Barbarino, Giovanni; Noferini, Vanni; Van Dooren, Paul

Role extraction for digraphs via neighborhood pattern similarity

Published in:
Physical Review E

DOI:
10.1103/PhysRevE.106.054301

Published: 01/11/2022

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Barbarino, G., Noferini, V., & Van Dooren, P. (2022). Role extraction for digraphs via neighborhood pattern similarity. *Physical Review E*, *106*(5), 1-11. Article 054301. https://doi.org/10.1103/PhysRevE.106.054301

Role extraction for digraphs via neighborhood pattern similarity

Giovanni Barbarino* and Vanni Noferini
Aalto University, Department of Mathematics and Systems Analysis, P.O. Box 11100, FI-00076 Aalto, Finland

Paul Van Dooren
Université catholique de Louvain, Department of Mathematical Engineering, Av. Lemaitre 4, B-1348 Louvain-la-Neuve, Belgium

(Received 10 May 2022; accepted 6 October 2022; published 1 November 2022)

We analyze the recovery of different roles in a network modeled by a directed graph, based on the so-called Neighborhood Pattern Similarity approach. Our analysis uses results from random matrix theory to show that, when assuming that the graph is generated as a particular stochastic block model with Bernoulli probability distributions for the different blocks, then the recovery is asymptotically correct when the graph has a sufficiently large dimension. Under these assumptions there is a sufficient gap between the dominant and dominated eigenvalues of the similarity matrix, which guarantees the asymptotic correct identification of the number of different roles. We also comment on the connections with the literature on stochastic block models, including the case of probabilities of order $\log(n)/n$ where n is the graph size. We provide numerical experiments to assess the effectiveness of the method when applied to practical networks of finite size.

DOI: 10.1103/PhysRevE.106.054301

I. INTRODUCTION

The analysis of large graphs frequently assumes that there is an underlying structure in the graph that allows us to represent it in a simpler manner. A typical example of this is the detection of communities, which are groups of nodes that have most of their connections with other nodes of the same group, and few connections with nodes of other groups. Various measures and algorithms have been developed to identify community structures [1] and many applications have also been found for these model structures [2–5]. Yet, many graph structures cannot be modeled using communities: for example, arrowhead and tree graph structures, which appear in overlapping communities, human protein-protein interaction networks, and food and web networks [3,6,7]. These more general types of network structures can be modeled as role structures, and the process of finding them is called the role extraction problem or block modeling [8–12]. The role extraction problem is a generalization of the community detection problem and it determines a representation of a network by a smaller structured graph, where nodes are grouped together based upon their interactions with nodes in either the same group or in different groups called roles. If no *a priori* information is available, one needs to verify all possible group and role assignments in order to determine the best role structure for the data, which leads to an NP-hard problem [9,13] for both the community detection problem and the more general role extraction problem.

There are many algorithms proposed for community detection, for both directed and undirected graphs [1,14–18], but they often do not state any conclusive results about the exact recovery of communities, because they make no statistical assumption about the underlying model of the graph. On the other hand, if one assumes that the adjacency matrix of the graph is a sample of a random matrix that follows certain rules, then the problem of recovering the correct underlying block structure may become tractable. The stochastic block model (SBM) is precisely such a model: the interactions between all nodes of a particular group with all nodes of another group follow exactly the same distribution [19]. There is a considerable literature on SBM [10,20,21], including variants that address diagonal scaling of the SBM [22].

To deal with this problem, researchers have proposed a variety of procedures, which vary greatly in their degrees of statistical accuracy and computational complexity. See, for example, modularity maximization [23], likelihood methods [24–28], Infomod methods [29,30], Monte Carlo methods [31,32], method of moments [33], belief propagation [34], convex optimization [35] and its variants [36,37], methods based on mixture models [38,39], the clique percolation method [40], spectral embeddings [41], and hierarchical clustering through minimum description length [42–44] or Bayesian model selection [45,46].

A class of algorithms that has been largely employed in the past years for such purpose are the so-called spectral methods [47–51]. Broadly speaking, a spectral method first performs an eigendecomposition of a symmetric matrix encoding the properties of the graph. Then the community membership is inferred by applying a clustering algorithm, typically K means, to the rows of the matrix formed by the first few leading eigenvectors. Spectral clustering is easier to implement and computationally less demanding than many other methods, which amount to computationally intractable combinatorial searches. From a theoretical standpoint, spectral clustering has been shown to enjoy good theoretical properties

*Corresponding author: giovanni.barbarino@aalto.fi

in stochastic block models [52–54]. In the computer science literature, spectral clustering is also a standard procedure for graph partitioning and for solving the planted partition model, a special case of the SBM [55].

As their first step requires the eigendecomposition of a symmetric matrix, spectral methods are commonly applied to undirected graphs. Moreover, when they do consider directed graphs, their analysis does not include the recovery of the underlying block structure [56].

In this paper we will show that a particular method, using the so-called Neighborhood Pattern Similarity (NPS) matrices [57,58], allows us to give a positive answer to the following question: Can we recover asymptotically the block structure of a general directed graph with a stochastic block model structure? A NPS matrix is a real symmetric positive semidefinite matrix, also for directed graphs, and therefore has real eigenvalues and eigenvectors. We then show that for sufficiently large graphs, the gap between dominant and dominated eigenvalues allows a convergent recovery of which nodes are associated with the different roles in the model. The nearest results available in the literature are the successful extraction of the correct roles in SBM for the community detection problem of undirected graphs [10,59], and the use of spectral clustering for the directed role extraction problem [60], in which a different type of stochastic block model is used. The present paper extends the asymptotic analysis of the general role modeling problem to specific symmetrizations of the standard SBM model for directed graphs, the NPS matrices, for which a correctness result is still missing in the literature. This is in particular one of the few existing results of correctness for the spectral clustering algorithm applied to directed graph with a SBM structure. Furthermore, our results can be seen as covering a whole class of methods, in the sense that our asymptotic analysis applies to all the admissible values of the scaling factor β and all NPS matrices S_k (including both any finite value of k and the limit $S = \lim_{k\to\infty} S_k$, which is the NPS matrix); see Sec. II C for the definitions of β and S_k. While in this paper we focus on the theoretical analysis of the method, determining optimal values of k and β for a practical implementation of the algorithm to analyze actual graphs is an interesting possible subject of future research.

In Sec. II we go over several preliminaries related to graphs, random matrices, stochastic block models and role modeling. Section III then yields the spectral bounds for the NPS matrix associated with the graph (as well as for the matrices S_k whose limit is the NPS matrix), and in Sec. IV we describe the asymptotic behavior of the clustering error. In Sec. V we give a few numerical experiments illustrating our theoretical analysis, and we conclude with a few final remarks in Sec. VI. Several technical proofs are moved to the Appendixes for the sake of readability.

II. PRELIMINARIES

A. Graph theory and role extraction

An unweighted *directed graph* or digraph, $G = (V, E)$, is an ordered pair of sets where the elements of $V = [n]$ are called *vertices* or nodes, while the elements of $E \subseteq V \times V$ are called *edges* or links. A walk of length ℓ on the digraph G from i_1 to $i_{\ell+1}$ is a sequence of vertices of the form $i_1, i_2, \ldots, i_{\ell+1}$ such that for all $j = 1, \ldots, \ell$, $(i_j, i_{j+1}) \in E$. G is said to be strongly connected if for all $i, j \in V$ there is a path on G from i to j.

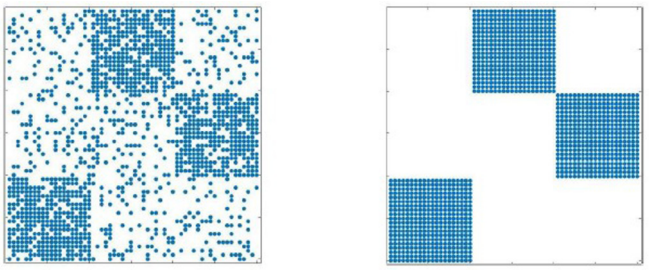

FIG. 1. Associating a regularly equivalent graph A_{id} to the permuted graph $P^T AP$.

The adjacency matrix of an unweighted digraph is defined as

$$A \in \mathbb{R}^{n\times n}, \quad A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise.} \end{cases}$$

In particular, A is a nonnegative matrix and $A \in \{0,1\}^{n\times n}$. It is well known that A is irreducible if and only if G is strongly connected; in that case, the Perron-Frobenius spectral theory for irreducible nonnegative matrices applies. Manifestly, there is a bijection between adjacency matrices and digraphs. Moreover, two digraphs are isomorphic (i.e., they coincide up to a relabeling of the vertices) if and only if their adjacency matrices are permutation similar.

Given a graph G with adjacency matrix A, the problem of role extraction consists in finding a positive integer $r \leqslant n$ and an *assignment* function $\xi : [n] \to [r]$ such that A can be well approximated by an ideal adjacency matrix E such that E_{ij} only depends on $\xi(i)$ and $\xi(j)$. Equivalently, if π is any permutation that reorders the nodes such that nodes of the same group are adjacent to each other, and P is the corresponding permutation matrix, then $P^T AP$ is approximately constant in the r blocks induced by the assignment. One can then associate with it a so-called *ideal* adjacency matrix A_{id} as illustrated in Fig. 1: for the blocks of A where 1 dominates, put all elements of the corresponding block in A_{id} equal to 1 and for the blocks where 0 dominates, put them all equal to 0 in A_{id}. In doing so, the nodes in each block of A_{id} are *regularly equivalent* [61], i.e., they have the same parents and the same children [57,58]. The above approximation problem for $P^T AP$ can thus be viewed as finding a nearby regularly equivalent graph to a given graph.

The role extraction problem can be also generalized into finding clusters of nodes $\mathcal{C}_1, \ldots, \mathcal{C}_q$, or equivalently an assignment function, such that for a given node x in cluster $\mathcal{C}_a$, the number of edges between x and the cluster C_b depends only on a and b. In the next section we will find that this problem is better formalized by the SBM.

B. Stochastic block model

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and consider the space of random variables $\Omega \to \{0,1\}^{a\times b}$. A random digraph $G(\omega)$, $\omega \in \Omega$, is a graph whose adjacency matrix $A(\omega)$ is one such

random variable. We denote the expectation of a random matrix A by $\mathbb{E}[A]$. We construct a random unweighted digraph as follows:

(1) The nodes are partitioned in q clusters of nodes, $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_q$, of size $m_1 n, m_2 n, \ldots, m_q n$ respectively.

(2) There is an edge between a node in cluster $\mathcal{C}_a$ to a node in cluster $\mathcal{C}_b$ with probability $p_{a,b} = f(n)\theta_{a,b}$, where $\theta_{a,b}$ depend only on a and b and $\max_{a,b} \theta_{a,b} = 1$.

Since $p_{a,b}$ are probabilities, necessarily $f(n) = O(1)$, but from classical information theory we know that exact recovery for the clusters requires [62,63]

$$nf(n) \to \infty, \tag{1}$$

and it is also more restrictive than the sufficient condition for clustering detection [64].

The adjacency matrix A_n of such a random graph is an $mn \times mn$ random matrix, where $m = \sum_{i=1}^{q} m_i$. Suppose that m_i may vary with n, but for every n, i we have $0 < m_{\min} \leqslant m_i \leqslant m_{\max}$ where $m_{\min}, m_{\max}, q, \theta_{a,b}$ do not depend on n. As a consequence m may vary, but it is always bounded between absolute constants $qm_{\min} \leqslant m \leqslant qm_{\max}$. Suppose $i \in \mathcal{C}_a$, $j \in \mathcal{C}_b$: then, A_{ij} is distributed as a Bernoulli variable centered on $\{0, 1\}$ with $\mathbb{P}(1) = p_{ab}$. In this section, we assume that the nodes of the same cluster are adjacent to each other, in order to simplify the notation. This does not affect the generality of our results.

Denoting by $\mathbf{1}_k \in \mathbb{R}^k$ the vector of all ones and by $M_n = \mathbb{E}[A_n]$, then

$$M_n = f(n) Z_n \Upsilon Z_n^T,$$

where

$$Z_n = \bigoplus_{i=1}^{q} \mathbf{1}_{m_i n} \in \mathbb{R}^{mn \times q},$$

$$\Upsilon \in \mathbb{R}^{q \times q}, \quad \Upsilon_{ab} = \theta_{ab} \; \forall \; 1 \leqslant a, b \leqslant q.$$

M_n is a deterministic matrix with precisely $s := \operatorname{rank}(f(n)\Upsilon) \leqslant q$ nonzero singular values: if $D = \operatorname{diag}(\sqrt{m_i}, \ldots, \sqrt{m_q})$, then $\widetilde{Z}_n := Z_n D^{-1}/\sqrt{n}$ has orthogonal columns and the nonzero singular values of M_n are those of $nf(n)D\Upsilon D$. We have in particular that

$$m_{\min} \leqslant \frac{\sigma_i(M_n)}{nf(n)\sigma_i(\Upsilon)} \leqslant m_{\max}, \quad i = 1, \ldots, s,$$
$$\sigma_i(M_n) = 0, \quad \forall \; i > s. \tag{2}$$

Analogously, $[M_n \; M_n^T]$ has precisely $r := \operatorname{rank}(f(n)[\Upsilon \; \Upsilon^T]) \leqslant q$ nonzero singular values with

$$m_{\min} \leqslant \frac{\sigma_i([M_n \; M_n^T])}{nf(n)\sigma_i([\Upsilon \; \Upsilon^T])} \leqslant m_{\max}, \quad i = 1, \ldots, r,$$
$$\sigma_i(M_n) = 0, \quad \forall \; i > r. \tag{3}$$

The above scenario is what arises in SBM theory, but in most references the matrix Υ is taken symmetric. In the following sections we will analyze the model described above, together with a spectral method designed to extract the clusters, which will be called roles, through the use of a similarity matrix S. For this reason, we report here a result we will need in our arguments about the matrix $Y_n := A_n - M_n$.

Theorem II.1. [[65], Remark 5.19] [[66], Corollary 2.3.5] Let E_N be $N \times N$ random matrices with independent, mean zero, and uniformly bounded entries. Suppose that σ^2 bounds the second moments of all entries, independently on N. In this case,

$$\limsup_{N \to \infty} \left\| \frac{1}{\sqrt{N}} E_N \right\| \leqslant 2\sigma$$

almost surely.

Since the entries of $E_n = Y_n/\sqrt{f(n)}$ have variance bounded by $\max_{i,j} \Upsilon_{i,j} = 1$, we get that

$$\|Y_n\|^2 \leqslant \delta^2 := 4mnf(n) \tag{4}$$

when n is big enough. In what follows, we will bound the norm of $[Y_n \; Y_n^T]$ with $\sqrt{2}\delta = 2\sqrt{2}\sqrt{mnf(n)}$, but the constant $2\sqrt{2}$ here is not sharp. In fact, both Theorem 4.1 in [67] and the experiments we will present suggest that the result holds with the tighter constant $1 + \sqrt{2}$, following the classical bound on the Marchenko-Pastur distribution. Since there is no such result in literature, we formulate it here as a conjecture.

Conjecture II.1. Let Z_N be $N \times N$ random matrices with independent, mean zero, and uniformly bounded entries. Suppose that σ^2 bounds the second moments of all entries, independently on N. If we call $X_N := [Z_N \; Z_N^T]$, then

$$\limsup_{N \to \infty} \left\| \frac{1}{\sqrt{2N}} X_N \right\| \leqslant \left(1 + \sqrt{\frac{1}{2}}\right)\sigma$$

almost surely. Moreover, if every entry has the same second moment σ^2, the bound is attained.

From now on, when we say "for any n big enough" or a similar formulation, we always implicitly mean that the result holds almost surely.

C. Role extraction via the similarity matrix S

In [57,58] it was proposed to solve the problem of role extraction for a digraph with adjacency matrix A by means of a Neighborhood Pattern Similarity matrix S, which is defined as the limit of the sequence of SPD matrices $(S_k)_{k \in \mathbb{N}}$ with

$$S_0 = 0, \quad S_{k+1} = \Gamma_A[I + \beta^2 S_k] = S_1 + \beta^2 \Gamma_A[S_k], \tag{5}$$

where the operator (depending on the matrix parameter W) Γ_W is defined as

$$\Gamma_W[X] = WXW^T + W^T XW.$$

It was shown in [57,58] that the sequence is convergent if and only if $\beta^2 < \rho(A \otimes A^T + A^T \otimes A)^{-1}$, and that the limit S satisfies $S = S_1 + \beta^2 \Gamma_A(S)$ or, equivalently,

$$\operatorname{vec} S = (I - \beta^2 A \otimes A^T - \beta^2 A^T \otimes A)^{-1} \operatorname{vec}(AA^T + A^T A).$$

It was also shown there that element (i, j) of the matrix S_k is the weighted sum of the walks of length up to k between nodes i and j, and that this can exploited to find nodes that should be associated with the same role. Note that S_1 is a known symmetrization for direct graphs called "Bibliometric Symmetrization" [17].

Throughout this paper, the parameter β in (5) is always assumed to satisfy $\beta^2 \|\Gamma_A\| < 1$, which is sufficient for the sequence $(S_k)_{k \in \mathbb{N}}$ to converge. Here and thereafter, we measure

the norm of the operator Γ_W induced by the spectral norm of its matrix argument. More concretely,

$$\|\Gamma_W\| = \sup_{X\neq 0} \frac{\|\Gamma_W[X]\|}{\|X\|}.$$

Lemma II.1. The norm of the linear matrix mapping $\Gamma_W : X \mapsto \Gamma_W[X]$ satisfies

$$\|\Gamma_W\| = \|[W \quad W^T]\|^2 \leqslant 2\|W\|^2.$$

From the previous result, whose proof can be found in the Appendixes, $\|\Gamma_W\| \leqslant 2\|W\|^2$, so that it is easy to compute a good enough β with very low computational effort. In fact, we can always choose, for example, $\beta^2 = 1/(4\|A\|^2)$ and obtain that necessarily $\beta^2\|\Gamma_A\| \leqslant 1/2$.

Consider the SBM described in Sec. II B. From now on, we drop for the sake of notational simplicity the suffixes emphasizing the dependence on the size n, so, for example, we simply write A, M, Z, Y for A_n, M_n, Z_n, Y_n. Given the random adjacency matrix A, suppose that $f(n)\Upsilon$ is a minimal role matrix, defined as follows.

Definition II.1. A square matrix B is a **minimal role** matrix if no two rows of the compound matrix $[B \quad B^T]$ are linearly dependent.

The matrix $M = \mathbb{E}[A] = f(n)Z\Upsilon Z^T$ is a deterministic block matrix, and the following result shows that it is possible to recover the original clusters by analyzing any of the matrices T_k generated as the S_k but replacing A with M.

Theorem II.2. [[58], Theorem 3.4] Let M, Υ be as in Sec. II B with minimal role matrix $f(n)\Upsilon$. If T_k is generated by the recurrence

$$T_0 = 0, \quad T_{k+1} = \Gamma_M[I + \beta^2 T_k] = T_1 + \beta^2\Gamma_M[T_k],$$

then it has rank $r = \operatorname{rank}(f(n)[\Upsilon \ \Upsilon^T]) \leqslant q$ and $T_k = Z\widehat{T}_k Z^T$ where $\widehat{T}_k$ is a SPD $q \times q$ matrix. Given $V_k \in \mathbb{R}^{mn\times r}$ the orthogonal matrix in the reduced SVD (or, equivalently, reduced eigendecomposition) of T_k, it follows that the set of the vectors of $\mathbb{R}^r$ that are a row of V_k has precisely q distinct elements. Moreover, the q original clusters of the graph coincide with the partition of $[mn]$ into the q subsets associated with the row indices that correspond to each distinct vector that is a row of V_k.

As a consequence, it is enough to perform an eigendecomposition of T_k, extract the reduced orthogonal matrix V_k and then identify the repeated q rows to recover the clustering. A natural thought is to try and apply the same method to the random symmetric matrix S_k generated by the recurrence (5), but some issues arise:

(1) T_k has rank $r \leqslant q$, while S_k is with high probability full rank, so we need a way to determine the truncation parameter r for the SVD.

(2) In the truncated eigendecomposition of S_k, the orthogonal matrix U_k has usually distinct rows. In order to retrieve the clusters, we thus need to estimate the number of roles q and perform a K-means algorithm on the rows.

There is method to do this, commonly referred to as Spectral Clustering of the matrix S_k. A detailed description is given in [58] and a concise description as pseudocode is given below:

Algorithm II.1.

Inputs: adjacency matrix A, number of roles q, scaling factor β, integer k.

Output: a partitioning of the nodes of the graph into q clusters.

Procedure:

(1) Compute the matrix X_1 whose columns are the q dominant singular vectors of $[A \quad A^T]$;

(2) For $h = 2, \ldots, k$ compute the matrix X_h whose columns are the q dominant singular vectors of $Y_h = [\beta AX_{h-1} \quad \beta A^T X_{h-1} \quad X_1]$;

(3) Apply the K-means algorithm to the rows of the matrix X_k.

The sparse singular value decomposition of $[A, A^T]$ can be computed using the Lanczos bidiagonalization procedure [68] and its complexity is $O(\mu q^2)$ because each matrix vector multiplication requires exactly 2μ flops, where μ is the number of edges in the graph, i.e., the number of nonzero entries of the $mn \times mn$ matrix A. For the same reason, the construction of the matrix Y_h requires exactly $2(\mu + mn)q$ flops. The singular value decomposition of the economy size singular value decomposition of the dense $mn \times 3q$ matrix Y_h, requires $O(mnq^2) + O(q^3)$ flops [69]. Altogether, we thus have a complexity of $O(kq[\mu + mn + mnq + q^2])$ to compute the low rank factor X_k, which scales well with μ. The subsequent clustering of the rows of X_k is then constrained to a q-dimensional space, and requires on average $O(mnq^2)$ flops per iteration of the K-means algorithm [70].

In the next sections, we show that the matrices S_k sport a clear gap between the eigenvalues $\lambda_r(S_k)$ and $\lambda_{r+1}(S_k)$ that lets us identify the rank r with high probability for big n. Moreover, when the matrix $f(n)[\Upsilon \ \Upsilon^T]$ is full-rank, so that $f(n)\Upsilon$ is minimal and $r = q$, we estimate the clustering relative error for the K-means algorithm on S_k, and show that it is proportional to $[nf(n)]^{-1}$.

III. SPECTRAL BOUNDS

We now consider the recurrence relation using the expected value M rather than A since this yields a good approximation for the S_k matrices. We denote these matrices as T_k and their recurrence is thus given by

$$T_0 = 0, \quad T_{k+1} = \Gamma_M[I_n + \beta^2 T_k], \ k \geqslant 0. \tag{6}$$

Note that in (6) the matrix parameter in the operator Γ is set to $M = \mathbb{E}[A]$, in contrast with (5) where it was set to A. Again, the parameter β^2 is chosen such that $\beta^2\|\Gamma_M\| < 1$, which is required for the sequence T_k to converge to $T = \lim_{k\to\infty} T_k$. In order to choose an appropriate β we need an estimation of $\|\Gamma_M\|$ depending only on the matrix A.

Lemma III.1. Let $\delta^2 = 4mnf(n)$. For n large enough, it holds

$$\|\Gamma_A - \Gamma_M\| \leqslant \delta^3\|[\Upsilon \ \Upsilon^T]\|/\sqrt{2} + 2\delta^2 \leqslant \|A\|^2.$$

Using the last result and Lemma II.1, we find that for n large enough, $\beta^2\|\Gamma_M\| \leqslant 3\beta^2\|A\|^2$ and $\beta^2\|\Gamma_A\| \leqslant 2\beta^2\|A\|^2$ so from now on, we always suppose that $\beta^2 \leqslant 1/6\|A\|^2$ and consequently

$$\gamma := \max\{\beta^2\|\Gamma_M\|, \beta^2\|\Gamma_A\|\} \leqslant \tfrac{1}{2}. \tag{7}$$

It was pointed out in [58] that the matrices S_k and T_k are all positive semidefinite, and that both sequences are ordered in the Loewner ordering:

$$0 = S_0 \preceq S_1 \preceq \cdots \preceq S, \quad 0 = T_0 \preceq T_1 \preceq \cdots \preceq T. \tag{8}$$

Moreover, as shown in Theorem II.2, if T_k were available then we would be able to recover exactly the original clustering that generated the random directed graph. Since we can only work on S_k, that is an approximation of T_k, it is essential to analyze the proximity between the two matrices more accurately. This will let us study how well the properties of T_k transfer to S_k and how effective is a spectral clustering algorithm applied to S_k.

Theorem III.1. For $k \geqslant 1$ it holds

$$\begin{aligned}\|S_k - T_k\| &\leqslant \|\Gamma_A - \Gamma_M\| \left(\sum_{i=0}^{k-1} \|\beta^2 \Gamma_A\|^i\right)\left(\sum_{i=0}^{k-1} \|\beta^2 \Gamma_M\|^i\right) \\ &\leqslant 4\|\Gamma_A - \Gamma_M\|,\end{aligned}$$

where the last inequality holds also for $\|S - T\|$.

A. Spectral gap

From Theorem II.2 we know that T_k has rank r, so it stands with reason to expect that its approximation S_k has r dominant singular values (which we refer to as the "signal") and $mn - r$ small singular values (which we refer to as the "noise"). Here we report estimations for the eigenvalues of S_k and T_k and then derive bounds on the respective gaps.

Lemma III.2. It holds

$$\lambda_r(T) \geqslant \lambda_r(T_k) \geqslant \left[\frac{\sigma_r([\Upsilon \ \Upsilon^T])}{4q} \frac{m_{\min}}{m_{\max}}\right]^2 \delta^4$$

for every $k \geqslant 1$ and n big enough.

Proof. Easy corollary of (8) and (3). □

Theorem III.2. It holds

$$\begin{aligned}\tfrac{1}{2}(1-\gamma^k)\|[\Upsilon \ \Upsilon^T]\|^2\delta^4 \geqslant \|S_k\| \geqslant \lambda_r(S_k) \geqslant \lambda_r(T_k)/2, \\ \tfrac{1}{2}\|[\Upsilon \ \Upsilon^T]\|^2\delta^4 \geqslant \|S\| \geqslant \lambda_r(S) \geqslant \lambda_r(T)/2, \\ 4(1-\gamma^k)\delta^2 \geqslant \lambda_{r+1}(S_k), \quad 4\delta^2 \geqslant \lambda_{r+1}(S)\end{aligned}$$

for every $k \geqslant 1$ and n big enough.

The gaps $\lambda_r(S_k) - \lambda_{r+1}(S_k)$ and $\lambda_r(S_k)/\lambda_{r+1}(S_k)$ between signal and noise, are expected to be large enough to allow for the correct truncation for the SVD of S_k, and a correct assignment of the different nodes in each "role," as we will show in the next section. This separation becomes more pronounced when the dimensions of the matrix and its subgroups increase, as we can see by applying Lemma III.2 and Theorem III.2:

(1) For the absolute gap,

$$\lambda_r(S_k) - \lambda_{r+1}(S_k) \geqslant \frac{\lambda_r(T_k)}{2} - 4(1-\gamma^k)\delta^2 = \Omega(\delta^4)$$

that is order of magnitudes greater than the following absolute gaps, since

$$\lambda_i(S_k) - \lambda_{i+1}(S_k) \leqslant \lambda_{r+1}(S_k) = O(\delta^2), \quad \forall i > r.$$

(2) For the relative gap,

$$\frac{\lambda_r(S_k)}{\lambda_{r+1}(S_k)} \geqslant \frac{\lambda_r(T_k)}{8(1-\gamma^k)\delta^2} = \Omega(\delta^2)$$

that is order of magnitudes greater than the previous relative gaps, since

$$\frac{\lambda_i(S_k)}{\lambda_{i+1}(S_k)} \leqslant \frac{\|S_k\|}{\lambda_r(S_k)} = O(1), \quad \forall i < r.$$

As a consequence, a comparison of the gaps between signal and noise with the other gaps is a clear indicator of the right rank r with which one should perform the truncated SVD in the algorithm. This holds also for the limit matrix S.

We can note that all the estimates get worse as $\sigma_r([\Upsilon \ \Upsilon^T])$ gets close to zero. This has to be expected since it is harder to compute the rank of an almost singular matrix. In fact, for example, in the case where all the probabilities $\theta_{a,b}$ are close to each other, it is harder to distinguish between different groups and clusters.

B. Dominant subspaces

In this subsection, we study the dominant subspace of a real symmetric matrix S_k (i.e., the invariant subspace associated with the largest r eigenvalues) and argue why, for sufficiently large n, it allows role extraction. Classically, distances between subspaces are measured via the concept of principal angles [71]: a multidimensional generalization of the acute angle between the unit vectors u, v, i.e., $0 \leqslant \theta(u,v) := \arccos|u^T v| \leqslant \pi/2$. More generally, if $\mathcal{U}, \mathcal{V}$ are subspaces whose orthonormal bases are given, respectively, as the columns of the matrices U, V, then the $\min\{\dim \mathcal{U}, \dim \mathcal{V}\}$ largest singular values of $U^T V$ are the sines of the principal angles between $\mathcal{U}^\perp$ (orthogonal complement of $\mathcal{U}$) and $\mathcal{V}$. Just as in the one-dimensional case, the principal angles between two subspaces are all zero if and only if the subspaces coincide, and more generally the smaller the principal angles the closer the subspaces.

To set up notation, fix $k \in \mathbb{N}$, and let E and F be the dominant subspaces of dimension r for S_k and T_k respectively. By classical results in geometry and linear algebra [72,73], the r largest singular values of the matrix

$$\sin\Theta := \Pi_E - \Pi_F \tag{9}$$

are the sines of the principal angles between the dominant subspaces of T_k and that of S_k, where Π_E and Π_F are the orthogonal projection matrices on the relative subspaces. Hence, the spectral norm of $\sin\Theta$ measures how well the dominant subspace of the similarity matrix S_k approximates the one of the ideal graph.

We rely on Davis-Kahan's $\sin(\Theta)$ theorem [72], in the form given by [[73], Theorem 5.3]. Call $\widehat{S}_k$ the best r-rank approximation of S_k. Since the rth eigenvalue of T_k is larger than the $(r+1)$-th eigenvalue of $\widehat{S}_k$ (which is 0), the assumptions of [[73], Theorem 5.3] apply and thus

$$\begin{aligned}\|\sin\Theta\| &\leqslant \frac{\|T_k - \widehat{S}_k\|}{\lambda_r(T_k)} \leqslant \frac{\|T_k - S_k\|}{\lambda_r(T_k)} + \frac{\|S_k - \widehat{S}_k\|}{\lambda_r(T_k)} \\ &\leqslant \frac{2\|T_k - S_k\|}{\lambda_r(T_k)}.\end{aligned}$$

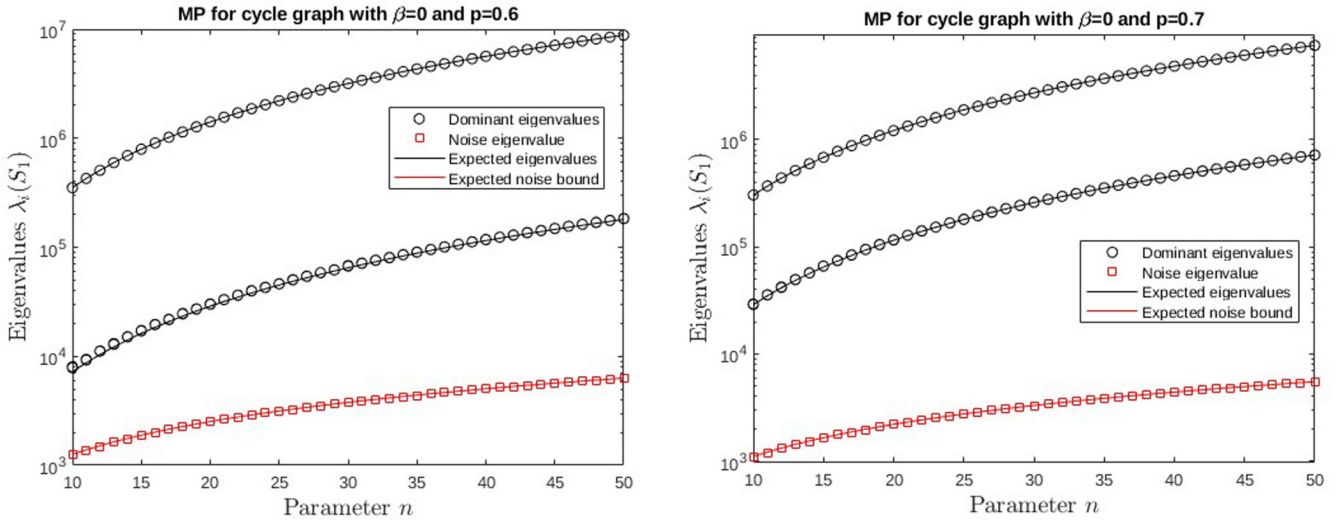

FIG. 2. Eigenvalues of S_1 and T_1 of a cycle graph for increasing n and varying probabilities. The matrix dimension is $30n$.

REMARK III.1 We could apply [[73], Theorem 5.3] reverting the roles of S_k and T_k, obtaining

$$\|\sin\Theta\| \leqslant \frac{\|S_k - T_k\|}{\lambda_r(S_k)}.$$

In this case, though, we prefer to deal with the deterministic quantity $\lambda_r(T_k)$ instead of the aleatory $\lambda_r(S_k)$, even if the estimation gets worse by a constant factor 2.

Using the results of the previous section, in turn this yields for sufficiently large n

$$\|\sin\Theta\| \leqslant \frac{4\sqrt{2}\delta^3\|[\Upsilon\ \Upsilon^T]\| + 16\delta^2}{\left[\frac{\sigma_r([\Upsilon\ \Upsilon^T])}{4q}\frac{m_{\min}}{m_{\max}}\right]^2\delta^4} = O(\delta^{-1}). \tag{10}$$

Therefore, we can state

Corollary III.1. Asymptotically as $n \to \infty$, the principal angles between the dominant subspaces of S_k and T_k tend to 0 at least as fast as δ^{-1}.

IV. CLUSTERING ERROR

In the previous sections we have estimated how close the matrix S_k is to the deterministic matrix T_k and how this influences their spectral properties and their dominant subspaces. Here we show that the same estimates can be used to bound the clustering error of the proposed method on S_k, under the technical hypothesis $r = q$ that is, the matrix $f(n)[\Upsilon\ \Upsilon^T]$ is full rank. Note that Υ is still allowed to be singular.

Recall that the model is generated by the clusters $\mathcal{C}_1, \ldots, \mathcal{C}_q$, where $\mathcal{C}_i$ has cardinality $n_i := m_i n$. Suppose that $\mathcal{T}_1, \ldots, \mathcal{T}_q$ are the resulting clusters from the algorithm operated on the similarity matrix S_k. Define the misclassification error $\widehat{f}$ as

$$\widehat{f} := \min_{\pi\in\mathcal{S}_q}\max_{i=1,\ldots,q}\frac{|\mathcal{T}_{\pi(i)} \triangle \mathcal{C}_i|}{|\mathcal{C}_i|},$$

where $\triangle$ is the symmetric difference of sets defined as the elements belonging to exactly one of the two sets, or equivalently $A \triangle B := (A \setminus B) \cup (B \setminus A)$. $\mathcal{S}_q$ is the qth symmetric

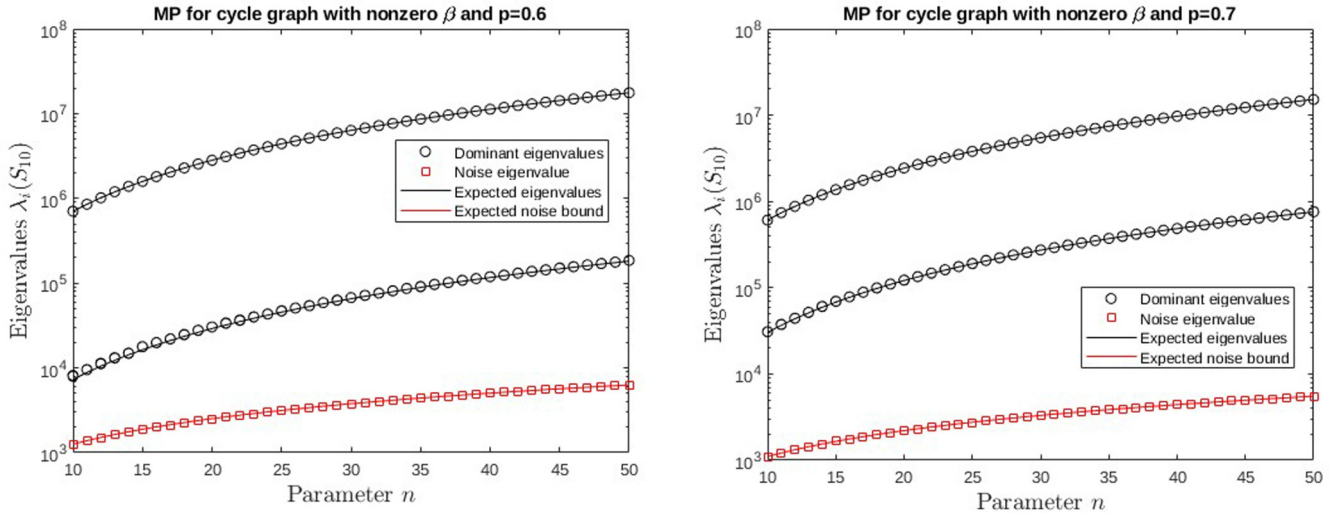

FIG. 3. Eigenvalues of S_{10} and T_{10} of a cycle graph for increasing n and varying probabilities. The matrix dimension is $30n$.

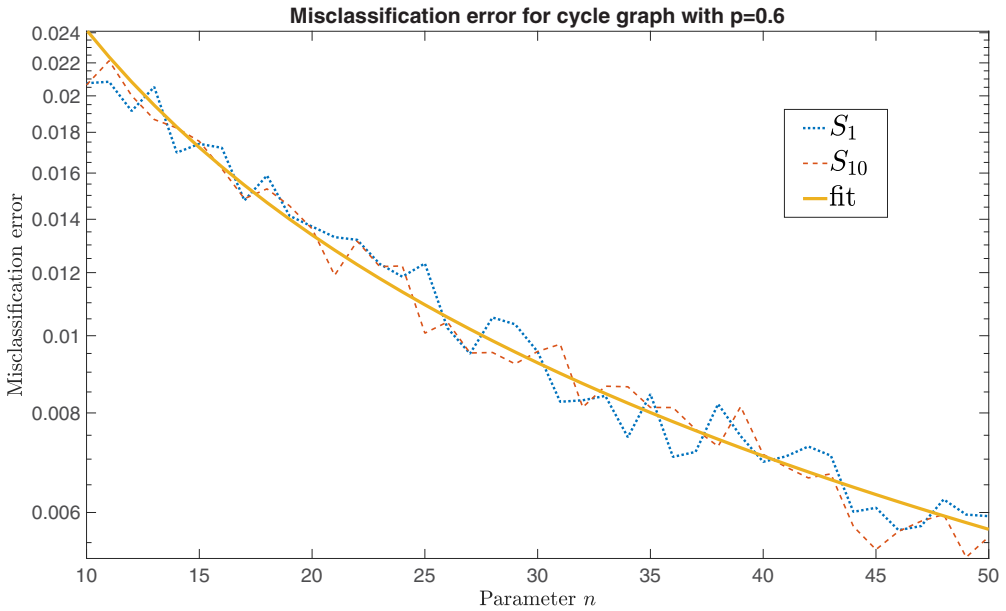

FIG. 4. Average misclassification error for the cycle graph relative to S_1 and S_{10}, and a fitted $O(1/n)$ function for comparison. The matrix dimension is $30n$.

group, that contains all the permutations on q elements. $\widehat{f}$ is thus a measure of the maximum rate of misclassified points over all clusters, up to the assignment of the correct clusters $\mathcal{C}_i$ to the $\mathcal{T}_i$ derived by the algorithm. In an Appendix, we give a proof for the following bound on $\widehat{f}$.

Theorem IV.1. There exists an absolute constant C such that asymptotically in n

$$\widehat{f} \leqslant Cq\frac{m_{\max}}{m_{\min}}\|\sin(\Theta)\|^2 \leqslant C\frac{q^5}{\delta^2}\frac{m_{\max}^5}{m_{\min}^5}\frac{\|[\Upsilon\ \Upsilon^T]\|^2}{\sigma_q([\Upsilon\ \Upsilon^T])^4} = O\left(\frac{1}{nf(n)}\right).$$

Note that the error goes asymptotically to zero as long as $nf(n) \to \infty$, which is exactly condition (1).

REMARK IV.1 The proof of the theorem follows the same steps as [60] and [74]. In particular, in the former we find a similar algorithm applied directly on the adjacency matrix A instead of S_k, but the analysis is limited to the case where $f(n)\Upsilon$ has full rank, while we work under the more general condition that $f(n)[\Upsilon\ \Upsilon^T]$ is full rank.

Observe moreover that all the results of Sec. III hold without any assumption on Υ, so we still have all the spectral bounds and the convergence of the dominant subspace of S_k to the one of T_k also in the general case.

Yet, for the algorithm to make sense, we need $f(n)\Upsilon$ to be a minimal role matrix as defined in Definition II.1. Moreover, since $r < q$, it is necessary to apply the K-means algorithm on S_k for $K = r, r+1, r+2, \ldots$ and look at the error in order to find the optimal number of clusters.

V. NUMERICAL EXAMPLES

In this section we illustrate the theoretical results of the paper using an example generated according to the rules of a SBM where only two different probabilities are used, namely p and $1-p$. We chose $q = 3$, $m_1 = m_2 = m_3 = 10$ and hence $m = 30$, and

$$\Upsilon = p\begin{bmatrix}0&1&0\\0&0&1\\1&0&0\end{bmatrix} + (1-p)\begin{bmatrix}1&0&1\\1&1&0\\0&1&1\end{bmatrix}. \quad (11)$$

We then ran simulations for matrices A with $n = 10{:}50$.

In Fig. 2 we took $\beta = 0$ which means that the sequences S_k and T_k are constant after one step, and hence that $S = S_1$ and $T = T_1$. The q dominant eigenvalues of S_1 are the circles in each plot (because of the structure of Υ there are two repeated ones). The full lines are their estimates obtained from the rank q matrix T_1, and clearly they are very accurate estimates as indicated in Theorem III.1. The squares correspond to the "noise" eigenvalue $\lambda_{r+1}(S_1)$ and the dashed line is its estimate $(3+\sqrt{8})p(1-p)mn$ according to Conjecture II.1 and $\lambda_{r+1}(S_1) = \sigma_{r+1}([A\ A^T])^2 \leqslant \|[Y\ Y^T]\|^2$. It is clear from these plots that this is also a very good estimate and that the ratio $\lambda_r(S_1)/\lambda_{r+1}(S_1)$ grows like $O(n)$. Moreover, the plots show that the gap $|\lambda_r(S_1) - \lambda_{r+1}(S_1)|$ shrinks with p getting closer to 0.5, which is expected since for $p = 0.5$ the rank of Υ drops to 1. This means that for p getting closer to 0.5, one has to require larger dimensions of the graph in order to recover an accurate enough grouping.

In Fig. 3 we performed the same experiment, but now with β chosen such that $\|\beta^2\Gamma_A(\cdot)\| \approx \frac{1}{2}$, which guarantees convergence of the method. In order to reduce the complexity of the method, we computed S_{10} and T_{10} rather than the limits S and T, since in 10 steps we should have reasonably good estimates of these limits. We can see from the plots that one has to wait for larger values of n to reach a sufficiently large gap $|\lambda_r(S_{10}) - \lambda_{r+1}(S_{10})|$ than for $|\lambda_r(S_1) - \lambda_{r+1}(S_1)|$ in Fig. 2.

In Fig. 4 we computed the misclassification error $\hat{f}$ of the clustering associated to the matrix Υ for $p = 0.6$. Using the same parameters m_i, n as before we show an averaged $\hat{f}$ over 60 000 instances for the clusters extracted from S_1 and

S_{10}, where we took $\beta = (2\|[A\ A^T]\|^2)^{-1}$. For comparison, we also plot the function $3/(10n+24)$ and note that it fits well both plots, thus confirming the bound $O(1/n)$ predicted by Theorem IV.1.

VI. CONCLUDING REMARKS

In this paper, we showed that the Neighborhood Pattern Similarity matrices S_k of a directed graph with adjacency matrix A have spectra that are well separated into two groups of eigenvalues, provided that the graph is sufficiently large and that it is generated according to the SBM with blocks where all elements in each block follow a Bernoulli distribution with the same probability $O(f(n))$ where $nf(n) \to \infty$.

The large eigenvalues are then associated with the nonzero eigenvalues of the expected value $\mathbb{E}[A]$, which is a low rank matrix, and the small eigenvalues are associated with the mean and variance of the random distribution used in the SBM. Moreover, the gap between the "large" eigenvalues and the "small" ones grows with n. It then follows that the recovery of the nodes grouping of the SBM, can be based on the dominant eigenspace of the matrices S_k.

This analysis was primarily based on the recurrences defining the matrices S and T and on the fact that the underlying adjacency matrix is generated according to a SBM. It is likely that our results can be extended for other types of distributions and that weighted graphs can also be dealt with, but our analysis here was limited to unweighted adjacency matrices for directed graphs.

We point out that the same analysis could in principle be conducted in the sparse limit case $f(n) = O(1/n)$, but since most of the results are formulated asymptotically in $nf(n)$ one has to explicitly compute all the implicit multiplicative constants. A technical work of refinement is also needed on each proposition to obtain the best constants and thus meaningful results. For these reasons, we postpone the limit sparse case analysis to future work.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for useful comments. G.B. and V.N. are supported by an Academy of Finland grant (Suomen Akatemian päätös 331240). G.B. thanks the Alfred Kordelinin säätiö for the financial support under Grant No. 210122. P.V.D. is supported by an Aalto Science Institute Visitor Programme.

APPENDIX A: OPERATOR Γ

1. Proof of Lemma II.1

Note that for all X

$$\|\Gamma_W[X]\| = \left\|[W \quad W^T]\begin{bmatrix}X & 0\\ 0 & X\end{bmatrix}\begin{bmatrix}W^T\\ W\end{bmatrix}\right\| \leqslant \|[W \quad W^T]\|^2\|X\|$$

and

$$\|[W \quad W^T]\|^2 = \left\|[W \quad W^T]\begin{bmatrix}W^T\\ W\end{bmatrix}\right\| = \|WW^T + W^TW\| \leqslant 2\|W\|^2.$$

Thus we have $\|\Gamma_W\| \leqslant [W \quad W^T]\|^2 \leqslant 2\|W\|^2$, whose first bound is satisfied for $X = I$ since

$$\|\Gamma_W[I]\|/\|I\| = \|WW^T + W^TW\| = \|[W \quad W^T]\|^2.$$

2. Proof of Lemma III.1

For any matrix X, if we rewrite $\Gamma_A[X] - \Gamma_M[X]$ as

$$[Y \quad Y^T]\begin{bmatrix}X & 0\\ 0 & X\end{bmatrix}\begin{bmatrix}M^T\\ M\end{bmatrix} + [M \quad M^T]\begin{bmatrix}X & 0\\ 0 & X\end{bmatrix}\begin{bmatrix}Y^T\\ Y\end{bmatrix} + [Y \quad Y^T]\begin{bmatrix}X & 0\\ 0 & X\end{bmatrix}\begin{bmatrix}Y^T\\ Y\end{bmatrix}$$

we readily see that

$$\|\Gamma_A - \Gamma_M\| = \sup_{X\neq 0}\frac{\|\Gamma_A[X] - \Gamma_M[X]\|}{\|X\|} \leqslant \|[Y \quad Y^T]\|^2 + 2\|[M \quad M^T]\|\|[Y \quad Y^T]\|.$$

Using (3) and (4), and recalling that $\delta^2 = 4mnf(n)$,

$$\|[Y \quad Y^T]\|^2 \leqslant 2\|Y\|^2 \leqslant 2\delta^2,$$

$$\|[M \quad M^T]\|^2 \leqslant \left(m_{\max}nf(n)\|[\Upsilon \ \Upsilon^T]\|\right)^2 \leqslant \delta^4\|[\Upsilon \ \Upsilon^T]\|^2/16.$$

Then the desired bound follows easily. As for the last bound, note that $\delta^2 \sim nf(n)$ and in virtue of (1), (2), and (4) we have

$$\|Y\| \leqslant \delta \ll \frac{\|\Upsilon\|}{4q}\frac{m_{\min}}{m_{\max}}\delta^2 \leqslant m_{\min}nf(n)\|\Upsilon\| \leqslant \|M\|.$$

If we call C the constant $\|\Upsilon\|m_{\min}/4qm_{\max}$, then

$$\|A\|^2 \geqslant (\|M\| - \|Y\|)^2 \geqslant C^2\delta^4 - 2C\delta^3 + \delta^2 \gg \delta^3\|[\Upsilon \ \Upsilon^T]\|/\sqrt{2} + 2\delta^2.$$

APPENDIX B: SPECTRAL BOUNDS

1. Proof of Theorem III.1

Denoting the increments $\Delta^S_{i+1} := S_{i+1} - S_i$ and $\Delta^T_{i+1} := T_{i+1} - T_i$ we obtain

$$S_{k+1} - T_{k+1} = \sum_{i=1}^{k+1}(\Delta^S_i - \Delta^T_i), \quad S_0 = T_0 = 0.$$

Observing that

$$\Delta^S_{k+1} = \beta^2\Gamma_A[\Delta^S_k], \quad \Delta^S_1 = \Gamma_A[I_n],$$
$$\Delta^T_{k+1} = \beta^2\Gamma_M[\Delta^T_k], \quad \Delta^T_1 = \Gamma_M[I_n],$$

and

$$\Delta^S_{k+1} - \Delta^T_{k+1} = \beta^2\Gamma_A[\Delta^S_k - \Delta^T_k] + \beta^2\Gamma_A[\Delta^T_k] - \beta^2\Gamma_M[\Delta^T_k]$$

we can estimate

$$\begin{bmatrix}\|\Delta^S_{k+1} - \Delta^T_{k+1}\|\\ \|\Delta^T_{k+1}\|\end{bmatrix} \leqslant N\begin{bmatrix}\|\Delta^S_k - \Delta^T_k\|\\ \|\Delta^T_k\|\end{bmatrix},$$

where

$$N := \beta^2\begin{bmatrix}\|\Gamma_A\| & \|\Gamma_A - \Gamma_M\|\\ 0 & \|\Gamma_M\|\end{bmatrix}$$

with initial conditions $\|\Delta_1^S - \Delta_1^T\| = \|\Gamma_A[I_n] - \Gamma_M[I_n]\| \leqslant \|\Gamma_A - \Gamma_M\|$ and $\|\Delta_1^T\| \leqslant \|\Gamma_M\|$. Hence, by induction on $k \geqslant 1$, it is not difficult to obtain the upper bound

$$\begin{bmatrix} \|\Delta_{k+1}^S - \Delta_{k+1}^T\| \\ \|\Delta_{k+1}^T\| \end{bmatrix} \leqslant N^k \begin{bmatrix} \|\Gamma_A - \Gamma_M\| \\ \|\Gamma_M\| \end{bmatrix} \leqslant \beta^{2k} \begin{bmatrix} \|\Gamma_A - \Gamma_M\| \sum_{i=0}^k \|\Gamma_A\|^i \|\Gamma_M\|^{k-i} \\ \|\Gamma_M\|^{k+1} \end{bmatrix},$$

which finally yields the bound

$$\|S_k - T_k\| \leqslant \|\Gamma_A - \Gamma_M\| \left(\sum_{i=0}^{k-1} \|\beta^2 \Gamma_A\|^i \right) \left(\sum_{i=0}^{k-1} \|\beta^2 \Gamma_M\|^i \right).$$

In virtue of (7), we can let k go to ∞ and find that

$$\sum_{i=0}^{\infty} \|\beta^2 \Gamma_A\|^i = \frac{1}{1 - \beta^2 \|\Gamma_A\|} \leqslant 2,$$

and the same holds for M, thus the desired bound follows.

2. Proof of Theorem III.2

Note first that by (1), (3), and (4), for n big enough,

$$\begin{aligned} \|[A\ A^T]\| &\leqslant \|[M\ M^T]\| + \|[Y\ Y^T]\| \\ &\leqslant \tfrac{1}{4} \|[\Upsilon\ \Upsilon^T]\| \delta^2 + \sqrt{2}\delta \leqslant \tfrac{1}{2} \|[\Upsilon\ \Upsilon^T]\| \delta^2. \end{aligned}$$

This can be used to bound $\|S_k\|$ since by the recurrence (5), Lemma II.1, condition (7), and induction we find

$$\begin{aligned} \|S_k\| &\leqslant \|\Gamma_A\|(1 + \beta^2 \|S_{k-1}\|) \leqslant \|\Gamma_A\| \frac{1 - \gamma^k}{1 - \gamma} \\ &\leqslant \tfrac{1}{2}(1 - \gamma^k) \|[\Upsilon\ \Upsilon^T]\|^2 \delta^4, \end{aligned}$$

and if we let $k \to \infty$ then $\|S\| \leqslant \|[\Upsilon\ \Upsilon^T]\|^2 \delta^4/2$. Moreover,

$$S_k = \Gamma_A[I + \beta^2 S_{k-1}] \preceq (1 + \beta^2 \|S_{k-1}\|) \Gamma_A[I] \preceq \frac{1 - \gamma^k}{1 - \gamma} S_1,$$

and by Weyl's theorem

$$\begin{aligned} \lambda_{r+1}(S_k) &\leqslant \frac{1 - \gamma^k}{1 - \gamma} \lambda_{r+1}(S_1) = \frac{1 - \gamma^k}{1 - \gamma} \sigma_{r+1}([A\ A^T])^2 \\ &\leqslant \frac{1 - \gamma^k}{1 - \gamma} \|[Y\ Y^T]\|^2 \leqslant 4(1 - \gamma^k)\delta^2, \end{aligned}$$

where again if we let $k \to \infty$ then $\lambda_{r+1}(S) \leqslant 4\delta^2$. Last, using (8) and Weyl's theorem,

$$\lambda_r(S) \geqslant \lambda_r(S_k) \geqslant \lambda_r(T_k) - \|S_k - T_k\|.$$

By Lemma III.1, Theorem III.1, and Lemma III.2, we know that $\lambda_r(T_k) = \Omega(\delta^4) \gg O(\delta^3) = \|S_k - T_k\|$ so for n big enough, $\lambda_r(S_k) \geqslant \lambda_r(T_k)/2$ and the same holds for $k \to \infty$.

APPENDIX C: CLUSTERING ERROR

Lemma C.1. [[75], Lemma 8, Appendix C] Let E, F be $a \times b$ matrices with orthonormal columns, and let Π_E, Π_F be the orthogonal projections on their respective ranges. Then there exists an orthogonal $b \times b$ matrix Q such that

$$\|E - FQ\|_F \leqslant \|\Pi_E - \Pi_F\|_F.$$

Theorem C.1. [[76], Theorem 1] Call $\mathbb{M}$ the set of $nm \times q$ matrices that have only q distinct rows. Let E, F be $mn \times q$ matrices, where $F \in \mathbb{M}$, whose rows $\mu_1, \ldots, \mu_q$ identify the clusters $\mathcal{C}_i$. Call $n_i := |\mathcal{C}_i|$ and

$$\Delta_i := \frac{1}{\sqrt{n_i}} \min\{\sqrt{k}\|E - F\|, \|E - F\|_F\}.$$

Suppose there exists $\rho \geqslant 100$ such that $\|\mu_i - \mu_j\| \geqslant \rho(\Delta_i + \Delta_j)$ for any $i \neq j$. Let $G \in \mathbb{M}$ be a 10-approximation of the K-means algorithm on E, that is,

$$\|E - G\|_F^2 \leqslant 10 \min_{N \in \mathbb{M}} \|E - N\|_F^2$$

and call $\nu_1, \ldots, \nu_k$ the rows of G. Partition the indices $1, \ldots, mn$ into q clusters $\mathcal{T}_i$ according to G and E as in $\mathcal{T}_r := \{i : \|E_{i,:} - \nu_r\| \leqslant \|E_{i,:} - \nu_s\|\ \forall s\}$. We have that there exists a permutation π and an absolute constant C such that $|\mathcal{C}_r \triangle \mathcal{T}_{\pi(r)}| \leqslant C n_r/\rho^2$ for every r.

1. Proof of Theorem IV.1

In order to analyze the method, we need first to better characterize the eigenvalue decomposition (EVD) of T_k. In fact, from Theorem II.1, we know that there exists a full-rank PSD matrix $\widehat{T}_k$ such that $T_k = Z\widehat{T}_k Z^T$. Recall now that $D = \mathrm{diag}(\sqrt{m_1}, \ldots, \sqrt{m_q})$, and that $\widetilde{Z} = ZD^{-1}/\sqrt{n}$ has orthonormal columns. If $D\widehat{T}_k D = U_k \Sigma_k U_k^T$ is its EVD, then

$$T_k = Z\widehat{T}_k Z^T = n\widetilde{Z} D \widehat{T}_k D \widetilde{Z}^T = n\widetilde{Z} U_k \Sigma_k U_k^T \widetilde{Z}^T,$$

where $\widetilde{W}_k := \widetilde{Z}U_k = ZD^{-1}U_k/\sqrt{n}$ has orthogonal columns, so that $\widetilde{W}_k n \Sigma_k \widetilde{W}_k^T$ is the q-reduced EVD of T_k. Note that $\widetilde{W}_k \in \mathbb{M}$ since its rows coincide with the ones of $D^{-1}U_k/\sqrt{n}$, that is a full rank $q \times q$ matrix. For the same reason, we have that $\widetilde{W}_k Q \in \mathbb{M}$ for every orthogonal $q \times q$ matrix Q. Moreover, the clustering induced by $\widetilde{W}_k$ and $\widetilde{W}_k Q$ are the same, and coincide with the original clustering $\mathcal{C}_1, \ldots, \mathcal{C}_q$. It follows that if W_k, the orthogonal matrix in the q-truncated SVD of S_k, is close to $\widetilde{W}_k Q$ for even one matrix Q, then it has good chance to generate a good clustering. Here we report two results formalizing the concept.

By Lemma C.1 and (9) there exists a $k \times k$ orthogonal matrix Q such that

$$\begin{aligned} \|\widetilde{W}_k Q - W_k\|_F &\leqslant \|\Pi_{\widetilde{W}_k} - \Pi_{W_k}\|_F \\ &\leqslant \sqrt{2q} \|\Pi_{\widetilde{W}_k} - \Pi_{W_k}\| = \sqrt{2q} \|\sin(\Theta)\|, \end{aligned}$$

where $\sin(\Theta)$ are the sines of the principal angles between the subspaces W_k and $\widetilde{W}_k$, thus

$$\begin{aligned} \Delta_i &= \frac{1}{\sqrt{m_i n}} \min\{\sqrt{q}\|W_k - \widetilde{W}_k Q\|, \|W_k - \widetilde{W}_k Q\|_F\} \\ &\leqslant \frac{\sqrt{2q}}{\sqrt{m_i n}} \|\sin(\Theta)\|. \end{aligned}$$

If we call $\mu_1, \ldots, \mu_q$ the distinct rows of $\widetilde{W}_k Q = ZD^{-1}U_k Q/\sqrt{n}$, then they are in the form $u_i/\sqrt{nm_i}$ where u_i

are the rows of $U_k Q$, that is an orthogonal matrix, so

$$\|\mu_i - \mu_j\|^2 = \left\| \frac{1}{\sqrt{nm_i}} u_i - \frac{1}{\sqrt{nm_j}} u_j \right\|^2 = \frac{1}{nm_i} + \frac{1}{nm_j} = \rho^2(\Delta_i + \Delta_j)^2,$$

where

$$\rho = \frac{\sqrt{\frac{1}{nm_i} + \frac{1}{nm_j}}}{\Delta_i + \Delta_j} \geqslant \frac{\sqrt{\frac{1}{m_i} + \frac{1}{m_j}}}{\frac{1}{\sqrt{m_i}} + \frac{1}{\sqrt{m_j}}} \frac{1}{\sqrt{2q}\|\sin(\Theta)\|} \geqslant \sqrt{\frac{m_{\min}}{m_{\max}}} \frac{1}{2\sqrt{q}\|\sin(\Theta)\|}.$$

By Corollary III.1, $\|\sin(\Theta)\| = O(\delta^{-1})$, so $\rho > 100$ for n big enough. The K-means algorithm applied to the matrix W_k outputs the clusters $\mathcal{T}_1, \ldots, \mathcal{T}_q$ and Theorem C.1 assures us that there is an absolute constant C for which

$$\widehat{f} = \min_{\pi \in \mathcal{S}_q} \max_{i=1,\ldots,q} \frac{|\mathcal{T}_{\pi(i)} \triangle \mathcal{C}_i|}{|\mathcal{C}_i|} \leqslant \frac{C}{\gamma^2} \leqslant 4Cq \frac{m_{\max}}{m_{\min}} \|\sin(\Theta)\|^2$$

We can finally conclude that by (10) and incorporating all the absolute constants into C,

$$\widehat{f} \leqslant C \frac{q^5}{\delta^2} \frac{m_{\max}^5}{m_{\min}^5} \frac{\|[\Upsilon \ \Upsilon^T]\|^2}{\sigma_q([\Upsilon \ \Upsilon^T])^4}.$$

[1] M. E. J. Newman and M. Girvan, Proc. Natl. Acad. Sci. U. S. A. **99**, 7821 (2002).
[2] S. Fortunato, Phys. Rep. **486**, 75 (2010).
[3] R. Guimerà, D. B. Stouffer, M. Sales-Pardo, M. Sales-Pardo, E. A. Leicht, M. E. J. Newman, and L. A. N. Amaral, Ecol. Soc. Am. **91**, 2941 (2010).
[4] M. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2010).
[5] M. A. Porter, J.-P. Onnela, and P. J. Mucha, Notices Am. Math. Soc. **56**, 1082 (2009).
[6] J. Kim, P. L. Krapivsky, B. Kahng, and S. Redner, Phys. Rev. E **66**, 055101(R) (2002).
[7] S. Pinkert, J. Reichardt, and J. Schultz, PLoS Comput. Biol. **6**, e1000659 (2010).
[8] M. Barahona, M. Beguerisse-Díaz, and B. Vangelov, in *2013 IEEE Global Conference on Signal and Information Processing* (IEEE, Piscataway, 2013), p. 937.
[9] V. Batagelj, P. Doreian, and A. Ferligoj, *Generalized Blockmodeling* (Cambridge University Press, Cambridge, 2004).
[10] J. Lei and A. Rinaldo, Ann. Stat. **43**, 215 (2015).
[11] J. Reichardt, *Structure in Complex Networks* (Springer, Berlin, 2009).
[12] J. Reichardt and D. R. White, Eur. Phys. J. B **60**, 217 (2007).
[13] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner, Internal Tech. Report 19, Faculty of Informatics, Universität Karlsruhe (2006).
[14] J. Duch and A. Arenas, Phys. Rev. E **72**, 027104 (2005).
[15] B.-Y. Jing, T. Li, N. Ying, and X. Yu, Stat. Sinica **32**, 1 (2022).
[16] X. Li, Y. Chen, and J. Xu, Stat. Sci. **36**, 2 (2021).
[17] S. Parthasarathy and V. Satuluri, in *Proceedings of the 14th International Conference on Extending Database Technology*, edited by A. Ailamaki, S. Amer-Yahia, J. Pate, T. Risch, P. Senellart, and J. Stoyanovich (Association for Computing Machinery, New York, NY, United States, 2011), p. 343.
[18] A. Lancichinetti and S. Fortunato, Phys. Rev. E **80**, 016118 (2009).
[19] P. W. Holland, K. B. Laskey, and S. Leinhardt, Soc. Netw. **5**, 109 (1983).
[20] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou, Ann. Stat. **46**, 2153 (2018).
[21] A. Y. Zhang and H. H. Zhou, Ann. Stat. **44**, 2252 (2016).
[22] T. Qin and K. Rohe, in *Advances in Neural Information Processing Systems*, edited by C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Advances in Neural Information Processing Systems 26 proceedings of the 27th Annual Conference on Neural Information Processing Systems, Vol. 26 (Curran Associates Inc., Red Hook, NY, United States, 2013), p. 3120.
[23] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[24] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech. (2008) P10008.
[25] P. J. Bickel and A. Chen, Proc. Natl. Acad. Sci. U. S. A. **106**, 21068 (2009).
[26] E. M. Airoldi, D. S. Choi, and P. J. Wolfe, Biometrika **99**, 273 (2012).
[27] A. A. Amini, P. J. Bickel, A. Chen, and E. Levina, Ann. Stat. **41**, 2097 (2013).
[28] A. Celisse, J.-J. Daudin, and L. Pierre, Electron. J. Stat. **6**, 1847 (2012).
[29] C. T. Bergstrom and M. Rosvall, Proc. Natl. Acad. Sci. U. S. A. **105**, 1118 (2008).
[30] M. Rosvall and C. T. Bergstrom, PLoS ONE **6**, e18209 (2011).
[31] T. Lane, C. Moore, J.-B. Rouquier, X. Yan, and Y. Zhu, in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, 2011), pp. 841–849.
[32] T. P. Peixoto, Phys. Rev. E **89**, 012804 (2014).
[33] A. Anandkumar, R. Ge, D. Hsu, and S. Kakade, in *Proceedings of the 26th Annual Conference on Learning Theory*, edited by S. Shalev-Shwartz and I. Steinwart, Vol. 30 (PMLR, Princeton, NJ, 2013), pp. 867–881.
[34] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Phys. Rev. E **84**, 066106 (2011).
[35] Y. Chen, S. Sanghavi, and H. Xu, in *Advances in Neural Information Processing Systems*, edited by F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Advances in Neural Information Processing Systems 25 proceedings of the 26th Annual Conference on Neural Information Processing Systems, Vol. 25 (Curran Associates, Red Hook, NY, United States, 2012).
[36] A. Coja-Oghlan, Comb. Probab. Comput. **19**, 227 (2010).
[37] K. Chaudhuri, F. Chung, and A. Tsiatas, in *Proceedings of the 25th Annual Conference on Learning Theory*, edited by S. Mannor, N. Srebro, and R. C. Williamson, JMLR Workshop and Conference Proceedings, Volume 23 Proceedings of the 25th

Annual Conference on Learning Theory (PMLR, Edinburgh, Scotland, 2012), pp. 35.1–35.23.
[38] E. A. Leicht and M. E. J. Newman, Proc. Natl. Acad. Sci. USA **104**, 9564 (2007).
[39] J. J. Ramasco and M. Mungan, Phys. Rev. E **77**, 036122 (2008).
[40] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2005).
[41] A. Athreya, V. Lyzinski, C. E. Priebe, D. L. Sussman, and M. Tang, Electron. J. Stat. **8**, 2905 (2014).
[42] T. P. Peixoto, Phys. Rev. Lett. **110**, 148701 (2013).
[43] T. P. Peixoto, Phys. Rev. X **4**, 011047 (2014).
[44] C. T. Bergstrom and M. Rosvall, Proc. Natl. Acad. Sci. USA **104**, 7327 (2007).
[45] E. Côme and P. Latouche, Stat. Model. **15**, 564 (2015).
[46] M. Mariadassou, S. Robin, and C. Vacher, Ann. Appl. Stat. **4**, 715 (2010).
[47] B. Karrer and M. E. J. Newman, Phys. Rev. E **83**, 016107 (2011).
[48] U. Von Luxburg, Stat. Comput. **17**, 395 (2007).
[49] S. Balakrishnan, A. Krishnamurthy, A. Singh, and M. Xu, in *Advances in Neural Information Processing Systems*, edited by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Advances in Neural Information Processing Systems 24 Proceedings of the 25th Annual Conference on Neural Information Processing Systems, Vol. 24 (Curran Associates, Red Hook, NY, United States, 2011).
[50] D. E. Fishkind, C. E. Priebe, D. L. Sussman, and M. Tang, J. Am. Stat. Assoc. **107**, 1119 (2012).
[51] D. E. Fishkind, C. E. Priebe, D. L. Sussman, M. Tang, and J. T. Vogelstein, SIAM J. Matrix Anal. Appl. **34**, 23 (2013).
[52] K. Rohe, S. Chatterjee, and B. Yu, Ann. Stat. **39**, 1878 (2011).
[53] J. Jin, Ann. Stat. **43**, 57 (2015).
[54] P. J. Bickel and P. Sarkar, Ann. Stat. **43**, 962 (2015).
[55] M. Jordan, A. Ng, and Y. Weiss, in *Advances in Neural Information Processing Systems*, 14 Proceedings of the 2001 Neural Information Processing Systems (NIPS) Conference, edited by T. Dietterich, S. Becker, and Z. Ghahramani, Vol. 14 (MIT Press, Cambridge, MA, United States, 2002), pp. 849–856.
[56] K. Nowicki and T. A. B. Snijders, J. Am. Stat. Assoc. **96**, 1077 (2001).
[57] A. Browet and P. Van Dooren, in *Proceedings of the 21st International Symposium on Mathematical Theory of Networks and Systems* (University of Groningen, Groningen, the Netherlands, 2014), p. 1412.
[58] M. Marchand, K. Gallivan, W. Huang, and P. Van Dooren, SIAM J. Math. Data Sci. **3**, 736 (2021).
[59] E. Abbe, J. Fan, K. Wang, and Y. Zhong, Ann. Stat. **48**, 1452 (2020).
[60] H. Qing and J. Wang, arXiv:2109.10319 (2021).
[61] S. P. Borgatti and M. G. Everett, J. Math. Sociol. **19**, 29 (1994).
[62] E. Abbe, A. S. Bandeira, and G. Hall, IEEE Trans. Inf. Theory **62**, 471 (2016).
[63] E. Mossel, J. Neeman, and A. Sly, in *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing (STOC'15)* (Association for Computing Machinery, New York, NY, United States, 2015), p. 69.
[64] L. Massoulié, in *Proceedings of the Forty-sixth Annual ACM Symposium on Theory of Computing (STOC'14)* (Association for Computing Machinery, New York, NY, United States, 2014), p. 694.
[65] Z. Bai and J. W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*, 2nd edition (Springer, New York, 2010), p. 552.
[66] T. Tao, *Topics in Random Matrix Theory* (American Mathematical Society. Providence, RI, 2012), p. 282.
[67] K. Hofmann-Credner and M. Stolz, Electron. Commun. Probab. **13**, 401 (2008).
[68] G. Golub and C. Van Loan, *Matrix Computations* (John Hopkins University Press, Baltimore, 1989), p. 642.
[69] D. Bau and N. Trefethen, *Numerical Linear Algebra* (IAM Publications, Philadelphia, 1997), p. 351.
[70] S. P. Lloyd, Least squares quantization in pcm, IEEE Trans. Inf. Theory **28**, 129 (1982).
[71] C. Jordan, Bull. Soc. Math. France **3**, 103 (1875).
[72] C. Davis and W. M. Kahan, SIAM J. Numer. Anal. **7**, 1 (1970).
[73] I. C. Ipsen, Linear Algebra Appl. **309**, 45 (2000).
[74] A. Joseph and B. Yu, Ann. Stat. **44**, 1765 (2016).
[75] A. A. Amini and Z. Zhou, J. Mach. Learn. Res. **20**, 1 (2019).
[76] O. Sheffet and P. Awasthi, in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, edited by A. Gupta, J. Rolim, K. Jansen, and R. Servedio (Springer, Berlin, 2012), p. 37.