arXiv:1307.2302v2 [stat.ML] 1 Aug 2013

THE BLESSING OF TRANSITIVITY IN SPARSE AND STOCHASTIC NETWORKS

KARL ROHE, TAI QIN

Department of Statistics, UW-Madison

Abstract. The interaction between transitivity and sparsity, two common features in empirical networks, implies that there are local regions of large sparse networks that are dense. We call this the blessing of transitivity and it has consequences for both modeling and inference. Extant research suggests that statistical inference for the Stochastic Blockmodel is more difficult when the edges are sparse. However, this conclusion is confounded by the fact that the asymptotic limit in all of the previous studies is not merely sparse, but also non-transitive. To retain transitivity, the blocks cannot grow faster than the expected degree. Thus, in sparse models, the blocks must remain asymptotically small.

Previous algorithmic research demonstrates that small "local" clusters are more amenable to computation, visualization, and interpretation when compared to "global" graph partitions. This paper provides the first statistical results that demonstrate how these small transitive clusters are also more amenable to statistical estimation. Theorem 2 shows that a "local" clustering algorithm can, with high probability, detect a transitive stochastic block of a fixed size (e.g. 30 nodes) embedded in a large graph. The only constraint on the ambient graph is that it is large and sparse–it could be generated at random or by an adversary–suggesting a theoretical explanation for the robust empirical performance of local clustering algorithms.

Introduction

Advances in information technology have generated a barrage of data on highly complex systems with interacting elements. Depending on the substantive area, these interacting elements could be metabolites, people, or computers. Their interactions could be represented in chemical reactions, friendship, or some type of communication. Networks (or graphs) describe these relationships. Therefore, the questions about the relationships in these data are questions regarding the structure of networks. Several of these questions are more naturally phrased as questions of inference; they are questions not just about the realized network, but about the mechanism that generated the network. To study questions in graph inference, it is essential to study algorithms under model parameterizations that reflect the fundamental features of the network of interest.

Sparsity and transitivity are two fundamental and recurring features. In sparse graphs, the number of edges in a network is orders of magnitude smaller than the number of possible

Research supported in part by NIH Grant EY09946, NSF Grant DMS-0906818, and grants from WARF.
E-mail address: `name[at]stat.wisc.edu, for karlrohe or qin`.

edges; the average element has 10s or 100s of relationships, even in networks with millions of other elements. Transitivity describes the fact that friends of friends are likely to be friends. The interaction of these simple and localized features has profound implications for determining the set of realistic statistical models. They imply that in large sparse graphs, there are local and dense regions. This is the blessing of transitivity.

One essential inferential goal is the discovery of communities or clusters of highly connected actors. These form essential feature in a multitude of empirical networks, and identifying these clusters helps answer vital scientific questions in many fields. A terrorist cell is a cluster in the communication network of terrorists; web pages that provide hyperlinks to each other form a community that might host discussions of a similar topic; a cluster in the network of biochemical reactions might contain metabolites with similar functions and activities. Several papers, that are briefly reviewed below, have proved theoretical results for various graph clustering algorithms under the Stochastic Blockmodel, a parametric model, where the model parameters correspond to a true partition of the nodes. Often, these estimators are also studied under the exchangeable random graph model, a non-parametric generalization of the Stochastic Blockmodel. The overarching goal of this paper is to show (1) how sparse and transitive models require a novel asymptotic regime and (2) how the blessing of transitivity makes edges become more informative in the asymptote, allowing for statistical inference even when cluster size and expected degrees do not grow with the number of nodes.

The first part of this paper studies how sparsity and transitivity interact in the Stochastic Blockmodel, and more generally, in the exchangeable random graph model. Interestingly, if a Stochastic Blockmodel is both sparse and transitive, then it has small blocks. The second part of this paper (1) introduces an intuitive and fast local clustering technique to find small clusters; (2) proposes the local Stochastic Blockmodel, which presumes a single stochastic block is embedded in a sparse and potentially adversarially chosen network; and (3) proves that if the proposed local clustering technique is initialized with any point in the stochastic block, then it returns the block with high probability. Figure 1 illustrates the types of clusters found by the proposed algorithm; in this case from a social network on epinions.com containing over 76,000 people.

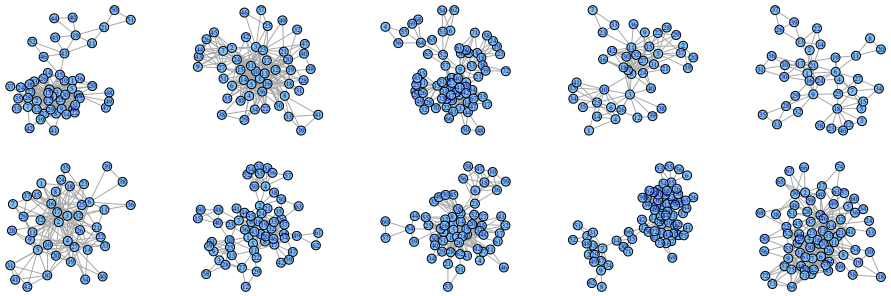

FIGURE 1. Local clusters from a sparse 76k node social network from epinions.com. Created with the igraph library in R [Csardi and Nepusz, 2006].

0.1. **Preliminaries.** Networks, or graphs, are represented by a vertex set and an edge set, $G = (V, E)$, where $V = \{1, \ldots, n\}$ contains the actors and

$$E = \{(i, j) : \text{ there is an edge from } i \text{ to } j\}.$$

The edge set can be represented by the adjacency matrix $A \in \{0, 1\}^{n \times n}$:

$$A_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

This paper only considers undirected graphs. That is, $(i, j) \in E \Rightarrow (j, i) \in E$. The adjacency matrix of such a graph is symmetric. Many of the results in the paper have a simple extension to weighted graphs, where $A_{ij} \geq 0$. For simplicity, we only discuss unweighted graphs. For $i \in V$, let $d_i = \sum_\ell A_{i\ell}$ denote the degree of node i. Define N_i as the neighborhood of node i,

$$N_i = \{j : (i, j) \in E\}.$$

Define the transitivity ratio of A as

$$trans(A) = \frac{\text{number of closed triplets in } A}{\text{number of connected triples of vertices in } A}.$$

Watts and Strogatz [1998] introduced an alternative measure of transitivity, the clustering coefficient. The local clustering coefficient, $C(i)$, is defined as the density of the subgraph induced by N_i, that is the number of edges between nodes in N_i divided by the total number of possible edges. The clustering coefficient for the entire network is the average of these values.

$$C = \frac{1}{n} \sum_i C(i)$$

This is related to the triangles in the graph because an edge (j, k) between two nodes in N_i makes a triangle with node i.

0.2. **Statistical models for random networks.** Suppose $A = \{A_{ij} : i, j \geq 1\}$ is an infinite array that is binary, symmetric, and random. If A is (jointly) exchangeable, that is

$$A \stackrel{d}{=} A_{\sigma,\sigma} = \{A_{\sigma(i),\sigma(j)} : i, j \geq 1\},$$

for any arbitrary permutation σ, then the Aldous-Hoover representation says there exists *i.i.d.* random variables $\xi_1, \xi_2, \ldots$ and an additional independent random variable α such that conditional on these variables, the elements of A are statistically independent [Hoover, 1979, Aldous, 1981, Kallenberg, 2005]. The global parameter α controls the edge density and for ease of notation, it is often dropped [Bickel et al., 2011]. One should think of this result as an extension of DeFinetti's Theorem to infinite exchangeable arrays.

While this representation has only been proven to be equivalent to exchangeability for infinite arrays, it is convenient to adopt this representation for finite graphs.

Definition 1. *Symmetric adjacency matrix $A \in \{0,1\}^{n\times n}$ follows the* ***exchangeable random graph model*** *if there exists i.i.d. random variables $\xi_1, \ldots, \xi_n$ such that probability distribution of A satisfies*

$$P(A|\xi_1, \ldots, \xi_n) = \prod_{i<j} P(A_{ij}|\xi_i, \xi_j).$$

For brevity, we will sometimes refer to this as the exchangeable model.

Independently of the research on infinite exchangeable arrays, Hoff et al. [2002] proposed the latent space model which assumes that (1) each person has a set of latent characteristics (e.g. past schools, current employer, hobbies, etc.) and (2) it is only these characteristics that produce the dependencies between edges. Specifically, conditional on the latent space characteristics, the relationships (or lack thereof) are independent. The Latent Space Model is equivalent to the exchangeable model in Definition 1.

The Stochastic Blockmodel is an exchangeable model that was first defined in Holland and Leinhardt [1983].

Definition 2. *The* ***Stochastic Blockmodel*** *is an exchangeable random graph model with $\xi_1, \ldots, \xi_n \in \{1, \ldots, K\}$ and*

$$P(A_{ij} = 1|\xi_i, \xi_j) = \Theta_{\xi_i,\xi_j}$$

for some $\Theta \in [0,1]^{K\times K}$.

In this model, the ξ_i correspond to group labels. The diagonal elements of Θ correspond to the probability of within-block connections. The off-diagonal elements correspond to the probabilities of between-block connections. When the diagonal elements are sufficiently larger than the off-diagonal elements, then the sampled network will have clusters that correspond to the blocks in the model.

The next subsection briefly reviews the existing literature that examines the consistency of various estimators for the partition created by the latent variables $\xi_1, \ldots, \xi_n$ in the Stochastic Blockmodel.

0.3. **Previous research.** This paper builds on an extensive body of literature examining various types of statistical estimators for the latent partition $\xi_1, \ldots, \xi_n$ in the Stochastic Blockmodel. These estimators fall into four different categories.

(1) Several have studied estimators that are solutions to discrete optimization problems (e.g. Bickel and Chen [2009], Choi et al. [2012], Zhao et al. [2011, 2012], Flynn and Perry [2012]). These objective functions are the likelihood function for the Stochastic Blockmodel or the Newman-Girvan modularity [Newman and Girvan, 2004], a measure that corresponds to cluster quality.

The works cited in this section give a sample of the previous literature on statistical inference for the Stochastic Blockmodel; it is not meant to be an exhaustive list. In particular, there are several highly relevant papers in the Computer Science literatures on (i) the planted partition model and (ii) the planted clique problem. The curious reader should consult the references in Ames and Vavasis [2010] and Chen et al. [2012b].

(2) Others have have studied various approximations to the likelihood that lead to more computationally tractable estimators. For example, Celisse et al. [2011] and Bickel et al. [2012] studied the variational approximation to the likelihood function and Chen et al. [2012a] studied the maximum pseudo-likelihood estimator.
(3) Building on on spectral graph theoretic results [Donath and Hoffman, 1973, Fiedler, 1973], several researchers have studied the statistical performance of spectral algorithms for estimating the partition in the Stochastic Blockmodel [McSherry, 2001, Dasgupta et al., 2004, Giesen and Mitsche, 2005, Coja-Oghlan and Lanka, 2009, Rohe et al., 2011, Rohe and Yu, 2012, Chaudhuri et al., 2012, Jin, 2012, Sussman et al., 2012, Fishkind et al., 2013]. Others have studied estimators that are solutions to semi-definite programs [Ames and Vavasis, 2010, Oymak and Hassibi, 2011, Chen et al., 2012b].
(4) More recently, Bickel et al. [2011], Channarond et al. [2011], Rohe and Yu [2012] have developed methods to stitch together network motifs, or simple "local" measurements on the network, in a way that estimates the partition in the Stochastic Blockmodel. Bickel et al. [2011] draws a parallel between this motif-type of estimator and method of moments estimation.

All of the previous results described above are sensitive to both (a) the population of the smallest block and (b) the expected number of edges in the graph; larger blocks and higher expected degrees lead to stronger conclusions. This limitation arrises because the proofs rely on some form of concentration of measure for a function of sufficiently many variables. Bigger blocks and more edges yield more variables, and thus, more concentration. This paper shows how transitivity leads to a different type of concentration of measure, where each edge becomes asymptotically more informative. As such, the results in this paper extend to blocks of fixed sizes and bounded expected degrees.

Section 2 proposes the `LocalTrans` algorithm that exploits the triangles built by network transitivity. As such, it is most similar to motif-type estimators in bullet (4). Our analysis of `LocalTrans` is the first to study whether a local algorithm (i.e. initialized from a single node) can estimate a block in the Stochastic Blockmodel. The emphasis on local structure aligns with the aims of network scan statistics; these compute a "local" statistic on the subgraph induced by N_i for all i and then return the maximum over all i. In the literature on network scan statistics, Rukhin and Priebe [2012] and Wang et al. [2013] have previously studied the anomaly detection properties under random graph models, including a version of the Stochastic Blockmodel.

1. Transitivity in sparse exchangeable random graph models

In this section, Proposition 1 and Theorem 1 show that previous parameterizations of the sparse exchangeable models and sparse Stochastic Blockmodels lack transitivity in the asymptote. That is, the sampled networks are asymptotically sparse, but they are not asymptotically transitive. Theorem 2 concludes the section by describing a parameterizations that produce sparse *and* transitive networks.

Define

$$p_{\max} = \max_{\xi_i, \xi_j} P(A_{ij} = 1 | \xi_i, \xi_j) \tag{2}$$

as the largest possible probability of an edge under the exchangeable model. In the statistics literature, previous parameterizations of sparse Stochastic Blockmodels, and sparse exchangeable models, have all ensured sparsity by sending $p_{\max} \to 0$.

Define

$$p_\Delta = P(A_{uv} = 1 | A_{iu} = A_{iv} = 1) \tag{3}$$

a population measure of the transitivity in the model. It is the probability of completing a triangle, conditionally on already having two edges.

By sending $p_{\max}$ to zero, a model removes transitivity.

Proposition 1. *Under the exchangeable random graph model* (1)

$$p_\Delta \leq p_{\max},$$

where these probabilities are defined in equations (2) *and* (3).

The next theorem gives conditions that imply the transitivity ratio of the sampled network converges to zero.

Theorem 1. *Under the exchangeable random graph model (Definition 1), define* $\lambda_n = E(d_i)$ *as the expected node degree. If* $\lambda_n \to \infty$, $\lambda_n = o(n)$, *and*

$$p_{\max} = o\left(\frac{P(A_{ij} = 1)}{P(A_{ij} = 1 | A_{i\ell} = 1)}\right), \tag{4}$$

where $p_{\max}$ *is defined in* (2), *then* $trans(A) \xrightarrow{P} 0$.

A proof of this theorem can be found in the appendix.

The denominator on the right hand side of Equation (4) quantifies how many edges are adjacent to the average edge, and thus controls how many 2-stars are in the graph. It can be crudely bounded with the maximum expected degree over the latent ξ_i. For

$$\lambda_n^{\max} = \max_{\xi_i} \; E(d_i | \xi_i),$$

it follows that

$$P(A_{ij} = 1 | A_{i\ell} = 1) \leq \lambda_n^{\max}/n.$$

Corollary 1. *Under the exchangeable random graph model (Definition 1), define* $\lambda_n = E(d_i)$ *as the expected node degree. If* $\lambda_n \to \infty$, $\lambda_n = o(n)$, *and*

$$p_{\max} = o\left(\frac{\lambda_n}{\lambda_n^{\max}}\right),$$

then $trans(A) \xrightarrow{P} 0$.

So, ensuring sparsity by sending $p_{\max}$ to zero removes transitivity both from the model and from the sampled network. The next subsection investigates the implications of restricting $p_{\max} > \epsilon > 0$ in sparse networks.

1.1. **Implications of non-vanishing $p_{\max}$ in the Stochastic Blockmodel.** It is easiest to consider a simplified parameterization of the Stochastic Blockmodel. The following parameterization is also called the planted partition model.

Definition 3. *The **four parameter Stochastic Blockmodel** is a Stochastic Blockmodel with K blocks, exactly s nodes in each block, $\Theta_{ii} = p$, and $\Theta_{ij} = r$ for $i \neq j$.*

In this model, (1) $n = Ks$, (2) the "in-block" probabilities are equal to p, and (3) the "out-of-block" probabilities equal to r. Moreover, the expected degree of each node is

$$\text{Expected degree under the four parameter model} = sp + (n-s)r. \tag{5}$$

Define this quantity as λ_n. Under the four parameter model, p is analogous to $p_{\max}$.

Note that $sp \leq \lambda_n$ and

$$n\frac{p}{\lambda_n} \leq K.$$

In sparse and transitive graphs, λ_n is bounded and p is non-vanishing. In this regime, K grows proportionally to n. The following proposition states this fact in terms of s, the population of each block.

Proposition 2. *Under the four parameter Stochastic Blockmodel, if p is bounded from below, then*

$$s = O(\lambda_n)$$

where s is the population of each block and λ_n is the expected node degree.

The following Theorem shows that graphs sampled from this parameterization are asymptotically transitive.

Theorem 2. *Suppose that A is the adjacency matrix sampled from the four parameter Stochastic Blockmodel (Definition 3) with n nodes. If $p > \epsilon > 0$, $r = O(n^{-1})$, and $s \geq 3$, then as $n \to \infty$*

$$trans(A) \xrightarrow{P} c > 0,$$

where c is a constant which depends on p, r, s.

Remark: If $r = \frac{c_0}{n}$, for some constant c_0, then

$$c \approx \frac{p^3 s^2}{p^2 s^2 + c_0^2 + 2spc_0}.$$

The appendix contains the proof for Theorem 2.

The fixed block size asymptotics in Theorem 2 align with two pieces of previous empirical research suggesting the "best" clusters in massive networks are small. Leskovec et al. [2009] found that in a large corpus of empirical networks, the tightest clusters (as judged by several popular clustering criteria) were no larger than 100 nodes, even though some of the networks had several million nodes. This result is consistent with findings in Physical Anthropology. Dunbar [1992] took various measurements of brain size in 38 different primates and found that the size of the neocortex divided by the size of the rest of the

For ease of exposition, this formula allows self-loops.

brain had a log-linear relationship with the size of the primates natural communities. In humans, the neocortex is roughly four times larger than the rest of the brain. Extrapolating the log-linear relationship estimated from the 38 other primates, Dunbar [1992] suggests that the average human does not have the social intellect to maintain a stable community larger than roughly 150 people (colloquially referred to as Dunbars number). Leskovec et al. [2009] found a similar result in several other networks that were not composed of humans. The research of Leskovec et al. [2009] and Dunbar [1992] suggests that the block sizes in the Stochastic Blockmodel should not grow asymptotically. Rather, block sizes should remain fixed (or grow very slowly).

1.2. **Implications for the exchangeable model.** The interaction between sparsity and transitivity also has surprising implications in the more general exchangeable random graph model. To see this, note that in the exchangeable model, it is sufficient to assume that $\xi_1, \ldots, \xi_n$ are *i.i.d.* Uniform$(0,1)$ [Kallenberg, 2005]. Then, the conditional density of ξ_i and ξ_j given $A_{ij} = 1$ is

$$c(\xi_i, \xi_j) = \frac{P(A_{ij} = 1 | \xi_i, \xi_j)}{P(A_{ij} = 1)} \tag{6}$$

When $p_{\max}$ does not converge to zero, there exist values of ξ_i^* and ξ_j^* such that $P(A_{ij} = 1|\xi_i, \xi_j)$ does not converge to zero. However, in a sparse graph, the edge density $P(A_{ij} = 1)$ (in the denominator of Equation (6)) converges to zero. So, $c(\xi_i^*, \xi_j^*)$ is asymptotically unbounded. For example, in the popular $P(A_{ij} = 1) = O(1/n)$ limit, $c(\xi_i^*, \xi_j^*)$ is proportional to n. In a sense, as a sparse and transitive network grows, each edge becomes more informative. This is the blessing of transitivity in sparse and stochastic networks.

This asymptotic setting, where $p_{\max}$ is bounded from below, makes for an entirely different style of asymptotic proof; the asymptotic power comes from the fact that each edge becomes increasingly informative in the asymptote. Previous consistency proofs rely on concentration of measure for functions of several independent random variables (i.e. several edges). In the sparse and transitive asymptotic setting, concentration follows from the blessing of transitivity, allowing asymptotic results with fixed block sizes and bounded degrees. For example, in Theorems 3 and 4 in the next section, neither the block size nor node degree grows in the asymptote.

2. Local (model + algorithm + results)

This section investigates clustering, or community detection, in sparse and transitive networks. Following the results of the last section, sparse and transitive communities are small. As such, this section is focused on finding small clusters of nodes. In an attempt to strip away as many assumptions as possible from the Stochastic Blockmodel, this section

(1) proposes a "localized" model with a small and transitive cluster embedded in a large and sparse graph (that could be chosen by an adversary),
(2) introduces a novel local clustering algorithm that explicitly leverages the graphs transitivity, and

(3) shows that this local algorithm will discover the cluster in the localized model with high probability.

Similarly to the last section, the interaction between sparsity and transitivity provides for these results, enabling both the fast algorithm and the fixed block asymptotics.

2.1. **The local Stochastic Blockmodel.** The "local" Stochastic Blockmodel (defined below) presumes that a small set of nodes S_* constitute a single block and the model parameterizes how these nodes relate to each other and how they relate to the rest of the network.

Definition 4. *Suppose $A \in \{0,1\}^{(n+s)\times(n+s)}$ is an adjacency matrix on $n+s$ nodes. If there is a set of nodes S_* with $|S_*| = s$ and*

(1) *$i, j \in S_*$ implies $P(A_{ij} = 1) \geq p_{in}$,*
(2) *$i \in S_*$ and $j \in S_*^c$ implies $P(A_{ij} = 1) \leq p_{out}$,*
(3) *the random variables $\{A_{ij} : \forall i \in S_* \text{ and } \forall j\}$ are both mutually independent and also independent of the rest of the graph*

*then A follows the **local Stochastic Blockmodel** with parameters S_*, p_{in}, p_{out}.*

The only assumption that this definition makes about edges outside of S_* (that is, (i,j) with $i, j \notin S_*$) is that they are independent of the edges that connect to at least one node in S_*. So, the edges outside of S_* could be chosen by an adversary, as long as the adversary does not observe the rest of the graph. The theorems below will add an additional assumption that the average degree (within S_*^c) must be not too large (i.e. it must be sparse).

2.2. **Local clustering with transitivity.** A local algorithm searches around a seed node for a tight community that includes this seed node. Several papers have demonstrated the computational advantages of local algorithms for massive networks [Priebe et al., 2005, Spielman and Teng, 2008]. In addition to fast running times and small memory requirements, the local results are often more easily interpretable [Priebe et al., 2005] and yield what appear to be "statistically regularized" results when compared to other, non-local techniques [Leskovec et al., 2009, Clauset, 2005, Liao et al., 2009]. Andersen et al. [2006], Andersen and Chung [2007], Andersen and Peres [2009] have studied the running times and given perturbation bounds showing that local algorithms can approximate the graph conductance. With the exception of Rukhin and Priebe [2012] and Wang et al. [2013], the previous literature has not addressed the statistical properties of local graph algorithms under statistical models.

Given an adjacency matrix A and a seed node i, this section defines an algorithm that finds a clusters around node i. This algorithm has a single tuning parameter *cut* that balances the size of the cluster with the tightness of the cluster. Smaller values of *cut* return looser clusters. The algorithm initializes the cluster with the seed node $S = \{i\}$. It then repeats the following step: For every edge between a node in S ($j \in S$) and a node not in S ($\ell \in S^c$), add ℓ to S if there are at least *cut* nodes that connect to both ℓ and j

(this ensures that (i, j) is contained in at least *cut*-many triangles). Stop the algorithm if all edges across the boundary of S are contained in fewer than *cut*-many triangles.

Algorithm 1 LocalTrans(A, i, cut)

1. Initialize set S to contain node i.
2. For each edge (i, ℓ) on the boundary of S ($i \in S$ and $\ell \notin S$) calculate $T_{i\ell}$:
$$T_{i\ell} = \sum_k A_{ik} A_{k\ell}.$$
3. If there exists any edge(s) (i, ℓ) on the boundary of S with $T_{i\ell} \geq cut$, then add the corresponding node(s) ℓ to S and return to step 2.
4. Return S.

Consider $\texttt{LocalTrans}(A, j, \tau)$ as a function that returns a set of nodes, then

$$i \in \texttt{LocalTrans}(A, j, cut) \implies j \in \texttt{LocalTrans}(A, i, cut).$$

Moreover, if $cut^+ > cut$ then,

$$\texttt{LocalTrans}(A, i, cut^+) \subset \texttt{LocalTrans}(A, i, cut).$$

This shows that the results of $\texttt{LocalTrans}(A, i, cut)$, for every node i and every parameter *cut*, can be arranged into a dendogram. `LocalTrans` only finds one branch of the tree. A simple and fast algorithm can find the entire tree.

To compute the entire dendogram, apply single linkage hierarchical clustering to the similarity matrix

$$T = (AA) \cdot A, \text{ where } \cdot \text{ is element-wise multiplication.} \tag{7}$$

The computational bottleneck of this algorithm is computing T, which can be computed in $O(|E|^{3/2})$. Techniques using fast matrix multiplication can slightly decrease this exponent [Alon et al., 1997].

When A contains no self loops, T_{ij} equals the number of triangles that contain both nodes i and j. We propose *single* linkage because it is the easiest to analyze and it yields good theoretical results. However, in some simulation, *average* linkage has performed better than single linkage. One could also state a local algorithm in terms of average linkage.

Proposition 3. *Viewing* `LocalTrans` *as a function that returns a set of nodes and* `GlobalTrans` *as functions that returns a set of sets,* $\texttt{LocalTrans}(A, i, \tau) \subset \texttt{GlobalTrans}(A, \tau)$. *Moreover,*

$$\bigcup_i \texttt{LocalTrans}(A, i, \tau) = \texttt{GlobalTrans}(A, \tau).$$

This is equivalent to finding the maximum spanning tree

Algorithm 2 GlobalTrans(A, τ)

1. Compute the similarity matrix $T = [AA] \cdot A$, where $\cdot$ is element-wise multiplication.
2. Run single linkage hierarchical clustering on similarity matrix T, i.e. grow a maximum spanning tree.
3. Cut the dendogram at level τ, i.e. delete any edges in the spanning tree with weight smaller than τ.
4. Return the connected components.

Proof. Nodes i and j are in the same cluster in both `LocalTrans`(A, i, τ) and `GlobalTrans`(A, τ) if and only if there exists a path from i to j such that every edge in the path is in at least τ triangles. □

2.3. **Local Inference.** The next theorem shows that `LocalTrans` estimates the local block in the Local Stochastic Blockmodel with high probability.

Theorem 3. *Under the local Stochastic Blockmodel (Definition 4), if*

$$\sum_{i,j \in S_*^c} A_{ij} \leq n\lambda,$$

then

(1) ***cut = 1:*** *for all $i \in S_*$, LocalTrans$(A, i, cut = 1) = S_*$ with probability greater than*

$$1 - \left(\frac{1}{2} s^2 (1 - p_{in}^2)^{s-2} + O(p_{out}^2 ns(s + \lambda)) \right).$$

(2) ***cut = 2:*** *for all $i \in S_*$, LocalTrans$(A, i, cut = 2) = S_*$ with probability greater than*

$$1 - \left(s^3 (1 - p_{in}^2)^{s-3} + O(p_{out}^3 ns(s + \lambda)^2) \right).$$

See the appendix for the proof of this theorem.

For the local Stochastic Blockmodel to create a transitive block with bounded expected degrees, it is necessary for p_{in} to be bounded from below and for s to be bounded from above. Then, $p_{out} = O(1/n)$ is a sufficient condition for bounded expected degrees. However, because of the inequality in bullet (2) of Definition 4, $p_{out} = O(1/n)$ is not a necessary condition for sparsity. As such, the restriction on p_{out} is particularly relevant. It is also fundamental to the bound in Theorem 3.

The approximation terms $O(p_{out}^2 ns(s+\lambda))$ and $O(p_{out}^3 ns(s+\lambda)^2)$ bound the probability of a connection in T across the boundary of S_*. The strength of these terms come from the fact that a connection across the boundary requires $cut + 1$ simultaneous edges across the boundary of S_*. Figure 2 gives a graphical explanation for the "+1". As such, p_{out} is raised to the *cut +1* power. When s and λ are fixed, then $p_{out} = o(n^{-1/2})$ and $p_{out} = o(n^{-1/3})$ make the approximation term asymptotically negligible for the case $cut = 1$ and $cut = 2$, respectively. Both of these settings allow the nodes in S_* have (potentially) large degrees. If $p_{out} = O(n^{-1})$, then the approximation terms become $O(n^{-1})$ and $O(n^{-2})$ respectively.

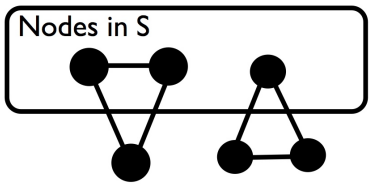

FIGURE 2. This figure illustrates the two types of triangles that contain nodes in both S_* and S_*^c. To make *one* triangle that crosses the boundary of S requires *two* edges to cross the boundary.

Theorem 2 and the algorithms `LocalTrans` and `GlobalTrans` all leverage the interaction between transitivity and sparsity, making the task of computing S and estimating S_* both algorithmically tractable and statistically feasible.

2.4. **Preliminary Data Analysis.** This section applies `GlobalTrans` to an online social network from the website slashdot.org, demonstrating the shortcomings of the proposed algorithms with input A and motivating the next section that uses the graph Laplacian as the input. The slashdot network contains 77 360 nodes with an average degree of roughly 12 [Leskovec et al., 2009]. This network is particularly interesting because it has a smaller transitivity ratio (.024) than the typical social network.

Figure 3 plots the size of the ten largest clusters returned by `GlobalTrans`(A, cut) as a function of *cut* (excluding the largest cluster that consists of the majority of the graph). The values of *cut* range from 3 to 500 and they are plotted on the $\log_e$ scale. Over this range of *cut*, only two times does a cluster exceed ten nodes. While we motivated the local techniques as searching for small clusters, these clusters are perhaps *too* small. It suggests that there are no clusters that are adequately described by the local Stochastic Blockmodel.

One potential reason for this failure is that under the Local Stochastic Blockmodel, the probability of a connection between a node in S_* and a node in S_*^c is uniformly bounded by some value, p_{out}. The slashdot social network, like many other empirical networks, has a long tailed degree distribution. A more realistic model might allow the nodes in S_* to be more highly connected to the high degree nodes in S_*^c. The next subsection (1) proposes a "degree-corrected" local Stochastic Blockmodel, (2) proves that `LocalTrans` with a simple adjustment can estimate S_* in the degree corrected model, and (3) demonstrates how this new version of the algorithm improves the results on the slashdot social network.

3. The Degree-corrected Local Stochastic Blockmodel

Inspired by Karrer and Newman [2011], the degree-corrected model in Definition 5 makes the probability of a connection between a node $i \in S_*$ and a node $j \notin S_*$ scale with the

This data can be downloaded at `http://snap.stanford.edu/data/soc-Slashdot0811.html`

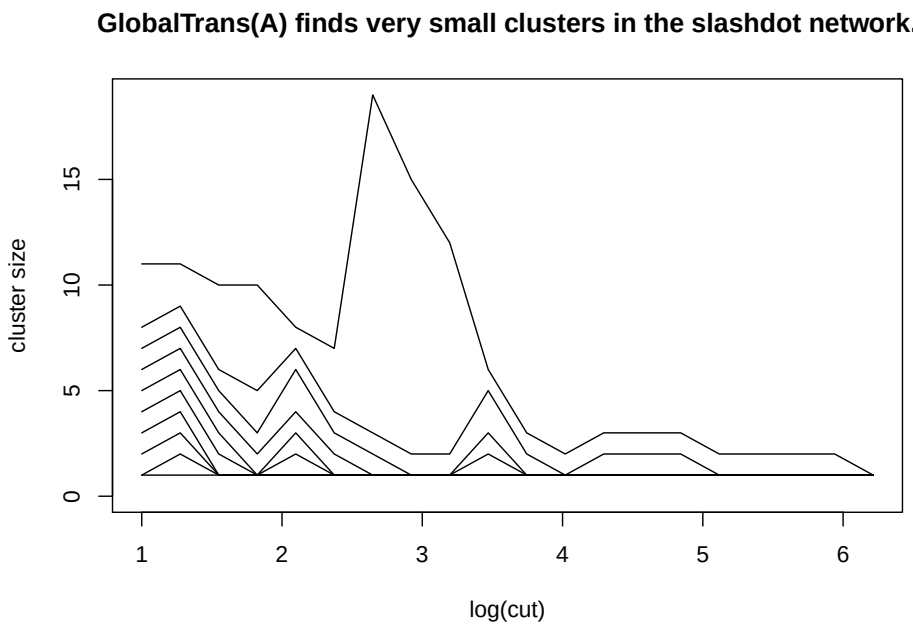

FIGURE 3. This plots the number of nodes in the largest ten clusters (ignoring a single giant cluster) found by `GlobalTrans`(A, cut) in the slashdot social network. These clusters are very small, and probably too small for many applications. Moreover, there are not that many of them.

degree of node j on the subgraph induced by S_*^c.

$$d_j^* = \sum_{\ell \in S_*^c} A_{\ell j} \tag{8}$$

For the following definition to make sense, we presume that d_j^* is fixed for all $j \in S_*^c$.

Definition 5. *Suppose $A \in \{0,1\}^{(n+s)\times(n+s)}$ is an adjacency matrix and S_* is a set of nodes with $|S_*| = s$. For $j \in S_*^c$, define d_j^* as in Equation* (8)*. If*

(1) *$i \in S_*$ and $j \in S_*^c$ implies*

$$P(A_{ij} = 1) \leq \frac{d_j^*}{n},$$

(2) *$i, j \in S_*$ implies $P(A_{ij} = 1) \geq p_{in}$,*
(3) *$\{A_{ij} : \forall j$ and $\forall i \in S_*\}$ are mutually independent*

then A follows the ***local degree-corrected Stochastic Blockmodel*** *with parameters S_*, p_{in}.*

The fundamental difference between the previous local model and this degree corrected version is the assumption that if $i \in S_*$ and $j \in S_*^c$, then

$$P(A_{ij} = 1) \leq \frac{d_j^*}{n}.$$

In the previous model, $P(A_{ij} = 1) \leq p_{out}$. This new condition can be interpreted as $P(A_{ij} = 1) \leq p_{out}\, d_j^*$ for $p_{out} = 1/n$. In this degree-corrected model, the nodes in S_* connect to more high degree nodes than they do under the previous local model.

The degree corrected model creates two types of problems for $\texttt{LocalTrans}(A, i, \tau)$. Because the high degree nodes in S_*^c create many connections to the nodes in S_*, it is more likely to create triangles with two nodes in S_*. Additionally, by definition, the high degree nodes outside of S_* have several neighbors outside of S_*. As such, it is more likely to create triangles with one node in S_* and two nodes outside of S_*. In essence, the high degree nodes create several triangles in the graph, washing out the clusters that $\texttt{LocalTrans}(A, i, \tau)$ can detect. To confront this difficulty, it is necessary to down weight the triangles that contain high degree nodes.

3.0.1. *The graph Laplacian.* Similarly to the adjacency matrix, the normalized graph Laplacian represents the graph as a matrix. In both spectral graph theory and in spectral clustering, the graph Laplacian offers several advantages over the adjacency matrix [Chung, 1997, Von Luxburg, 2007]. The spectral clustering algorithm uses the eigenvectors of the normalized graph Laplacian, not the adjacency matrix, because the normalized Laplacian is robust to high degree nodes [Von Luxburg, 2007].

For adjacency matrix A, define the diagonal matrix D and the normalized graph Laplacian L, both elements of $R^{n\times n}$, in the following way

$$\begin{aligned} D_{ii} &= d(i) \\ L_{ij} &= [D^{-1/2}AD^{-1/2}]_{ij} = \tfrac{A_{ij}}{\sqrt{D_{ii}D_{jj}}}. \end{aligned} \tag{9}$$

Some readers may be more familiar defining L as $I - D^{-1/2}WD^{-1/2}$. For our purposes, it is necessary to drop the $I-$.

The last section utilized the matrix $T = [AA] \cdot A$ to find the triangles in the graph. To confront the degree corrected model, the next theorem uses $[LL] \cdot L$ instead. The interpretation of this matrix is similar to T. It differs because it down weights the contribution of each triangle by the inverse product of the node degrees. For example, where a triangle between nodes i, j, k would add 1 to element T_{ij}, it would add $(d(i)d(j)d(k))^{-1}$ to the i, jth element of $[LL] \cdot L$.

Some versions of spectral clustering use the random walk graph Laplacian, an alternative form of the normalized graph Laplacian.

$$L_{RW} = D^{-1}A$$

While the algorithmic results from spectral clustering can be depend on the choice of graph Laplacian, `LocalTrans` returns exactly the same results with L as it does with L_{RW}. To see this, first imagine that if the graph is directed, then A is asymmetric, and for T to

correspond to *directed* cycles of length three, it is necessary to take the transpose of the final A, that is $[AA] \cdot A^T$. Since L_{RW} is asymmetric, it is reasonable to use the additional transpose from the directed formulation. It is easy to show that

$$[LL] \cdot L = [L_{RW} L_{RW}] \cdot L_{RW}^T. \tag{10}$$

Chaudhuri et al. [2012] and Chen et al. [2012a] have recently proposed a "regularized" graph Laplacian. Chaudhuri et al. [2012] propose replacing D with $D_\tau = D + \tau I$, where $\tau > 0$ is a regularization constant. They show that a spectral algorithm with

$$L_\tau = D_\tau^{-1/2} A D_\tau^{-1/2}$$

has superior performance on sparse graphs. Similarly, it will help to use L_τ with `LocalTrans`. (Note that the equivalence in Equation (10) still holds with the regularized versions of the Laplacians.)

The next theorem shows that under the local degree-corrected model — with the regularized graph Laplacian, a specified choice of tuning parameter *cut*, and $i \in S_*$ — the estimate $\texttt{LocalTrans}(L_\tau, i, cut) = S_*$ with high probability. Importantly, using L_τ instead of A allows for reasonable results under the degree-corrected model.

Theorem 4. *Let A come from the local degree-corrected Stochastic Blockmodel. Define λ such that*

$$\sum_{i,j \in S_*^c} A_{ij} \leq n\lambda. \tag{11}$$

Set $cut = [2(s-1)p_{in} + 2\lambda + \tau]^{-3}$. If

$$n \geq 3\left(2(s-1)p_{in} + 2\lambda + \tau\right)^{3/\epsilon} \tau^{-1/\epsilon},$$

and $s \geq 3$, then for any $i \in S_*$,

$$\texttt{LocalTrans}(L_\tau, i, cut) = S_*$$

with probability at least

$$1 - \left({}^1\!/_2\ s^2 (1 - p_{in}^2)^{s-2} + s \exp\left(-{}^1\!/_4\ s p_{in} + \lambda\right) + O(n^{3\epsilon - 1}) \right).$$

A proof of Theorem 4 can be found in the Appendix.

Because simple summary statistics (of sparsity and transitivity) on empirical networks contradict the types of models studied in the literature, Theorem 4 tries to minimizes the assumptions on the "global" structure of the graph. It only assumes that the graph outside of S_*, i.e. the induced subgraph on S_*^c, is sparse. There are no other assumptions on this part of the graph.

This result is asymptotic in $n = |S_*^c|$, with S_* fixed and containing nodes with bounded expected degree; the assumption in Equation 11 and the definition of d_j^* imply that the nodes in S_* have expected degree less than $s + \lambda$.

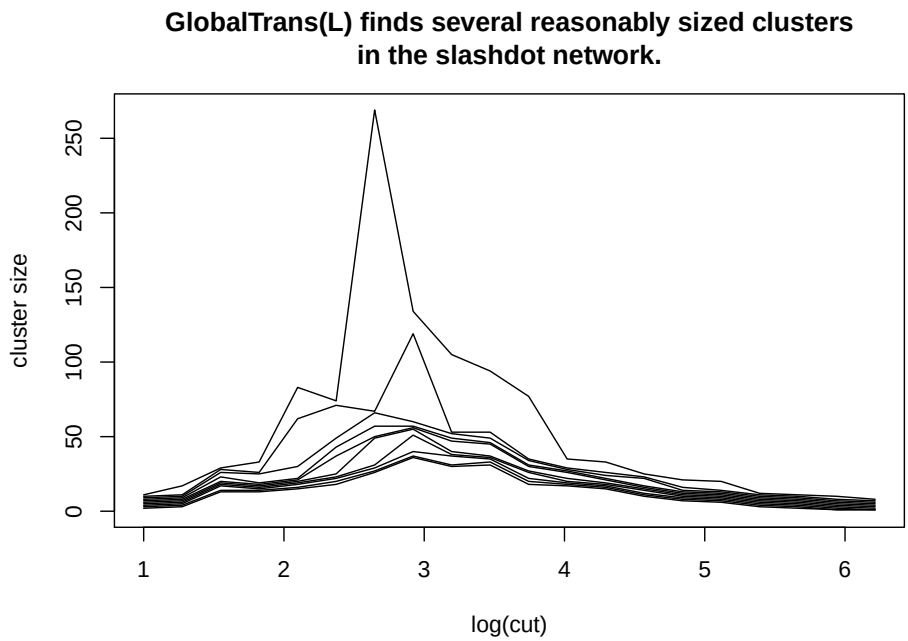

FIGURE 4. A plot of the number of nodes in the largest ten clusters (ignoring one very large cluster) found by `GlobalTrans`(L_τ, cut) in the slashdot social network. `GlobalTrans` with L_τ instead of A finds much larger clusters.

3.1. **Preliminary Data Analysis.** Recall that Figure 3 illustrates how `GlobalTrans`(A, cut) fails to find any clusters larger than twenty nodes in the slashdot social network. Figure 4 shows that using L_τ instead of A corrects for the problems observed in Figure 3. It plots the size of the largest ten clusters in the slashdot social network found by `GlobalTrans`(L_{12}, cut) for values of cut between $3 * 10^{-6}$ and $500 * 10^{-6}$. It finds several clusters that exceed twenty nodes. In this analysis, and all other analyses using L_τ, the regularization constant τ is set equal to the average node degree (as suggested in Chaudhuri et al. [2012]). In this case, $\tau \approx 12$.

Figure 5 shows some of the clusters from the slashdot social network. Specifically, it plots twenty-four of the induced subgraphs from `GlobalTrans`$(L_{\tau=12}, cut = 32 * 10^{-6})$. Because the clusters are not so large, the sub-graphs are easily visualized and it is easy to see how these clusters have several different structures. Some are nearly planar; others appear as densely connected, "clique-like" sub-graphs; other clusters are a collection of several smaller clusters, weakly strung together. Figure 1 in the introduction gives a similar plot for the epinions social network. These visualizations were created using the graph visualization tool in the igraph package in R [Csardi and Nepusz, 2006].

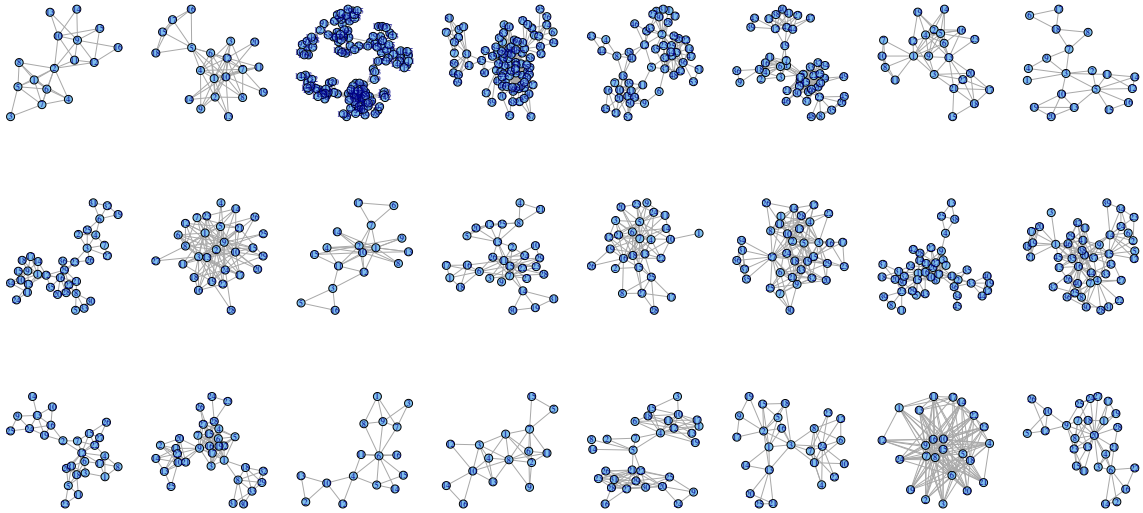

FIGURE 5. Twenty-four small clusters from the slashdot data set. Because `GlobalTrans` discovers small clusters, one can easily plot and visualize the clusters with a standard graph visualization tool [Csardi and Nepusz, 2006]. The point of this figure is to show the variability in cluster structures; some are tight, clique-like clusters; others are small lattice-like clusters; others are "stringy" collections of three or four tight clusters. This highlights the ease of visualizing the results of local clustering.

Figure 6 illustrates how `LocalTrans`$(L_{\tau=12}, i, cut)$ changes as a function of *cut* for a certain node in the epinions social network. Each of the four panels displays the network for $cut = 58 * 10^{-6}$. In each of the four panels, solid nodes are the nodes that are included in `LocalTrans`$(L_{\tau=12}, i, cut)$ for four different values of *cut*. This seed node was selected because the local cluster is slowly growing as *cut* decreases and you can see in this in Figure 6. The left most panel displays the results for the largest value of *cut*. This returns the smallest cluster and not surprisingly, the igraph package plots these nodes in the center of the larger graph. Moving to the right, the clusters grow larger and the additional nodes start to extend to the periphery of the visualization. While the clusters for this node grow slowly, for many other nodes, the transitions are abrupt. For example, the nodes that join the cluster in the last panel in Figure 6 jump from cluster sizes of one or two into this bigger cluster. Then, decreasing *cut* a little bit more, this cluster becomes part of a giant component.

In particular, it was chosen as the "slowest growing" from a randomly chosen set of 200 nodes.

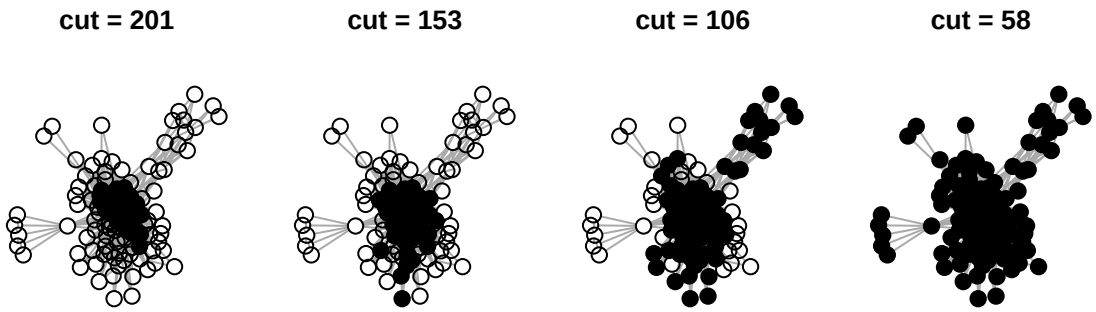

FIGURE 6. Starting from a seed node, this figure demonstrates how `LocalTrans`$(L_{\tau=12}, i, cut)$ grows as *cut* decreases. In each panel, the graph is drawn for the smallest value of *cut*, and the solid nodes correspond to the nodes returned by `LocalTrans`$(L_{\tau=12}, i, cut)$, where the value of *cut* is given above the graph in the units 10^{-6}. Moving from left to right, the clusters grow larger, and the additional nodes start to extend to the periphery of the visualization.

4. Discussion

The tension between transitivity and sparsity in networks that implies that there are local regions of the graph that are dense and transitive. This leads to the blessing of dimensionality, which says that edges (in sparse and transitive graphs) become asymptotically more informative. For example, under the exchangeable model, if the model is sparse and transitive, then the conditional density of the latent variables ξ_i, ξ_j, given $A_{ij} = 1$, is asymptotically unbounded, concentrating on the values of ξ_i, ξ_j that are consistent with the local structure in the model. This has important implications for statistical models, methods, and estimation theory.

In sparse and non-transitive Stochastic Blockmodels, the block structure is *not* revealed in the local structure of the network. Rather, the blocks are revealed by comparing the edge density of various partitions. However, under transitive models, the local structure of the network can reveal the block structure. As such, these blocks can be estimated by fast local algorithms. Theorems 3 and 4 show that `LocalTrans` performs well under a *local* Stochastic Blockmodel that makes minimal assumptions on the nodes outside of the true cluster; this is the first statistical result to demonstrate how local clustering algorithms can be robust to vast regions of the graph.

This paper studies small clusters because (1) they can create sparse *and* transitive Stochastic Blockmodels, (2) they are relatively easy to find, both computationally and statistically, and (3) they are easy to plot and visualize. In future research, we will study how these ideas can be used to find large partitions in networks. Sparse and transitive models do not preclude large partitions, as long as some type of local structure exists

within each partition. It is not yet clear how global algorithms like spectral clustering might leverage this transitive structure in a stochastic model; this is one area for future research.

References

David J Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

Noga Alon, Raphael Yuster, and Uri Zwick. Finding and counting given length cycles. *Algorithmica*, 17(3):209–223, 1997.

Brendan PW Ames and Stephen A Vavasis. Convex optimization for the planted k-disjoint-clique problem. *arXiv preprint arXiv:1008.2814*, 2010.

Reid Andersen and Fan Chung. Detecting sharp drops in pagerank and a simplified local partitioning algorithm. *Theory and Applications of Models of Computation*, pages 1–12, 2007.

Reid Andersen and Yuval Peres. Finding sparse cuts locally using evolving sets. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, pages 235–244. ACM, 2009.

Reid Andersen, Fan Chung, and Kevin Lang. Local graph partitioning using pagerank vectors. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 475–486. IEEE, 2006.

P. Bickel, D. Choi, X. Chang, and H. Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Arxiv preprint arXiv:1207.0865*, 2012.

P.J. Bickel and A. 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.

P.J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):38–59, 2011.

Alain Celisse, J-J Daudin, and Laurent Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *arXiv preprint arXiv:1105.3288*, 2011.

Antoine Channarond, Jean-Jacques Daudin, and Stéphane Robin. Classification and estimation in the stochastic block model based on the empirical degrees. *arXiv preprint arXiv:1110.6517*, 2011.

K. Chaudhuri, F. Chung, and A. Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research*, 2012: 1–23, 2012.

A. Chen, A.A. Amini, P.J. Bickel, and E. Levina. Fitting community models to large sparse networks. *Arxiv preprint arXiv:1207.2340*, 2012a.

Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering sparse graphs. *arXiv preprint arXiv:1210.3335*, 2012b.

D.S. Choi, P.J. Wolfe, and E.M. Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284, 2012.

Fan RK Chung. *Spectral Graph Teory*, volume 92. Amer Mathematical Society, 1997.

Aaron Clauset. Finding local community structure in networks. *Physical Review E*, 72(2): 026132, 2005.

Amin Coja-Oghlan and André Lanka. Finding planted partitions in random graphs with general degree distributions. *SIAM Journal on Discrete Mathematics*, 23(4):1682–1714, 2009.

Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `http://igraph.sf.net`.

Anirban Dasgupta, John E Hopcroft, and Frank McSherry. Spectral analysis of random graphs with skewed degree distributions. In *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*, pages 602–610. IEEE, 2004.

W.E. Donath and A.J. Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5):420–425, 1973.

R.I.M. Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493, 1992.

M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2): 298–305, 1973.

Donniell E Fishkind, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Applications*, 34(1):23–39, 2013.

C.J. Flynn and P.O. Perry. Consistent biclustering. *Arxiv preprint arXiv:1206.6927*, 2012.

Joachim Giesen and Dieter Mitsche. Reconstructing many partitions using spectral techniques. In *Fundamentals of Computation Theory*, pages 433–444. Springer, 2005.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

P.W. Holland and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5 (2):109–137, 1983.

Douglas N Hoover. Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, 1979.

Jiashun Jin. Fast network community detection by score. *arXiv preprint arXiv:1211.5803*, 2012.

Olav Kallenberg. *Probabilistic symmetries and invariance principles.* Springer Science+ Business Media, 2005.

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

Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

Chung-Shou Liao, Kanghao Lu, Michael Baym, Rohit Singh, and Bonnie Berger. Isorankn: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 25 (12):i253–i258, 2009.

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

Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

Samet Oymak and Babak Hassibi. Finding dense clusters via" low rank+ sparse" decomposition. *arXiv preprint arXiv:1104.5186*, 2011.

Carey E Priebe, John M Conroy, David J Marchette, and Youngser Park. Scan statistics on enron graphs. *Computational & Mathematical Organization Theory*, 11(3):229–247, 2005.

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

Karl Rohe and Bin Yu. Co-clustering for directed graphs; the stochastic co-blockmodel and a spectral algorithm. *arXiv preprint arXiv:1204.2296*, 2012.

Andrey Rukhin and Carey E Priebe. On the limiting distribution of a graph scan statistic. *Communications in Statistics-Theory and Methods*, 41(7):1151–1170, 2012.

Daniel A Spielman and Shang-Hua Teng. A local clustering algorithm for massive graphs and its application to nearly-linear time graph partitioning. *arXiv preprint arXiv:0809.3232*, 2008.

D.L. Sussman, M. Tang, D.E. Fishkind, and C.E. Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

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

H. Wang, M. Tang, Y. Park, and C. E. Priebe. Locality statistics for anomaly detection in time series of graphs. *ArXiv e-prints*, June 2013.

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

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108(18):7321–7326, 2011.

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40 (4):2266–2292, 2012.

Appendix A. Proofs for Section 1

A.1. **Proof of Theorem 1.**

Proof. Recall that the transitivity ratio of A is

$$trans(A) = \frac{\text{number of closed triplets in } A}{\text{number of connected triples of vertices in } A}.$$

Both the numerator and the denominator of the transitivity ratio have other formulations that suggest how they can be computed.

$$\begin{aligned}\text{number of closed triplets in } A &= 6 \times \text{Number of triangles in } A \\ &= \text{trace}(AAA)\end{aligned}$$

$$\begin{aligned}\text{number of connected triples of vertices in } A &= 2 \times \text{Number of 2-stars in } A \\ &= 2\sum_j \binom{d_j}{2} \\ &= \sum_j d_j^2 - d_j.\end{aligned}$$

For ease of notation, define $X_n = trace(AAA)$ and $Y_n = \sum_i d_i^2 - d_i$. So, $trans(A) = X_n/Y_n$. To show that transitivity converges to zero, use

$$P\left(\frac{X_n}{Y_n} > \epsilon\right) \leq \frac{E(X_n/Y_n)}{\epsilon}$$

and the following Lemma.

Lemma 1. *If $\lambda_n = o(n)$, then there exists a sequence f_n such that $E(X_n) = o(f_n)$ and*

$$P(Y_n \geq f_n) \to 1.$$

Using Lemma 1 and fact that $X_n/Y_n \leq 1$ a.s.,

$$\begin{aligned}E\frac{X_n}{Y_n} &\leq E\left(\frac{X_n}{f_n}1\{Y_n > f_n\} + 1\{Y_n < f_n\}\right) \\ &\leq \frac{E(X_n)}{f_n} + P(Y_n < f_n) \\ &\to 0.\end{aligned}$$

Now, to prove Lemma 1. For ease of notation, define $d = \sum_i d_i$. From Bickel, Levina, Chen, define $\hat{\rho} = \frac{d}{2n(n-1)}$. They show that $\hat{\rho}/\rho_n \xrightarrow{P} 1$, where $\rho_n = P(A_{12} = 1)$. So, this converges to zero:

$$P\left(\frac{d}{2n(n-1)\rho_n} < 1/2\right) = P\left(d < n(n-1)\rho_n\right)$$

Define $M_n = n(n-1)\rho_n$. Then, $P\left(d > M_n\right) \to 1$. Define $f_n = M_n^2/n - M_n$. Notice that

$$min_{\sum_i d_i = m} \sum_i d_i^2 - d_i \geq n((m/n)^2 - m/n) = m^2/n - m.$$

Putting these pieces together,

$$\begin{aligned}
P(Y_n \geq f_n) &= \int_\Omega 1\{Y_n \geq f_n\} dP \\
&\geq \int_{d>M_n} 1\{\sum_i d_i^2 - d \geq f_n\} dP \\
&= \sum_{m>M_n} \int_{d=m} 1\{\sum_i d_i^2 - d \geq f_n\} dP \\
&\geq \sum_{m>M_n} \int_{d=m} 1\{m^2/n - m \geq f_n\} dP \\
&= \sum_{m>M_n} \int_{d=m} dP \\
&= P(d > M_n) \to 1.
\end{aligned}$$

The last piece is to show that $E(X_n) = o(f_n)$. From the definition of f_n and the fact that $\rho_n = \lambda_n/n$,

$$\begin{aligned}
f_n &= \frac{(n(n-1)\rho_n)^2}{n} - n(n-1)\rho_n \\
&= n(n-1)^2(\lambda_n/n)^2 - n(n-1)\lambda_n/n \\
&= \lambda_n(n-1)\left(\frac{n-1}{n}\lambda_n - 1\right) \\
&\to \lambda_n^2 n
\end{aligned}$$

Define

$$S_{12} = \sum_i A_{i1}A_{2i}$$

as the number of two stars with nodes 1 and 2 as end points. Then, under the assumption that

$$p_{\max} = o\left(\frac{P(A_{13}=1)}{P(A_{13}=1|A_{23}=1)}\right),$$

it follows that

$$\begin{aligned}
E(X_n) &= p_{ct} n(n-1) E(S_{12}) \\
&\leq p_{\max} n^2 (nE(A_{13}A_{23})) \\
&= p_{\max} n^3 E(A_{13}|A_{23}=1) E(A_{23}=1) \\
&= o(\rho_n^2 n^3) \\
&= o(f_n).
\end{aligned}$$

□

A.2. **Proof of Theorem 2:**

Proof. Let $r = c_0/n$ and let p be fixed.

Number of triangles: Let Δ_n denote the number of triangles. Notice that there are three types of triangles: (1) let Δ^i denote the number of triangles with all nodes in block i; (2) let Δ_{21} denote the number of triangles with 2 nodes in the same block and one node in a separate block; (3) let Δ_{111} denote the number of triangles with nodes in three separate blocks.

$$E(\Delta_{21}) = K(K-1)\binom{s}{2}spr^2 = K(K-1)\binom{s}{2}spc_0^2/(s^2K^2) \leq p(s-1)/2,$$

$$E(\Delta_{111}) = \binom{K}{3}s^3r^3 \leq K^3s^3c_0^3/(6s^3K^3) = c_0^3/6.$$

By the Markov inequality, $\Delta_{21}/K \xrightarrow{P} 0, \Delta_{111}/K \xrightarrow{P} 0$. Finally, Δ^i are iid. So, by LLN, their average converges in probability to their expectation. Putting these pieces together with Slutsky's theorem, the number of triangles over K is,

$$\frac{1}{K}\Delta_n \xrightarrow{P} E(\Delta^i) = \binom{s}{3}p^3.$$

Number of two stars: Let S_n be the number of two-stars. Define the events $B = \{|S_n - E(S_n)| > t\}$ and $A = \{\text{Maximum Degree} \leq M\}$

$$P(B) = P(BA) + P(BA^c) \leq P(BA) + P(A^c)$$

Apply the bounded difference inequality within the set BA. Define $A_i \in \{0,1\}^{n-i}$ for $i = 1, \ldots, n-1$ as the ith row of the upper triangle of the adjacency matrix A. To bound the bounded difference constant, first notice that $c_1 \geq c_i$ for all i. Moreover, we have

$$c_1 \leq 3M^2.$$

This is because node 1 belongs to at most $3\binom{M}{2}$ triplets. By changing the edges of node 1, S_n can increase or decrease by at most $3\binom{M}{2}$. By the bounded difference inequality,

$$P(BA) \leq 2\exp\left(-\frac{t^2}{n9M^4}\right).$$

Choose $M = \log n$ and $t = K\epsilon$ for any $\epsilon > 0$, we have $P(BA) \to 0$. More over, by concentration inequality,

$$P(A^c) = P(\cup_i\{d_i \geq M\}) \leq nP(d_1 \geq M) \to 0$$

Therefore, we have

$$S_n/K \xrightarrow{P} E(S_n)/K.$$

Notice that $E(S_n)/K$ is equal to the expected number of two-stars whose center is in the first block. So,

$$\begin{aligned} E(S_n)/K &= s\left(\binom{s-1}{2}p^2 + (s-1)(n-s)pr + \binom{n-s}{2}r^2\right) \\ &\to s\left(\binom{s-1}{2}p^2 + (s-1)pc_0 + c_0^2/2\right) \end{aligned}$$

Finally,

$$TranRatio(A) = \frac{3 \times \text{number of triangles}}{\text{number of 2-stars}} \xrightarrow{P} \frac{3E\Delta_n}{ES_n}.$$

□

Appendix B. Proofs for Section 2 and 3

B.1. **Proof of Theorem 3:**

Proof. Define the following events

$$\begin{aligned} B_\alpha &= \{S_* \text{ and } S_*^c \text{ are separated with cutting level } \alpha\}, \\ C_\alpha &= \{S_* \text{ is clustered within one block with cutting level } \alpha\}, \\ D_\alpha &= \{\text{Every pair of nodes in } S_* \text{ have at least } \alpha \text{ common neighbor.}\} \end{aligned}$$

If both events B_α and C_α are satisfied, then for any $i \in S_*$, LocalTrans(A, i, α) recovers block S_* correctly. Events D_α implies that S_* is clustered within one block with cutting level α, that is $D_\alpha \in C_\alpha$. To see this, assume the contrary, then there exists a partition $S_* = S_1 \cup S_2$, such that for any $j \in S_1, k \in S_2, T(j,k) < \alpha$. However, D_α implies that S_* is connected, hence there exists $u \in S_1, v \in S_2$, such that $A(u,v) = 1$. Moreover, D_α also implies that u, v have at least α common neighbors. Hence, $T(u,v) \geq \alpha$. This is a contradiction.

The following lemma leads to the desired results.

Lemma 2. *Under the conditions above,*

$$\begin{aligned} P(B_1^c) &= O(p_{out}^2 ns(s+\lambda)), \\ P(B_2^c) &= O(p_{out}^3 ns(s+\lambda)^2), \\ P(D_\alpha^c) &\leq \frac{1}{2}s^2 \sum_{k=0}^{\alpha-1} \binom{s-2}{k} (1-p_{in}^2)^{s-2-k}. \end{aligned}$$

By Lemma 2, we have

$$\begin{aligned}
P(\text{correctly clustering } S_* \text{ with cutting level 1}) &\geq P(B_1 \cap C_1) \\
&= 1 - P(B_1^c \cup C_1^c) \\
&\geq 1 - P(B_1^c) + P(C_1^c) \\
&\geq 1 - P(B_1^c) + P(D_1^c) \\
&= 1 - \left(\frac{1}{2}s^2(1-p_{in}^2)^{s-2} + O(p_{out}^2 ns(s+\lambda))\right)
\end{aligned}$$

$$\begin{aligned}
P(\text{correctly clustering } S_* \text{ with cutting level 2}) &\geq P(B_2 \cap C_2) \\
&= 1 - P(B_2^c \cup C_2^c) \\
&\geq 1 - P(B_2^c) + P(C_2^c) \\
&\geq 1 - P(B_2^c) + P(D_2^c) \\
&= 1 - \left(s^3(1-p_{in}^2)^{s-3} + O(p_{out}^3 ns(s+\lambda)^2)\right)
\end{aligned}$$

□

Proof of Lemma 2:

Proof.

$$\begin{aligned}
P(B_1^c) &= P(\text{There exists at least two nodes } i \in S_* \text{ and } j \in S_*^c, \text{ such that } T_{ij} \geq 1) \\
&= P(\bigcup_{i \in S_*} \bigcup_{j \in S_*^c} \bigcup_{k \in S_* \cup S_*^c, k \neq i,j} \{A_{ij}A_{jk}A_{ki} = 1\}) \\
&\leq sn[P(\bigcup_{k \in S_*} \{A_{ij}A_{jk}A_{ki} = 1\}) + P(\bigcup_{k \in S_*^c} \{A_{ij}A_{jk}A_{ki} = 1\})] \\
&\leq sn(sp_{out}^2 + np_{out}^2 \frac{\lambda}{n}) \\
&= O(p_{out}^2 ns(s+\lambda))
\end{aligned}$$

$$\begin{aligned}
P(B_2^c) &= P(\text{There exists at least two nodes } i \in S_* \text{ and } j \in S_*^c, \text{ such that } T_{ij} \geq 2) \\
&= P(\bigcup_{i \in S_*} \bigcup_{j \in S_*^c} \bigcup_{k < l \in S_* \cup S_*^c, k,l \neq i,j} \{A_{ij}A_{jk}A_{ki}A_{jl}A_{li} = 1\}) \\
&\leq snP(\bigcup_{k,l \in S_*} \cup \bigcup_{k \in S_*, l \in S_*^c} \cup \bigcup_{k,l \in S_*^c} \{A_{ij}A_{jk}A_{ki}A_{jl}A_{li} = 1\}) \\
&\leq sn(\frac{1}{2}s^2p_{out}^3 + \frac{1}{2}n^2p_{out}^3(\frac{\lambda}{n})^2 + nsp_{out}^3\frac{\lambda}{n}) \\
&= O(p_{out}^3 ns(s+\lambda)^2)
\end{aligned}$$

$$\begin{aligned} P(D_\lambda^c) &= P(\text{There exists at least two nodes } i,j \in S_*, \text{ such that } i \text{ and } j \text{ has less than } \alpha \text{ neighbors}) \\ &= P\left(\bigcup_{i,j\in S_*} \{i \text{ and } j \text{ has less than } \alpha \text{ neighbors}\}\right) \\ &\leq \frac{1}{2}s^2 \sum_{k=0}^{\alpha-1} \binom{s-2}{k} (1-p_{in}^2)^{s-2-k} \end{aligned}$$

□

Proof of Theorem 4:

Proof. In the proof, we assume that $i,j \in S_*$, $P(A_{ij}=1)=p_{in}$. The proof can be easily extended to the case where $P(A_{ij}=1) \geq p_{in}$. First, we prove that $P(\max_{i\in S_*} D_{ii} \geq 2\mathbb{E}D_{11})$ is well bounded with some non-vanishing probability. $\forall i \in S_*$,

$$\mathbb{E}D_{ii} = \mathbb{E}D_{11} = \sum_{j\in S_*, j\neq i} \mathbb{E}A_{ij} + \sum_{j\in S_*^c} \mathbb{E}A_{ij} = (s-1)p_{in} + \lambda.$$

$\forall i \in S_*,\ \forall \epsilon > 0$,

$$P(D_{ii} \geq \mathbb{E}D_{ii} + \epsilon) \leq \exp\left\{-\frac{\epsilon}{2(\mathbb{E}D_{ii}+\epsilon/3)}\right\}.$$

Take $\epsilon = \mathbb{E}D_{11}$, and take union bound for all $i \in S_*$, we have

$$P(\max_{i\in S_*} D_{ii} \geq 2\mathbb{E}D_{11}) \leq s\exp\left\{-\frac{3}{8}\mathbb{E}D_{11}\right\} = s\exp\{-\frac{3}{8}[(s-1)p_{in}+\lambda]\}.$$

Let O denote the set $\{D_{ii} \leq 2\mathbb{E}D_{11}, \forall i \in S_*\}$. Then within the set O, by the same argument from the proof of Theorem 3, we have that with probability at least $\frac{1}{2}s^2(1-p_{in}^2)^{s-2}$, S_* is clustered within one block by LocalTrans(L_τ, i, cut) for any $i \in S_*$ with cut

$$cut = (2\mathbb{E}D_{11}+\tau)^{-3} = (2(s-1)p_{in}+2\lambda+\tau)^{-3}.$$

Second part proves that $\forall i \in S_*, j \in S_*^c, P(T_{ij} \geq cut)$ is $o(1)$. Notice that for any $j \in S_*^c$, the (j,j)th element of D_τ (denote as D_{jj}^τ) is $d_j^* + \sum_{i\in S_*} A_{ij} + \tau$, so we have

$$D_{jj}^\tau \geq d_j^* + \tau, \quad \forall j \in S_*^c$$

For any $i \in S_*$, we have

$$D_{ii}^\tau \geq \tau, \quad \forall i \in S_*$$

$\forall i \in S_*,\ j \in S_*^c$,

$$\begin{aligned}
P(T_{ij} \geq cut) &\leq \frac{d_j^*}{n} P(T_{ij} \geq \ cut | A_{ij} = 1) \\
&= \frac{d_j^*}{n} P\left(\frac{1}{D_{ii}^\tau D_{jj}^\tau} \sum_{k=1}^n \frac{A_{ik}A_{kj}}{D_{kk}^\tau} \geq cut \right) \\
&\leq \frac{d_j^*}{n} P\left(\frac{1}{\tau(d_j^* + \tau)} \sum_{k=1}^n \frac{A_{ik}A_{kj}}{D_{kk}^\tau} \geq cut \right) \\
&\leq \frac{d_j^*}{n} \left[P\left(\sum_{k \in S_*} A_{ik}A_{kj} \geq \tau^2 (d_j^* + \tau) cut/2 \right) + \right. \\
&\qquad \left. P\left(\sum_{k \in S_*^c} \frac{A_{ik}A_{kj}}{d_k + \tau} \geq \tau (d_j^* + \tau) cut/2 \right) \right]
\end{aligned}$$

For the first term, when n large,

$$\begin{aligned}
P(\sum_{k \in S_*} A_{ik}A_{kj} \geq \tau^2 (d_j^* + \tau) cut/2) &\leq P(\sum_{k \in S_*, k \neq i} A_{ik}A_{kj} > 0) \\
&\leq 1 - (1 - p_{in} \frac{d_j^*}{n})^{s-1} \\
&\leq p_{in}(s-1) \frac{d_j^*}{n}
\end{aligned}$$

On the other hand, notice that $\{A_{ik}A_{kj}, k \in S_*/\{i\}\}$ are independent random variables with $A_{ik}A_{kj} \sim Ber(p_{in}\frac{d_j^*}{n})$ by the assumption. $E[\sum_{k \in S_*, k \neq i} A_{ik}A_{kj}] = p_{in}(s-1)\frac{d_j^*}{n}$. By concentration inequality, when n is sufficiently large (independent of j), we have

$$\begin{aligned}
P(\sum_{k \in S_*} A_{ik}A_{kj} \geq \tau^2 (d_j^* + \tau) cut/2) &\leq P\left(\sum_{k \in S_*} A_{ik}A_{kj} \geq p_{in}(s-1)\frac{d_j^*}{n} + \right. \\
&\qquad \left. d_j^* \left(\tau^2 cut/2 - \frac{p_{in}(s-1)}{n} \right) \right) \\
&\leq \exp\left(-\frac{c_1^2 (d_j^*)^2}{2c_2 d_j^*/n + 2c_1 d_j^*/3} \right) \\
&= \exp(-c_1 d_j^*)
\end{aligned}$$

where $c_1 = \tau^2 cut/3$.

For the second term, without loss of generality, assume that $A_{1j} = A_{2j} = ... = A_{d_j^*, j} = 1, A_{k,j} = 0, \forall k > d_j^*$. Notice that $\{A_{ik}, k = 1, 2, ..., n, k \neq i\}$ are independent random

variables with $A_{ik} \sim Ber(\frac{d_k^*}{n})$. Applying concentration inequality on the sequence $\{A_{ik}, k = 1, 2, ..., d_j^*\}$, with $a_k = \frac{1}{d_k^*}$. Define $X = \sum_{k=1}^{d_j^*} a_k A_{ik}$, then $\mathbb{E}X = \frac{d_j^*}{n}$, and

$$v = \sum_{k=1}^{d_j^*} a_k^2 \mathbb{E}A_{ik} = \frac{1}{n}\sum_{k=1}^{d_j^*} \frac{1}{d_j^*} \leq \frac{d_j^*}{n}.$$

When n is sufficiently large, we have,

$$\begin{aligned}
P\left(\sum_{k\in S_*^c} \frac{A_{ik}A_{kj}}{d_k+\tau} \geq \tau(d_j^*+\tau)cut/2\right) &\leq P\left(\sum_{k=1}^{d_j^*} \frac{A_{ik}}{d_k} \geq \tau(d_j^*+\tau)cut/2\right) \\
&\leq P\Bigg(\sum_{k=1}^{d_j^*} \frac{A_{ik}}{d_k} \geq \frac{d_j^*}{n} + d_j^*(\tau cut/2 - \frac{1}{n})\Bigg) \\
&\leq \exp\left(-\frac{c^2(d_j^*)^2}{2(v+cd_j^*/3)}\right) \\
&\leq \exp\left(-\frac{c^2(d_j^*)^2}{2(d_j^*/n+cd_j^*/3)}\right) \\
&= \exp(-cd_j^*)
\end{aligned}$$

where $c = \tau cut/3$.

On the other hand,

$$\begin{aligned}
P\left(\sum_{k\in S_*^c} \frac{A_{ik}A_{kj}}{d_k+\tau} \geq \tau(d_j^*+\tau)cut/2\right) &\leq P\left(\sum_{k=1}^{d_j^*} \frac{A_{ik}}{d_k+\tau} \geq \tau(d_j^*+\tau)cut/2\right) \\
&\leq P\left(\sum_{k=1}^{d_j^*} A_{ik} > 0\right) \\
&\leq 1 - \left(1 - \frac{d_j^*}{n}\right)^{d_j^*} \\
&\leq \frac{(d_j^*)^2}{n}
\end{aligned}$$

To sum up,when n is sufficiently large, we have that $\forall i \in S_*,\ j \in S_*^c$

$$P(T_{ij} \geq cut) \leq \frac{d_j^*}{n} \min\left\{\frac{(d_j^*)^2}{n} + p_{in}(s-1)\frac{d_j^*}{n},\quad 2e^{-cd_j^*}\right\}.$$

where $cut = (2\mathbb{E}D_{11} + \tau)^{-3} = (2(s-1)p_{in} + 2\lambda + \tau)^{-3}$, and $c = \tau cut/3$.

Next we show that, for any $i \in S_*, j \notin S_*$, $P(T_{ij} \geq cut) = O(n^{3\epsilon-2})$, where $\epsilon > \log_n(\frac{1}{c})$.

Case 1: $d_j^* < n^\epsilon$.

$$\begin{aligned} nP(T_{ij} \geq cut) &\leq \min\left\{\frac{(d_j^*)^3}{n} + p_{in}s\frac{(d_j^*)^2}{n},\ 2d_j^* e^{-cd_j^*}\right\} \\ &\leq \frac{(d_j^*)^3}{n} + p_{in}s\frac{(d_j^*)^2}{n} \\ &\leq \frac{n^{3\epsilon}}{n} + p_{in}s\frac{n^{2\epsilon}}{n} \\ &= O(n^{3\epsilon-1}) \end{aligned}$$

Case 2: $d_j^* \geq n^\epsilon$. Notice that the derivate of $u(x) = 2d_j^* \exp(-cd_j^*)$ is $(1-cx)\exp(-cx)$. So, if $n^\epsilon > 1/c$ then $u(d_j^*) \leq u(n^\epsilon) = 2n^\epsilon \exp(-cn^\epsilon)$.

$$\begin{aligned} nP(T_{ij} \geq cut) &\leq \min\left\{\frac{(d_j^*)^3}{n} + p_{in}s\frac{(d_j^*)^2}{n},\ 2d_j^* \exp(-cd_j^*)\right\} \\ &\leq 2d_j^* \exp(-cd_j^*) \\ &\leq 2n^\epsilon \exp(-cn^\epsilon) \\ &= o(n^{-1}) \end{aligned}$$

So, independent of d_j^*,

$$P(T_{ij} \geq cut) = O(n^{3\epsilon-2})$$

Putting the pieces together,

$$\begin{aligned} P(S_* \text{ and } S_*^c \text{ are not separated }) &\leq s\sum_{j\in S_*^c}\frac{d_j^*}{n}\min\left\{\frac{(d_j^*)^2}{n} + p_{in}s\frac{d_j^*}{n},\ 2e^{-cd_j^*}\right\} \\ &\leq snO(n^{3\epsilon-2}) \\ &= O(n^{3\epsilon-1}) \end{aligned}$$

Finally, recall that $O = \{D_{ii} \leq 2\mathbb{E}D_{11}, \forall i \in S_*\}$, we have that for any $i \in S_*$,

$$\begin{aligned} P(\{\text{LocalTrans}(L_\tau, i, cut) \text{ returns } S_*\}) &\geq 1 - \frac{1}{2}s^2(1-p_{in}^2)^{s-2} - O(n^{3\epsilon-1}) - P(O^c) \\ &\geq 1 - \Bigg(\frac{1}{2}s^2(1-p_{in}^2)^{s-2} + \\ &\qquad s\exp\{-\frac{3}{8}[(s-1)p_{in} + \lambda)\} + O(n^{3\epsilon-1})\Bigg). \end{aligned}$$

□