Journal of Machine Learning Research 24 (2023) 1-58 Submitted 3/19; Revised 11/23; Published 11/23

Higher-Order Spectral Clustering Under Superimposed Stochastic Block Models

Subhadeep Paul PAUL.963@OSU.EDU
Department of Statistics
The Ohio State University
Columbus, OH 43210, USA

Olgica Milenkovic MILENKOV@ILLINOIS.EDU
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

Yuguo Chen YUGUO@ILLINOIS.EDU
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL 61820, USA

Editor: Jie Peng

Abstract

Higher-order motif structures and multi-vertex interactions are becoming increasingly important in studies of functionalities and evolution patterns of complex networks. To elucidate the role of higher-order structures in community detection over networks, we introduce a Superimposed Stochastic Block Model (SupSBM). The model is based on a random graph framework in which certain higher-order structures or subgraphs are generated through an independent hyperedge generation process and then replaced with graphs superimposed with edges generated by an inhomogeneous random graph model. Consequently, the model introduces dependencies between edges which allow for capturing more realistic network phenomena, namely strong local clustering in a sparse network, short average path length, and community structure. We then proceed to rigorously analyze the performance of a recently proposed higher-order spectral clustering method on the SupSBM. In particular, we prove non-asymptotic upper bounds on the misclustering error of higher-order spectral community detection for a SupSBM setting in which triangles are superimposed with undirected edges. We assess the model fit of the proposed model and compare it with existing random graph models in terms of observed properties of real network data obtained from diverse domains by sampling networks from the fitted models and a nonparametric network cross-validation approach.

Keywords: Higher-order structures, Hypergraphs, Network data, Spectral community detection, Superimposed random graph model

1. Introduction

Network data science has traditionally focused on studies capturing two-way interactions or connections between pairs of vertices or agents in networks. It has by now become apparent that many aspects of the relational organization, functionality, and the evolving structure of

a complex network can only be understood through higher-order subgraph (motif) interactions involving more than two vertices (Milo et al., 2002; Shen-Orr et al., 2002; Mangan and Alon, 2003; Honey et al., 2007; Alon, 2007; Porter et al., 2009; Benson et al., 2016; Yaveroğlu et al., 2014; Chen and Chen, 2018). Certain subgraphs in networks function as fundamental units of control and regulation of network communities and dynamics: for example, network motifs are crucial regulators in brain networks (Sporns and Kötter, 2004; Park and Friston, 2013; Battiston et al., 2017), transcriptional regulatory networks (Mangan and Alon, 2003), food webs (Paulau et al., 2015; Li and Milenkovic, 2017), social networks (Girvan and Newman, 2002; Snijders, 2001) and air traffic networks (Rosvall et al., 2014; Benson et al., 2016). Traditionally, statistical and algorithmic work on network motifs has been concerned with discovering and counting the frequency of over-expressed subgraphs (which are usually determined in comparison with some statistical null model) in various real-world networks (Alon, 2007; Klusowski and Wu, 2018). Indeed, frequency distributions or spectra of motifs have been shown to provide useful information about the regulatory and dynamic organization of networks obtained from disparate sources. Network motifs have also recently been used to perform learning tasks such as community detection (Benson et al., 2016; Li and Milenkovic, 2017; Tsourakakis et al., 2017). A parallel line of work has focused on identifying communities in hypergraphs and was reported in Zhou et al. (2006), Angelini et al. (2015), Kim et al. (2017), Ghoshdastidar and Dukkipati (2017), and Chien et al. (2018).

Simultaneously, over the last three decades of research on applications involving networks, it has been observed that many real-world networks display certain properties. These properties include low average path length, strong local clustering, highly heterogeneous vertex degree distribution, core-periphery or hub structure, and modular organization (Newman, 2003; Barabási and Albert, 1999; Watts and Strogatz, 1998). Local clustering refers to an overabundance of triangles and other relevant higher-order structures in an otherwise sparse network. The hub structure implies most of the communication among entities in the network is passed through a number of hubs or influential entities, while the modular organization means that the network is divided into a number of clusters of communities. Unfortunately, existing random graph models with community structures based on Erdös-Rényi (ER) random graphs (Erdös and Rényi, 1960), such as the Stochastic Block Models (SBMs) (Holland et al., 1983; Snijders and Nowicki, 1997; Bickel and Chen, 2009; Choi et al., 2012; Rohe et al., 2012; Celisse et al., 2012; Rohe et al., 2011; Qin and Rohe, 2013; Jin, 2015; Lei and Rinaldo, 2015; Decelle et al., 2011; Hajek et al., 2016; Abbe and Sandon, 2015; Gao et al., 2017), their degree-corrected versions (Karrer and Newman, 2011; Zhao et al., 2012), and other extensions fail to produce graphs with strong local clustering, i.e., with over-abundant triangles and other relevant higher-order structures. As Bollobás et al. (2011) pointed out, many real-world networks contain the number of edges and triangles roughly of the same asymptotic order in the number of nodes. A model with conditional independence in the edges cannot model such networks because it cannot produce the same density of higher-order structures as that of edges.

To address the aforementioned problem, a number of more realistic network models with some of the desired motif structures have been proposed in the literature. However, most such models are not mathematically tractable in general or in the context of community detection due to dependencies among the edges (Bollobás et al., 2011). Notable exceptions

include the mathematically tractable random graph model with local clustering and dependences among edges proposed in Bollobás et al. (2011). There, the authors constructed random graphs by superimposing small subgraphs and edges, thereby introducing dependencies among subsets of vertices. More specifically, they constructed an inhomogeneous random hypergraph with conditionally independent hyperedges and then replaced each hyperedge with a complete graph over the same set of vertices. A similar model, termed the Subgraph Generation Model (SUGM), was proposed in Chandrasekhar and Jackson (2014, 2016).

More recently, Hajek and Sankagiri (2018) analyzed a variation of the preferential attachment model with community structure and proposed a message-passing algorithm to recover the communities. In parallel, a geometric block model that uses Euclidean latent space geometric graphs instead of the usual Erdoš-Reńyi graphs for the mixture components was introduced in Galhotra et al. (2017, 2018). Although all these models capture some aspects of real-life networks and introduce controlled dependencies among the edges in the graphs, they fail to provide a general approach for combining dependent motif structures and analytical techniques that highlight if communities should be identified through pairwise or higher-order interactions.

Our contributions are two-fold. First, we propose a new Superimposed Stochastic Block Model (SupSBM), a random graph model for networks with community structure obtained by generalizing the framework of Chandrasekhar and Jackson (2014) and Bollobás et al. (2011) to account for communities akin to the classical SBM. SupSBM captures the most relevant aspects of the higher-order organization of the network, e.g., it incorporates triangles and other motifs, but couples them through edges that may be viewed as noise in the motif-based graphs. The community structure of interest may be present either at a higher-order structural level only or both at the level of higher-order structures and edges. Drawing parallels with the classical SBM, which is a mixture of Erdös-Rényi graphs, SupSBM may be viewed as a mixture of superimposed inhomogeneous random graphs generated according to the process described in Chandrasekhar and Jackson (2014) and Bollobás et al. (2011). We develop an estimation strategy where the communities are first estimated using a spectral clustering algorithm. The model parameters are later estimated using an approximate generalized method of moments. We show the proposed SupSBM fits the various aspects of real network data obtained from disparate application domains very well. For this purpose, we sample a large number of networks from the fitted SupSBM and three other competing models, and create bootstrap distributions of a number of network properties which are then compared with the observed value of the network property in question. Further, we also implement the network cross-validation approach in Li et al. (2020b) to select between the two low rank models, namely, SBM and SupSBM, in terms of performance in the task of predicting presence of edges and triangles for the datasets we consider.

Second, we derive theoretical performance guarantees for higher-order spectral clustering methods (Benson et al., 2016; Tsourakakis et al., 2017) applied to the SupSBM. The main difference between our analysis and previous lines of work on spectral algorithms for the SBM (Rohe et al., 2011; Lei and Rinaldo, 2015; Gao et al., 2017; Chin et al., 2015; Vu and Lei, 2013), and hypergraph SBM (Ghoshdastidar and Dukkipati, 2017; Kim et al., 2017; Chien et al., 2018) is that the elements of the analogs of adjacency matrices in our analysis are dependent and cannot be rewritten as sums of independent random variables.

We derive several non-asymptotic upper bounds of the spectral norms of such generalized adjacency matrices, and these results are of independent interest in other areas of network analysis. For this purpose, we notice that even though the terms in the sums are dependent, any given term is dependent only on a small fraction of other terms. We exploit this behavior to carefully control the effects of such dependence on the functions of interest. We use Chernoff-style concentration inequalities under limited dependence (Warnke, 2017) to complete our analysis. In addition, we derive corollaries implying performance guarantees for the non-uniform hypergraph SBM. The analysis of the non-uniform hypergraph SBM reveals interesting insights regarding the benefit of using ordinary versus higher-order spectral clustering methods on non-uniform hypergraphs.

Since the first online posting of the work, several new and related directions on the subject of clustering and community detection based on motifs and hypergraph partitioning were reported in Li et al. (2019b), Li et al. (2020a), Chien et al. (2020), Li et al. (2019a), and Underwood et al. (2020). These deal both with spectral clustering and correlation clustering models and adapt the methods to account for motifs such as triangles and special geometric structures. Nevertheless, none of the works use the concept of superimposed random graphs, nor do they perform a statistical analysis of the ultimate performance limits of community detection on superimposed random graph and motif models.

The remainder of the article is organized as follows. Section 2 defines superimposed random graph models and then develops the SupSBM. Section 3 describes the higher-order spectral clustering method, while Section 4 presents a non-asymptotic analysis of the misclustering rate of the method under the SupSBM. Section 5 presents methods for the estimation of model parameters and assessing model fit. Some real-world network examples are discussed in Section 6. The Appendix contains proofs of all the theorems and many auxiliary lemmas used in the derivations.

2. Superimposed random graph and block models

We start our analysis by defining what we refer to as an *inhomogeneous superimposed random graph model,* which is based on the random graph models described in Bollobás et al. (2011) and Chandrasekhar and Jackson (2014). We then proceed to introduce a natural extension of the SBM in which the community components are superimposed random graphs. Our main focus is on models that superimpose edges and triangles, as these are prevalent motifs in real social and biological networks (Alon, 2007; Benson et al., 2016; Li and Milenkovic, 2017; Laniado et al., 2016). However, as discussed in subsequent sections, the superimposed SBM can be easily extended to include other superimposed graph structures.

Formally, the proposed random graph model, denoted by $G_s(n, P^e, \mathbb{P}^t)$, is a superimposition of a classical dyadic (edge-based) random graph $G_e(n, P^e)$ and a triadic (triangle-based) random graph $G_t(n, \mathbb{P}^t)$. In this setting, n denotes the number of vertices in the graph, P^e denotes an $n \times n$ matrix whose (i, j)th entry equals the probability of an edge in G_e between the vertices i and j, and $\mathbb{P}^t$ denotes a 3-way (3rd order) $n \times n \times n$ tensor whose (i, j, k)th element equals the probability of a triangle involving the vertices (i, j, k) in G_t.

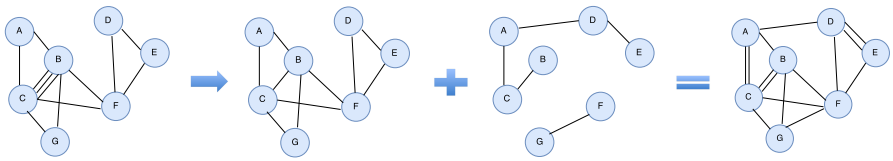
A random graph from the model $G_s(n, P^e, \mathbb{P}^t)$ is generated as follows. One starts with n unconnected vertices. The $G_t(n, \mathbb{P}^t)$ graph is generated by creating triangles ($3-$hyperedges)

for each of the $\binom{n}{3}$ 3-tuples of vertices (i, j, k) according to the outcome of *independent* Bernoulli random variables T_{ijk} with parameter $p^t_{ijk} = (\mathbb{P}^t)_{ijk}$. The hyperedges are consequently viewed as triangles in a graph, which results in a loss of their generative identity. Note that this process may lead to multi-edges between pairs of vertices i and j if these are involved in more than one triangle. The multi-edges in the graph G_t are collapsed into single edges such that there are no multi-edges in the graph. However, all pairs of vertices (i, j) still remain within all their constituent triangles as before the merging procedure. Next, the graph $G_e(n, P^e)$ is generated by placing edges between the $\binom{n}{2}$ pairs of vertices (i, j) according to the outcomes of independent Bernoulli random variables E_{ij} with parameter $p^e_{ij} = (P^e)_{ij}$. Note this is simply the usual inhomogeneous random graph model (Bollobás et al., 2007) that may be viewed as a generalization of the Erdös-Rényi model in which the probabilities of individual edges are allowed to be unequal. The two independently generated graphs are then superimposed to arrive at $G_s(n, P^e, \mathbb{P}^t)$.

The graph generation process is depicted by an example in Figure 1. Observe that the superimposed graph is allowed to contain multi-edges (or, more precisely, exactly two edges) between two vertices if and only if those vertices are involved in both at least one triangle in G_t and an edge in G_e. A practical justification for this choice of a multi-edge model comes from the fact that pair-wise and triple-wise affinities often provide complementary information. For example, Laniado et al. (2016) studied gender patterns in dyadic and triadic ties in an online social network and found different degrees of gender homophily in different types of ties. Hence instead of duplicating evidence from the same source, we retain two parallel edges in the graph only if they reinforce the information provided by each other. This way, we capture the *diversity* of interactions two nodes are involved in, but not the number of interactions of each type, which could be modeled as weights. Clearly, the resulting graph G_s has dependencies among its edges and strong local clustering properties for properly chosen matrices $\mathbb{P}^t$ due to the increased presence of triangles.

Furthermore, we would like to point out that this inhomogeneous superimposed random graph model differs in a number of important ways from non-uniform random hypergraph models on which the non-uniform hypergraph SBM, analyzed by Ghoshdastidar and Dukkipati (2017), Chien et al. (2018) and others, is based. First, our model captures networks in which we cannot differentiate between an "ordinary" edge and a hyperedge, as hyperedges simply appear as higher-order structures in the graph. In contrast, the non-uniform hypergraph SBM is a model for networks in which different types of hyperedges are distinguishable during the observation process and labeled. Hence, a major technical difficulty of analyzing methods under the SupSBM is dealing with edge dependencies that are not present in the non-uniform hypergraph SBM. Second, we collapse all multi-edges generated in the hyperedge generation process into single edges which are more realistic as observable network interaction models. We do, however, allow for double edges if there is complementary evidence of both dyadic and triadic ties.

In the simplest incarnation of the model, one may choose $(P^e)_{ij} = p^e$ for all i, j and $(\mathbb{P}^t)_{ijk} = p^t$ for all i, j, k. In this case, the graph G_e is a classical Erdös-Rényi dyadic random graph, while G_t before multi-edges collapsing may be thought of as a generalization of Erdös-Rényi graphs to the triadic setting. We refer to this model as Superimposed Erdös-



(a) (b) (c) (d)

Figure 1: (a) A realization of the graph G_t with $n = 7$ vertices, before multi-edge collapsing; (b) the collapsed graph G_t; (c) the dyadic graph G_e, and (d) the superimposed graph G_s.

Rényi (SupER) model. Note this model is identical to the model in Chandrasekhar and Jackson (2014) except for the collapsing of multi-edges in the graph G_t. The collapsing of multi-edges step leads to significantly fewer multi-edges in our model, making it more suitable for modeling network data. We describe next the SupSBM based on G_s graphs.

2.1 Superimposed stochastic block models

Our superimposed stochastic block model (SupSBM) is based on the inhomogeneous superimposed random graph framework defined in the previous section. We consider two types of SupSBMs. In the first case, "community signals" are present both in the higher-order structures and the dyadic edges, while in the second case, the "community signals" are present only in the higher-order structures but not in the dyadic edges. Drawing a parallel with the classical SBM, where intra- and inter-community edges are generated via Erdös-Rényi graphs, both the intra- and inter-community edges in SupSBM are generated by superimposed random graph models (G_s) as defined in the previous section.

We formally define a graph with n vertices and k communities generated from a SupSBM as follows. Each vertex of the graph is assigned a community label vector of length k, which takes the value of 1 at the position corresponding to its community and 0 at all other positions. To organize the labels, we define an $n \times k$ community assignment matrix C whose ith row C_i is the community label vector for the ith vertex. Given the community assignments for all the vertices in the graph, the triangle hyperedge indicators T_{ijk} involving three distinct vertices i, j, k are (conditionally) independent, and they follow a Bernoulli distribution with a parameter that depends only on the community assignments, i.e.,

$$P(T_{ijk} = 1 | C_{ip} = 1, C_{jq} = 1, C_{kl} = 1) = \pi^t_{pql}, \quad p, q, l \in \{1, \ldots, k\},$$

where π^t is a 3-way $k \times k \times k$ tensor of parameters. The triangle hyperedges naturally reduce to a triangle, and as before, multi-edges are collapsed to form the graph G_t.

An edge between two vertices i and j is generated independently of other edges and hyperedges following a Bernoulli distribution with a parameter that also depends on the community assignments so that the edge indicator variable E_{ij} satisfies

$$P(E_{ij} = 1 | C_{ip} = 1, C_{jq} = 1) = \pi^e_{pq}, \quad p, q \in \{1, \ldots, k\},$$

where π^e is a $k \times k$ matrix of model parameters. For the case that the community structure is present only in the higher-order structures and not at the level of dyadic edges, this

parameter equals p^e irrespective of the communities that the vertices i and j belong to. The desired graph is obtained by superimposing G_t and G_e following the process described in the previous section.

3. Estimation of community structure

The community assignments can be obtained using variants of the spectral clustering procedure. In particular, we can use the usual spectral clustering of the edge-based adjacency matrix (McSherry, 2001; Ng et al., 2002; Von Luxburg, 2007; Rohe et al., 2011; Lei and Rinaldo, 2015), the recently proposed higher order spectral clustering with triangle motif-adjacency matrix (Benson et al., 2016; Tsourakakis et al., 2017; Li and Milenkovic, 2017) or a spectral clustering on edge-triangle weighted adjacency matrix.

Spectral clustering methods that use network motifs or hyperedges, also known as higher-order spectral clustering methods, have been studied in a number of recent papers (Zhou et al., 2006; Benson et al., 2016; Tsourakakis et al., 2017; Li and Milenkovic, 2017). In particular, Benson et al. (2016) introduced a method that creates a "motif adjacency matrix" for each motif structure of interest. In a motif adjacency matrix, the (i, j)th element represents the number of motifs that include vertices i and j. Spectral clustering is applied to the motif adjacency matrix in a standard form in order to find communities of motifs.

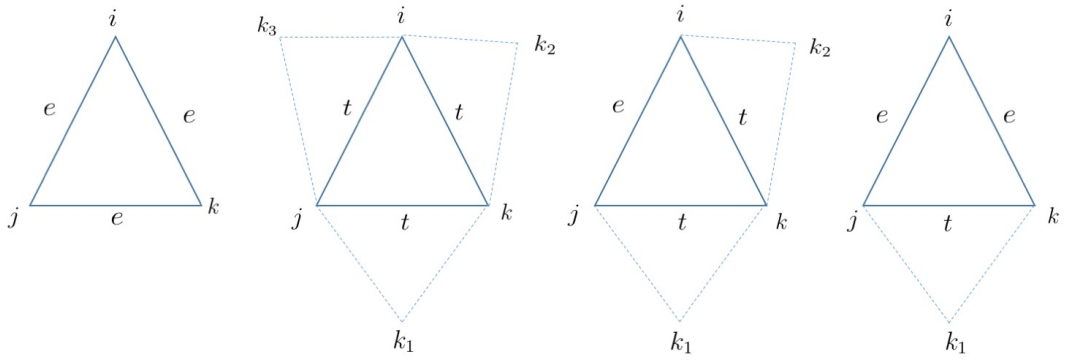
Given an observed network, we obtain two motif adjacency matrices involving edges (A_E) and triangles (A_T), such that $(A_E)_{ij}$ represents the number of observed edges between vertices i and j, while $(A_T)_{ij}$ represents the number of observed triangles including both i and j as vertices. The ordinary spectral clustering proceeds using the edge-based adjacency matrix A_E, while the higher-order spectral clustering uses the triangle motif adjacency matrix A_T. In both cases, the algorithm computes the k eigenvectors corresponding to the k largest (in absolute value) eigenvalues of the corresponding motif adjacency matrix. The algorithm subsequently performs the greedy clustering algorithm in Gao et al. (2017, Algorithm 2) on the rows of the $n \times k$ matrix of eigenvectors, which runs in polynomial time.

3.1 Motif adjacency matrices and superimposed random graphs

Let $G \sim G_s(n, P^e, \mathbb{P}^t)$ be a graph generated from the inhomogeneous superimposed edge-triangle random graph model. We define the edge and triangle adjacency matrices A_E and A_T respectively, as explained in the previous section. Note these matrices are not the motif adjacency matrices of G_e and G_t, since there are edges in G_t that contribute to A_E and triangles from G_e that contribute to A_T. In addition, many "incidentally generated" or imposed triangles (Chandrasekhar and Jackson, 2014) may arise due to superimposition, which also contributes to A_T. The different scenarios are depicted in Figure 2. Accordingly, for our analysis, we introduce the following six matrices.

(a) A_{E^2}: the adjacency matrix of edges in G_e; here, $(A_{E^2})_{ij} = E_{ij}$.

(b) A_{T^2}: the adjacency matrix of triangle motifs in G_t; here, $(A_{T^2})_{ij} = \sum_k T_{ijk}$.



(a) (b) (c) (d)

Figure 2: Imposed triangles generated through the superimposition of edges and triangles: (a) E^3, (b) T^3, (c) T^2E, and (d) TE^2.

(c) A_{E^3}: the motif adjacency matrix of all triangles formed by random edges from G_e. The generative indicator random variable for a triangle of this class reads as:

$$E^3_{ijk} = E_{ij}E_{jk}E_{ik},$$

and $(A_{E^3})_{ij} = \sum_k E_{ij}E_{jk}E_{ik}$.

(d) A_{T^3}: the motif adjacency matrix of all triangles formed by three intersecting triangles from G_t. The generative indicator random variable for a triangle of this class reads as:

$$T^3_{ijk} = 1\Big(\sum_{k_1\neq k} T_{ijk_1} > 0\Big)1\Big(\sum_{k_2\neq i} T_{jkk_2} > 0\Big)1\Big(\sum_{k_3\neq j} T_{ikk_3} > 0\Big),$$

and $(A_{T^3})_{ij} = \sum_{k\neq(i,j)} T^3_{ijk}$.

(e) A_{T^2E}: the motif adjacency matrix of all triangles formed by two triangles from G_t and one edge from G_e. The generative indicator random variable for a triangle of this class reads as:

$$\begin{aligned}(T^2E)_{ijk} =& 1\Big(\sum_{k_1\neq k} T_{ijk_1} > 0\Big)1\Big(\sum_{k_2\neq i} T_{jkk_2} > 0\Big)E_{ik}\\ &+ 1\Big(\sum_{k_1\neq k} T_{ijk_1} > 0\Big)1\Big(\sum_{k_2\neq j} T_{ikk_2} > 0\Big)E_{jk}\\ &+ 1\Big(\sum_{k_1\neq i} T_{jkk_1} > 0\Big)1\Big(\sum_{k_2\neq j} T_{ikk_2} > 0\Big)E_{ij},\end{aligned}$$

and $(A_{T^2E})_{ij} = \sum_k (T^2E)_{ijk}$.

(f) A_{TE^2}: the motif adjacency matrix of all triangles formed by one triangle from G_t and two edges from G_e. The generative indicator random variable for a triangle of this class reads as:

$$\begin{aligned}(TE^2)_{ijk} =& 1\Big(\sum_{k_1\neq k} T_{ijk_1} > 0\Big)E_{jk}E_{ik} + 1\Big(\sum_{k_1\neq k} T_{jkk_1} > 0\Big)E_{ij}E_{ik} \\ &+ 1\Big(\sum_{k_1\neq k} T_{ikk_1} > 0\Big)E_{jk}E_{ij},\end{aligned}$$

and $(A_{TE^2})_{ij} = \sum_k (TE^2)_{ijk}$.

We call the first two types of structures model-generated, while the last four types of motifs as incidentally generated. Note that except for case (c), an incidental triangle involving vertices (i, j, k) arises only if there is no model-generated triangle involving (i, j, k) already present. Hence, we multiply each of the random variables T^3, T^2E, and TE^2 by the factor $(1 - T_{ijk})$ that indicates this dependence. For case (c), since we allow a multi-edge between two vertices that are both involved in a triangle hyperedge and an edge, it is possible to have an incidental triangle in addition to a model-generated triangle on the same triple of vertices.

With these definitions, we have the number of triangles on the vertex triple (i, j, k) as

$$\begin{aligned}&T_{ijk} + E^3_{ijk} + (1 - T_{ijk})T^3_{ijk} + (1 - T_{ijk})(1 - E^3_{ijk})(1 - T^3_{ijk}) \max((T^2E)_{ijk}, (TE^2)_{ijk}) \\ &= T_{ijk} + \Psi_{ijk}.\end{aligned}$$

The above implies we may observe a maximum of two triangles among the (i, j, k) tuple. If (i, j, k) does not have a triangle of type T, then we may observe an incidentally generated triangle of type T^3_{ijk}. If (i, j, k) does not have a triangle of either T or E^3 type, then an additional (only one) incidentally generated triangle is possible if any of the two indicators, T^2E_{ijk} and TE^2_{ijk}, is 1. The triangle adjacency matrix reads as

$$(A_T)_{ij} = \sum_k (T_{ijk} + \Psi_{ijk}),$$

capturing both model-based and incidental triangles. Obviously, we only observe the matrices A_E and A_T and not their specific constituents, as in real networks, we do not have labels describing how an interaction is formed. Hence, even though the community structure is most explicitly described by A_{T^2}, we need to analyze how this matrix reflects on A_T and what the properties of the latter matrix are based on A_{T^2}.

4. Analysis of higher-order spectral clustering

We analyze the higher-order spectral clustering method under the triangle-edge supSBM model. The primary goal of our analysis is to provide a theoretical guarantee of the accuracy of detecting the community structure of a graph generated from the SupSBM using the higher-order spectral clustering method. We will consider both versions of SupSBM, namely, one with community structure present only at the triangle level and the other with

community structure present both at the triangle and edge levels. In what follows, we first prove a number of concentration results on the spectral norm for certain motif adjacency matrices under the more general inhomogeneous superimposed random graph model. Subsequently, we specialize our analysis to the SupSBMs.

We start with some notation. Let

$$p^e_{\max} = \max_{i,j} p^e_{ij} \quad \text{and} \quad p^t_{\max} = \max_{i,j,k} p^t_{ijk}$$

denote the maximum probability of edge inclusion in G_e and triangle hyperedge inclusion in G_t, respectively. It is well-known for the usual edge-based adjacency matrix A_{E^2} that the spectral norm $\|A_{E^2} - E[A_{E^2}]\|_2$ is bounded by $c_1\sqrt{\Delta_e}$ with probability at least $1 - n^{-r}$ (Lei and Rinaldo, 2015; Gao et al., 2017; Chin et al., 2015), where $\Delta_e = np^e_{\max}$ and $p^e_{\max} > c_0 \log n$ where c_0, c_1, r are some constants. The quantity Δ_e can be interpreted as the maximum expected degree of a vertex in the graph. The following five results, summarized in Lemmas 1 to 5, provide non-asymptotic error bounds that hold in general settings, as described in the statements of the respective lemmas. Note that we make repeated use of the symbols c or r to represent different generic constants as needed in the proofs in order to avoid notational clutter. The proofs of all theoretical results are delegated to the Appendix.

4.0.1 Bounds for component matrices

Lemma 1 *Let $G_t(n, \mathbb{P}^t)$ be a 3-uniform hypergraph in which each possible 3-hyperedge is generated according to a Bernoulli random variable T_{ijk} with parameter p^t_{ijk}, independent of all other 3-hyperedges. Let A_{T^2}, as before, stand for the triangle-motif adjacency matrix. Furthermore, let $\Delta_t = n^2 p^t_{\max}$ and assume $p^t_{\max} > c\frac{\log n}{n^2}$ for some constant $c > 0$. Then, for some constant $r > 0$, there exists a constant $c_1(c, r) > 0$ such that with probability at least $1 - n^{-r}$, one has*

$$\|A_{T^2} - E[A_{T^2}]\|_2 \leq c_1\sqrt{\Delta_t}.$$

Note that in the above bound, Δ_t may be interpreted as the maximum expected "triangle degree" of vertices in G_t. Drawing a parallel with adjacency matrices of graphs, one may define the "degree" of a row of an arbitrary matrix as the sum of the elements in that row. Then, Δ_t is an upper bound on the degree of a row in the matrix $E[A_{T^2}]$, much like Δ_e is an upper bound for the degrees of the rows in $E[A_{E^2}]$. The above result for triangle-motif adjacency matrix is hence an analogue of a similar result for standard adjacency matrices described in Lei and Rinaldo (2015), Gao et al. (2017), and Chin et al. (2015). The arguments used to prove the result in the cited papers are based on an $\epsilon-$net analysis of random regular graphs laid out in Friedman et al. (1989) and Feige and Ofek (2005). We extend these arguments to the case of triangle hyperedges; due to the independence of the random variables corresponding to the hyperedges involved in all sums of interest, we do not require new concentration inequalities to establish the claim. This is not the case for the results to follow.

For bounding the spectral norm of the other four relevant matrices, namely, A_{E^3}, A_{T^3}, A_{T^2E}, A_{TE^2}, we use the following property of the spectral norm of a square symmetric matrix. For any $n \times n$ square symmetric matrix X, define the spectral norm of X as

$\|X\|_2 = \sigma_{\max}(X)$, the largest singular value of X, the 1-norm as $\|X\|_1 = \max_j \sum_i |X_{ij}|$, and the $\infty-$norm as $\|X\|_\infty = \max_i \sum_j |X_{ij}|$. Now assume X is an $n \times n$ symmetric matrix whose elements are non-negative random variables. Let the entries of its expectation, $E[X]$, also be non-negative. Then,

$$\begin{aligned}\|X - E[X]\|_2 &\leq \sqrt{\|X - E[X]\|_1 \|X - E[X]\|_\infty} \\ &= \|X - E[X]\|_1 \\ &= \max_i \sum_j |X_{ij} - E[X]_{ij}| \\ &\leq \max_i \sum_j X_{ij} + \max_i \sum_j E[X]_{ij}, \end{aligned} \tag{4.1}$$

where the first inequality is Corollary 2.3.2 in Golub and Van Loan (2012), and the second equality follows since $X - E[X]$ is a symmetric matrix by assumption. Note the first term in the final sum is the *degree* of row i of the matrix X. Hence, a high-probability bound on the maximum degree will allow us to upper bound this quantity. The second term equals the maximum expected degree of X, which is a deterministic quantity. Importantly, in Lemmas 2-5 that follow, we show that $\max_i \sum_j |X_{ij}|$ is bounded by a constant multiple of $\max_i \sum_j E[X]_{ij}$ with high probability and for all four matrices. While these bounds are not the strongest possible concentration inequalities, they suffice to prove our subsequent theorems as we only need to bound the spectral norms of the matrices instead of the norms of the deviations. This is the case since under the allowed range of growth rates for $p^t_{\max}$ and $p^e_{\max}$, the expected maximum degrees $\max_i \sum_j E[X]_{ij}$ are generally of moderate size (i.e., the bounds are adequate for sparse but loose for dense graphs).

Let $\tau_{\max} = \max\{n(p^e_{\max})^2, \log n\}$, $\Delta_{E^3} = \max\{n^2(p^e_{\max})^3, (\log n)^2\}$ and assume $np^e_{\max} > \log n$. Then we have the following result.

Lemma 2 *Let $G_e(n, P^e)$ be an inhomogeneous edge-based random graph in which each edge is independently generated by a Bernoulli random variable E_{ij} with parameter p^e_{ij}, $i, j = 1, \ldots, n$. Let $\Delta_{E^3} = \max\{n^2(p^e_{\max})^3, (\log n)^2\}$ and assume $np^e_{\max} > \log n$. Then, with probability at least $1 - n^{-\frac{1}{4}} - n^{-\frac{1}{11}}$,*

$$\max_i \sum_j (A_{E^3})_{ij} \leq 9\Delta_{E^3}.$$

For the next three lemmas in this section, we additionally assume $n^2 p^t_{\max} > (\log n)^2$.

Lemma 3 *Let $G \sim G_s(n, P^e, \mathbb{P}^t)$ be a graph generated by the superimposed random graph model. Let*

$$\Delta_{T^3} = \max\{n^5(p^t_{\max})^3, (\log n)^4\},$$

and assume $n^2 p^t_{\max} > (\log n)^2$. Then with probability at least $1 - n^{-\frac{1}{15}} - n^{-\frac{1}{4}} - n^{-\frac{1}{7}}$, one has

$$\max_i \sum_j (A_{T^3})_{ij} \leq 25\,\Delta_{T^3}.$$

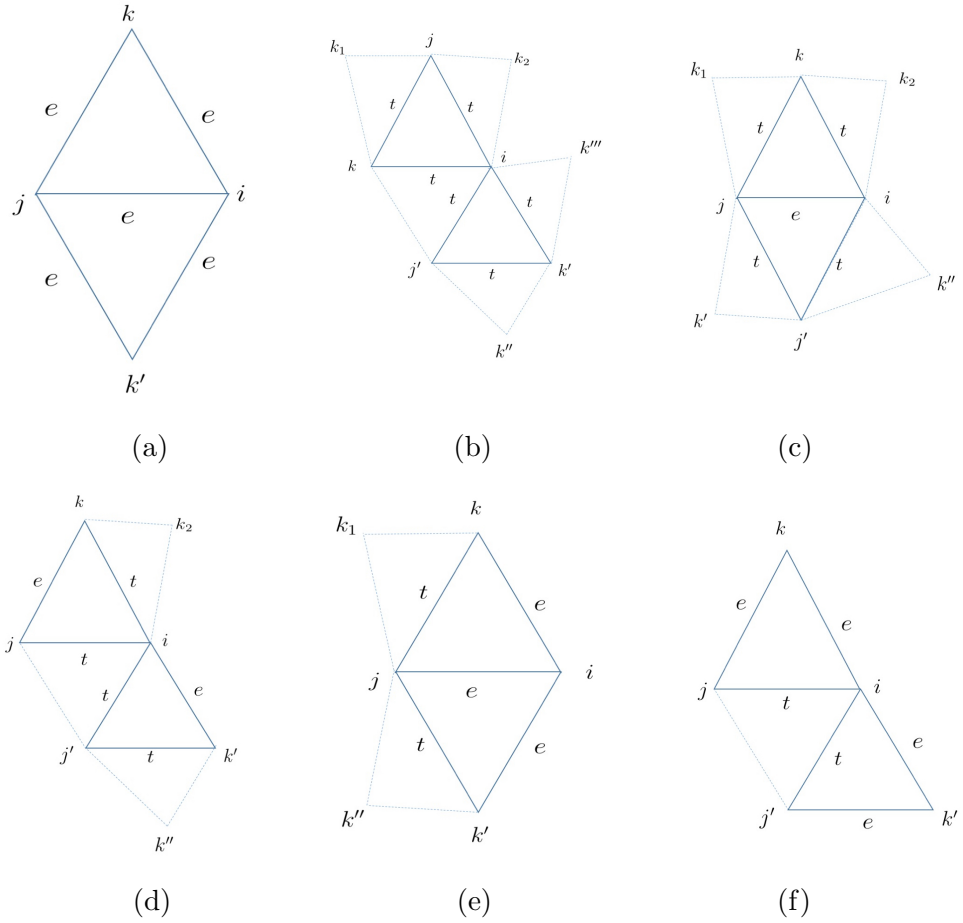


Figure 3: Dependence among the random variables of incidental triangles that include vertex i, (a) E^3, (b) T^3, (c) T^2E of type 1, (d) T^2E of type 2, (e) TE^2 of type 1, and (f) TE^2 of type 2.

Lemma 4 *Let $G \sim G_s(n, P^e, \mathbb{P}^t)$ be a graph generated by the superimposed random graph model. Let*

$$\Delta_{T^2E} = \max\{n^4 (p^t_{\max})^2 p^e_{\max},\ (\log n)^4\}.$$

Assume $n^2 p^t_{\max} > (\log n)^2$ and $np^e_{\max} > \log n$. Then with probability at least $1 - n^{-\frac{29}{700}} - n^{-\frac{1}{4}} - 2n^{-\frac{1}{7}}$, one has

$$\max_i \sum_j (A_{T^2E})_{ij} \le 28\, \Delta_{T^2E}.$$

Lemma 5 *Let $G \sim G_s(n, P^e, \mathbb{P}^t)$ be a graph generated by the superimposed random graph model. Let*

$$\Delta_{TE^2} = \max\{n^3 p^t_{\max} (p^e_{\max})^2,\ (\log n)^3\}.$$

Assume $n^2 p_{\max}^t > (\log n)^2$ *and* $np_{\max}^e > \log n$. *Then, with probability at least* $1 - n^{-\frac{1}{80}} - 2n^{-\frac{1}{7}} - n^{-1}$, *one has*

$$\max_i \sum_j (A_{TE^2})_{ij} \leq 10\, \Delta_{TE^2}.$$

As we noted earlier, the usual ϵ-net approach cannot be applied directly to prove upper bounds on the spectral norm of the motif adjacency matrices: A_{E^3}, A_{T^3}, A_{T^2E}, A_{TE^2}. This is because the elements of these adjacency matrices are dependent, and consequently, the sums of the random variables used in the $\epsilon-$net approach include dependent variables. Instead, we take a different strategy. The proofs of all the above results follow a similar outline. In each case, the degree of a row i is a sum of *dependent* triangle-indicator random variables for triples that include vertex i. However, in each case, we carefully characterize the events that lead to two such *indicator random variables* to be dependent. We then show that the number of *realized triangle indicators* that any indicator is dependent on is limited with high probability. This allows us to apply Theorem 9 of Warnke (2017), reproduced below as a proposition, in an iterative manner to obtain concentration results on the respective sums.

Proposition 1 (Theorem 9 of Warnke (2017)) *Let* (Y_i), $i \in \mathcal{I}$ *be a collection of non-negative random variables with* $\sum_{i\in\mathcal{I}} E(Y_i) \leq \mu$. *Assume that* $\sim$ *is a symmetric relation on* $\mathcal{I}$ *such that each* Y_i *with* $i \in \mathcal{I}$ *is independent of* $\{Y_j : j \in \mathcal{I},\, j \nsim i\}$. *Let* $Z_C = \max \sum_{i\in\mathcal{J}} Y_i$, *where the maximum is taken over all sets* $\mathcal{J} \subset \mathcal{I}$ *such that* $\max_{j\in\mathcal{J}} \sum_{i\in\mathcal{J}, i\sim j} Y_i \leq C$. *Then for all* $C, t > 0$ *we have*

$$P(Z_C \geq \mu + t) \leq \min\Big\{ \exp\left(-\frac{t^2}{2C(\mu + t/3)}\right), \left(1 + \frac{t}{2\mu}\right)^{-t/2C} \Big\}.$$

While we relegate technically involved rigorous proofs to the Appendix, we graphically illustrate all the events leading to the dependencies (the relations $\sim$ in the proposition above) among the indicators within the various collections of indicators in Figure 3. For the result on triangle-indicators of type E^3, let $I_i = \{E^3_{ijk} = E_{ij}E_{jk}E_{ik},\ (j,k) = \{1,\dots,n\}^2,\ (j,k) \neq i\}$, denote the collection of indicator random variables for the presence of triangles of type E^3 attached to vertex i. The key observation is that two indicators E^3_{ijk} and $E^3_{ijk'}$ are dependent if and only if they share an edge indicator E_{ij} (see Figure 3(a)). In the notation of the proposition, we have $E^3_{ijk} \sim E^3_{ij'k}$ for all $j' \neq (i,k)$, since E^3_{ijk} and $E^3_{ij'k}$ share an edge indicator random variable E_{ik}, while $E^3_{ijk} \sim E^3_{ijk'}$ for all $k' \neq (i,j)$, since E^3_{ijk} and $E^3_{ijk'}$ share an edge indicator random variable E_{ij}. Then we define a "good event" Γ, under which

$$\max_{(i,j,k)\in I_i} \sum_{(i,j',k')\in I_i,\, (i,j,k)\sim(i,j',k')} E^3_{(i,j',k')},$$

i.e., the number of indicators of type E^3 that are realized (i.e., $E^3_{ij'k'} = 1$) and dependent on E^3_{ijk}, is bounded by $8\tau_{\max}$. Finally, we show the event Γ, which states that for a vertex

pair (i, j), there are at most $4\tau_{\max}$ vertices k' such that the vertex pairs (i, k') and (j, k') are connected by edges from G_e, occurs with high probability.

The proofs for the other lemmas follow a similar strategy. For the family of random variables $I_i = \{(T^3)_{ijk},\ j = \{1, \ldots, n\},\ k = \{1, \ldots, n\}\}$, two indicators $(T^3)_{ijk}$ and $(T^3)_{ij'k'}$ are dependent if and only if one of the triangle indicators from G_t responsible for the ik or ij "sides" of $(T^3)_{ijk}$, i.e, the set $\{T_{ikk_3},\ k_3 \neq j\}$ or $\{T_{ijk_1},\ k_1 \neq k\}$ includes j' or k' as a vertex and is consequently part of the indicator $(T^3)_{ij'k'}$ (see Figure 3(b)). Next, let $I_i = \{(T^2E)_{ijk},\ j = \{1, \ldots, n\},\ k = \{1, \ldots, n\}\}$ denote the set of all indicator variables for incidentally generated triangles of type T^2E that includes the vertex i. Two random variables in the family may be dependent on two scenarios. One possibility is that the edge indicator E_{ik} is common between $(T^2E)_{(i,j,k)}$ and $(T^2E)_{ikj'}$ for some j' (see Figure 3 (c)). The other possibility is that one of the triangle indicators in the sets $\{T_{ijk_1}, k_1 \neq k\}$ or $\{T_{jkk_2}, k_2 \neq i\}$ is also involved in creating $(T^2E)_{ij'k'}$ for some j' and k' (see Figure 3 (d)). Finally, let $I_i = \{(TE^2)_{ijk},\ j = \{1, \ldots, n\},\ k = \{1, \ldots, n\}\}$ denote the set of all indicator variables for incidentally generated triangles of type TE^2 including the vertex i. Let $(TE^2)_{ijk}$ be a representative indicator random variable from this set. Consider another element $(TE^2)_{ij'k'}$ in I_i. This element is dependent on $(TE^2)_{ijk}$ in two ways. First, one of the indicators from G_e, say E_{ik}, in TE^2_{ijk}, may also be a side in the incidental triangle characterized by $(TE^2)_{ij'k}$ for some j' (see Figure 3(e)). Second, one of the sides ij may have been created by a triangle indicator from G_t, with the same triangle indicator being involved in creating the incidental triangle characterized by $(TE^2)_{ij'k'}$ for some j' and k' (see Figure 3(f)). For each case, we define suitable good events that hold with high probability and show that under those events, the number of realized indicator variables in the family that one indicator variable depends on is bounded as required by the proposition.

4.0.2 Concentration bound for A_T

As noted by Chandrasekhar and Jackson (2014), in the superimposed random graph framework, the generative probabilities summarized in $\mathbb{P}^t$ and P^e must satisfy certain conditions in order to ensure that the imposed triangles do not significantly outnumber the generative triangles. Accordingly, we impose the following asymptotic growth conditions on $p^t_{\max}$ and $p^e_{\max}$:

$$c_1 \frac{\log n}{n} \leq p^e_{\max} \leq c_2 \frac{n^{2/5-\eta}}{n}, \tag{4.2}$$

$$c_3 \frac{(\log n)^8}{n^2} < p^t_{\max} < c_4 \frac{n^{2/5-\epsilon}}{n^2}, \tag{4.3}$$

for some $\epsilon > 0, \eta > 0$ and constants c_1, c_2, c_3, c_4 independent of n. In the second part of the following theorem we will impose an additional assumption to further simplify the result:

$$p^t_{\max} > c_5 n^2 (p^e_{\max})^6. \tag{4.4}$$

Note that assumptions (4.3) and (4.4) together imply the upper bound on $p^e_{\max}$ in (4.2). We can see this by noting the upper bounds in (4.3) and (4.4) imply $n^6(p^e_{\max})^6 < \frac{c_4}{c_5} n^{12/5-\epsilon}$, and consequently, $np^e_{\max} < (\frac{c_4}{c_5})^{1/6} n^{2/5-\epsilon/6}$.

We want to remind the reader that $\Delta_e = np^e_{\max}$ and $\Delta_t = n^2 p^t_{\max}$ are the maximum expected degrees of a node in the dyadic and triadic components of the superimposed graph respectively. Therefore if Δ_e and Δ_t are asymptotically comparable, the superimposed graph will have an asymptotically comparable density of edges coming from the dyadic and the triadic components. Typical examples are the following two sets of growth rates: $p^e_{\max} = O(\frac{(\log n)^8}{n})$, $p^t_{\max} = O(\frac{(\log n)^8}{n^2})$ and $p^e_{\max} = O(\frac{n^{1/4}}{n})$, $p^t_{\max} = O(\frac{n^{1/4}}{n^2})$.

In the next theorem we combine the previous results to arrive at a concentration bound for the matrix A_T under the assumptions made on $p^e_{\max}$ and $p^t_{\max}$ in (4.2) and (4.3).

Theorem 1 *Let A_T denote the triangle-motif adjacency matrix of a random graph G generated by the inhomogeneous superimposed random graph model $G_s(n, P^e, \mathbb{P}^t)$. Let $\Delta_t = n^2 p^t_{\max}$ and $\Delta_{E^3} = \max\{n^2(p^e_{\max})^3, (\log n)^2\}$, and assumptions (4.2) and 4.3) on $p^e_{\max}$ and $p^t_{\max}$ hold, then with probability at least $1 - o(1)$, one has*

$$\|A_T - E[A_T]\|_2 \leq \tilde{c}(\sqrt{\Delta_t} + \Delta_{E^3}),$$

where $\tilde{c}$ is a constant independent of n. If in addition, the assumption (4.4) holds, then with probability at least $1 - o(1)$, one has

$$\|A_T - E[A_T]\|_2 \leq \tilde{c}_1\sqrt{\Delta_t}.$$

Note in the above theorem, we can also make the definition of Δ_{E^3} to be just $n^2(p^e_{\max})^3$, and drop the $(\log n)^2$ term, since by the assumptions on $p^e_{\max}$ and $p^t_{\max}$, the $(\log n)^2$ term can be absorbed in $\sqrt{\Delta_t}$.

We also note the similarity of the upper bound of this concentration inequality with that obtained for A_{T^2} in Lemma 1. The above result tells us that under the assumed conditions, the effect of the incidental triangles on the concentration of A_T is limited, and the rate in the upper bound is predominantly determined by the rate for A_{T^2}. This suggests that while the superimposition process induces dependencies between the edges in G_s through the presence of triangles from G_t, the model, under suitable sparsity conditions, is still mathematically tractable. The influence of the incidental triangles can be analyzed and controlled.

In the SBM literature, trimmed adjacency matrices are often used for sparse graphs with bounded maximum expected degrees to remove the $O(\log n)$ minimum degree requirements since trimmed adjacency matrices have better concentration properties. Such analysis techniques cannot be directly applied in our model settings. While the bound in Lemma 1 can be improved by removing nodes with triangle degrees greater than $c\Delta_t$ for some constant c, we run into difficulty attempting to do so with the bound on A_T in Theorem 1. Typically Δ_t would be much larger than the expected maximum triangle degrees of the other motif-adjacency matrices of the incidental triangles. For Lemmas 2-5, we have used a more loose technique that bounds the operator norm of the adjacency-type matrices with the maximum expected degree instead of the square root of the maximum expected degree. This loose bound suffices for graphs with denser triangle densities since all the maximum expected degrees Δ_{E^3}, Δ_{T^3}, Δ_{T^2E}, Δ_{TE^2} are smaller than $\sqrt{\Delta_t}$. However, we were not able to remove the poly-log terms in the upper bounds by using the trimmed versions of those adjacency matrices. This will be an important future research direction.

4.1 Higher-order spectral clustering under the SupSBM

Next, we turn our attention to analyzing random graphs generated by SupSBMs, and focus in particular on quantifying the misclustering error rate under the higher-order spectral clustering algorithm. Let $\hat{C}$ denote the $n \times k$ matrix of eigenvectors corresponding to the k largest absolute-value eigenvalues of the triangle motif adjacency matrix A_T. To obtain the community assignments for the vertices, we use the greedy clustering algorithm in Gao et al. (2017, Algorithm 2) on the rows of $\hat{C}$, which runs in polynomial time. As noted in Gao et al. (2017), the more commonly used $(1+\epsilon)$-approximate k-means clustering (Kumar et al., 2004; Lei and Rinaldo, 2015) provided an inferior approximation for growing k since the factor ϵ is proportional to k. Let $\mu > 0$ be a small constant such that the critical radius $r = \mu\sqrt{k/n}$ in Algorithm 2 of Gao et al. (2017). We define the misclustering error rate R as follows. Let $\bar{e}$ and $\hat{e}$ denote the vectors containing the true and estimated community labels of all the vertices in V. Then we define

$$R = \inf_{\Pi} \frac{1}{n} \sum_{i=1}^{n} 1(\bar{e}_i \neq \Pi(\hat{e}_i)),$$

where the infimum is taken over all permutations $\Pi(\cdot)$ of the community labels.

Theorem 2 *Let $G \sim G_s(C, \pi^e, \pi^t)$ be a graph generated from the n-vertex k-block SupSBM with parameters C, π^e, π^t as defined in Section 2.1. Let A_T be the triangle motif adjacency matrix as defined earlier and $\lambda_{\min}(E[A_T])$ denote the minimum in absolute value non-zero eigenvalue of the matrix $E[A_T]$. If assumptions (4.2) and 4.3) hold, then with probability at least $1 - o(1)$, the misclustering rate of community detection using the higher-order spectral clustering method satisfies*

$$R_T \leq \frac{128\tilde{c}^2(\Delta_t + \Delta_{E^3}^2)}{\mu^2(\lambda_{\min}(E[A_T])^2}.$$

While the above result holds for general SupSBMs, we can evaluate the quantity $\lambda_{\min}(E[A_T])$ under a special case to gain further insight on the result. For the special case, we first define the generation of the triangle hyperedges in the following manner:

$$P(T_{ijk} = 1 | C_i, C_j, C_k) = \begin{cases} \frac{a_t}{n^2}, & \text{if } C_i = C_j = C_k, \\ \frac{b_t}{n^2}, & \text{otherwise,} \end{cases}$$

so that the probability of a triangle hyperedge equals a_t/n^2 if the three vertices involved are in the same community, and b_t/n^2 if at least one of the vertices is in a different community than the other two. The dyadic edges are generated according to the following rule: the probability of an edge is a_e/n if both the end points belong to the same community and b_e/n if they belong to different communities. We further assume that all communities are of the same size, leading to balanced n-vertex k-block SupSBMs, denoted by $G_s(C, n, k, a_e, b_e, a_t, b_t)$, in which all the k communities have n/k vertices. We use the notations $\asymp$ and $\lesssim$ to mean asymptotically of the same order and asymptotically less, respectively.

Theorem 3 *Let $G \sim G_s(C, n, k, a_e, b_e, a_t, b_t)$ be a graph generated from the balanced n-vertex k-block SupSBM. If assumptions (4.2) and (4.3) hold, then with probability at least $1-o(1)$, the misclustering rate of community detection using the higher-order spectral clustering method satisfies*

$$R_T \lesssim \frac{a_t + \frac{a_e^6}{n^2}}{\left(\frac{a_t - b_t}{k^2} + \frac{(kb_e^2 + a_e^2 + a_e b_e - 2b_e^2)(a_e - b_e)}{k^2 n}\right)^2},$$

as $n \to \infty$. If we further assume $a_e \asymp b_e$ and $a_t \asymp b_t$, then the above simplifies to

$$R_T \lesssim \frac{a_t + \frac{a_e^6}{n^2}}{\left(\frac{a_t - b_t}{k^2} + \frac{b_e^2(a_e - b_e)}{kn}\right)^2}.$$

4.1.1 Examples of consistent community detection

Now we consider a few example growth rates to understand what conditions in the upper bound lead to consistent community detection. Note in the balanced n-vertex k-block SupSBM, the number of triangles is $O(na_t)$ while the number of edges is $O(na_e)$. Since in real networks often the number of edges and the number of triangles are of the same order (Bollobás et al., 2011), it is natural to assume the asymptotic setup that $a_t \asymp a_e$. Let us assume $a_t = m_t D$, $b_t = s_t D$, $a_e = m_e D$, $b_e = s_e D$ for constants m_t, s_t, m_e, s_e, and D is a function of n. We consider three scenarios with D being $O((\log n)^8)$, $O(n^{1/4})$, and $O(n^{2/5-\epsilon})$. Then, ignoring the constants the above result becomes

$$R_T \lesssim \frac{D + \frac{D^6}{n^2}}{\frac{D^2}{k^4} + \frac{D^6}{k^2 n^2}}.$$

In each of the three scenarios, the numerator is $O(D)$ and the denominator is greater than $O(\frac{D^2}{k^4})$. In the first scenario, we have, $R_T \lesssim \frac{k^4}{(\log n)^8}$, and consistent community detection is possible as long as $k = O(\log n)^2$. In the second scenario, $R_T \lesssim \frac{k^4}{n^{1/4}}$, and consistent community detection is possible as long as $k = O(n^{1/16})$. Finally, in the third growth scenario, $R_T \lesssim \frac{k^4}{n^{2/5}}$, and consistent community detection is possible as long as $k = O(n^{1/10})$.

We consider another scenario where $a_e = b_e$, and therefore the community structure in SupSBM is expressed purely through the triadic graph. In this case the upper bound in Theorem 3 boils down to

$$R_T \lesssim \frac{a_t + \frac{a_e^6}{n^2}}{(\frac{a_t - b_t}{k^2})^2}.$$

Consistent community detection is still possible in this scenario with the same set of conditions on the growth rate of k as in the previous paragraph.

4.2 Uniform and non-uniform hypergraph SBMs

In what follows, we analyze the performance of the higher-order spectral clustering under the uniform and non-uniform hypergraph SBMs (Ghoshdastidar and Dukkipati, 2017; Chien

et al., 2018; Ahn et al., 2018). The balanced n-vertex k-block *3-uniform hypergraph* SBM $G_t(C, n, k, a_t, b_t)$ is defined in the following way. All the k communities have an equal number of vertices $s = n/k$, and the probability of forming a triangle hyperedge equals a_t/n^2 if all three vertices belong to the same community, while the probability of forming a triangle hyperedge equals b_t/n^2 if one of the vertices belongs to a different community than the other two.

Non-uniform hypergraphs involve hyperedges connecting varying number of vertices. We consider a model for non-uniform hypergraphs with two types of hyperedges: edges and triangles. As mentioned earlier, the supSBM is a model for graphs and is distinct from such non-uniform hypergraph SBMs. The observations are labeled as two-way and three-way interactions between entities in the later case. Hence, in non-unifrom hypergraph, we have a way to differentiate between an edge and a triangle hyperedge. The n-vertex k-block balanced non-uniform hypergraph SBM $G_H(C, n, k, a_e, b_e, a_t, b_t)$ is defined in the same way as a SupSBM, except that we do not replace the generated triangle hyperedges with three ordinary edges and we do not collapse multiedges.

If we assume a hypergraph is generated from a uniform hypergraph SBM on triangle hyperedges, then spectral clustering of the motif adjacency matrix is equivalent to spectral clustering based on A_{T^2} only. Let $\hat{C}^{(T^2)}$ be the matrix of eigenvectors corresponding to the k largest absolute eigenvalues of the matrix A_{T^2}. Then, using the bound for A_{T^2} in Lemma 1, we arrive at the following result.

Corollary 1 *Let G_t be a triangle hypergraph generated from the k-block uniform triangle hypergraph SBM with parameters C, n, k, a_t, b_t. Then, with probability at least $1 - n^{-c}$, the misclustering rate of the community assignments R_{T^2} obtained using the higher-order spectral clustering algorithm applied to the triangle motif adjacency matrix equals*

$$R_{T^2} \leq \frac{c\|A_{T^2} - E[A_{T^2}]\|_2^2}{\mu^2(\lambda_{\min}(E[A_{T^2}]))^2} \lesssim \frac{k^4 a_t}{(a_t - b_t)^2}.$$

The above corollary has important implication for non-uniform hypergraph SBMs. Assume that we are given a non-uniform hypergraph generated from the n-vertex k-block balanced non-uniform hypergraph SBM $G_H(C, n, k, a_e, b_e, a_t, b_t)$. The question of interest is: Given a_e, b_e, a_t, b_t, with $a_e \asymp b_e$ and $a_t \asymp b_t$, should one use the edge-based adjacency matrix, the triangle-based adjacency matrix, or a combination thereof? Let

$$a_t \asymp \frac{a_e}{\delta}, \quad a_t - b_t = m\frac{a_e - b_e}{\delta}, \quad a_e \asymp b_e, \tag{4.5}$$

so that asymptotically, the probabilities a_e/n and b_e/n are $n\delta$ times the probabilities a_t/n^2 and b_t/n^2, while the difference between the probabilities $(a_e - b_e)/n$ is $n\delta/m$ times that of the difference $(a_t - b_t)/n^2$. Clearly, δ captures the asymptotic difference between the densities of triangle hyperedges and dyadic edges, while m captures the difference in the "communal" qualities between these two types of hyperedges. Note that the notation for asymptotic equivalence ignores all constants.

Remark 1 *Let $G \sim G_H(C, k, a_e, b_e, a_t, b_t)$ be a graph generated from the non-uniform hypergraph SBM. Assume the relationships between the probabilities a_e, b_e, a_t, b_t are as in (4.5).*

Then, spectral clustering based on a triangle adjacency matrix has a lower error rate than spectral clustering based on an edge adjacency matrix if $\frac{k^2\delta}{m^2} \lesssim 1$*, and a higher error rate if* $\frac{k^2\delta}{m^2} \gtrsim 1$*.*

The above results also allow us to bound the error rate of spectral clustering of a weighted motif adjacency matrix under the non-uniform hypergraph SBM. Let $A_W = A_{E^2} + wA_{T^2}$ be the weighted sum of adjacency matrices of edges and triangle hyperedges with known relative weight $w > 0$. Clearly, $E[A_W] = E[A_{E^2}] + wE[A_{T^2}]$ and the smallest non-zero eigenvalue of $E(A_W)$ is $\lambda_{\min}(E[A_W]) = \frac{1}{k}\{(a_e - b_e) + \frac{w}{k}(a_t - b_t)\}$. Then, with probability at least $1 - o(1)$ we have

$$\|A_W - E[A_W]\|_2 \leq \|A_{E^2} - E[A_{E^2}]\|_2 + w\|A_{T^2} - E[A_{T^2}]\|_2 \lesssim \sqrt{\Delta} + w\sqrt{\Delta_t},$$

and the error rate is upper bounded according to

$$R_W \lesssim \frac{k^2(\sqrt{a_e} + w\sqrt{a_t})^2}{((a_e - b_e) + \frac{w}{k}(a_t - b_t))^2}.$$

When the asymptotic relationships of (4.5) hold, we can further simplify this expression to

$$R_W \lesssim \frac{k^2(1 + \frac{w}{\sqrt{\delta}})^2}{(1 + \frac{mw}{k\delta})^2} \frac{a_e}{(a_e - b_e)^2}. \tag{4.6}$$

While Remark 1 suggests that depending upon the values of δ and m, either the edge-based or triangle-based adjacency matrix has a lower error rate, in practice it might be beneficial for numerical stability to use a weighted average of both of them. The result in (4.6) provides a bound for any weighted sum of these two hyperedge adjacency matrices.

5. Estimation of model parameters and model fit

In this section, we discuss an estimation method for the parameters of the SupSBM once the community assignments have been obtained. We also present two strategies for model assessment and comparison, one through parametric bootstrap and the other through network cross-validation.

5.1 Estimating the model parameters

Once we have obtained the community assignments, the parameters can be estimated using an approximate generalized method of moments approach similar to Chandrasekhar and Jackson (2014). We work with the following set of $(k^2 + k^3)$ sample moments

$$S^e_{pq} = \sum_{C_{ip=1}, C_{jq}=1} (A_E)_{ij}, \qquad S^t_{pql} = \sum_{C_{ip=1}, C_{jq}=1, C_{kl}=1} \Psi_{ijk}.$$

Let S^e and S^t denote the vectors of the sample moments defined above. Next we need to write down the corresponding population moments, i.e., $E[S^e]$ and $E[S^t]$, under the SupSBM. However, in the general k-block supSBM, it is quite difficult to exactly compute the population moments. This is because the incidental triangles on a vertex triple can be

generated by triangles involving vertices that are in communities different from the original three vertices, making it difficult to enumerate probabilities of such triangles. Therefore to make an approximation, we first define the following quantities.

$$\bar{\pi}^e = \frac{\sum_{p,q} n^e_{pq}\pi^e_{pq}}{\binom{n}{2}}, \qquad \bar{\pi}^t = \frac{\sum_{p,q,l} n^t_{pql}\pi^t_{pql}}{\binom{n}{3}},$$

where n^e_{pq} and n^t_{pql} denote the total number of possible edges and the total number of possible triangles between communities p and q. Then we approximate the k^2 edge-based population moments as follows:

$$E[S^e_{pq}] = \pi^e_{pq} + 1 - (1-\bar{\pi}^t)^{(n-2)},$$

and the k^3 triangle-based population moments as follows:

$$E[S^t_{pql}] = \pi^t_{pql} + (\bar{\pi}^e)^3 + (1-\pi^t_{pql})\left(\binom{n-3}{3}(\bar{\pi}^t)^3 + \binom{n-3}{2}(\bar{\pi}^t)^2\bar{\pi}^e + \binom{n-3}{1}\bar{\pi}^t(\bar{\pi}^e)^2\right).$$

Next, we estimate the parameters by minimizing the following constrained optimization problem:

$$[\pi^e, \pi^t] = \underset{0\le\pi^e\le 1,\ 0\le\pi^t\le 1}{\arg\min} \{(S^e - E[S^e])^T(S^e - E[S^e]) + (S^t - E[S^t])^T(S^t - E[S^t])\}.$$

5.2 Model fit through parametric bootstrap

We formulate the following parametric bootstrap scheme to assess model fit and compare it with other random graph models. We repeatedly sample networks from the fitted models and form bootstrap distributions of key network properties. The network properties we consider are the average path length (L), the clustering coefficient or transitivity (C), the maximized modularity score (M), and the distribution of vertex degrees (Newman, 2018; Bullmore and Sporns, 2009). The average path length is defined as the average of the shortest paths between pairs of vertices in the whole network. A short average path length indicates that a vertex in the network can be reached from another vertex in relatively few hops. The clustering coefficient or transitivity is defined as three times the ratio of the number of triangles (i.e., closed triples) with the number of connected triples in the network. The modularity score for a given community assignment is a quality function that measures the difference between the observed number of intra-community edges and what would be expected from a null model with the same degree distribution. The maximum of this modularity score indicates how modular or partitioned into communities a network is for an optimal (in the sense of maximizing this modularity score) community assignment (Girvan and Newman, 2002). Many real networks are known to exhibit a small average path length and yet a high clustering coefficient, a property known as the "small-world" property Watts and Strogatz (1998). In each case, the fit of a model to a dataset is assessed by comparing the observed value of the property with that of a bootstrap distribution of the property formed through repeatedly sampling networks from the fitted models. If the observed value of the property is within the histogram of the property in sampled graphs, then we determine the model to be able to generate graphs with that property well. We

use this procedure to assess the fit of SupSBM in comparison to the ER, the SBM, and the SupER models. For the SBM and the SupSBM models, we use the ordinary edge-based spectral clustering method to estimate the communities for a fairer comparison. We use the constrained optimization approach described in the previous section to estimate the parameters of the SupER and the SupSBM models.

5.3 Model selection through network cross-validation

We also develop a model validation and selection strategy between the two low-rank models, the SBM and the SupSBM, using the recently proposed network cross-validation through edge sampling method (Li et al., 2020b). While we can easily obtain in-sample model fit using various metrics on the whole network data, obtaining an estimate of the test or generalization error is a more challenging problem for any metric. The s-fold network cross-validation approach of Li et al. (2020b) randomly splits the pairs of vertices in the network into s groups. The training data is formed with $s-1$ sets, and the remaining set is used as test data. A matrix completion method is used to complete the unobserved entries and form a full adjacency matrix. It was shown in Li et al. (2020b) that this approach is valid as long as a low-rank assumption on the probability matrix can be made. This assumption is valid for the SBM and the SupSBM that we compare. One additional issue is that the method Li et al. (2020b) used for matrix completion yields an estimated adjacency matrix with continuous values. Since our models require the actual binary graph and cannot be fitted on a matrix with continuous values, we obtained a binary graph by thresholding the elements of the estimated adjacency matrix at a threshold. The thresholding operation is such that if an element of the adjacency matrix is above the threshold, then we replace the element with 1, and we replace it with 0 otherwise. The threshold we chose is the average value of the elements of the *estimated adjacency matrix* for the subset of pairs of vertices that have a link in the original graph. The metric we use for model comparison is the average squared error in predicting the existence of an edge and the number of triangles between a pair of vertices. We could have also used the area under the ROC curve (AUC) for the comparison in terms of edge existence, but it cannot be used for comparison in terms of the number of triangles. While AUC can be used as an accuracy metric for predicting the existence of a triangle in vertex triples, the network cross-validation method splits a network into training and test datasets on the basis of pairs of vertices, and not triples of vertices, making it difficult to adopt the metric for out of sample accuracy. Therefore, We use the average squared error metric in both the edge and triangle prediction cases to remain consistent. Further the paper Li et al. (2020b), where the method was proposed, also recommended the use of the mean squared error as metric. We estimate the two models using the training data and predict the expected values of observing an edge and the number of triangles between a pair of vertices.

Finally, we also propose to choose the number of communities K in the SupSBM using a modification of the cross-validation method described above from Li et al. (2020b). The problem of estimating the number of communities in general SBMs has been studied extensively in the literature (Bickel and Sarkar, 2016; Chen and Lei, 2018; Cerqueira and Leonardi, 2020; Yan et al., 2018; Le and Levina, 2022; Li et al., 2020b). Both the methods in Chen and Lei (2018) and Li et al. (2020b) are based on cross-validation techniques. We

choose the method in Li et al. (2020b) for estimating K in our model setting since it is easy to implement and naturally fits with our method of selecting between SBMs and SupSBMs. The method was shown to be consistent for selecting the number of communities in networks generated from SBMs (Li et al., 2020b). We fit the SupSBM for a set of candidate K values and compute the cross-validation error via the mean squared error in predicting the number of edges and triangles. We choose K as the value that minimizes the cross-validation error.

6. Real Data analysis

This section analyzes four well-known and widely-studied network datasets using the model developed here. We study the fit of the SupSBM to these datasets and compare with three other random graph models: the ER, the SBM, and the SupER, through the parametric bootstrap technique outlined in Section 5. We also use the cross-validation approach to select a model among the two competing models.

6.1 Model fit on data from diverse domains

The four datasets we analyze come from disparate application areas and are described below:

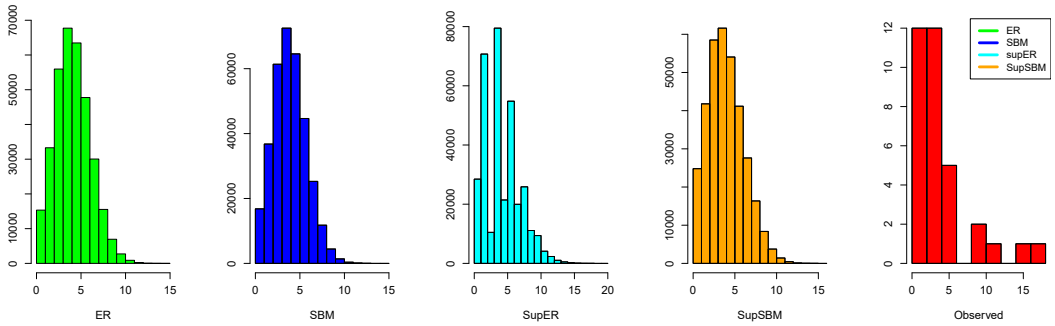
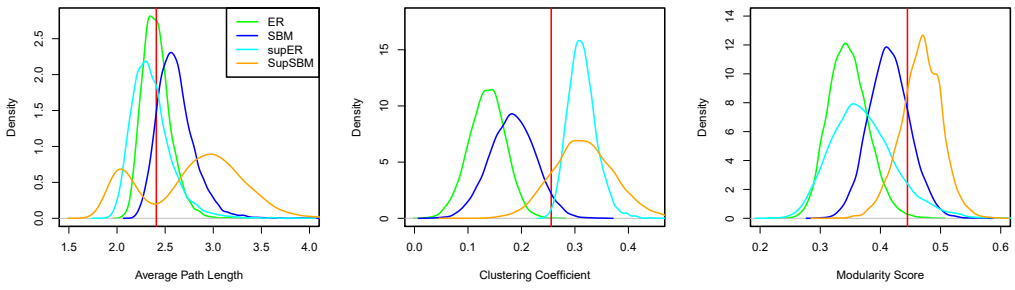


Figure 4: The observed degree distribution in karate club network along with histograms of degrees in simulated networks from various models fitted to the network.

Friendship network: karate club data. The karate club data (Zachary, 1977) is a frequently used benchmark dataset for network community detection (Newman and Girvan, 2004; Bickel and Chen, 2009; Jin, 2015). The network describes friendship patterns of 34 members of a karate club.

Animal social network: dolphin data. This dataset describes an undirected social network involving 62 dolphins in Doubtful Sound, New Zealand, curated by Lusseau et al. (2003). Over the course of the study, the group split into two due to departure of a "well connected" dolphin.

Biological network: neuronal network of *C. Elegans*. This dataset contains the entire connectome or "wiring diagram" of the nervous system of a small nematode called *Caenorhabditis Elegans* (Chen et al., 2006; White et al., 1986; Sohn et al., 2011; Vershynin, 2010). The vertices of the network are the neurons and the edges are synaptic connections among the neurons. We convert the network into an undirected network by assigning an



(a) (b) (c)

Figure 5: Model fit in karate club data: densities of the various metrics in graphs generated from the fitted models, namely, (a) Average path length, (b) Clustering coefficient, (c) Modularity score. The vertical line represents the observed value of the property in the graph.

edge between two vertices if there is an edge between them in either direction. The resulting network contains 297 nodes and 2151 connections.

Web hyperlink network: political blogs data. The political blogs dataset (Adamic and Glance, 2005), collected during the 2004 US presidential election, comprise 1490 political blogs with hyperlinks between them, giving rise to directed edges. This benchmark dataset has been analyzed by a number of authors (Karrer and Newman, 2011; Amini et al., 2013; Qin and Rohe, 2013; Joseph and Yu, 2016; Jin, 2015; Gao et al., 2017; Paul and Chen, 2016) in order to test community detection algorithms. Following previous approaches, we first convert directed edges into undirected edges using the same method described above for *C. Elegans* data, and consider the largest connected component of the resultant graph, which contains 1222 vertices.

In the karate club dataset, all models do equally well in correctly predicting the skewed degree distribution, with SupER also correctly predicting a heavier tail stretching beyond 15. (Figure 4). All models predict the average path length adequately. In terms of clustering coefficients, the ER and SBM generate graphs with lower clustering coefficients, while SupSBM and SupER generate graphs with comparable or higher clustering coefficients. Comparing the densities of the clustering coefficient for graphs generated from various models, the SupSBM appears most appropriate. Finally, in terms of modularity, both ER and SupER models generate graphs with significantly lower modularity than observed, while the SBM and the SupSBM appear to be matching the observed data (Figure 5). The behaviors in terms of the clustering coefficient and modularity are along expected lines since the superimposed models can generate networks with a higher number of triangles, while the models with community structure can generate networks with higher modularity scores.

Next, we investigate the ability of the models to fit the dolphin data in terms of the average path length, clustering coefficient, and modularity in Figure 6. An additional figure containing the degree distributions can be found in the Appendix. Only the SupSBM with $k = 3$ is able to produce graphs with clustering coefficients comparable to the observed

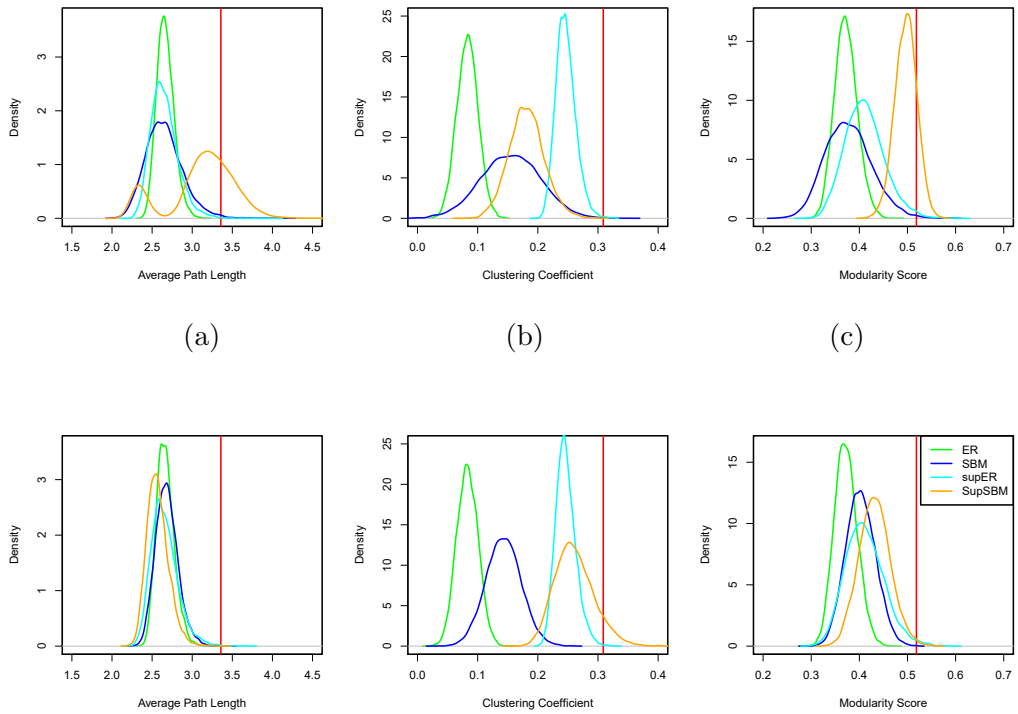
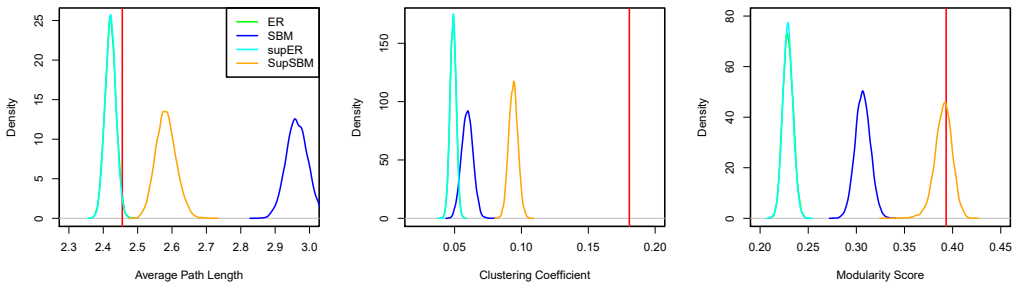


Figure 6: Model fit in dolphin social network data: densities of the various metrics in graphs generated from the fitted models, namely, (a) Average path length, (b) Clustering coefficient, (c) Modularity score. The vertical line represents the observed value of the property in the graph. The first row presents results with $k = 2$ and the second row with $k = 3$ for SBM and SupSBM.

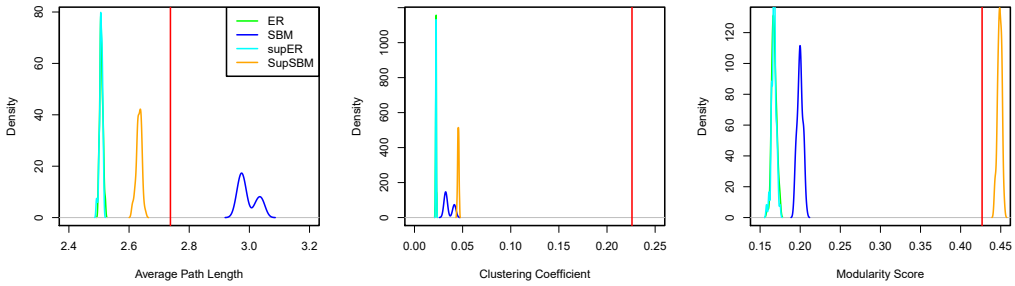
clustering coefficient. The clustering coefficients of graphs from the SupER model are generally closer (even though still lower) to the observed clustering coefficient compared to those from ER and SBM. This observation validates the fact that SupER and SupSBM can account for local motif structures and local clustering better due to the superimposition process. The observed modularity value is predicted very well by the SupSBM with $k = 2$, while the SupSBM with $k = 3$ predicts slightly lower than the observed value. The SupER model produces graphs with modularities that are lower than the observed modularity, as would be expected due to not modeling the community structure. Overall we notice that the SupSBM fits the clustering coefficient and modularity better than other models. We further note that the SupSBM is able to do so without increasing the average path length significantly, which is important for the widely observed network small-world property described earlier.

For the *C. Elegans* neuronal network in Figure 7, we note that SupSBM generates graphs with modularity comparable to observed modularity. The SupSBM also generates graphs



(a) (b) (c)

Figure 7: Model fit in *C. Elegans*: densities of the various metrics in graphs generated from the fitted models, namely, (a) Average path length, (b) Clustering coefficient, (c) Modularity score. The vertical line represents the observed value of the property in the graph. The SBM and SupSBM are fitted with $k = 2$.



(a) (b) (c)

Figure 8: Model fit in political blogs data: densities of the various metrics in graphs generated from the fitted models, namely, (a) Average path length, (b) Clustering coefficient, (c) Modularity score. The vertical line represents the observed value of the property in the graph. The SBM and SupSBM are fitted with $k = 2$.

with both clustering coefficient and average path length closer to the observed values than the SBM. This once again shows the SupSBM is able to model small-world property in networks very well. An additional figure containing the degree distributions can be found in the Appendix.

The political blogs data is perhaps the most challenging dataset for all four models. This is because it is known to have highly skewed and heterogeneous degree distribution that none of the competing models can fit well. We notice this in the degree distribution plots given in the Appendix. We note in Figure 8 that while none of the models fit clustering coefficient well, the SupSBM generates networks with modularity very close to the observed modularity. The SupSBM is also close to a good fit in terms of average path length.

6.1.1 Cross-validation

For each of our datasets, we compare the SBM and SupSBM with a 10-fold cross-validation error. We compute the error with the average of the squared error in predicting the existence of edges and the number of triangles between pairs of vertices. For the purpose of the comparison, we set the number of communities $K = 2$ for both SBM and SupSBM. The results are presented in Table 1. Overall from the table it appears that both SBM and SupSBM are reasonably close in the edge prediction task, but SBM consistently fails in the task of predicting triangles. In particular, while for the task of edge prediction, the error from SupSBM is within 10% to 20% of the error from SBM, for the task of triangle prediction, the error from SBM is often 300% to 2000% higher than the error from SupSBM.

Table 1: 10-fold cross validation mean squared error in predicting existence of edge and number of triangles between an unobserved vertex pair for SBM and SupSBM using low rank network CV method of Li et al. (2020b).

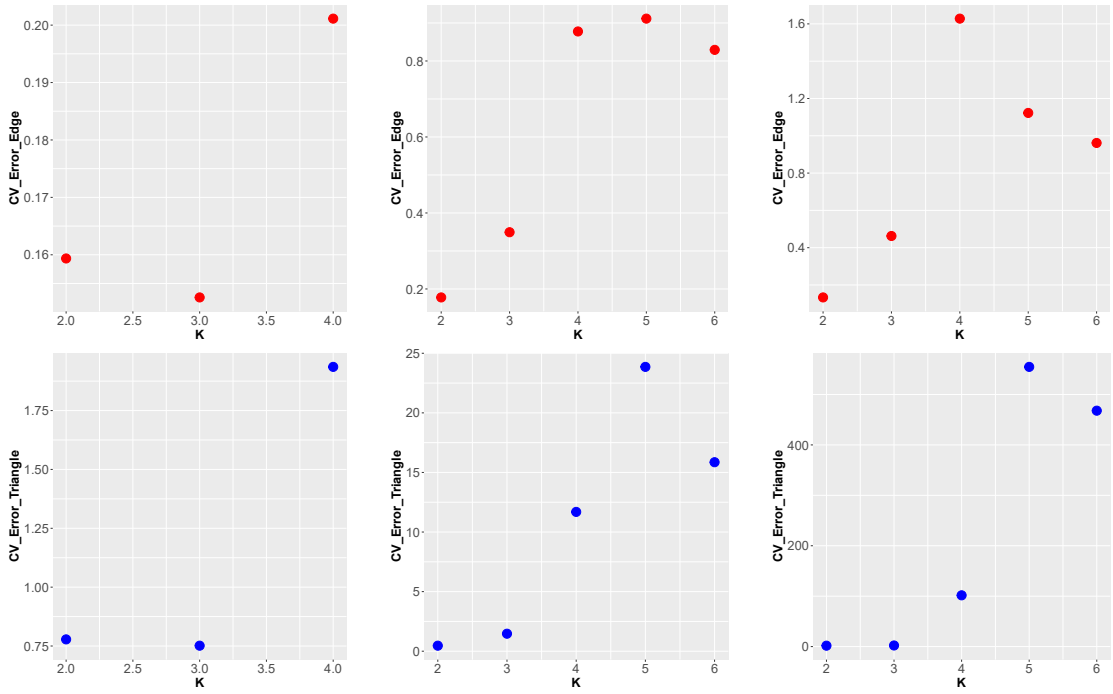
Model	Karate club	Dolphin	*C. Elegans*	Political blogs
Edge				
SBM	0.1400	0.1490	0.1163	0.0690
SupSBM	0.1545	0.1715	0.1463	0.0773
Triangle				
SBM	0.9463	1.8591	46.8680	129.8989
SupSBM	0.7377	0.4642	1.8014	14.6117

6.1.2 Choosing the number of communities

Next, we show how one can choose the number of communities K by using the previously described cross-validation method on real datasets. In Figure 9 we plot the cross-validation error for the edge and triangle prediction metrics for different values of K in the karate club, dolphin and *C. Elegans* datasets. We note that our metrics have the smallest cross-validation error for $K = 3$ (karate club), $K = 2$ (dolphin) and $K = 2$ (*C. Elegans*).

6.1.3 Performance of various spectral clustering algorithms

We test the effectiveness of spectral clustering using a weighted sum of adjacency and Laplacian matrices for higher-order structures on three benchmark network datasets. In particular, we choose to work with a uniformly weighted edge-triangle adjacency matrix, $A_W = A_E + A_T$, where A_E and A_T are the observed edge and triangle adjacency matrices defined earlier. The normalized Laplacian matrix is obtained as $L_w = D_w^{-1/2} A_W D_W^{-1/2}$, where D_W is a diagonal matrix such that $(D_W)_{ii} = \sum_j (A_W)_{ij}$. We compare the performance of various known forms of spectral clustering methods based on edge-based matrices, namely those using adjacency matrices (spA), normalized Laplacian matrices (spL), and regularized normalized Laplacian matrices (rspL) (Sarkar and Bickel, 2015; Chin et al., 2015; Qin and Rohe, 2013) with their weighted higher-order structure counterparts, hospA, hospL and horspL, respectively. In all six instances of the spectral clustering, the eigenvectors are row-



(a) Karate club data (b) Dolphin data (c) *C. Elegans* data

Figure 9: Selecting the number of communities K: cross-validation of edge and triangle errors for different values of K for three datasets. Based on these results, we selected $K = 3$ for karate club data, $K = 2$ for the dolphin data and $K = 2$ for the *C. Elegans* data.

normalized before applying the k-means algorithm. Table 2 summarizes the performance of the methods with respect to known community structures on three of network datasets described earlier.

Table 2: The number of misclustered vertices for various spectral community detection algorithms that use different forms of weighted higher-order matrices. Performance is evaluated based on a known ground truth model.

Dataset	spA	hospA	spL	hospL	rspL	horspL
Political blogs	63	71	588	59	64	64
Karate club	0	0	1	0	0	0
Dolphin	2	2	2	1	2	1

In the political blogs data, we note the hospA and horspL are competitive with the corresponding edge based methods spA and rspL, respectively. However, for spectral clustering based on the normalized Laplacian matrix, the edge-based method spL completely fails to detect the community structure due to well-documented reasons described in Qin and Rohe (2013), Jin (2015), Joseph and Yu (2016), and Gao et al. (2017). On the other hand, hospL succeeds in splitting the graph into two communities with only 59 misclustered vertices. In

the karate club dataset, the method spL misclusters one vertex, while all other methods manage to recover the communities in an error-free manner. In the dolphin dataset, only hospL and horspL miscluster one dolphin, while all the remaining methods miscluster two dolphins.

7. Conclusion and future directions

We proposed and analyzed a superimposed stochastic block model, which is a random graph model that produces networks with properties similar to that observed in real networks. In particular, it can generate sparse networks with short average path lengths, high clustering coefficient, and community structure. Therefore the model produces graphs with the small-world property. To produce the strong clustering property, the model allows for dependencies among the edges yet remains mathematically suitable for the analysis of algorithms. We have extensively tested the fit of the model and compared it to a number of existing random graph models on four datasets from diverse application domains. The model performs better than ER and SBM with respect to several metrics for most of the datasets. However, further extensive simulation and testing on real data is needed to compare the utility of the model when compared to other more sophisticated and potentially harder-to-analyze models for network data, including specialized preferential attachment and latent space models. Our model should be viewed as a step towards creating a more realistic network model while maintaining the relative ease of theoretical analysis of edge random graph models. While not pursued here, a degree correction to the model, similar to that of degree-corrected SBM, may be expected to produce more realistic networks with highly heterogeneous degree distribution and hub nodes, while retaining the aforementioned properties. We hope to extend the model in this direction in our future work.

We have also analyzed the performance of the higher-order spectral clustering algorithm under the proposed SupSBM. This analysis showed that the method is consistent for estimating the community structure in a graph generated from the SupSBM. The consistency property continues to hold even when the community structure is expressed only through the triadic component of the model and not through the dyadic component.

Acknowledgments

We thank Prof. Lutz Warnke of Georgia Institute of Technology for explaining how his work may be applied in parts of the analysis. We also thank Prof. Ji Zhu and Dr. Xianshi Yu from the University of Michigan for their helpful comments and suggestions. The work was supported by the National Science Foundation grants CCF-1527636, DMS-1406455, DMS-1830412, and DMS-2015561, the Center of Science of Information NSF STC center, and an NIH U01 grant for targeted software development.

Appendix

Proof of Lemma 1

Proof We follow and extend the arguments in the proof of a similar result for standard adjacency matrices in Lei and Rinaldo (2015); Gao et al. (2017), and Chin et al. (2015) to the case of triangle-motif adjacency matrices. The arguments in all of the above mentioned papers rely on the use of $\epsilon-$nets on random regular graphs (Friedman et al., 1989; Feige and Ofek, 2005).

Let S denote the unit sphere in the n dimensional Euclidean space. An $\epsilon-$net of the sphere is defined as follows:

$$\mathcal{N} = \{x = (x_1, \ldots, x_n) \in S : \forall i,\, \epsilon\sqrt{n}x_i \in \mathbb{Z}\},$$

where $\mathbb{Z}$ denotes the set of integers. Hence, $\mathcal{N}$ is a set of grid points of size $\frac{1}{\epsilon\sqrt{n}}$ spanning all directions within the unit sphere. For our analysis we only use $\epsilon = 1/2-$nets of spheres and henceforth use $\mathcal{N}$ to denote such nets.

Next, we recall Lemma 2.1 of Lei and Rinaldo (2015) which established that for any $W \in \mathbb{R}^{n\times n}$, one has $\|W\|_2 \leq 4\sup_{x,y\in\mathcal{N}} |x^T W y|$. Hence, a constant-approximation upper bound for $\|A_{T^2} - E[A_{T^2}]\|_2$ may be found by optimizing $|x^T(A_{T^2} - E[A_{T^2}])y|$ over all possible pairs $(x, y) \in \mathcal{N}$. In addition, note that

$$x^T(A_{T^2} - E[A_{T^2}])y = \sum_{i,j} x_i y_j (A_{T^2} - E[A_{T^2}])_{ij} = \sum_{i,j}\sum_{k\neq i,j} x_i y_j (T_{ijk} - E[T_{ijk}]). \quad (7.1)$$

We now divide the pairs (x_i, y_j) into two sets, the set of *light pairs* L and the set of *heavy pairs* H, according to

$$L = \{(i,j) : |x_i y_j| \leq \frac{\sqrt{\Delta_t}}{n}\},$$
$$H = \{(i,j) : |x_i y_j| > \frac{\sqrt{\Delta_t}}{n}\},$$

where Δ_t is as defined in the statement of the theorem.

We bound the term $x^T(A_{T^2} - E[A_{T^2}])y$ separately for the light and heavy pairs, as summarized in the following two lemmas.

Lemma 6 *(Light pairs) For some constant $r_1 > 0$, there exists a constant $c_2(r_1) > 0$, such that with probability at least $1 - \exp(-r_1 n)$,*

$$\sup_{x,y\in T} |\sum_{(i,j)\in L}\sum_k x_i y_j (T_{ijk} - E[T_{ijk}])| < c_2(r_2)\sqrt{\Delta_t}.$$

Whenever clear from the context, we suppress the dependence of the constants on other terms (e.g., $c_2(r_2) = c_2$.)

To obtain a similar bound for heavy pairs, we first note that

$$\sup_{x,y\in T} |\sum_{(i,j)\in H}\sum_k x_i y_j w_{ijk}| \leq \sup_{x,y\in T} |\sum_{(i,j)\in H}\sum_k x_i y_j a_{ijk}| + \sup_{x,y\in T} |\sum_{(i,j)\in H}\sum_k x_i y_j p_{ijk}|. \quad (7.2)$$

The second term can be easily bounded as follows:

$$|\sum_{(i,j)\in H}\sum_{k} x_i y_j p_{ijk}| \leq \sum_{(i,j)\in H}\sum_{k}\frac{x_i^2 y_j^2}{|x_i y_j|}p_{ijk} \leq \frac{n}{\sqrt{\Delta_t}}\sum_{k}\max_{i,j,k}(p_{ijk})\sum_{i,j}x_i^2 y_j^2 \leq \frac{n}{\sqrt{\Delta_t}}\frac{\Delta_t}{n} = \sqrt{\Delta_t}.$$

How to bound the first term is described in the next Lemma 7.

Lemma 7 *For some constant $r_2 > 0$, there exists a constant $c_3(r_2) > 0$ such that with probability at least $1 - n^{-r_2}$, $\sum_{(i,j)\in H}\sum_k x_i y_j T_{ijk} \leq c_3\sqrt{\Delta_t}$.*

Combining the results for the light and heavy pairs, we find that with probability at least $1 - n^{-r}$,

$$\|A_{T^2} - E[A_{T^2}]\|_2 \leq 4\sup_{x,y\in T}|x^T(A_{T^2} - E[A_{T^2}])y| \leq c_1\sqrt{\Delta_t}.$$

This completes the proof of Lemma 1. ■

Proof of Lemma 2

Proof The proof of this result and those of Lemmas 3, 4 and 5 will repeatedly use Theorem 9 of Warnke (2017), which has been reproduced in Proposition 1.

We define the key quantities needed to apply this proposition. Let $I_i = \{E^3_{ijk} = E_{ij}E_{jk}E_{ik},\ (j,k) = \{1,\ldots,n\}^2,\ (j,k) \neq i\}$, denote the collection of indicator random variables for the presence of triangles of type E^3 attached to vertex i. We have an upper bound on the expectation of the sum of these indicator variables as follows:

$$E[\sum_{I_i} E^3_{ijk}] = E[\sum_j\sum_k E_{ij}E_{jk}E_{ik}] \leq n^2(p^e_{\max})^3 \leq \Delta_{E^3}.$$

Clearly, two indicator variables in the set I_i are independent if they do not share any edge indicator random variable. Following the notation of the proposition, we have $E^3_{ijk} \sim E^3_{ij'k}$ for all $j' \neq (i,k)$, since E^3_{ijk} and $E^3_{ij'k}$ share an edge indicator random variable E_{ik}, while $E^3_{ijk} \sim E^3_{ijk'}$ for all $k' \neq (i,j)$, since E^3_{ijk} and $E^3_{ijk'}$ share an edge indicator random variable E_{ij}. We will show that

$$\max_{(i,j,k)\in I_i}\sum_{(i,j',k')\in I_i,\,(i,j,k)\sim(i,j',k')} E^3_{(i,j',k')}, \tag{7.3}$$

i.e., the number of indicators of type E^3 that are realized (i.e., $E^3_{ij'k'} = 1$) and dependent on E^3_{ijk}, is bounded when a "good event" occurs with high probability.

Now we define the "good event" Γ:

$\Gamma = \{$For a vertex pair (i,j), there are at most $C = 4\tau_{\max}$ vertices k'
such that the vertex pairs (i,k') and (j,k') are connected by edges from $G_e\}$,

where $\tau_{\max} = \max\{n(p^e_{\max})^2, \log n\}$ as defined before. Therefore, for any E^3_{ijk}, the good event Γ restricts the number of indicators of type E^3 in the set I_i, which are 1 and are dependent on E^3_{ijk}, to $2C$ as follows. Under the good event Γ, we have

$$\max_{(i,j,k)\in I_i} \sum_{(i,j',k')\in I_i:(i,j',k')\sim(i,j,k)} E^3_{i,j',k'} = \sum_{k'} E^3_{i,j,k'} + \sum_{j'} E^3_{i,j',k} \leq 2C = 8\tau_{\max}.$$

For $t = 8\Delta_{E^3}$, $\mu = \Delta_{E^3}$, Proposition 1 implies

$$\begin{aligned}
P(\sum_{i,j,k\in I_i} E^3_{ijk} \geq 9\Delta_{E^3}) &\leq \min\left\{ \exp\left(-\frac{64\Delta^2_{E^3}}{16\tau_{\max}(\Delta_{E^3} + 8\Delta_{E^3}/3)}\right), \left(1 + \frac{8\Delta_{E^3}}{2\Delta_{E^3}}\right)^{-\frac{8\Delta_{E^3}}{16\tau_{max}}} \right\} \\
&= \min\left\{ \exp\left(-\frac{12\Delta_{E^3}}{11\tau_{\max}}\right),\ 5^{-\Delta_{E^3}/2\tau_{\max}} \right\} \\
&\leq \exp\left(-\frac{12}{11}\log n\right) \\
&= n^{-\frac{12}{11}},
\end{aligned}$$

where the last inequality is a consequence of the following argument. If $\tau_{\max} = n(p^e_{\max})^2$, then $\frac{\Delta_{E^3}}{\tau_{\max}} \geq np^e_{\max} > \log n$ by assumption on $p^e_{\max}$, and if $\tau_{\max} = \log n$, then $\frac{\Delta_{E^3}}{\tau_{\max}} \geq \log n$ by definition of Δ_{E^3}.

Next, from Bernstein inequality and union bound we have,

$$\begin{aligned}
P(\Gamma^C) &\leq n^2 P(\tau_{ij} > 4\tau_{max}) \\
&\leq n^2 P\Big(\sum_{k\neq i,j} (E_{ik}E_{jk} - p^e_{ik}p^e_{jk}) > 3\tau_{max}\Big) \\
&\leq n^2 \exp\left(-\frac{9\tau^2_{\max}}{2\sum_k p^e_{ik}p^e_{jk}(1 - p^e_{ik}p^e_{jk}) + \frac{6}{3}\tau_{\max}}\right) \\
&\leq n^2 \exp\left(-\frac{9\tau^2_{\max}}{2\tau_{\max} + 2\tau_{\max}}\right) \\
&\leq n^2 \exp\left(-\frac{9}{4}\tau_{\max}\right) \\
&\leq \exp\left(-\frac{1}{4}\log n\right) \\
&= n^{-\frac{1}{4}},
\end{aligned}$$

where the last inequality holds since $\tau_{\max} \geq \log n$ by definition.

Then using union bound over all n vertices results in a bound for $\max_i \sum_j (A_{E^3})_{ij}$ with high probability as follows,

$$P(\max_i \sum_j (A_{E^3})_{ij} \geq 9\Delta_{E^3}) \leq n.n^{-\frac{12}{11}} + P(\Gamma^C) \leq n^{-\frac{1}{11}} + n^{-\frac{1}{4}}.$$

This completes the proof of the theorem. ■

Proof of Lemma 3

Proof Recall the definition of the triangle indicator random variable T^3_{ijk}:

$$T^3_{ijk} = 1(\sum_{k_1 \neq k} T_{ijk_1} > 0)1(\sum_{k_2 \neq i} T_{jkk_2} > 0)1(\sum_{k_3 \neq j} T_{ikk_3} > 0). \tag{7.4}$$

For any vertex i, define the degree of i in matrix A_{T^3} according to

$$(d_{T^3})_i = \sum_{j \neq i} \sum_{k \neq (i,j)} T^3_{ijk}.$$

The expectation of the degree may be bounded as

$$\begin{aligned}
E[(d_{T^3})_i] &= E[\sum_{j \neq i} \sum_{k \neq (i,j)} 1(\sum_{k_1 \neq k} T_{ijk_1} > 0)1(\sum_{k_2 \neq i} T_{jkk_2} > 0)1(\sum_{k_3 \neq j} T_{ikk_3} > 0)] \\
&\leq \sum_j \sum_k P(\sum_{k_1 \neq k} T_{ijk_1} > 0)P(\sum_{k_2 \neq i} T_{jkk_2} > 0)P(\sum_{k_3 \neq j} T_{ikk_3} > 0) \\
&\leq \sum_j \sum_k (np^t_{\max})^3 \\
&\leq n^5 (p^t_{\max})^3 \\
&\leq \Delta_{T^3},
\end{aligned}$$

where the second inequality follows since

$$P(\sum_{k_1 \neq k} T_{ijk_1} > 0) \leq P(\cup_{k_1 \neq k} \{T_{ijk_1} = 1\}) \leq \cup_{k_1 \neq k} P(\{T_{ijk_1} = 1\}) \leq np^t_{\max}.$$

Let $I_i = \{(T^3)_{ijk}, j = \{1, \ldots, n\}, k = \{1, \ldots, n\}\}$ denote the set of all triangle indicator random variables incident to vertex i and generated incidentally by three other triangle indicator random variables in G_t according to definition (7.4). Consequently, in the set I_i, two indicators $(T^3)_{ijk}$ and $(T^3)_{ij'k'}$ are dependent if and only if one of the triangle indicators from G_t responsible for the ik or ij "sides" of $(T^3)_{ijk}$, i.e, the sets $\{T_{ikk_3},\, k_3 \neq j\}$ or $\{T_{ijk_1},\, k_1 \neq k\}$ includes j' or k' as a vertex and is consequently part of the indicator $(T^3)_{ij'k'}$ (see Figure 3(b)). We refer to an event corresponding to the above described scenario as TC. Note that this event also accounts for the dependence between $T^3_{ij'k}$ and T^3_{ijk} by letting $k' = k$ and between $T^3_{ijk'}$ and T^3_{ijk} by letting $j' = j$. As in Proposition 1, we use the notation $(i, j, k) \sim (i, j', k')$ to mean the random variable indexed by (i, j, k) is dependent on that indexed by (i, j', k').

We will show that

$$\max_{(i,j,k) \in I_i} \sum_{(i,j',k') \in I_i,\, (i,j,k) \sim (i,j',k')} T^3_{(i,j',k')}, \tag{7.5}$$

i.e., the number of incidentally generated triangle indicators that are realized (i.e., $T^3_{ij'k'} = 1$) and dependent on T^3_{ijk}, is bounded, provided that certain "good events" occur with

high probability. Note any indicator variable $T^3_{ij'k'}$, which is dependent on T^3_{ijk}, can be equivalently written as the following indicator:

$$\{T^3_{ij'k'}|(i,j,k)\sim(i,j',k')\} = T_{ij'j}1(\sum_{k''\neq i} T_{j'k'k''}>0)1(\sum_{k'''\neq j'} T_{ik'k'''}>0) + T_{ij'k}1(\sum_{k''\neq i} T_{j'k'k''}>0)1(\sum_{k'''\neq j'} T_{ik'k'''}>0).$$

Further, define

$$V_{ij'k'} = 1(\sum_{k''\neq i} T_{j'k'k''}>0)1(\sum_{k'''} T_{ik'k'''}>0).$$

Consequently, the inner sum in (7.5) can be written as

$$\sum_{(i,j',k')\in I_i,\,(i,j,k)\sim(i,j',k')} T^3_{(i,j',k')} = 2\sum_{j'k'} T_{ij'j}V_{ij'k'} = 2\sum_{j'} T_{ij'j}\sum_{k'} V_{ij'k'}.$$

Next, we define a "good event" as $\Gamma = \Gamma_1 \cap \Gamma_2$, where Γ_1 and Γ_2 are two events that for any i, j, k may be described as follows:

$$\begin{aligned}\Gamma_1 = \{&\text{For a vertex pair } (i,j), \text{ there are at most } 5V_{\max} \text{ vertices } k'\\ &\text{such that the edges } ik' \text{ and } jk' \text{ are introduced by triangles from } G_t\},\\ \Gamma_2 = \{&\text{The number of triangles in } G_t \text{ sharing an edge } ij \text{ is at most } 3W_{\max}\},\end{aligned}$$

where $V_{\max} = \max\{n^3(p^t_{\max})^2,\ (\log n)^2\}$ and $W_{\max} = \max\{np^t_{\max},\ \log n\}$.

Hence, the event Γ_2 essentially asserts that there are at most $3W_{\max}$ choices for the value of j'. For any choice of j', the event Γ_1 asserts that there are $V_{\max}$ choices for a k'. Consequently, under the "good event" Γ the above sum is upper bounded by $6V_{\max}W_{\max}$.

Recall that the event TC describes the only setting for which two random variables in the set I_i are dependent on each other. Therefore in the notation of Proposition 1, we have $J = I_i$ under the good event Γ. Then

$$\max_{(i,j,k)\in I_i} \sum_{(i,j',k')\in I_i,\,(i,j,k)\sim(i,j',k')} T^3_{(i,j',k')} \le 30V_{\max}W_{\max}, \quad E[\sum_{(i,j,k)\in I_i} T^3_{(i,j,k)}] \le \Delta_{T^3}.$$

Consequently, $\max_J \sum_{\alpha\in J} T^3_\alpha = (d_{T^3})_i$. Applying Proposition 1 for $t = 24\Delta_{T^3}$ leads to

$$\begin{aligned}&P(\sum_{(i,j,k)\in I_i} T^3_{(i,j,k)} \ge 25\Delta_{T^3})\\ \le&\min\left\{\exp\left(-\frac{576\Delta^2_{T^3}}{60V_{\max}W_{\max}(\Delta_{T^3}+24\Delta_{T^3}/3)}\right),\ \left(1+\frac{24\Delta_{T^3}}{2\Delta_{T^3}}\right)^{-\frac{\Delta_{T^3}}{60V_{\max}W_{\max}}}\right\}\\ =&\min\left\{\exp\left(-\frac{576\Delta_{T^3}}{540V_{\max}W_{\max}}\right), 13^{-\Delta_{T^3}/60V_{\max}W_{\max}}\right\}\\ \le&\exp\left(-\frac{576}{540}\log n\right)\\ =&n^{-\frac{16}{15}}.\end{aligned}$$

The last inequality may be established through the following argument. If $W_{\max} = np^t_{\max}$, then $np^t_{\max} \geq \log n$, which implies

$$n^3(p^t_{\max})^2 \geq n^3 \left(\frac{\log n}{n}\right)^2 = n(\log n)^2.$$

Then, $V_{\max} = n^3(p^t_{\max})^2$, and consequently

$$\frac{\Delta_{T^3}}{V_{\max}W_{\max}} = \max\left\{n,\ \frac{(\log n)^4}{n^4(p^t_{\max})^3}\right\} \geq n.$$

On the other hand, if $W_{\max} = \log n$, then $np^t_{\max} < \log n$. Now, either $V_{\max} = (\log n)^2$, in which case $W_{\max}V_{\max} = (\log n)^3$ and $\frac{\Delta_{T^3}}{V_{\max}} \geq \log n$. Or, $V_{\max} = n^3(p^t_{\max})^2$, and consequently $V_{\max}W_{\max} = n^3(p^t_{\max})^2 \log n$. Then

$$\frac{\Delta_{T^3}}{V_{\max}W_{\max}} = \max\left\{\frac{n^2p^t_{\max}}{\log n},\ \frac{(\log n)^4}{n^3(p^t_{\max})^2 \log n}\right\} \geq \log n,$$

since $n^2p^t_{\max} > (\log n)^2$ by assumption.

Next, we need to show that the probability of the "bad event" (i.e., complement of the good event) is exponentially small. For that, we note

$$P(\Gamma^C) = P(\Gamma_1^C \cup \Gamma_2^C) \leq P(\Gamma_1^C) + P(\Gamma_2^C).$$

The last term $P(\Gamma_2^C)$ can be easily bounded using Bernstein's inequality as follows. Let $W_{ij} = \sum_k T_{ijk}$. Then W_{ij} counts the number of triangles in G_t sharing an edge ij. The event Γ_2 asserts that the number of triangles in G_t sharing an edge is at most $3W_{\max} = 3\max\{np^t_{\max}, \log n\}$. From Bernstein's inequality and the union bound we consequently have

$$\begin{aligned}
P(\Gamma_2^C) &\leq n^2 P(W_{ij} > 3W_{max}) \\
&\leq n^2 \exp\left(-\frac{9W^2_{\max}}{2\sum_k p^t_{ijk}(1-p^t_{ijk}) + \frac{6}{3}W_{\max}}\right) \\
&\leq n^2 \exp\left(-\frac{9W^2_{\max}}{2W_{\max} + 2W_{\max}}\right) \\
&\leq n^2 \exp\left(-\frac{9}{4}W_{\max}\right) \\
&\leq \exp\left(-\frac{1}{4}\log n\right) \\
&= n^{-\frac{1}{4}}.
\end{aligned}$$

We now turn our attention to the event Γ_1, which is a bound on $\sum_{k'} V_{ij'k'}$ with i and j' being fixed. Looking at the definition, the sum $\sum_{k'} V_{ij'k'}$ includes dependent random variables; two random variables in the sum, say $V_{ij'k'}$ and $V_{ij'k''}$, are dependent if and only if their expressions contain a common indicator $T_{ik'k''}$ from G_t (i.e., has both ik' and ik'' as tuples, see Figure 10(a)). First, we define $I_{ij'}$ to be the collection of all $V_{ij'k'}$ with fixed

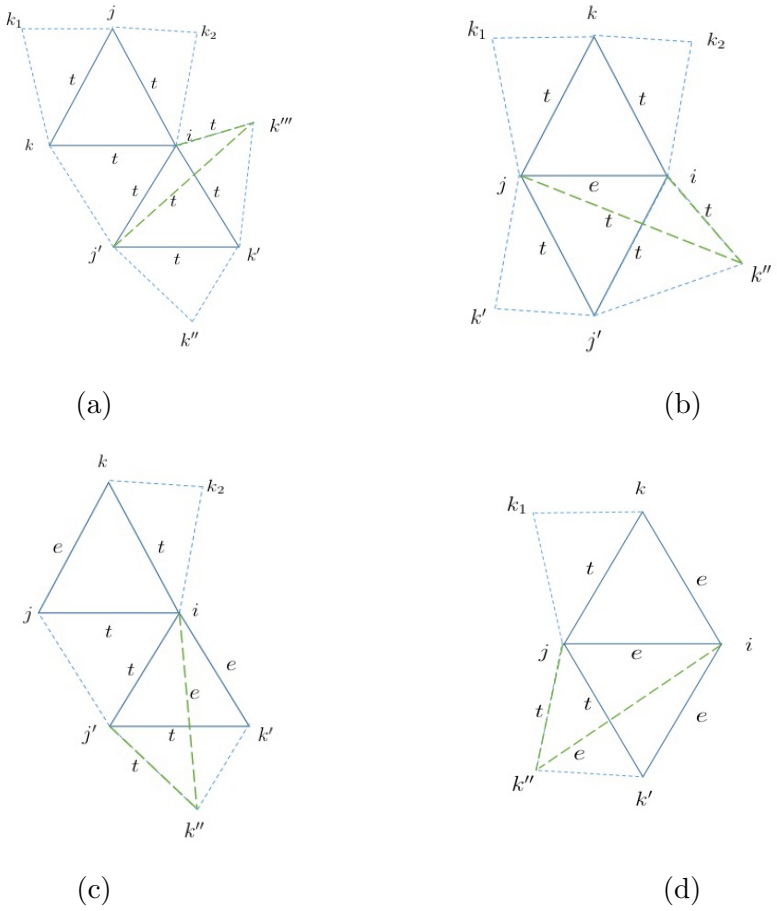


Figure 10: Second-order dependencies that need to be taken into account in the concentration inequalities for "good events": (a) Γ_1 for T^3, (b) Γ_1 for T^2E, (c) Γ_3 for T^2E, and (d) Γ_3 for TE^2.

i and j'. In the notation of Proposition 1, this $I_{ij'}$ is our set $\mathcal{J}$. To apply Proposition 1 to $\sum_{k'} V_{j'k'}$, we first observe that one may upper bound the relevant expectation as

$$E[\sum_{k'} 1(\sum_{k'' \neq i} T_{j'k'k''} > 0) 1(\sum_{k''' \neq j'} T_{ik'k'''} > 0)] \leq n.(np^t_{\max})^2 \leq V_{\max}.$$

Since an indicator $V_{ij'k''}$ in the sum is 1 if we have a k'' such that $T_{ik'k''} = 1$. The number of triangles in G_t with ik' as a side can be bounded by referring to the event Γ_2. Therefore, under the good event Γ_2, the sum over k'' is upper bounded by $3W_{\max}$:

$$\max_{(i,j',k') \in I_{ij'}} \sum_{(i,j',k'') \in I_{ij'}, (i,j',k'') \sim (i,j',k')} V_{ij'k''} \leq 3W_{\max}.$$

Then, with $t = 4V_{\max}$ and $\mu = V_{\max}$, we have

$$P\Big(\sum_{(i,j'k') \in I_{ij'}} V_{ij'k'} \geq 5V_{\max} \Big)$$

$$
\begin{aligned}
&\leq \min\left\{ \exp\left(-\frac{16V_{\max}^2}{6W_{\max}(V_{\max}+4V_{\max}/3)}\right),\ \left(1+\frac{4V_{\max}}{2V_{\max}}\right)^{-\frac{4V_{\max}}{6W_{\max}}} \right\} \\
&= \min\left\{ \exp\left(-\frac{48V_{\max}}{42W_{\max}}\right),\ 3^{-\frac{2V_{\max}}{3W_{\max}}} \right\} \\
&\leq \exp\left(-\frac{8}{7}\log n\right) \\
&= n^{-\frac{8}{7}}.
\end{aligned}
$$

The last inequality holds due to the following argument. If $W_{\max} = np^t_{\max}$, then $p^t_{\max} \geq \frac{\log n}{n}$, and consequently, $n^2 p^t_{\max} \geq n\log n$. Then $\frac{V_{\max}}{W_{\max}} \geq n^2 p^t_{\max} > \log n$. If $W_{\max} = \log n$, then $\frac{V_{\max}}{W_{\max}} \geq \log n$, since $V_{\max} \geq (\log n)^2$.

Now, since there are at most n choices for j', for any i, the union bound leads to

$$P(\Gamma_1^C) \leq nP(V_{ij'} \geq 5V_{\max}) \leq n^{-\frac{1}{7}}.$$

Combining the results we have

$$P((d_{T^3})_i \geq 2\Delta_{T^3}) \leq n^{-\frac{16}{15}} + n^{-\frac{1}{4}} + n^{-\frac{1}{7}}.$$

Invoking the union bound, now over all i, and noting that the event Γ does not depend on i, we can show that $\max_i (d_{T^3})_i \leq c_1\Delta_{T^3}$ with probability at least $1 - n^{-\frac{1}{15}} - n^{-\frac{1}{4}} - n^{-\frac{1}{7}}$. By Equation (4.1), the claimed result holds. ■

Proof of Lemma 4

Proof Triangles of type T^2E are generated by two triangles from G_t and one edge from G_e. Without loss of generality, we may assume that in $(T^2E)_{ijk}$, the sides ij and jk are generated by triangles from G_t and that the side ik is generated by an edge from G_e. Then the corresponding indicator variable for this type of incidental triangle can be written as:

$$T^2E_{ijk} = 1\Big(\sum_{k_1\neq k} T_{ijk_1} > 0\Big)1\Big(\sum_{k_2\neq i} T_{jkk_2} > 0\Big)E_{ik}.$$

Then, we have

$$
\begin{aligned}
E\Big[\sum_j (A_{T^2E})_{ij}\Big] &= E\Big[\sum_j\sum_k (T^2E)_{ijk}\Big] \\
&\leq \sum_j\sum_k P\Big(\sum_{k_1\neq k} T_{ijk_1} > 0\Big)P\Big(\sum_{k_2\neq i} T_{jkk_2} > 0\Big)P(E_{ik}=1) \\
&\leq \sum_j\sum_k (np^t_{\max})^2(p^e_{\max}) \\
&\leq n^4(p^t_{\max})^2 p^e_{\max} \\
&\leq \Delta_{T^2E}.
\end{aligned}
$$

Let the set $I_i = \{(T^2E)_{ijk},\ j = \{1,\dots,n\},\ k = \{1,\dots,n\}\}$ denote the set of all indicator variables for incidentally generated triangles of type T^2E that includes the vertex i. Two random variables in the family may be dependent on two scenarios. One possibility is that the edge indicator E_{ik} is common between $(T^2E)_{(i,j,k)}$ and $(T^2E)_{ikj'}$ for some j' (see Figure 3(c)). The other possibility is that one of the triangle indicators in the sets $\{T_{ijk_1}, k_1 \neq k\}$ or $\{T_{jkk_2}, k_2 \neq i\}$ is also involved in creating $(T^2E)_{ij'k'}$ for some j' and k' (see Figure 3(d)). We refer to these two types of dependencies as TC_1 and TC_2, respectively.

We proceed as in the proof of the previous theorem and describe "good events" under which the sum of random variables that a random variable depends on can be upper bounded with high probability. For this purpose, we characterize TC_1 and TC_2 using indicator variables. First, a $(T^2E)_{ij'k}$ which is dependent on $(T^2E)_{ijk}$ through TC_1 can be represented as

$$\{(T^2E)_{ij'k}|(i,j,k) \overset{TC_1}{\sim} (i,j'k)\} = Q_{j'} = 1\bigg(\sum_{k'\neq k} T_{ij'k'} > 0\bigg)1\bigg(\sum_{k''\neq i} T_{j'kk''} > 0\bigg).$$

With regards to the event TC_2, a $(T^2E)_{ij'k'}$ which is dependent on $(T^2E)_{ijk}$ through TC_2 can be represented as

$$(T^2E)_{ij'k'}|(i,j,k) \overset{TC_2}{\sim} (i,j'k') = R_{j'k'} = T_{ijj'}1\bigg(\sum_{k''\neq i} T_{j'k'k''} > 0\bigg)1(E_{ik'} = 1).$$

At this time, we define the random variable U_{ijk} as follows:

$$U_{ijk} = 1\bigg(\sum_{k'\neq i} T_{jkk'} > 0\bigg)1(E_{ik} = 1).$$

Define a "good event" as $\Gamma = \Gamma_1 \cap \Gamma_2 \cap \Gamma_3$, where Γ_1 and Γ_2 are defined as before and Γ_3 is defined as:

$$\Gamma_3 = \{\text{For a vertex pair } (i,j),\ \text{there are at most } 4U_{\max} \text{ vertices } k,\ \text{such that}$$
$$\text{the edge } ik \text{ arises from } G_e \text{ and edge } jk \text{ arises from a triangle in } G_t,\ \text{i.e., } U_{ijk} = 1\},$$

where $U_{\max} = \max\{n^2 p^t_{\max} p^e_{\max}, (\log n)^2\}$.

Then under Γ_1,

$$\max_{(i,j,k)\in I_i} \sum_{(i,j',k)\in I_i,\,(i,j,k)\overset{TC_1}{\sim}(i,j',k)} (T^2E)_{ij'k} = \sum_{j'} Q_{j'} \leq 5V_{\max}, \tag{7.6}$$

and under Γ_2 and Γ_3,

$$\max_{(i,j,k)\in I_i} \sum_{(i,j',k)\in I_i,\,(i,j,k)\overset{TC_2}{\sim}(i,j',k)} (T^2E)_{ij'k} = 2\sum_{j'}\sum_{k'} R_{j'k'} \leq 30W_{\max}U_{\max}. \tag{7.7}$$

We once again apply Proposition 1 to $\sum_j (A_{T^2E})_{ij}$ under the good event Γ with $J = I_i$ as follows. The upper bound C may be found from

$$\max_{(i,j,k)\in I_i} \sum_{(i,j',k)\in I_i,\,(i,j,k)\sim(i,j',k)} (T^2E)_{ij'k} \leq 5V_{\max} + 30W_{\max}U_{\max}$$

$$\leq 35 \max\{n^3(p^t_{\max})^2,\ n^2 p^t_{\max} p^e_{\max} \log n,\ (\log n)^3\}$$
$$= 35 C_1.$$

Then, $E[\sum_{(i,j,k)\in I_i} T^2 E_{ijk}] \leq \Delta_{T^2E}$, and with $t = 27\Delta_{T^2E}$,

$$\begin{aligned}
&P\Big(\sum_{(i,j,k)\in I_i} T^2 E_{ijk} \geq 28\Delta_{T^2E}\Big) \\
\leq &\min\left\{ \exp\left(-\frac{729\Delta^2_{T^2E}}{70C_1(\Delta_{T^2E} + 27\Delta_{T^2E}/3)}\right),\ \left(1 + \frac{27\Delta_{T^2E}}{2\Delta_{T^2E}}\right)^{\frac{-27\Delta_{T^2E}}{70C_1}} \right\} \\
= &\min\left\{ \exp\left(-\frac{729\Delta_{T^2E}}{700C_1}\right),\ \left(\frac{29}{2}\right)^{-27\Delta_{T^2E}/70C_1} \right\} \\
\leq &\exp\left(-\frac{729}{700}\log n\right) \\
= &n^{-\frac{729}{700}},
\end{aligned}$$

where the last inequality holds due to the following argument. If $C_1 = n^3(p^t_{\max})^2$, then $\frac{\Delta_{T^2E}}{C_1} \geq np^e_{\max}$ which, by assumption, is greater than $\log n$. If $C_1 = n^2 p^t_{\max} p^e_{\max} \log n$, then $\frac{\Delta_{T^2E}}{C} \geq \frac{n^2 p^t_{\max}}{\log n}$ which, by assumption, is greater than $\log n$. Finally, if $C_1 = (\log n)^3$, then $\frac{\Delta_{T^2E}}{C_1} \geq \log n$.

In our previous proofs, we already established upper bounds for $P(\Gamma_1^C)$ and $P(\Gamma_2^C)$. To complete the proof of the claimed result, we only need to determine an upper bound on $P(\Gamma_3^C)$.

Note the event Γ_3 occurs if $\sum_k U_{ijk} \leq U_{\max}$ for any fixed i, j. We further note that the sum $\sum_k U_{ijk}$ includes dependent random variables. An upper bound on the expectation of this sum reads as

$$E(\sum_k U_{ijk}) \leq E[\sum_k 1(\sum_{k'\neq i} T_{jkk'} > 0) 1(E_{ik} = 1)] \leq n^2 p^t_{\max} p^e_{\max} \leq U_{\max}.$$

Fix i and j and define I_{ij} to be the collection of all random variables $U_{ijk}, k = \{1, \ldots, n\}$. Given i and j are fixed, a random variable in the sum $\sum_{k'} U_{ijk'}$ is 1 and also dependent on the indicator U_{ijk} if and only if there is a triangle indicator $T_{jkk'}$ from G_t generates an edge for both the incidental triangles characterized by U_{ijk} and $U_{ijk'}$ (see Figure 4(c)). The set Γ_2 essentially limits the frequency of such observed triangles $T_{jkk'}$s in G_t which has jk as one of the edges. Under the event Γ_2,

$$\max_{(i,j,k)\in I_{ij}} \sum_{(i,j,k')\in I_{ij}:(i,j,k)\sim(i,j,k')} U_{ijk'} \leq 3W_{\max},$$

and for $t = 4U_{\max}$,

$$P(\max \sum_{(ijk)\in I_{ij}} U_{ijk} \geq 5U_{\max})$$

$$
\begin{aligned}
&\leq \min\left\{ \exp\left(-\frac{16U_{\max}^2}{6W_{\max}(U_{\max}+4U_{\max}/3)}\right),\ \left(1+\frac{4U_{\max}}{2U_{\max}}\right)^{\frac{-4U_{\max}}{6W_{\max}}} \right\} \\
&= \min\left\{ \exp\left(-\frac{48U_{\max}}{42W_{\max}}\right),\ 3^{-2U_{\max}/3W_{\max}} \right\} \\
&\leq \exp\left(-\frac{8}{7}\log n\right) \\
&= n^{-\frac{8}{7}},
\end{aligned}
$$

where the last inequality follows since if $W_{\max} = np_{\max}^t$, then $\frac{U_{\max}}{W_{\max}} \geq np_{\max}^e$ which, by assumption, is greater than $c_2 \log n$; and, if $W_{\max} = \log n$, then $\frac{U_{\max}}{W_{\max}} \geq \log n$. Combining the previous results we obtain

$$
P(\max_i \sum_j (A_{T^2E})_{ij}) \geq 28\Delta_{T^2E}) \leq n^{-\frac{29}{700}} + n^{-\frac{1}{4}} + 2n^{-\frac{1}{7}}.
$$

Applying the union bound over all indices i we can bound $\max_i (d_{T^2E})_i \leq c_1 \Delta_{T^2E}$ with probability at least $1 - n^{-c''}$. Then, from Equation (3.3) we arrive at the result claimed in the theorem. ■

PROOF OF LEMMA 5

Proof For incidental triangles of type TE^2, the generating class consists of one triangle from G_t and two edges from G_e. Recall the indicator variable corresponding to TE^2 is

$$
TE^2_{ijk} = 1\Big(\sum_{k_1 \neq k} T_{ijk_1} > 0 \Big) E_{jk} E_{ik},
$$

Consequently, we have

$$
\begin{aligned}
E[(d_{TE^2})_i] &= E\Big[\sum_j \sum_k TE^2_{ijk} \Big] \\
&\leq \sum_j \sum_k P\Big(\sum_{k_1 \neq k} T_{ijk_1} > 0 \Big) P(E_{jk} = 1) P(E_{ik} = 1) \\
&\leq \sum_j \sum_k np_{\max}^t (p_{\max}^e)^2 \\
&\leq n^3 p_{\max}^t (p_{\max}^e)^2 \\
&\leq \Delta_{TE^2}.
\end{aligned}
$$

Next, let $I_i = \{(TE^2)_{ijk}, j = \{1, \ldots, n\}, k = \{1, \ldots, n\}\}$, denote the set of all indicator variables for incidentally generated triangles of type TE^2 including the vertex i. Let $(TE^2)_{ijk}$ be a representative indicator random variable from this set. For another $(TE^2)_{ij'k'}$ in I_i is dependent on $(TE^2)_{ijk}$ in two ways. First, one of the indicators from G_e, say E_{ik}, in TE^2_{ijk}, may also be a side in the incidental triangle characterized by $(TE^2)_{ij'k}$ for some j' (see Figure 3(e)). Second, one of the sides ij may have been created by a triangle indicator

from G_t, with the same triangle indicator being involved in creating the incidental triangle characterized by $(TE^2)_{ij'k'}$ for some j' and k' (see Figure 3(f)). We refer to these two types of dependencies as TC_1 and TC_2, respectively.

With regards to dependencies of type TC_1, define the following random variable:

$$\{(TE^2)_{ij'k}|(i,j,k) \overset{TC_1}{\sim} (i,j'k)\} = K_{j'} = 1\Big(\sum_{k''\neq i} T_{kj'k''} > 0\Big)E_{ik}E_{ij'}.$$

Each $K_{j'}$ characterizes an incidentally generated triangle in I_i which is dependent on $(TE^2)_{ijk}$ through dependency of type TC_1, and, therefore the sum of such indicator random variables is $2\sum_{j'} K_{j'}$ (Figure 3(e)).

With regards to dependencies of type TC_2, define the random variable

$$\{(TE^2)_{ij'k'}|(i,j,k) \overset{TC_2}{\sim} (i,j'k')\} = S_{j'k'} = T_{ijj'}E_{ik'}E_{j'k'}.$$

Each $S_{j'k'}$ characterizes an incidentally generated triangle in I_i with a dependency of type TC_2 with $(TE^2)_{ijk}$. Then, the sum of indicator random variables with TC_2 type of dependency with $(TE^2)_{ijk}$ is given by $\sum_{j'}\sum_{k'} S_{j'k'}$ (Figure 3(f)).

Define a "good event" as $\Gamma = \Gamma_2 \cap \Gamma_3 \cap \Gamma_4$, where Γ_2 and Γ_3 are defined as before and we define Γ_4 as follows:

$$\Gamma_4 = \{\text{Two vertices } \{i,j\} \text{ have at most } 4\tau_{\max} \text{ common neighbors}\{k'\}\},$$

where $\tau_{\max} = \max\{n(p^e_{\max})^2, (\log n)\}$.

We will apply Proposition 1 to $\sum_j (A_{TE^2})_{ij}$ under the good event Γ and obtain an upper bound on $P(\Gamma^C)$. Under the event Γ_3, it holds that

$$\max_{(i,j,k)\in I_i} \sum_{(i,j',k)\in I_i,\,(i,j,k)\overset{TC_1}{\sim}(i,j',k)} (TE^2)_{ij'k} = 2\sum_{j'} K_{j'} \leq 2\sum_{j'} 1\Big(\sum_{k''\neq i} T_{kj'k''} > 0\Big)E_{ij'} \leq 8U_{\max}.$$

Furthermore, under the events Γ_4 and Γ_2, we have

$$\max_{(i,j,k)\in I_i} \sum_{(i,j',k')\in I_i,\,(i,j,k)\overset{TC_2}{\sim}(i,j',k')} (TE^2)_{ij'k} = \sum_{j'}\sum_{k'} S_{j'k'} \leq \sum_{j'} T_{ijj'} \sum_{k'} E_{ik'}E_{j'k'} \leq 12\tau_{\max}W_{\max}.$$

Therefore the upper bound C needed for the proposition may be found according to

$$C = 8U_{\max}+12\tau_{\max}W_{\max} \leq 20\max\{n^2p^t_{\max}p^e_{\max},\ np^t_{\max}\log n,\ n(p^e_{\max})^2\log n,\ (\log n)^2\} = 20C_2.$$

Then, for $t = 9\Delta_{TE^2}$,

$$\begin{aligned}
&P(\max_{(i,j,k)\in I_i} (TE^2)_{ijk} \geq 10\Delta_{TE^2}) \\
&\leq \min\left\{\exp\left(-\frac{81\Delta^2_{TE^2}}{20C_2(\Delta_{TE^2}+9\Delta_{TE^2}/3)}\right),\ \left(1+\frac{9\Delta_{TE^2}}{2\Delta_{TE^2}}\right)^{\frac{-9\Delta_{TE^2}}{40C_2}}\right\} \\
&= \min\left\{\exp\left(-\frac{81\Delta_{TE^2}}{80C_1}\right),\ \left(\frac{11}{2}\right)^{-9\Delta_{TE^2}/40C_2}\right\}
\end{aligned}$$

$$\leq \exp\Big(-\frac{81}{80}\log n\Big)$$
$$= n^{-\frac{81}{80}},$$

where the last inequality follows since if $C = n^2 p^t_{\max} p^e_{\max}$, then $\frac{\Delta_{TE^2}}{C} \geq np^e_{\max}$, which is by assumption greater than $c_2 \log n$; if $C = np^t_{\max} \log n$, then $\frac{\Delta_{TE^2}}{C} \geq \frac{(np^e_{\max})^2}{\log n} \geq \log n$; and if $C = n(p^e_{\max})^2 \log n$, then $\frac{\Delta_{TE^2}}{C} \geq \frac{n^2 p^t_{\max}}{\log n} \geq \log n$. Finally, if $C = (\log n)^2$, then $\frac{\Delta_{TE^2}}{C} \geq \log n$.

We bounded the probability $P(\Gamma_2^C)$ in the proof of Lemma 3 and the probability $P(\Gamma_3^C)$ in the proof of Lemma 4, while a bound on $P(\Gamma_4^C)$ is given in the proof of Lemma 2.

Combining the expressions for all previously evaluated bounds, we obtain

$$P(\max_i \sum_j (A_{TE^2})_{ij} \geq 10\Delta_{TE^2}) \leq n^{-\frac{1}{80}} + 2n^{-\frac{1}{7}} + n^{-1}.$$

Taking the union bound over all i, we can show that $\max_i (d_{TE^2})_i \leq c_1 \Delta_{TE^2}$ holds with probability at least $1 - n^{-c''}$. The claimed result then follows from Equation (3.3). ■

Proof of Theorem 1

Proof We start by noting that combining the results of Lemmas 1 through 5 we have,

$$\begin{aligned}
\|A_T - E[A_T]\|_2 &\leq \|A_{T^2} - E[A_{T^2}]\|_2 + \|A_\Psi - E[A_\Psi]\|_2 \\
&\leq \|A_{T^2} - E[A_{T^2}]\|_2 + \max_i \sum_j \{(A_{E^3})_{ij} + (A_{T^3})_{ij} + (A_{T^2E})_{ij} + (A_{TE^2})_{ij}\} \\
&\quad + \max_i \sum_j E[(A_{E^3})_{ij} + (A_{T^3})_{ij} + (A_{T^2E})_{ij} + (A_{TE^2})_{ij}] \\
&\leq c(\sqrt{\Delta_t} + \Delta_{E^3} + \Delta_{T^3} + \Delta_{T^2E} + \Delta_{TE^2}),
\end{aligned}$$

for a large enough constant c with probability at least $1 - o(1)$.

Now under the given assumptions (4.3) and (4.2) on $p^e_{\max}$ and $p^t_{\max}$, we have the following results:

$$\begin{aligned}
\Delta_{T^3} = \max\{n^5 (p^t_{\max})^3, (\log n)^4\} &\leq \max\{\sqrt{\Delta_t} n^4 (p^t_{\max})^{5/2}, \sqrt{\Delta_t}\} \\
&\leq \max\{\sqrt{\Delta_t} n^{-\frac{5}{2}\epsilon}, \sqrt{\Delta_t}\} \\
&= \sqrt{\Delta_t},
\end{aligned}$$

$$\begin{aligned}
\Delta_{T^2E} = \max\{n^4 (p^t_{\max})^2 p^e_{\max}, (\log n)^4\} &\leq \max\{\sqrt{\Delta_t} n^3 (p^t_{\max})^{3/2} p^e_{\max}, \sqrt{\Delta_t}\} \\
&\leq \max\{\sqrt{\Delta_t} n^{-\frac{5}{2}\epsilon}, \sqrt{\Delta_t}\} \\
&= \sqrt{\Delta_t},
\end{aligned}$$

$$\Delta_{TE^2} = \max\{n^3 (p^t_{\max})(p^e_{\max})^2, (\log n)^3\} \leq \max\{\sqrt{\Delta_t} n^2 (p^t_{\max})^{1/2} (p^e_{\max})^2, \sqrt{\Delta_t}\},$$

$$\leq \max\{\sqrt{\Delta_t} n^{-\frac{5}{2}\epsilon}, \sqrt{\Delta_t}\}$$
$$= \sqrt{\Delta_t}.$$

Consequently,

$$\|A_T - E[A_T]\|_2 \leq \tilde{c}(\sqrt{\Delta_t} + \Delta_{E^3}),$$

with probability at least $1-o(1)$, where $\tilde{c}$ is the maximum of all constants used for bounding the individual matrix terms. If in addition, we assume relationship (4.4), we have

$$\Delta_{E^3} = \max\{n^2(p^e_{\max})^3, (\log n)^2\} \leq \sqrt{\Delta_t},$$

and consequently,

$$\|A_T - E[A_T]\|_2 \leq \tilde{c}_1\sqrt{\Delta_t},$$

with probability at least $1 - o(1)$. ■

Proof of Theorem 2

Proof We use the well-known Davis-Kahan Theorem (Davis and Kahan, 1970; Stewart and Sun, 1990) that characterizes the influence of perturbations on the eigenvectors of a matrix. For a symmetric matrix X, let $\lambda_{\min}(X)$ stand for its smallest (in absolute value) non-zero eigenvalue. Since $\hat{C}_{n\times k}$ is the matrix of eigenvectors it has orthonormal columns, and hence we have the following bound

$$\|\hat{C} - C(C^TC)^{-1/2}\mathcal{O}\|_F^2 \leq 8\frac{k\|A_T - E[A_T]\|_2^2}{(\lambda_{\min}(E[A_T])^2},$$

where $\mathcal{O}$ is an arbitrary orthogonal matrix (Lei and Rinaldo, 2015). Next, from the analysis in Gao et al. (2017), we have the following result relating the misclustering rate of the polynomial time greedy clustering algorithm with the difference between A_T and its expectation:

$$R \leq 64\frac{\|A_T - E[A_T]\|_2^2}{\mu^2(\lambda_{\min}(E[A_T])^2}, \tag{7.8}$$

where $\mu > 0$ is a small constant as in Gao et al. (2017).

■

Proof of Theorem 3

Proof We derive a lower bound on $\lambda_{\min}(E[A_T])$ under this special case. We start by computing the expectations of the motif adjacency matrices A_{T^2}, and A_{E^3} under the SupSBM. In both the cases, these expectations are of the form $C((g-h)I_k + h1_k1_k^T)C^T$, where as before C denotes the community assignment matrix, I_k is the k-dimensional identity matrix, 1_k is the k-dimensional vector of all 1s, and g and h are functions of the parameters n, k, a_e, b_e, a_t, b_t.

For matrices of the form $C((g-h)I_k + h1_k1_k^T)C^T$, with $g > h > 0$, 1_k is an eigenvector corresponding to the eigenvalue $\frac{n}{k}(g-h)+nh$, and the remaining non-zero eigenvalues are of the form $\frac{n}{k}(g-h)$, where the values of g and h differ for the different matrices (Rohe et al., 2011). Since $nh > 0$, the smallest non-zero eigenvalue equals $\frac{n}{k}(g-h)$.

Next, we note that the expected value of A_{T^2} equals $E[A_{T^2}]_{ij} = \sum_{k\neq i,j} p^t_{ijk}$. When $C_i = C_j$, i.e., when the vertices i and j are in the same community, then

$$E[A_{T^2}]_{ij} = \left(\frac{n}{k} - 2\right)\frac{a_t}{n^2} + (k-1)\frac{n}{k}\frac{b_t}{n^2},$$

while when $C_i \neq C_j$,

$$E[A_{T^2}]_{ij} = (n-2)\frac{b_t}{n^2}.$$

The difference between the two above entities equals

$$\left(\frac{n}{k} - 2\right)\frac{a_t}{n^2} + (k-1)\frac{n}{k}\frac{b_t}{n^2} - (n-2)\frac{b_t}{n^2} = \left(\frac{n}{k} - 2\right)\frac{a_t - b_t}{n^2}.$$

Hence,

$$E[A_{T^2}] = C\left(\left(\frac{n}{k} - 2\right)\frac{a_t - b_t}{n^2}I_k + (n-2)\frac{b_t}{n^2}1_k1_k^T\right)C^T.$$

Consequently,

$$\lambda_{\min}(E[A_{T^2}]) = \frac{n}{k}\left(\frac{n}{k} - 2\right)\frac{a_t - b_t}{n^2} = \left(\frac{n}{k} - 2\right)\frac{a_t - b_t}{nk}. \tag{7.9}$$

To determine $E[A_{E^3}]$, we first note that

$$E[A_{E^3}]_{ij} = \sum_{k\neq i,j} p_{ij}p_{jk}p_{ik} = p_{ij}\sum_{k\neq i,j} p_{jk}p_{ik}.$$

When $C_i = C_j$,

$$E[A_{E^3}]_{ij} = \frac{a_e}{n}\left\{\left(\frac{n}{k} - 2\right)\frac{a_e^2}{n^2} + (k-1)\frac{n}{k}\frac{b_e^2}{n^2}\right\},$$

while when $C_i \neq C_j$,

$$E[A_{E^3}]_{ij} = \frac{b_e}{n}\left\{2\left(\frac{n}{k} - 1\right)\frac{a_eb_e}{n^2} + (k-2)\frac{n}{k}\frac{b_e^2}{n^2}\right\}.$$

The difference between the above two probabilities equals

$$\frac{b_e^2(a_e - b_e)}{n^2} + \frac{\left(a_e^2 + a_eb_e - 2b_e^2\right)(a_e - b_e)}{kn^2} - 2\frac{a_e(a_e + b_e)(a_e - b_e)}{n^3}.$$

Hence,

$$E[A_{E^3}] = Z\left(\left(\frac{b_e^2(a_e - b_e)}{n^2} + \frac{\left(a_e^2 + a_eb_e - 2b_e^2\right)(a_e - b_e)}{kn^2} - 2\frac{a_e(a_e + b_e)(a_e - b_e)}{n^3}\right)I_k\right.$$

$$+\frac{b_e}{n}\left(2\Big(\frac{n}{k}-1\Big)\frac{a_e b_e}{n^2}+(k-2)\frac{n}{k}\frac{b_e^2}{n^2}\right)1_k 1_k^T\Bigg)Z^T.$$

Consequently, the smallest non-zero eigenvalue equals

$$\lambda_{\min}(E[A_{E^3}])=\frac{(kb_e^2+a_e^2+a_eb_e-2b_e^2)(a_e-b_e)}{k^2n}-2\frac{a_e(a_e+b_e)(a_e-b_e)}{kn^2}. \tag{7.10}$$

Now note that $E[A_T]=E[A_{T^2}]+E[A_{E^3}]+E[A_{T^2E}]+E[A_{T^3}]+E[A_{TE^2}]$, and all matrices in the sum under the SupSBM model may be written in the form $C((g-h)I_k+y1_k1_k^T)C^T$. Consequently $E[A_T]$ can also be written in the form $C((g-h)I_k+y1_k1_k^T)C^T$. Then, we have $\lambda_{\min}(E[A_T])=\frac{n}{k}(g-h)$ for some g and h. Now note that the $(g-h)$ term in $E[A_T]$ is the sum of the corresponding $(g-h)$ terms in the component matrices, all of which are positive due to the community structure of the SupSBM. Hence, the $(g-h)$ term of $E[A_T]$ is going to be greater than the $(g-h)$ term of $E[A_{T^2}]$, so that $\lambda_{\min}(E[A_T])\geq(\lambda_{\min}(E[A_{T^2}])+\lambda_{\min}(E[A_{E^3}]))$. This implies that we can replace $\lambda_{\min}(E[A_T])$ with $\lambda_{\min}(E[A_{T^2}])+\lambda_{\min}(E[A_{E^3}])$ in the upper bound from Theorem 2.

Next, we note $\Delta_t=n^2p^t_{\max}$. Under the n-vertex k-block balanced SupSBM model, $p^t_{\max}=\frac{a_t}{n^2}$. Therefore, $\Delta_t=a_t$. Similarly, $\Delta_{E^3}=n^2(p^e_{\max})^3$. Under the n-vertex k-block balanced SupSBM model, $p^e_{\max}=\frac{a_e}{n}$. Therefore, $\Delta_{E^3}=\frac{a_e^3}{n}$. Further as $n\to\infty$, we have $\lambda_{\min}(E[A_{T^2}])\asymp\frac{(a_t-b_t)}{k^2}$ and $\lambda_{\min}(E[A_{E^3}])\asymp\frac{(kb_e^2+a_e^2+a_eb_e-2b_e^2)(a_e-b_e)}{k^2n}$. Therefore, we can write the upper bound from Theorem 2 as

$$R_T\lesssim\frac{a_t+\frac{a_e^6}{n^2}}{\left(\frac{(a_t-b_t)}{k^2}+\frac{(kb_e^2+a_e^2+a_eb_e-2b_e^2)(a_e-b_e)}{k^2n}\right)^2}.$$

If we further assume $a_e\asymp b_e$ and $a_t\asymp b_t$, then the above simplifies to

$$R_T\lesssim\frac{a_t+\frac{a_e^6}{n^2}}{\left(\frac{(a_t-b_t)}{k^2}+\frac{b_e^2(a_e-b_e)}{kn}\right)^2}.$$

■

Proof of Corollary 1

Proof The first inequality can be obtained in an analogous manner as (7.8) in the proof of Theorem 2. This inequality relates the misclustering rate R_{T^2} with $\|A_{T^2}-E[A_{T^2}]\|_2$ and $\lambda_{\min}(E[A_{T^2}])$ through the Davis-Kahan Theorem and the analysis of the greedy algorithm in Gao et al. (2017). The second inequality is obtained by replacing the numerator with the bound from Lemma 1 and the denominator with the result computed in the proof of Theorem 3. ■

Proof of Remark 1

Proof We start by analyzing $E[A_{E^2}]$. Clearly,

$$E[A_{E^2}] = C\left(\frac{(a_e - b_e)}{n} I_k + \frac{b_e}{n} 1_k 1_k^T\right) C^T,$$

so that $\lambda_{\min}(E[A_{E^2}]) = \frac{a_e - b_e}{k}$. This implies the error rate for spectral clustering with edges (using the bound from Lei and Rinaldo (2015)) is

$$R_E \lesssim \frac{k^2 a_e}{(a_e - b_e)^2}.$$

Therefore we have the following asymptotic relationship between the two error rates:

$$\frac{k^4 a_t}{(a_t - b_t)^2} \asymp \frac{k^4 a_e/\delta}{\frac{m^2(a_e - b_e)^2}{\delta^2}} \asymp \frac{k^2\delta}{m^2} \frac{a_e}{(a_e - b_e)^2}.$$

Hence, the error rate obtained by using the information about edges is $\frac{k^2\delta}{m^2}$ times that of using triangles. Consequently, the error rate is lower for triangle hyperedges if $\frac{k^2\delta}{m^2} \lesssim 1$ and higher otherwise. ■

Proofs of auxiliary lemmas

Proof of Lemma 6

Proof Define $u_{ij} = x_i y_j 1((i,j) \in L) + x_j y_i 1((j,i) \in L)$ for all $i, j = 1, \ldots, n$. Then,

$$\sum_{(i,j)\in L} \sum_k x_i y_j (T_{ijk} - E[T_{ijk}]) = \sum_{i<j} \sum_k (T_{ijk} - p_{ijk}) u_{ij}.$$

Note that each term in the above sum is a zero-mean random variable bounded in absolute value, $|(T_{ijk} - p_{ijk})u_{ij}| \le 2\sqrt{\Delta_t}/n$. By applying Bernstein's inequality we have

$$\begin{aligned}
&P\left(\Big|\sum_{i<j}\sum_{k\neq(i,j)} (T_{ijk} - p_{ijk})u_{ij}\Big| \ge c_2\sqrt{\Delta_t}\right) \\
\le& 2\exp\left(-\frac{\frac{1}{2}c_2^2\Delta_t}{\sum_{i<j}\sum_{k\neq(i,j)} p_{ijk}(1-p_{ijk})u_{ij}^2 + \frac{1}{3}2\frac{\sqrt{\Delta_t}}{n}c\sqrt{\Delta_t}}\right) \\
\le& 2\exp\left(-\frac{\frac{1}{2}c_2^2\Delta_t}{\max_{i,j}(\sum_{k\neq(i,j)} p_{ijk})\sum u_{ij}^2 + \frac{2}{3}c_2\frac{\Delta_t}{n}}\right) \\
\le& 2\exp\left(-\frac{\frac{1}{2}c_2^2\Delta_t}{\frac{\Delta_t}{n}(2 + \frac{2c_2}{3})}\right) \\
\le& 2\exp\left(-\frac{c_2^2}{4 + \frac{4c_2}{3}} n\right),
\end{aligned}$$

where the third inequality follows as a consequence of two observations. First, since $\Delta_t \geq n^2 \max_{i,j,k} p_{ijk}$, we have

$$\max_{i,j}(\sum_{k\neq(i,j)} p_{ijk}) \leq n \max_{i,j,k} p_{ijk} \leq \frac{\Delta_t}{n}.$$

Second,

$$\sum_{i,j} u_{ij}^2 \leq 2\sum_{i,j}(x_i^2 y_j^2) \leq 2\|x\|_2^2\|y\|_2^2 \leq 2.$$

From Lemma 5 in Vershynin (2010) regarding the covering number of a sphere, we have $|\mathcal{N}| \leq \exp(n\log 5)$. Hence, taking the union bound over all possible x and y we obtain

$$P\Big(\sup_{x,y\in\mathcal{N}}\Big|\sum_{(i,j)\in L}\sum_k x_i y_j(T_{ijk}-E[T_{ijk}])\Big| \geq c_2\sqrt{\Delta_t}\Big) \leq \exp\Big(\Big(-\frac{c_2^2}{4+\frac{4c_2}{3}}+\log 5\Big)n\Big).$$

The claimed result now follows from selecting a sufficiently large constant c_2 and $r_1 = \Big(-\frac{c_2^2}{4+\frac{4c_2}{3}}+\log 5\Big)$. ■

Proof of Lemma 7

Proof We first address the subset of heavy pairs $H_1 = \{(i,j) \in H : x_i > 0, y_j > 0\}$. The other cases may be analyzed similarly.

Define the following two families of sets:

$$I_1 = \Big\{\frac{2^{-1}}{\sqrt{n}} \leq x_i \leq \frac{1}{\sqrt{n}}\Big\}, \quad I_s = \Big\{\frac{2^{s-1}}{2\sqrt{n}} < x_i \leq \frac{2^s}{2\sqrt{n}}\Big\}, \quad s = 2,3,\ldots,\lceil\log_2 2\sqrt{n}\rceil,$$

$$J_1 = \Big\{\frac{2^{-1}}{\sqrt{n}} \leq y_i \leq \frac{1}{\sqrt{n}}\Big\}, \quad J_t = \Big\{\frac{2^{t-1}}{2\sqrt{n}} < y_i \leq \frac{2^t}{2\sqrt{n}}\Big\}, \quad t = 2,3,\ldots,\lceil\log_2 2\sqrt{n}\rceil.$$

Next, for two arbitrary sets I and J of vertices, also define

$$e(I,J) = \begin{cases} \sum_{i\in I}\sum_{j\in J}\sum_{k\neq(i,j)} T_{ijk}, & I\cap J = \emptyset, \\ \sum_{(i,j)\in I\times J\setminus(I\cap J)^2}\sum_{k\neq(i,j)} T_{ijk} + \sum_{(i,j)\in(I\cap J)^2, i<j}\sum_{k\neq(i,j)} T_{ijk}, & I\cap J \neq \emptyset, \end{cases}$$

$$\mu(I,J) = E[e(I,J)], \quad \bar{\mu} = |I||J|n\max_{i,j,k} p_{ijk} \leq |I||J|\frac{\Delta_t}{n},$$

Finally, let $\bar{\mu}_{st} = \bar{\mu}(I_s, J_t)$, $\lambda_{st} = e(I_s,J_t)/\bar{\mu}_{st}$, $\alpha_s = |I_s|2^{2s}/n$, $\beta_t = |J_t|2^{2t}/n$, and $\sigma_{st} = \lambda_{st}\sqrt{\Delta_t}2^{-(s+t)}$.

We have the following two results establishing relationships between the previously introduced entities.

Lemma 8 *Let $d_{t,i} = \sum_j\sum_{k\neq i,j} T_{ijk}$ denote the triangle-degree of vertex i. Then, for all i, and a constant $r_3 > 0$, there exists a constant $c_4(r_3) > 0$ such that $d_{t,i} \leq c_4\Delta_t$ with probability at least $1 - n^{-r_3}$.*

Lemma 9 *For a constant $r_4 > 0$, there exists constants $c_5(r_4), c_6(r_4) > 1$ such that for any pair of vertex sets $I, J \subseteq \{1, \ldots, n\}$ such that $|I| \leq |J|$, with probability at least $1 - 2n^{-r_4}$, at least one of the following statements holds:*

(a) $\frac{e(I,J)}{\bar{\mu}(I,J)} \leq e\, c_5$,

(b) $e(I,J) \log \frac{e(I,J)}{\bar{\mu}(I,J)} \leq c_6\, |J|\, \log \frac{n}{|J|}$.

Now, we use the result of the two previous lemmas to complete the proof of the claimed result for the heavy pairs. We note

$$\sum_{(i,j)\in H_1} x_i y_j \sum_{k\neq(i,j)} T_{ijk} \leq 2 \sum_{(s,t):2^{(s+t)}\geq\sqrt{\Delta_t}} e(I_s, J_t) \frac{2^s}{2\sqrt{n}} \frac{2^t}{2\sqrt{n}} \leq \frac{\sqrt{\Delta_t}}{2} \sum_{(s,t):2^{(s+t)}\geq\sqrt{\Delta_t}} \alpha_s \beta_t \sigma_{st}.$$

We would like to bound the right-hand-side of the inequality by a constant multiple of $\sqrt{\Delta_t}$. To this end, first note the following two facts:

$$\sum_s \alpha_s \leq 4(1/2)^{-2} = 1, \quad \sum_t \beta_t \leq 1.$$

Following the approach of Lei and Rinaldo (2015) and Chin et al. (2015), we split the set of pairs $C : \{(s,t) : 2^{(s+t)} \geq \sqrt{\Delta_t}, |I_s| \leq |J_t|\}$ into six parts and show that desired invariant for each part is bounded.

- $C_1 : \{(s,t) \in C, \sigma_{st} \leq 1\}$:

$$\sum_{(s,t)} \alpha_s \beta_t \sigma_{st} 1\{(s,t) \in C_1\} \leq \sum_{s,t} \alpha_s \beta_t \leq 1.$$

- $C_2 : \{(s,t) \in C \backslash C_1, \lambda_{st} \leq e\, c_5\}$:
  Since
  $$\sigma_{st} = \lambda_{st} \sqrt{\Delta_t} 2^{-(s+t)} \leq \lambda_{st} \leq e\, c_5,$$
  consequently
  $$\sum_{(s,t)} \alpha_s \beta_t \sigma_{st} 1\{(s,t) \in C_2\} \leq e\, c_5 \sum_{s,t} \alpha_s \beta_t \leq e\, c_5.$$

- $C_3 : \{(s,t) \in C \backslash (C_1 \cup C_2), 2^{s-t} \geq \sqrt{\Delta_t}\}$:
  By Lemma 8, $e(I_s, J_t) \leq c_4 |I_s| \Delta_t$. Hence,
  $$\lambda_{st} = e(I_s, J_t)/\bar{\mu}_{st} \leq c_4 \frac{|I_s|\Delta_t}{|I_s||J_t|\Delta_t/n} \leq c_4 \frac{n}{|J_t|},$$
  and consequently,
  $$\sigma_{st} \leq c_4 \sqrt{\Delta_t} 2^{-(s+t)} \frac{n}{|J_t|} \leq c_4 2^{-2t} \frac{n}{|J_t|},$$
  for $(s,t) \in C_3$. Then,
  $$\sum_{(s,t)} \alpha_s \beta_t \sigma_{st} 1\{(s,t) \in C_3\} \leq \sum_s \alpha_s \sum_t \beta_t c_1\, 2^{-2t} \frac{n}{|J_t|}$$

$$\leq \sum_s \alpha_s \sum_t 2^{2t} \frac{|J_t|}{n} c_4\, 2^{-2t} \frac{n}{|J_t|}$$
$$\leq c_4 \sum_s \alpha_s$$
$$\leq c_4.$$

- $C_4 : \{(s,t) \in C \backslash (C_1 \cup C_2 \cup C_3), \log \lambda_{st} > \frac{1}{4}[2t \log 2 + \log(1/\beta_t)]\}$:
  From part (b) of Lemma 9, we have,

  $$\lambda_{st} \log \lambda_{st} \frac{|I_s||J_t|\Delta_t}{n} \leq \frac{e(I_s,J_t)}{\bar{\mu}(I_s,J_t)} \log \frac{e(I_s,J_t)}{\bar{\mu}(I_s,J_t)} \bar{\mu}(I_s,J_t) \leq c_6\, |J_t| \log \frac{2^{2t}}{|J_t|},$$

  which is equivalent to

  $$\sigma_{st}\alpha_s \leq c_6 \frac{1}{\log \lambda_{st}} \frac{2^{s-t}}{\sqrt{\Delta_t}} \{2t \log 2 + \log(1/\beta_t)\} \leq 4\, c_6 \frac{2^{s-t}}{\sqrt{\Delta_t}}.$$

  Then,

  $$\sum_{(s,t)} \alpha_s \beta_t \sigma_{st} 1\{(s,t) \in C_4\} = \sum_t \beta_t \sum_s \sigma_{st}\alpha_s 1\{(s,t) \in C_4\}$$
  $$\leq 4\, c_6 \sum_t \beta_t \sum_s \frac{2^{s-t}}{\sqrt{\Delta_t}} 1\{(s,t) \in C_4\}$$
  $$\leq 8\, c_6 \sum_t \beta_t$$
  $$\leq 8\, c_6.$$

- $C_5 : \{(s,t) \in C \backslash (C_1 \cup C_2 \cup C_3 \cup C_4), 2t \log 2 \geq \log(1/\beta_t)]\}$:
  First, note that since $(s,t) \notin C_4$, we have $\log \lambda_{st} \leq \frac{1}{4}[2t \log 2 + \log(1/\beta_t)] \leq t \log 2$ and hence $\lambda_{st} \leq 2^t$. Next, $\sigma_{st} = \lambda_{st}\sqrt{\Delta_t} 2^{-(s+t)} \leq 2^{-s}\sqrt{\Delta_t}$, and hence $\sigma_{st}\alpha_s \leq 4c_6 \frac{2^{s-t}}{\sqrt{\Delta_t}} 4t \log 2$. Therefore,

  $$\sum_{(s,t)} \alpha_s \beta_t \sigma_{st} 1\{(s,t) \in C_5\} \leq \sum_t \beta_t \sum_s 4\, c_6 \frac{2^{s-t}}{\sqrt{\Delta_t}} 4t \log 2 \leq 2\, c_6 \log 2 \sum_t \beta_t \leq 2\, c_6.$$

- $C_6 : \{(s,t) \in C \backslash (C_1 \cup C_2 \cup C_3 \cup C_4 \cup C_5)\}$:
  Since $2t \log 2 < \log(1/\beta_t)$, we have $\log \lambda_{st} \leq t \log 2 \leq \log(1/\beta_t)/2$. This observation, along with the fact $\lambda_{st} \geq 1$, implies that $\lambda_{st} \leq 1/\beta_t$. As a result,

  $$\sum_{(s,t)} \alpha_s \beta_t \sigma_{st} 1\{(s,t) \in C_6\} \leq \sum_s \alpha_s \sum_t 2^{-(s+t)} \sqrt{\Delta_t} \{(s,t) \in C_6\} \leq \sum_s \alpha_s \leq 2.$$

In a similar fashion, the set of pairs $C : \{(s,t) : 2^{(s+t)} \geq \sqrt{\Delta_t}, |I_s| > |J_t|\}$ is split into six categories in order to bound $\sum_{(s,t)} \alpha_s \beta_t \sigma_{st}$. The derivations are omitted.

Collecting all the previously obtained terms, we arrive at the claimed result for heavy pairs: for some constant $r_2 > 0$, there exists a constant $c_3(r_2) > 0$ such that with probability at least $1 - 2n^{-r_2}$, one has

$$\sum_{(i,j)\in H} \sum_k x_i y_j T_{ijk} \leq c_3 \sqrt{\Delta_t}.$$

■

Proof of Lemma 8

Proof We note $d_{t,i} = \sum_j \sum_k T_{ijk}$ is a sum of independent random variables, each bounded in absolute value by 1. Therefore, Bernstein's inequality gives

$$\begin{aligned} P(d_{t,i} \geq c_4 \Delta_t) &\leq P\Big(\sum_j \sum_k w_{ijk} \geq (c_4 - 1)\Delta_t \Big) \\ &\leq \exp\Big(- \frac{\frac{1}{2}(c_4-1)^2 \Delta_t^2}{\sum_j \sum_k p_{ijk}(1-p_{ijk}) + \frac{1}{3}(c_4-1)\Delta_t} \Big) \\ &\leq \exp\Big(- \Delta_t \frac{3(c_4-1)^2}{2c_4+4} \Big) \\ &\leq n^{-c_7}, \end{aligned}$$

where the last inequality follows since $\Delta_t \geq c \log n$. Taking the union bound over all values of i we obtain that $\max_i d_{t,i} \leq c_4 \Delta_t$ with probability at least $1 - n^{-r_3}$, where c_4 is a function of the constant r_3. ■

Proof of Lemma 9

Proof If $|J| > n/e$, then the result of Lemma 8 implies

$$\frac{e(I,J)}{\Delta_t |I||J|/n} \leq \frac{\sum_{i\in I} \max_i d_{t,i}}{\Delta_t |I|/e} \leq \frac{|I| c_2 \Delta_t}{\Delta_t |I|/e} \leq c_2\, e,$$

and consequently, (a) holds for this case.

If $|J| < n/e$, let $S(I,J) = \{(i,j), i \in I, j \in J\}$. We next invoke Corollary A.1.10 of Alon and Spencer (2004), described below.

Proposition 2 *For independent Bernoulli random variables $X_u \sim Bern(p_u), u = 1, \ldots, n$ and $p = \frac{1}{n}\sum_u p_u$, we have*

$$P(\sum_u (X_u - p_u) \geq a) \leq \exp(a - (a + pn)\log(1 + a/pn)).$$

Using the above result, for $l \geq 8$, we have

$$P(e(I,J) \geq l\bar{\mu}(I,J)) \leq P\Big(\sum_{(i,j)\in S(I,J)} \sum_{k\neq(i,j)} (T_{ijk} - p_{ijk}) \geq l\bar{\mu}(I,J) - \sum_{(i,j)\in S(I,J)} \sum_{k\neq(i,j)} p_{ijk} \Big)$$

$$\leq P\Big(\sum_{(i,j)\in S(I,J)} \sum_{k\neq(i,j)} w_{ijk} \geq (l-1)\bar{\mu}(I,J) \Big)$$
$$\leq \exp\big((l-1)\bar{\mu}(I,J) - l\bar{\mu}(I,J)\log l \big)$$
$$\leq \exp\Big(-\frac{1}{2} l \log l \bar{\mu}(I,J) \Big).$$

For a constant $c_5 > 0$, let

$$t(I,J)\log t(I,J) = \frac{c_5|J|}{\bar{\mu}(I,J)} \log \frac{n}{|J|},$$

and let $l(I,J) = \max\{8, t(I,J)\}$. Then, from the previous calculations, we have

$$P(e(I,J) \geq l(I,J)\bar{\mu}(I,J)) \leq \exp(-\frac{1}{2}\bar{\mu}(I,J)l(I,J)\log l(I,J)) \leq c_3|J| \log \frac{n}{|J|}.$$

From this point onwards identical arguments as those used in Lei and Rinaldo (2015) can be invoked to complete the proof of Lemma 9. ■

7.1 Additional degree distribution figures

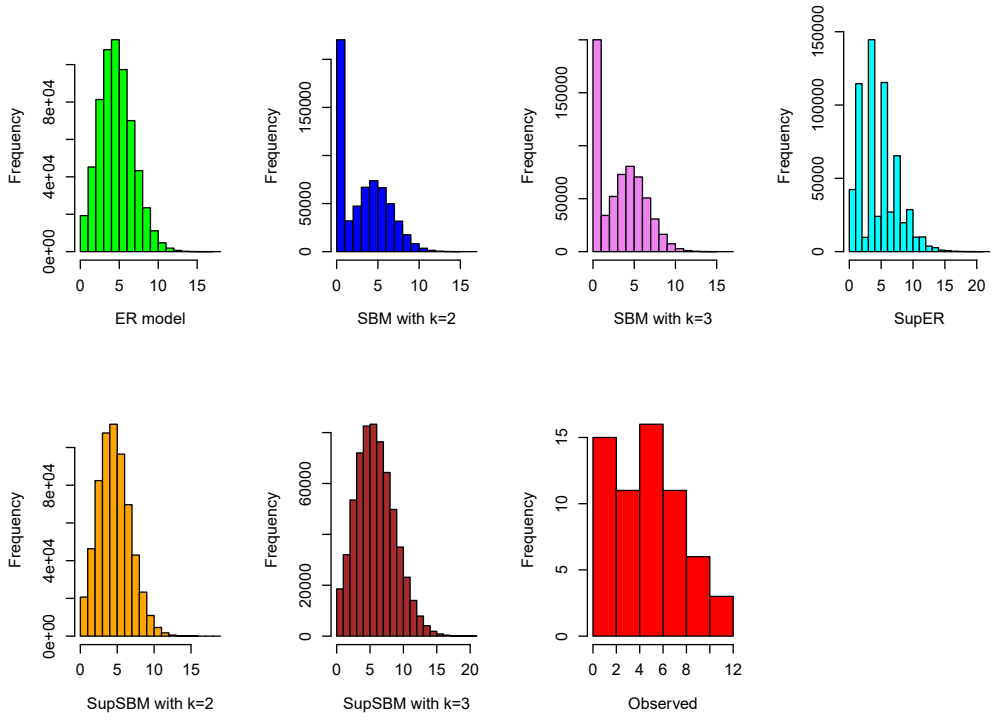


Figure 11: Degree distribution in the dolphin social network.

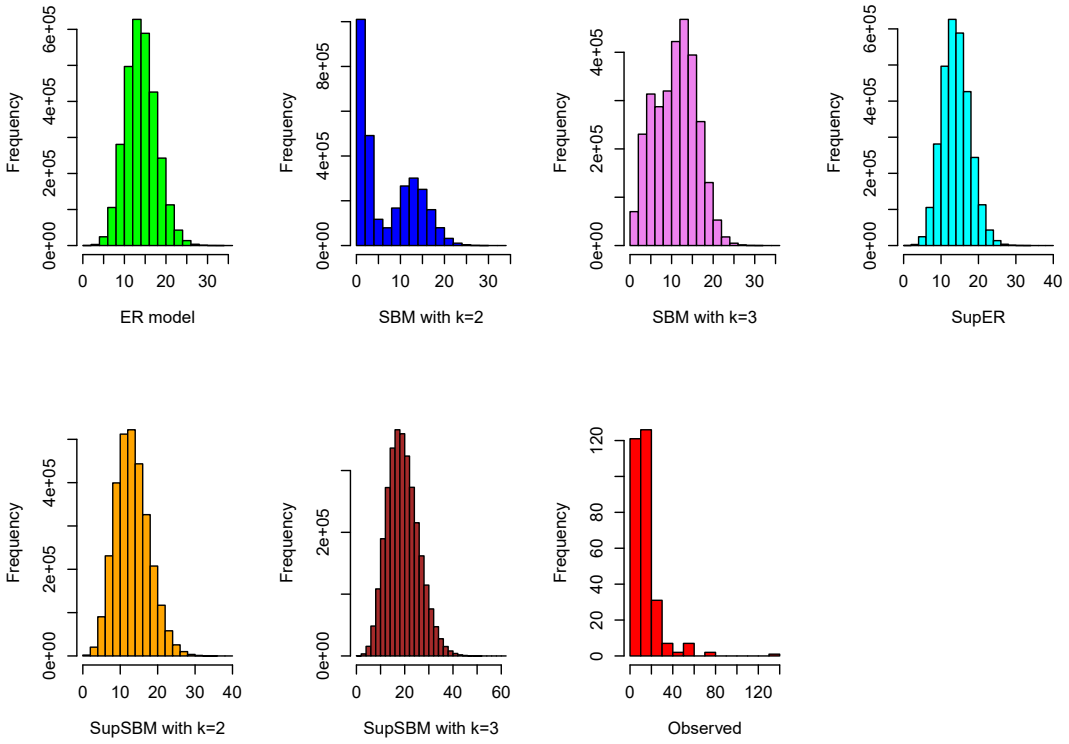


Figure 12: Degree distribution in the *C. Elegans* neuronal network.

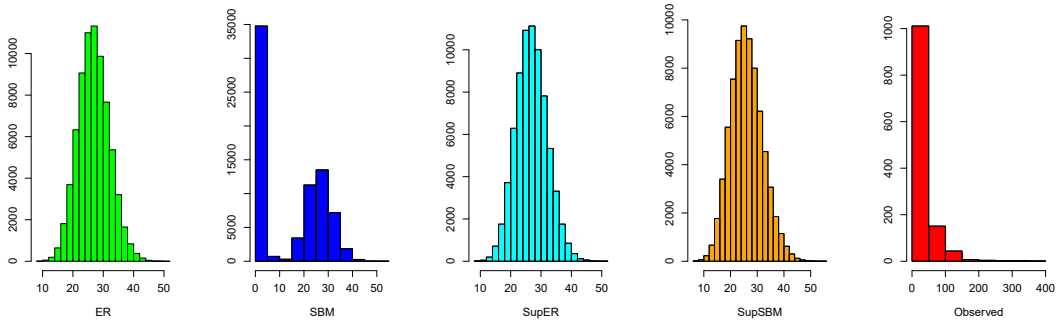


Figure 13: Degree distribution in the political blogs network.

References

Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 670–688. IEEE, 2015.

Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, pages 36–43. ACM, 2005.

Kwangjun Ahn, Kangwook Lee, and Changho Suh. Hypergraph spectral clustering in the weighted stochastic block model. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):959–974, 2018.

Noga Alon and Joel H. Spencer. *The Probabilistic Method*. John Wiley & Sons, 2004.

Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007.

Arash A. Amini, Aiyou Chen, Peter J. Bickel, and Elizaveta Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41 (4):2097–2122, 2013.

Maria Chiara Angelini, Francesco Caltagirone, Florent Krzakala, and Lenka Zdeborov. Spectral detection on sparse hypergraphs. In *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing*, pages 66–73. IEEE, 2015.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

Federico Battiston, Vincenzo Nicosia, Mario Chavez, and Vito Latora. Multilayer motif analysis of brain networks. *Chaos*, 27(4):047404, 2017.

Austin R. Benson, David F. Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

Peter J. Bickel and Aiyou Chen. A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106 (50):21068–21073, 2009.

Peter J. Bickel and Purnamrita Sarkar. Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society, Series B*, 78(1):253–273, 2016.

Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

Béla Bollobás, Svante Janson, and Oliver Riordan. Sparse random graphs with clustering. *Random Structures & Algorithms*, 38(3):269–323, 2011.

Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

Alain Celisse, Jean-Jacques Daudin, and Laurent Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899, 2012.

Andressa Cerqueira and Florencia Leonardi. Estimation of the number of communities in the stochastic block model. *IEEE Transactions on Information Theory*, 66(10):6403–6412, 2020.

Arun Chandrasekhar and Matthew O Jackson. A network formation model based on subgraphs. *arXiv preprint arXiv:1611.07658*, 2016.

Arun G. Chandrasekhar and Matthew O. Jackson. Tractable and consistent random graph models. Technical report, National Bureau of Economic Research, 2014.

Beth L. Chen, David H. Hall, and Dmitri B. Chklovskii. Wiring optimization can relate neuronal structure and function. *Proceedings of the National Academy of Sciences*, 103 (12):4723–4728, 2006.

Kehui Chen and Jing Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 113(521):241–251, 2018.

Yinghan Chen and Yuguo Chen. An efficient sampling algorithm for network motif detection. *Journal of Computational and Graphical Statistics*, 27(3):503–515, 2018.

Eli Chien, Antonia Tulino, and Jaime Llorca. Active learning in the geometric block model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3641–3648, 2020.

I Chien, Chung-Yi Lin, and I-Hsiang Wang. Community detection in hypergraphs: Optimal statistical limit and efficient algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 871–879, 2018.

Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference On Learning Theory (COLT)*, pages 391–423, 2015.

D. S. Choi, P. J. Wolfe, and E. M. Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99:273–284, 2012.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

Paul Erdös and Alfréd Rényi. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(1):17–60, 1960.

Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

Joel Friedman, Jeff Kahn, and Endre Szemeredi. On the second eigenvalue of random regular graphs. In *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, pages 587–598. ACM, 1989.

Sainyam Galhotra, Arya Mazumdar, Soumyabrata Pal, and Barna Saha. The geometric block model. *arXiv preprint arXiv:1709.05510*, 2017.

Sainyam Galhotra, Arya Mazumdar, Soumyabrata Pal, and Barna Saha. Connectivity in random annulus graphs and the geometric block model. *arXiv preprint arXiv:1804.05013*, 2018.

Chao Gao, Zongming Ma, Anderson Y. Zhang, and Harrison H. Zhou. Achieving optimal misclassification proportion in stochastic block models. *The Journal of Machine Learning Research*, 18(60):1–45, 2017.

Debarghya Ghoshdastidar and Ambedkar Dukkipati. Consistency of spectral hypergraph partitioning under planted partition model. *The Annals of Statistics*, 45(1):289–315, 2017.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*, volume 3. JHU Press, 2012.

Bruce Hajek and Suryanarayana Sankagiri. Recovering a hidden community in a preferential attachment graph. *arXiv preprint arXiv:1801.06818*, 2018.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788–2797, 2016.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: some first steps. *Social Networks*, 5:109–137, 1983.

Christopher J. Honey, Rolf Kötter, Michael Breakspear, and Olaf Sporns. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proceedings of the National Academy of Sciences*, 104(24):10240–10245, 2007.

Jiashun Jin. Fast community detection by score. *The Annals of Statistics*, 43(1):57–89, 2015.

Antony Joseph and Bin Yu. Impact of regularization on spectral clustering. *The Annals of Statistics*, 44(4):1765–1791, 2016.

Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

Chiheon Kim, Afonso S. Bandeira, and Michel X. Goemans. Community detection in hypergraphs, spiked tensor models, and sum-of-squares. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 124–128. IEEE, 2017.

Jason M. Klusowski and Yihong Wu. Counting motifs with graph sampling. In *Conference On Learning Theory (COLT)*, pages 1966–2011, 2018.

Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions. In *In Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 454–462. IEEE, 2004.

David Laniado, Yana Volkovich, Karolin Kappler, and Andreas Kaltenbrunner. Gender homophily in online dyadic and triadic relationships. *EPJ Data Science*, 5(1):19, 2016.

Can M. Le and Elizaveta Levina. Estimating the number of communities by spectral methods. *Electronic Journal of Statistics*, 16(1):3315–3342, 2022.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

Pan Li and Olgica Milenkovic. Inhomogoenous hypergraph clustering with applications. In *Advances in Neural Information Processing Systems*, pages 2305–2315, 2017.

Pan Li, Gregory J. Puleo, and Olgica Milenkovic. Motif and hypergraph correlation clustering. *IEEE Transactions on Information Theory*, 66(5):3065–3078, 2019a.

Pei-Zhen Li, Ling Huang, Chang-Dong Wang, and Jian-Huang Lai. EdMot: An edge enhancement approach for motif-aware community detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 479–487, 2019b.

Pei-Zhen Li, Ling Huang, Chang-Dong Wang, Jian-Huang Lai, and Dong Huang. Community detection by motif-aware label propagation. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(2):1–19, 2020a.

Tianxi Li, Elizaveta Levina, and Ji Zhu. Network cross-validation by edge sampling. *Biometrika*, 107(2):257–276, 2020b.

David Lusseau, Karsten Schneider, Oliver J. Boisseau, Patti Haase, Elisabeth Slooten, and Steve M. Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405, 2003.

Shmoolik Mangan and Uri Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985, 2003.

Frank McSherry. Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 529–537. IEEE, 2001.

Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594): 824–827, 2002.

M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2): 167–256, 2003.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.

Mark Newman. *Networks*. Oxford university press, 2018.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 2:849–856, 2002.

Hae-Jeong Park and Karl Friston. Structural and functional brain networks: from connections to cognition. *Science*, 342(6158):1238411, 2013.

Subhadeep Paul and Yuguo Chen. Orthogonal symmetric non-negative matrix factorization under the stochastic block model. *arXiv preprint arXiv:1605.05349*, 2016.

Pavel V. Paulau, Christoph Feenders, and Bernd Blasius. Motif analysis in directed ordered networks and applications to food webs. *Scientific Reports*, 5:11926, 2015.

Mason A. Porter, Jukka-Pekka Onnela, and Peter J. Mucha. Communities in networks. *Notices of the American Mathematical Society*, 56(9):1082–1097, 2009.

Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128, 2013.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Karl Rohe, Tai Qin, and Haoyang Fan. The highest dimensional stochastic blockmodel with a regularized estimator. *Statistica Sinica*, 39(4):1878–1915, 2012.

Martin Rosvall, Alcides V. Esquivel, Andrea Lancichinetti, Jevin D. West, and Renaud Lambiotte. Memory in network flows and its effects on spreading dynamics and community detection. *Nature Communications*, 5:4630, 2014.

Purnamrita Sarkar and Peter J. Bickel. Role of normalization in spectral clustering for stochastic blockmodels. *The Annals of Statistics*, 43(3):962–990, 2015.

Shai S. Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of Escherichia Coli. *Nature Genetics*, 31(1):64–68, 2002.

Tom A. B. Snijders. The statistical evaluation of social network dynamics. *Sociological Methodology*, 31(1):361–395, 2001.

Tom A. B. Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classificaion*, 14:75–100, 1997.

Yunkyu Sohn, Myung-Kyu Choi, Yong-Yeol Ahn, Junho Lee, and Jaeseung Jeong. Topological cluster analysis reveals the systemic organization of the Caenorhabditis elegans connectome. *PloS Computational Biology*, 7(5):e1001139, 2011.

Olaf Sporns and Rolf Kötter. Motifs in brain networks. *PLoS Biology*, 2(11):e369, 2004.

Gilbert W. Stewart and Ji-Guang Sun. *Matrix Perturbation Theory*. Academic Press, Boston, MA., 1990.

Charalampos E. Tsourakakis, Jakub Pachocki, and Michael Mitzenmacher. Scalable motif-aware graph clustering. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1451–1460, 2017.

William G. Underwood, Andrew Elliott, and Mihai Cucuringu. Motif-based spectral clustering of weighted directed networks. *Applied Network Science*, 5(1):1–41, 2020.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, 2007.

Vincent Q. Vu and Jing Lei. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

Lutz Warnke. Upper tails for arithmetic progressions in random subsets. *Israel Journal of Mathematics*, 221(1):317–365, 2017.

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 1998.

J. G. White, E. Southgate, J. N. Thomson, and S. Brenner. The structure of the nervous system of the nematode caenorhabditis elegans: the mind of a worm. *Philosophical Transactions of the Royal Society of London, Series B*, 314:1–340, 1986.

Bowei Yan, Purnamrita Sarkar, and Xiuyuan Cheng. Provable estimation of the number of blocks in block models. In *International Conference on Artificial Intelligence and Statistics*, pages 1185–1194. PMLR, 2018.

Ömer Nebil Yaveroğlu, Noël Malod-Dognin, Darren Davis, Zoran Levnajic, Vuk Janjic, Rasa Karapandza, Aleksandar Stojmirovic, and Nataša Pržulj. Revealing the hidden language of complex networks. *Scientific Reports*, 4:4547, 2014.

Wayne W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, pages 452–473, 1977.

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40: 2266–2292, 2012.

Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems*, pages 1601–1608, 2006.