right),$$

and similarly for $\mathcal{C}(X_2)$. Moreover, multiplying the first equation of (B.2) (resp., the second equation) by $(X^2)^T$ (resp., by $(X^1)^T$) gives

$$\begin{aligned} \lambda (X^2)^T S \;&=\; -(X^2)^T W_\tau X^1 - (X^2)^T D X^1 \Gamma^1, \\ \lambda (X^1)^T S \;&=\; -(X^1)^T W_\tau X^2 - (X^1)^T D X^2 \Gamma^2. \end{aligned}$$

Because $(X^1)^T W_\tau X^2 = (X^2)^T W_\tau X^1$, we have

$$\begin{aligned} &\operatorname{Tr}\left(\lambda S^T \left(X^1 - X^2 \right) \right) \\ &= \operatorname{Tr}\left(-(X^1)^T D X^2 \Gamma^2 + (X^2)^T D X^1 \Gamma^1 \right). \end{aligned}$$

Therefore,

$$\begin{aligned} \mathcal{C}(X^1) - \mathcal{C}(X^2) \;&=\; \operatorname{Tr}\left(-(X^1)^T D X^1 \Gamma^1 + (X_2)^T D X^1 \Gamma^1 \right) \\ &\quad - \operatorname{Tr}\left(-(X^2)^T D X^2 \Gamma^2 + (X^1)^T D X^2 \Gamma^2 \right). \end{aligned} \tag{B.3}$$

Moreover, because Γ^1 and D are symmetric, we have

$$\begin{aligned}
&\operatorname{Tr}\left(-(X^1)^T D X^1 \Gamma^1 + (X_2)^T D X^1 \Gamma^1\right) \\
&= \sum_{k\in[K]} \Gamma^1_{kk} \left(\left(-(X^1)^T D X^1\right)_{kk} + \left((X_2)^T D X^1\right)_{kk}\right) \\
&= \sum_{k\in[K]} \Gamma^1_{kk} \left(-(X^1_{\cdot k})^T D X^1_{\cdot k} + (X^2_{\cdot k})^T D X^1_{\cdot k}\right).
\end{aligned}$$

The constraints $X^T_{\cdot k} D X_{\cdot k} = 1$ verified by both X_1 and X_2 leads to

$$\begin{aligned}
&-(X^1_{\cdot k})^T D X^1_{\cdot k} + (X^2_{\cdot k})^T D X^1_{\cdot k} \\
&= -\frac{(X^1_{\cdot k})^T D X^1_{\cdot k} + (X^2_{\cdot k})^T D X^2_{\cdot k}}{2} + (X^2_{\cdot k})^T D X^1_{\cdot k} \\
&= -\|X^1_{\cdot k} - X^2_{\cdot k}\|_D,
\end{aligned}$$

where the last equality uses $(X^2_{\cdot k})^T D X^1_{\cdot k} = (X^1_{\cdot k})^T D X^2_{\cdot k}$. This ensures that

$$\begin{aligned}
&\operatorname{Tr}\left(-(X^1)^T D X^1 \Gamma^1 + (X_2)^T D X^1 \Gamma^1\right) \\
&= -\sum_{k\in[K]} \Gamma^1_{kk} \|X^1_{\cdot k} - X^2_{\cdot k}\|_D. \qquad \text{(B.4)}
\end{aligned}$$

Similarly, we can establish that

$$\begin{aligned}
&\operatorname{Tr}\left(-(X^2)^T D X^2 \Gamma^2 + (X^1)^T D X^2 \Gamma^2\right) \\
&= -\sum_{k\in[K]} \Gamma^2_{kk} \|X^1_{\cdot k} - X^2_{\cdot k}\|_D. \qquad \text{(B.5)}
\end{aligned}$$

We finish the proof by combining (B.4) and (B.5) with (B.3). □

REFERENCES

[1] S. Fortunato and M. E. Newman, "20 years of network community detection," *Nature Physics*, vol. 18, no. 8, pp. 848–850, 2022.

[2] P. Holme and J. Saramäki, "Temporal networks," *Physics Reports*, vol. 519, no. 3, pp. 97–125, 2012, temporal Networks.

[3] S. Paul and Y. Chen, "Consistent community detection in multi-relational data through restricted multi-layer stochastic block-model," *Electronic Journal of Statistics*, vol. 10, no. 2, pp. 3807–3870, 2016.

[4] K. Avrachenkov, M. Dreveton, and L. Leskelä, "Community recovery in non-binary and temporal stochastic block models," *arXiv e-prints*, pp. arXiv–2008, 2020.

[5] S. Bhattacharyya and S. Chatterjee, "General community detection with optimal recovery conditions for multi-relational sparse networks with dependent layers," *arXiv preprint arXiv:2004.03480*, 2020.

[6] J. Lei and K. Z. Lin, "Bias-adjusted spectral clustering in multi-layer stochastic block models," *Journal of the American Statistical Association*, pp. 1–13, 2022.

[7] A. Ghasemian, P. Zhang, A. Clauset, C. Moore, and L. Peel, "Detectability thresholds and optimal algorithms for community structure in dynamic networks," *Physical Review X*, vol. 6, no. 3, p. 031005, 2016.

[8] C. Matias and V. Miele, "Statistical clustering of temporal networks through a dynamic stochastic block model," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 79, no. 4, pp. 1119–1141, 2017.

[9] Y. Chi, X. Song, D. Zhou, K. Hino, and B. L. Tseng, "Evolutionary spectral clustering by incorporating temporal smoothness," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 153–162.

[10] F. Liu, D. Choi, L. Xie, and K. Roeder, "Global spectral clustering in dynamic networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 5, pp. 927–932, 2018.

[11] L. Dall'Amico, R. Couillet, and N. Tremblay, "Community detection in sparse time-evolving graphs with a dynamical bethe-hessian," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7486–7497, 2020.

[12] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *Science*, vol. 328, no. 5980, pp. 876–878, 2010.

[13] A. R. Pamfil, S. D. Howison, R. Lambiotte, and M. A. Porter, "Relating modularity maximization and stochastic block models in multilayer networks," *SIAM Journal on Mathematics of Data Science*, vol. 1, no. 4, pp. 667–698, 2019.

[14] G. Rossetti and R. Cazabet, "Community discovery in dynamic networks: A survey," *ACM Comput. Surv.*, vol. 51, no. 2, Feb. 2018.

[15] K. Avrachenkov and M. Dreveton, *Statistical Analysis of Networks*. Now Publishers, 2022.

[16] P. Barucca, F. Lillo, P. Mazzarisi, and D. Tantari, "Disentangling group and link persistence in dynamic stochastic block models," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2018, p. 123407, 12 2018.

[17] P. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, pp. 109–137, 1983.

[18] B. Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," *Phys. Rev. E*, vol. 83, p. 016107, Jan 2011.

[19] K. Avrachenkov, M. Dreveton, and L. Leskelä, "Recovering communities in temporal networks using persistent edges," in *Computational Data and Social Networks (CSoNet 2021)*, ser. Lecture Notes in Computer Science, D. Mohaisen and R. Jin, Eds., vol. 13116. Springer, 2021, pp. 243–254.

[20] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, p. 026113, 2004.

[21] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," *Physical Review E*, vol. 74, no. 1, p. 016110, 2006.

[22] M. E. Newman, "Equivalence between modularity optimization and maximum likelihood methods for community detection," *Physical Review E*, vol. 94, no. 5, p. 052315, 2016.

[23] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner, "On finding graph clusterings with maximum modularity," in *International Workshop on Graph-Theoretic Concepts in Computer Science*. Springer, 2007, pp. 121–132.

[24] M. E. J. Newman, "Spectral methods for community detection and graph partitioning," *Phys. Rev. E*, vol. 88, p. 042822, Oct 2013.

[25] M. Rosvall, A. V. Esquivel, A. Lancichinetti, J. D. West, and R. Lambiotte, "Memory in network flows and its effects on spreading dynamics and community detection," *Nature Communications*, vol. 5, no. 1, p. 4630, 2014.

[26] J. Xu, T. L. Wickramarathne, and N. V. Chawla, "Representing higher-order dependencies in networks," *Science Advances*, vol. 2, no. 5, p. e1600028, 2016.

[27] L. Eldén and H. Park, "A procrustes problem on the stiefel manifold," *Numerische Mathematik*, vol. 82, no. 4, pp. 599–619, 1999.

[28] L.-H. Zhang, W. H. Yang, C. Shen, and J. Ying, "An eigenvalue-based method for the unbalanced procrustes problem," *SIAM Journal on Matrix Analysis and Applications*, vol. 41, no. 3, pp. 957–983, 2020.

[29] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin, "Detecting communities and their evolutions in dynamic social networks—a Bayesian approach," *Machine Learning*, vol. 82, no. 2, pp. 157–189, 2011. [Online]. Available: https://doi.org/10.1007/s10994-010-5214-7

[30] K. S. Xu and A. O. Hero, "Dynamic stochastic blockmodels for time-evolving social networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 4, pp. 552–562, 2014.

[31] M. Pensky and T. Zhang, "Spectral clustering in the dynamic stochastic block model," *Electronic Journal of Statistics*, vol. 13, no. 1, pp. 678–709, 2019. [Online]. Available: https://doi.org/10.1214/19-EJS1533

[32] K. Xu, "Stochastic block transition models for dynamic networks," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, G. Lebanon and S. V. N. Vishwanathan, Eds., vol. 38, 2015, pp. 1079–1087. [Online]. Available: https://proceedings.mlr.press/v38/xu15.html

[33] J. Fournet and A. Barrat, "Contact patterns among high school students," *PLoS ONE*, vol. 9, no. 9, pp. 1–17, 09 2014.

[34] R. Mastrandrea, J. Fournet, and A. Barrat, "Contact patterns in a high school: A comparison between data collected using wear-

able sensors, contact diaries and friendship surveys," *PLoS ONE*, vol. 10, no. 9, pp. 1–26, 09 2015.
[35] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina *et al.*, "High-resolution measurements of face-to-face contact patterns in a primary school," *PloS one*, vol. 6, no. 8, p. e23176, 2011.
[36] M. Génois, C. L. Vestergaard, J. Fournet, A. Panisson, I. Bonmarin, and A. Barrat, "Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers," *Network Science*, vol. 3, no. 3, pp. 326–347, 2015.
[37] M. Génois and A. Barrat, "Can co-location be used as a proxy for face-to-face contacts?" *EPJ Data Science*, vol. 7, no. 1, pp. 1–18, 2018.
[38] A. Kumar and R. Kannan, "Clustering with spectral norm and the k-means algorithm," in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, 2010, pp. 299–308.
[39] W. Gander, G. H. Golub, and U. Von Matt, "A constrained eigenvalue problem," *Linear Algebra and its Applications*, vol. 114, pp. 815–839, 1989.

Konstantin Avrachenkov received his Master's degree in Control Theory from St. Petersburg State Polytechnic University (1996), Ph.D. degree in Mathematics from the University of South Australia (2000) and a Habilitation (Doctor of Science) from the University of Nice Sophia Antipolis (2010). Currently, he is a Director of Research at INRIA Sophia Antipolis, France. He is an associate editor of Probability in the Engineering and Informational Sciences, ACM TOMPECS, Stochastic Models, IEEE Network Magazine and on the Advisory Board of the International Journal of Performance Evaluation. He has won 5 best paper awards. His main research interests are Markov chains, Markov decision processes, stochastic games and singular perturbations. He applies these methodological tools to the modelling and control of networks and to design data mining and machine learning algorithms.

Maximilien Dreveton received his Bachelor's and Master's degrees, in the field of Physics, from the Ecole Normale Supérieure de Lyon, France in 2013 and 2015. He obtained his PhD in computer science from Inria Sophia Antipolis in 2022 and is currently working as a postdoctoral fellow at EPFL in Lausanne. His research interests include statistical analysis of random graphs, and more particularly community detection in networks.

Lasse Leskelä received the MSc and DSc degrees in mathematics from Helsinki University of Technology, Finland, in 1999 and 2005, respectively. He also holds a title of Docent at the University of Jyväskylä, Finland. Currently he is an Associate Professor of Applied Mathematics at Aalto University, Finland. His research interests include network statistics, random graphs, probability theory, and stochastic processes.