REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
17-10-2019	Final Report	1-May-2016 - 30-Sep-2019

4. TITLE AND SUBTITLE
Final Report: RESEARCH AREA 10: NETWORK SCIENCE: Local Algorithms for Large Informatics Graphs

5a. CONTRACT NUMBER
W911NF-16-1-0285

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
International Computer Science Institute
1947 Center Street
Suite 600
Berkeley, CA 94704 -1159

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
69041-NS.4

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Michael Mahoney
UU	UU	UU	UU		19b. TELEPHONE NUMBER 510-666-2900

RPPR Final Report
as of 06-Nov-2019

Agency Code:

Proposal Number: 69041NS **Agreement Number: W911NF-16-1-0285**
INVESTIGATOR(S):

Name: Michael Mahoney
Email: mmahoney@ICSI.Berkeley.EDU
Phone Number: 5106662900
Principal: Y

Organization: **International Computer Science Institute**
Address: 1947 Center Street, Berkeley, CA 947041159
Country: USA
DUNS Number: 187909478 EIN: 943024996
Report Date: 31-Dec-2019 Date Received: 17-Oct-2019
Final Report for Period Beginning 01-May-2016 and Ending 30-Sep-2019
Title: RESEARCH AREA 10: NETWORK SCIENCE: Local Algorithms for Large Informatics Graphs
Begin Performance Period: 01-May-2016 **End Performance Period:** 30-Sep-2019
Report Term: 0-Other
Submitted By: Michael Mahoney Email: mmahoney@ICSI.Berkeley.EDU
Phone: (510) 666-2900
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 0 **STEM Participants:** 0

Major Goals: A serious problem with many existing machine learning and data analysis tools in the complex networks area is that they are often very brittle and/or do not scale well to larger networks. As a consequence, analysts often develop intuition on small networks, with 10^2 or 10^3 nodes, and then try to apply these methods on larger networks, with 10^5 or 10^7 or more nodes. Larger networks, however, often have \emph {very} different static and dynamic properties than smaller networks. These properties are often deeply counterintuitive and surprisingly subtle to identify and test, but they often
have consequences that are very practically relevant. At root, the problem is that existing algorithmic/statistical methods fail to capture properly the ``coupling''
between very local structure and very global structure in these networks. For example, local structure might be the properties of ego networks in large social graphs, and global structure might be the diameter and other large-scale properties of the entire network. These failings clearly indicate that the metrics that are intuitive to people, based on their experience with small-scale networks, and that are used to control inference or obtain domain-specific insight, are often simply inappropriate for very large-scale graphs. Relatedly, existing algorithmic/statistical tools make local-global assumptions that are strongly violated by realistic networks. This has very detrimental consequences for our ability to extract domain insight from algorithmic tools applied to these networks. The proposed research will develop and implement existing and novel local and locally-biased algorithms on large networks to evaluate how those algorithms perform for downstream tasks of interest on realistic social/information networks to develop insights for use in downstream network analysis. In particular, primary research directions that will be considered include the following.
\begin{itemize}
\item
Local algorithms for other network problems: these algorithms can be applied to a small part of a very large graph, and they can come with improved algorithmic and/or statistical guarantees.
\item
Implementations on large-scale graphs: the behavior of these local algorithms depends strongly on implementation details, for graphs consisting of hundreds of thousands up to hundreds of millions of~nodes.
\item
Relationship with meaningful metrics of domain interest: these algorithms often have implicit embedding properties that can meaningfully be interpreted in terms of metrics of interest to the domain scientist.
\item
Intermediate-scale structure: as the size of the local region grows, these algorithms can be used to characterize intermediate-scale structure, e.g., whether a small contagion will propagate to the entire~graph.
\item

Algorithmic-statistical tradeoffs: the improved algorithmic/statistical guarantees of these algorithms are sometimes synergistic and sometimes not, and characterizing these tradeoffs is essential for their use.
\end{itemize}
The research work will focus on these methodological developments with an eye toward solving important practical problems such as scalable inference on graphs, understanding and controlling contagion and viral propagation on networks, and developing tools that are useful to the domain scientist at extracting interpretable and actionable insight from noisy and complex networked data.

Accomplishments: We have results along four major directions so far. One is developing a novel local clustering algorithm that is a hybrid of spectral and flow and that does particularly well in poorly structured graphs. A second is an initial version of a social discrete choice model. A third is results on algorithmic-statistical radios on the related problem of sketched ridge regression. A fourth is initial results on using better graph embedding ideas from random walks to do better on various graph clustering tasks. We have initial version of publications reporting all four directions. We also are nearly complete a long publication to send to a venue such as SIAM Review on flow-based improvement methods. We also have an open-source repo of code for people to use these methods.

Training Opportunities: Several postdocs and graduate students were supported on this project. In addition, postdoc has the opportunity to mentor visiting students (not funded on this grant) to help them develop a project. This involved training for the postdocs, who have an interest in an academic career. One postdoc has started a position as an assistant professor at the University of Waterloo.

Results Dissemination: Flow-based Algorithms for Improving Clusters
K. Fountoulakis, M. Liu, D. F. Gleich and M. W. Mahoney
(To be uploaded to arXiv soon, then to be submitted to SIREV.)
(This paper provides a review/overview of flow-based improvement methods.)

Limit theorems for out-of-sample extensions of the adjacency and Laplacian spectral embeddings,
K. Levin, F. Roosta, M. Tang, M. W. Mahoney, C. E. Priebe,
Technical Report, Preprint: arXiv:1910.00423 (2019)
(Submitted to JMLR.)
(This paper provides bounds for related adjacency and Laplacian spectral embeddings in the global case; we are still working on analogous bounds for the local case.)

Statistical guarantees for local graph clustering,
W. Ha, K. Fountoulakis, and M. W. Mahoney,
Technical Report, Preprint: arXiv:1906.04863
(Short version submitted to NeurIPS 2019, and long version to be then submitted to JMLR.)
(This paper provides a statistical analysis of strongly local spectral methods.)

Parallel and Communication Avoiding Least Angle Regression,
S. Das, J. Demmel, K. Fountoulakis, L. Grigori, and M. W. Mahoney,
Technical Report, Preprint: arXiv:1905.11340 (2019)
(Submitted to SISC.)
(This paper develops communication avoiding methods for the least angle regression problem, and the homotopy-based methods can be used in the arXiv:1906.04863 paper.)

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Michael Mahoney
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:

International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Kimon Fountoulakis
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Benjamin Erichson
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

ARTICLES:

Publication Type: Journal Article Peer Reviewed: N **Publication Status:** 4-Under Review
Journal: arXiv
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 8/2/17 12:00AM Date Published:
Publication Location:
Article Title: Social Discrete Choice Models
Authors: Danqing Zhang, Kimon Fountoulakis, Junyu Cao, Mogeng Yin, Michael Mahoney, Alexei Pozdnoukhov
Keywords: Discrete Choice Models
Abstract: Human decision making underlies data generating process in multiple application areas, and models explaining and predicting choices made by individuals are in high demand. Discrete choice models are widely studied in economics and computational social sciences. As digital social networking facilitates information flow and spread of influence between individuals, new advances in modeling are needed to incorporate social information into these models in addition to characteristic features affecting individual choices. In this paper, we propose the first latent class discrete choice model that incorporates social relationships among individuals represented by a given graph. We add social regularization to represent similarity between friends, and we introduce latent classes to account for possible preference discrepancies between different social groups. Training of the social discrete choice models is performed using a specialized Monte Carlo expectation maximization algorithm. Scalabili

Acknowledged Federal Support: **Y**

RPPR Final Report
as of 06-Nov-2019

Statistical guarantees for local graph clustering

Wooseok Ha[*,†] Kimon Fountoulakis[*,‡] Michael W. Mahoney[§]

June 13, 2019

Abstract

Local graph clustering methods aim to find small clusters in very large graphs. These methods take as input a graph and a seed node, and they return as output a good cluster in a running time that depends on the size of the output cluster but that is independent of the size of the input graph. In this paper, we adopt a statistical perspective on local graph clustering, and we analyze the performance of the ℓ_1-regularized PageRank method (a popular local graph clustering method) for the recovery of a single target cluster, given a seed node inside the cluster. Assuming the target cluster has been generated by a random model, we present two results. In the first, we show that the optimal support of ℓ_1-regularized PageRank recovers the full target cluster, with bounded false positives. In the second, we show that if the seed node is connected solely to the target cluster then the optimal support of ℓ_1-regularized PageRank recovers exactly the target cluster. We also show that the solution path of ℓ_1-regularized PageRank is monotonic. From a computational perspective, this permits the application of the forward stagewise algorithm, which in turn permits us to approximate the entire solution path of the local cluster in a running time that does not depend on the size of the entire graph.

1 Introduction

In many data applications, one is interested in finding small-scale structure in a very large data set. As an example, consider the following version of the so-called *local graph clustering problem*: given a large graph and a seed node in that graph, quickly find a good small cluster that includes that seed node. From an algorithmic perspective, one typically considers worst-case input graphs, and one may be interested in running time guarantees, e.g, to find a good cluster in a time that depends linearly or sub-linearly on the size of the entire graph. From a statistical perspective, such a local graph clustering problem can be understood as a recovery problem. One assumes that there exists a target cluster in a given large graph, where the graph is assumed to have been generated by a random model, and the objective is to recover the target cluster from one node inside the cluster.

In this paper, we consider the so-called *ℓ_1-regularized PageRank algorithm* [17], a popular algorithm for the local graph clustering problem, and we establish statistical recoverability guarantees for it. Previous theoretical analysis on local graph clustering, e.g., [6, 51], is based on the notion of conductance (a cluster quality metric that considers the internal versus external connectivity of a cluster) and considers running time performance for worst-case input graphs. In contrast, our goal

[*]Equal contribution

[†]Department of Statistics, University of California at Berkeley, Berkeley, CA, USA. E-mail: haywse@gmail.com.

[‡]School of Computer Science, University of Waterloo, Waterloo, ON, Canada. E-mail: kfountou@uwaterloo.ca.

[§]ICSI and Department of Statistics, University of California at Berkeley, Berkeley, CA, USA. E-mail: mmahoney@stat.berkeley.edu.

will be to study the average-case performance of the ℓ_1-regularized PageRank algorithm, under a certain type of a local random graph model. This model concerns the target cluster and its adjacent nodes, and it encompasses the stochastic block model [25, 1] and the planted clustering model [4, 8] as special cases.

Within this random graph model, we provide theoretical guarantees for the unique optimal solution of the ℓ_1-regularized PageRank optimization problem (see Theorem 1 and Theorem 2). Observe that our statistical perspective is more aligned with statistical guarantees for the sparse regression problem (and the lasso problem [42]), where the objective is to recover the true parameter and/or support from noisy data. Given this connection, we also establish a result for the exact support recovery of ℓ_1-regularized PageRank (see Theorem 3).

We are also interested in the computational performance of local graph clustering algorithms, i.e., that the running time depends on the size of the output but not on the size of the entire graph. To that end, we will show that the solution path of our algorithm for the ℓ_1-regularized PageRank problem is monotonic (see Theorem 4). This means that, as the ℓ_1 norm regularization parameter becomes smaller, the individual coordinates of the solution are monotonically increasing, and thus the number of nonzero nodes in the optimal solution increases. This is a crucial property that allows us to use popular statistical tools, in particular, the forward stagewise algorithm [14, 18, 43], to obtain an approximate solution of the entire ℓ_1-regularized PageRank solution path. This, in combination with known results about the locality of gradient-based algorithms [17], means that, by terminating our algorithm early, we can find the approximate path in a running time that does not depend on the size of the entire graph. This makes the forward stagewise algorithm a good candidate for recovering an approximate and partial solution path of the ℓ_1-regularized problem for very large graphs.

Literature review. There is a large body of related work, the most relevant of which is: work in theoretical computer science on local graph algorithms; work in statistics on stochastic graph models; and work in statistics on solution path algorithms. We discuss each in turn.

The origins of local graph clustering are with the work of Spielman and Teng [41]. Subsequent to their original results, there has been a great deal of follow-up work on local graph clustering procedures, including with random walks [6], local Lanczos spectral approximations [39], evolving sets [7], seed expansion methods [27], optimization-based approaches [17, 16], and local flow methods [46]. There also exist local higher-order clustering [47], linear algebra approaches [39], spectral methods based on Heat Kernel PageRank [26], newer seed set selection techniques for local flow methods [44], and parallel local spectral approaches [40]. In all of these cases, given a seed node, or a seed set of nodes, the goal of existing local graph clustering approaches is to compute a cluster "nearby" the seed that is related to the "best" cluster nearby the seed. Here, "best" and "nearby" are intentionally left under-specified, as they can be formalized in one of a few different but related ways. For example, "best" is usually related to a clustering score such as conductance. In fact, many existing methods for local graph clustering with theoretical guarantees are motivated through the problem of finding a cluster that is near the seed node and that also has small conductance value [41, 6, 17, 44, 46].

There are also numerous papers in statistics on partitioning random graphs. Arguably, the stochastic block model (SBM) is the most commonly employed random model for graph partitioning, and it has been extensively studied [2, 3, 48, 30, 32, 33, 31, 35, 5, 1]. Recent work has also generalized the SBM to a degree-corrected block model, to capture degree heterogeneity of the network [13, 21, 50, 19]. The literature in this area is too extensive to cover in this paper, but we refer the readers to excellent survey papers on the graph partitioning problem [1]. We should emphasize that the traditional graph partitioning problem is quite different than the local graph clustering problem. Among other things, while the former partitions all the vertices of a graph into different

clusters, for the latter problem our objective is to find a single cluster given a seed node in the cluster.

Solution path algorithms are designed to solve the optimization problem over a full range of regularization parameter values, or to find a subset of the full solution path when the algorithm is terminated early. Since the seminal work of [14], the idea of designing path algorithms has gained much attention in the sparse regression literature [52, 22, 9], which have rendered the exploration of full regression coefficient paths and characterization of bias-variance tradeoff very efficient. Unlike regression setting, however, this type of path algorithm has been less studied in local graph clustering. In [20], the authors propose a method to generate an approximate solution path associated with a PageRank diffusion to reveal cluster structures at multiple scale. In this work, we exploit the fact that the solution path of the ℓ_1-regularized PageRank problem is monotonic, which allows us to apply forward stagewise algorithm to *provably* approximate the entire ℓ_1 regularization solution path.

Notation. We write $[n] = \{1, \ldots, n\}$ for any $n \geq 1$. Throughout the paper we assume we have a connected, undirected graph $G = (V, E)$, where V denotes the set of nodes, with $|V| = n$, and $E \subset (V \times V)$ denotes the set of edges. We denote by A the adjacency matrix of G, i.e.,

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

For an unweighted graph, w_{ij} is set to 1 for all $(i, j) \in E$. We denote by D the diagonal degree matrix of G, i.e., $D_{ii} \coloneqq d_i = \sum_{j:(i,j)\in E} w_{ij}$, where d_i is the weighted degree of node i. In this case, $d = (d_i) \in \mathbb{R}^n$ denotes the degree vector, and the volume of a subset of nodes is define as $\mathrm{Vol}(B) = \sum_{i\in B} d_i$ for $B \subseteq V$. We denote by $L = D - A$ the graph Laplacian; and $Q \coloneqq \alpha D + \frac{1-\alpha}{2} L$.

For given sets of indexes $B_1, B_2 \subseteq [n]$, we write M_{B_1,B_2} to denote the submatrix of M indexed by B_1 and B_2. If $B_1 = \{i\}$ is a singleton, we use M_{i,B_2} to indicate the i-th row of M whose columns are indexed by B_2. Analogously, we use $M_{B_1,j}$ to indicate the j-th column of M whose rows are indexed by B_1. We denote by $B_1 \backslash B_2$ a set difference between B_1 and B_2, and denote by $B_1^c = [n] \backslash B_1$ the complement of B_1.

2 ℓ_1-regularized PageRank

PageRank [34, 11] is a popular approach for ranking the nodes of a graph. It is defined as the stationary distribution of a Markov chain, which is encoded by a convex combination of the input distribution $\mathbf{s} \in \mathbb{R}^n$ and the (lazy) random walk operator W:

$$p^{\mathrm{PR}} = \alpha \mathbf{s} + (1 - \alpha) W p^{\mathrm{PR}}, \tag{1}$$

where $W = (\mathbf{I} + AD^{-1})/2$ and where $\alpha \in (0, 1)$ is the teleportation parameter. To measure the ranking or importance of the nodes of the "whole" graph, PageRank is often computed by setting the input vector $\mathbf{s}$ to be a uniform distribution over $[n]$.

For local graph clustering, where the aim is to identify a target cluster, given a seed node in the cluster, the input distribution $\mathbf{s}$ is set to be equal to one for the seed node and zero everywhere else. This "personalized" PageRank [24] measures the closeness or similarity of the nodes to the given seed node, and it outputs a ranking of the nodes that is "personalized" with respect to the seed node (as opposed to the original PageRank, which considers the entire graph). From an operational point of view, the underlying dynamic process in (1) defining personalized PageRank "teleports" a random walker back to the original seed node with probability α.

The personalized PageRank vector can be obtained by solving the linear system (1), but this can be prohibitively expensive—especially when there is a single seed node or a small seed set of seed nodes, and when one is interested in very small clusters in a very large graph. In the seminal work of [6], the authors propose an iterative algorithm, called *Approximate Personalized PageRank (APPR)*, to solve this running time problem. They do so by approximating the personalized PageRank vector, while running in time *independent* of the size of the entire graph. APPR was developed from an algorithmic (or "theoretical computer science") perspective, but it is equivalent to applying a coordinate descent type algorithm to the linear system (1) with a particular scheme of early stopping. Motivated by this, [17] recently proposed the ℓ_1*-regularized PageRank optimization problem*. Unlike APPR, the solution method for the ℓ_1-regularized PageRank optimization problem is purely optimization-based. It uses an ℓ_1 norm regularization to set automatically to be zero nodes dissimilar to the seed node, thereby resulting in a highly sparse output. In this manner, ℓ_1-regularized PageRank can estimate the personalized ranking, while maintaining the most relevant nodes at the same time. Prior work [17] also showed that proximal gradient descent (ISTA) can solve the ℓ_1-regularized PageRank problem, with access to only a small portion of the entire graph, i.e., without even touching the entire graph, thereby allowing the method to easily scale to very large-scale graphs.

In this paper, we investigate the statistical performance of ℓ_1-regularized PageRank by reformulating the local graph clustering into the problem of sparse recovery. Here is a more precise definition of the ℓ_1-regularized PageRank optimization problem from [17] that we consider.

Definition 1 (**ℓ_1-regularized PageRank**)**.** Given a graph $G = (V, E)$, with $|V| = n$, and a seed vector $\mathbf{s} \in \mathbb{R}^n$, the ℓ_1-regularized PageRank [17] on the graph is defined as

$$\widehat{x} = \underset{x \in \mathbb{R}^n}{\arg\min} \left\{ \underbrace{\frac{1}{2} x^\top Q x - \alpha x^\top \mathbf{s}}_{:= f(x)} + \rho\alpha \|Dx\|_1 \right\}, \tag{2}$$

where recall $Q = \alpha D + \frac{1-\alpha}{2} L$, and where $\rho > 0$ is a user-specified parameter that controls the amount of the regularization.

To see the intuition behind (2), observe that if we set $\rho = 0$ and $\widehat{p}^{\text{PR}} = D\widehat{x}$, then we can see that it recovers the original PageRank solution of (1). In other words, the optimization problem in (2) adds an additional ℓ_1 norm regularization to the quadratic objective of the linear system (1) (and set $x = D^{-1} p^{\text{PR}}$) in order to keep the nodes most relevant to the seed node and to set the rest to be zero.

Proximal gradient descent (ISTA) and the locality property. Since the ℓ_1-regularized PageRank problem (2) is convex, there are numerous ways to solve it using convex optimization techniques. Prior work [17] applies proximal gradient descent (ISTA) to the ℓ_1-regularized PageRank optimization problem. This proceeds by alternating between locally updating the PageRank vector (gradient descent step), and applying soft-thresholding step. That is, for $t \geq 0$,

$$x^{(t+1)} = \mathsf{prox}_{\eta\rho\alpha\|\cdot\|_1} \left(x^{(t)} - \eta \nabla f(x^{(t)}) \right), \tag{3}$$

where $\mathsf{prox}_{\eta\rho\alpha\|\cdot\|_1}(z)$ is the soft-thresholding operator, given by

$$\mathsf{prox}_{\eta\rho\alpha\|\cdot\|_1}(z) = \text{sign}(z) \cdot \max\{|z| - \eta\rho\alpha, 0\}.$$

Importantly, it has been shown in [17] that proximal gradient descent, when initialized at zero, solves the ℓ_1-regularized PageRank optimization problem with running time that depends only on

the cardinality of the support and its neighbors of the optimal solution. Therefore, when the size of the support set of the solution is small and it has small external connectivity, the proximal gradient descent finds the local cluster with access to only a small portion of the graph. This property, also called the *strong locality property*, is a key feature of any local graph clustering method, enabling them to handle very large/massive graphs [40].

If we set step size $\eta = 1/d$ (so each component has step size $\eta_i = 1/d_i$) in the updating equation (3), then using $\nabla f(x) = Qx - \alpha \mathbf{s}$ and $x = D^{-1}p$, the proximal gradient descent can be rewritten as

$$D^{-1}p^{(t+1)} = \mathsf{prox}_{\eta\rho\alpha\|\cdot\|_1}\left(D^{-1}\left((1-\alpha)Wp^{(t)} + \alpha\mathbf{s}\right)\right).$$

This shows that ℓ_1-regularized PageRank is the stationary distribution of the process consisting of the original diffusion (1) followed by soft-thresholding. The soft-thresholding then promotes sparsity structure on the unique stationary distribution of ℓ_1-regularized PageRank diffusion.

Additional properties. Here, we state additional properties of ℓ_1-regularized PageRank that will be useful for our analysis. The proof of these lemmas can be found in Appendix B.1.

The following lemma guarantees that the ℓ_1-regularized PageRank vector is non-negative.

Lemma 1. *Let $\widehat{x}$ be the vector given in* (2)*. Then $\widehat{x}$ is non-negative, i.e., $\widehat{x}_j \geq 0$ for all $j \in V$.*

The following lemma guarantees that the gradient of f at the optimal solution $\widehat{x}$ cannot be positive.

Lemma 2. *Let* $\mathrm{support}(\widehat{x}) := \{i \in [n] \mid \widehat{x}_i \neq 0\}$ *be the support set of the optimal solution. Then*

$$\nabla_j f(\widehat{x}) = (Q\widehat{x})_j - \alpha \mathbf{s}_j = \begin{cases} -\rho\alpha d_j & \widehat{x}_j > 0, \\ [-\rho\alpha d_j, 0] & \widehat{x}_j = 0 \text{ and } j \text{ is a neighbor of nonzero node}, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Lemma 2 gives the optimality condition for (2), which we frequently use in the proof of our results.

3 Statistical guarantees under random model

In this section, we introduce a random model that we consider for generating a target cluster, and then we provide recovery guarantees for ℓ_1-regularized PageRank. Our results show that the optimal support of ℓ_1-regularized PageRank recovers the target cluster with bounded false-positives. Under additional assumptions, we show that exact recovery is also possible.

3.1 Random graph model

We will assume the graph is generated according to the following model.

Definition 2 (Local random model). Given a graph $G = (V, E)$ that has n vertices, let $K \subset V$ be a target cluster inside the graph, and let K^c denote the complement of K. If two vertices i and j belong to K, then we draw an edge between i and j with probability p, independently of all other edges; if $i \in K$ and $j \in K^c$, then we draw an edge with probability q, independently of all other edges; and otherwise, we allow any (deterministic or random) model to generate edges among vertices in K^c.

Definition 2 says that the adjacency matrix $A \in \mathbb{R}^{n\times n}$ is symmetric, and for any $i, j \in V$, we have that A_{ij} is an independent draw from a Bernoulli distribution with probability p if $i, j \in K$, and from a Bernoulli distribution with probability q if $i \in K$ and $j \in K^c$. For the rest of the graph, i.e., when both i and j belong to K^c, A_{ij} can be generated from an arbitrary fixed model. Under this definition, we can also naturally define the average graph, which is the graph induced by the average adjacency matrix $\mathbb{E}[A]$, where the expectation is taken with respect to the distribution defined by Definition 2. That is, the average graph is an undirected graph $\bar{G} = (V, E)$ whose adjacency matrix is $\mathbb{E}[A]$, where

$$\mathbb{E}[A_{ij}] = \begin{cases} p & \text{if } i \in K \text{ and } j \in K, \\ q & \text{if } i \in K \text{ and } j \in K^c, \\ \text{Any value} & \text{if } i \in K^c \text{ and } j \in K^c. \end{cases} \tag{5}$$

The average degree matrix is similarly denoted by $\mathbb{E}[D]$ and the average graph Laplacian is defined as $\mathbb{E}[L] = \mathbb{E}[D] - \mathbb{E}[A]$. The model in Definition 2 allows us to formulate the problem of local graph clustering as the recovery of a target cluster. Since we are interested in recovering a single target cluster, it is natural to make assumptions only for nodes in the target cluster and nodes adjacent to the target cluster, and to leave the interactions between other nodes unspecified.

This model is also fairly general and covers several popular random graph models appearing in the literature, including the stochastic block model (SBM) [25, 1] and the planted clustering model [4, 8, 12]. For instance, if the subgraph with the vertices within K^c is generated from the SBM, then the entire graph $G = (V, E)$ follows the SBM. On the other hand, if the subgraph of K^c is generated from the classical Erdős-Rényi model with probability q, the entire graph $G = (V, E)$ follows the Planted Densest Subgraph (in this case nodes in K^c do not belong to any clusters). Hence, the results we obtain here for our model holds more broadly across these different random graph models.

Before we move on to our results, we need additional notation. We write $S \subseteq K$ to denote a singleton of the given seed node. Let $k = |K|$ denote the cardinality of the target cluster. According to our local model, any node in the target cluster has the same average degree, $\mathbb{E}[d_i] = p(k-1) + q(n-k)$, which we denote by $\bar{d}$. For the nodes ℓ outside K, we write $\mathbb{E}[d_\ell]$ to denote its average degree, where the expectation is taken with respect to any distribution. For graphs generated according to Definition 2, the following parameter plays a crucial role in determining the behavior of ℓ_1-regularized PageRank for local graph clustering:

$$\gamma := \frac{p \cdot (k-1)}{\bar{d}} \in (0, 1). \tag{6}$$

Intuitively, one can think of this ratio as the ratio of the random walker staying inside K under the average graph. (Note $\bar{d}$ is the average degree of the target cluster and $p \cdot (k-1)$ is the degree of the target cluster when restricted to the subgraph $\mathbb{E}[A_{K,K}]$.) Thus, we can expect that the performance of any random walk-based methods will depend strongly on the number γ. In the extreme scenario where $\gamma = 1$, we have $q = 0$, while for $\gamma = 0$, we have $p = 0$. With this definition, we can also write $p \cdot (k-1) = \gamma\bar{d}$ and $q(n-k) = (1-\gamma)\bar{d}$.

3.2 Recovery of target cluster with bounded false positives

Here, we investigate the performance of ℓ_1-regularized PageRank on the graph generated by the local random model in Definition 2, and we state two of our main theorems.

Our first main result guarantees full recovery of the target cluster for an appropriate choice of the regularization parameter. In particular, if we set ρ to be less than $\mathcal{O}\left(\frac{\gamma p}{d^2}\right)$, then the optimal solution (2) fully recovers the target cluster K, as long as the seed node S is initialized inside K. The proof of Theorem 1 is given in Section A.4.

Theorem 1 (**Full recovery**). *Suppose that* $p^2 k \geq \mathcal{O}\left(\frac{\log k}{\delta^2}\right)$. *If we set*

$$\rho \leq \left(\frac{1-\alpha}{1+\alpha}\right)^2 \left(\frac{1-\delta}{1+\delta}\right)^2 \frac{\gamma p}{(1+\delta)\bar{d}^2}, \tag{7}$$

then with probability at least $1 - 6\exp(-\mathcal{O}\left(\delta^2 p^2 k\right))$,[1] *the solution to Problem* (2) *fully recovers the cluster* K, *i.e.,*

$$K \subseteq \operatorname{support}(\widehat{x}).$$

Our next main result provides an upper bound on the false positives present in the support set of the ℓ_1-regularized PageRank vector. By "false positives," we mean the nonzero nodes that belong to K^c. We measure the size of false positives using a notion of volume, where we recall the volume of a subset of vertices $B \subset V$ is given by $\operatorname{Vol}(B) = \sum_{i \in B} d_i$. The proof of Theorem 2 is given in Section A.5.

Theorem 2 (**Bounds on false positives**). *Suppose the same conditions as Theorem 1. If we set*

$$\rho = \left(\frac{1-\alpha}{1+\alpha}\right)^2 \left(\frac{1-\delta}{1+\delta}\right)^2 \frac{\gamma p}{(1+\delta)\bar{d}^2}, \tag{8}$$

then with probability at least $1 - 6\exp(-\mathcal{O}\left(\delta^2 p^2 k\right))$,[2] *we have*

$$\operatorname{Vol}(FP) \leq \operatorname{Vol}(K)\big[\underbrace{\left(\frac{1+\alpha}{1-\alpha}\right)^2 \left(\frac{1+\delta}{1-\delta}\right)^3 \frac{1}{\gamma^2} - 1}_{=\mathcal{O}\left(\frac{1}{\gamma^2}\right)-1}\big], \tag{9}$$

where $FP = \{i \in \operatorname{support}(\widehat{x}) : i \in K^c\}$ *is the collection of false positive nodes.*

The above results, Theorem 1 and Theorem 2, show several regimes where ℓ_1-regularized PageRank can fully recover the target cluster with nonvanishing probability. In particular, when $p = \mathcal{O}(1)$, the size of the target cluster k is required to be larger than $\mathcal{O}(\log k)$, which includes the constant size $k = \mathcal{O}(1)$. This is often the regime of interest for local graph clustering, where the goal is to find small- and meso-scale clusters in massive graphs [28, 29]. In addition, Theorem 1 indicates that if γ is small, then we need to set ρ to be small to recover the entire cluster. Intuitively, more mass will leak out to K^c for small γ, so we need to run more steps of random walk (ρ smaller in our optimization framework) to find the right cluster. However, this means that the ℓ_1-regularized PageRank vector will also pick up many nonzero nodes in K^c, resulting in many false positives in the support set. Indeed, Theorem 2 shows that the volume of false positives grows quadratically as $1/\gamma$, so we need γ to be bounded to get a meaningful recovery from local clustering. In the case of $p = \mathcal{O}(1), k = \mathcal{O}(1)$, this amounts to requiring that $q = \mathcal{O}\left(\frac{1}{n}\right)$ in order for the target cluster to keep high mass inside K.

[1] The precise statement is as follows: assume $(1-\delta)p^2 k \geq c_0^{-1}\delta^{-2}\log k$ for a fixed constant $c_0 > 0$, then with probability at least $1 - 6e^{-c_0\delta^2(1-\delta)p^2 k}$, the statement in the theorem holds.

[2] The same probability bound as Theorem 1.

Several other comments are worth making regarding these results. First, the current bound we obtain in (9) may not be tight with respect to α and other constants, and the factor $(\frac{1+\alpha}{1-\alpha})^2$ may be an artifact of our proof (see also the proof of Theorem 2 and Lemma 10). Based on our empirical results, ℓ_1-regularized PageRank performs well across a broad range of α values, and we have not seen much difference in terms of performance among different α's. The role of α in ℓ_1-regularized PageRank is closely tied to the regularization parameter ρ, and we leave the question of selecting optimal α for future work. On the other hand, we think the rate $\mathcal{O}\left(1/\gamma^2\right)$ is still tight, and we demonstrate this through the simulation study (see Section 5).

3.3 Exact recovery of target cluster with no false positives

Next, we study the scenarios under which ℓ_1-regularized PageRank can exhibit a stronger recovery guarantee. Specifically, under some additional conditions, we show that the support set of the optimal solution (2) identifies the target cluster exactly, without making any false positives. For this stronger exact recovery result, we require the following assumption about the parameters of the model.

Assumption 1. *We assume $p = \mathcal{O}(1)$ and $k = \mathcal{O}(1)$, i.e., the within-cluster connectivity and the size of the target cluster do not scale with the size of the graph n. Also, we assume $q = c/n$ for a fixed numerical constant $c > 0$.*

As we noted above, the setting $k = \mathcal{O}(1)$ is often the case of interest for local graph clustering, where we would like to identify small- and medium-scale structure in large graphs [28, 29]. In this case, Assumption 1 requires $p = \mathcal{O}(1)$, so that the underlying "signal" of the problem does not vanish as the size of the graph grows, $n \to \infty$. As discussed earlier, this means q must also scale as $\mathcal{O}\left(n^{-1}\right)$ for the local clustering algorithm to find the target without making many false positives.

Now we turn to the statements of exact recovery guarantees for ℓ_1-regularized PageRank when applied to the noisy graph generated from Definition 2. In particular, the fact that $q = \mathcal{O}(1/n)$ from Assumption 1, allows that with nonvanishing probability there is a node in the target cluster that is solely connected to K. This node will serve as a "good" seed node input in the ℓ_1-regularized PageRank. With this choice of seed node, we now give conditions under which the optimal solution $\widehat{x}$ has no false positives with nonvanishing probability. The proof of Theorem 3 is given in Section A.6.

Theorem 3 (No false positives). *Suppose the same conditions as Theorem 1, and assume also that Assumptions 1 holds. If $k \geq 2(c+3)$, and*

$$\rho \geq \left(\frac{1-\alpha}{1+\alpha}\right)^2 \left(\frac{1-\delta}{1+\delta}\right)^2 \frac{\gamma p}{(1+\delta)\bar{d}^2},\ \alpha \in [0.1, 0.9],\ \delta \geq 0.1, \tag{10}$$

then for n sufficiently large, with probability at least $1 - 6\exp(-\mathcal{O}\left(\delta^2 p^2 k\right)) - (1-\exp(-1.5c))^k - \mathcal{O}\left(n^{-1}\right)$,[3] there is a good starting node in K such that ℓ_1-regularized PageRank parameterized with that node as a seed node satisfies

$$\operatorname{support}(\widehat{x}) \subseteq K,$$

as long as

$$\frac{C(0.5c+1)}{\gamma p} = \mathcal{O}\left(\frac{1}{\gamma p}\right) < d_j \text{ for all node } j \in K^c \text{ adjacent to } K, \tag{11}$$

where $C > 0$ is a universal constant.

[3]The precise statement is as follows: assume $(1-\delta)p^2 k \geq c_0^{-1}\delta^{-2}\log k$ for a fixed constant $c_0 > 0$, then with probability at least $1 - 6e^{-c_0\delta^2(1-\delta)pk} - (1-\exp(-1.5c))^k - \mathcal{O}\left(n^{-1}\right)$, the statement in the theorem holds.

While Theorem 3 guarantees no false positives in the solution of ℓ_1-regularized PageRank, when combined with Theorem 1, it establishes that ℓ_1-regularized PageRank recovers the target cluster exactly, even when the target cluster is constant-sized. We require $\alpha \in [0.1, 0.9]$ and $\delta \leq 0.1$ in the condition (10) to avoid overly complicated constants; while this simplifies the statements of the theorem, it is not difficult to show that a similar result holds more generally.

Some sort of condition like (11) about the realized degree seems necessary in order that the ℓ_1-regularized PageRank has no false positives. The optimization program (2) assigns less weights to low degree nodes in the ℓ_1 penalty, so any nodes adjacent to K will become active unless the ℓ_1-regularized PageRank penalizes them with nontrivial weights. Unlike Theorem 1 and Theorem 2, condition (11) rules out some specific models to which Theorem 3 can be applied. For example, planted clustering model with $p = \mathcal{O}(1)$ and $q = \mathcal{O}(1/n)$ does not satisfy this condition because the degrees in K^c do not concentrate. For the stochastic block model, this condition is still satisfied if nodes adjacent to the target cluster belong to the clusters with degree larger than $\mathcal{O}(1/\gamma p)$. In practice, condition (11) may not be always applicable for every node adjacent to K, in which case the nodes that violate this condition may enter the model as false positives. We require the condition here though, since our model is essentially local and we do not have control outside K beyond its neighbors.

4 Stagewise PageRank and solution paths

In this section, we show that the stagewise algorithm can be used for our ℓ_1-regularized PageRank problem to approximate the whole regularization path. This is possible because we show that ℓ_1 regularization path is monotonic. Furthermore, we show that the stagewise algorithm requires only local operations per iteration. This means that the computational complexity per iteration depends only on the current nonzero nodes and its neighbors. The local operations, in combination with the monotonicity of the stagewise algorithm, allows us to implement the path algorithm without touching the whole graph, thus making the algorithm strongly local and scalable to large-scale graph analysis.

Forward stagewise algorithm is a popular path algorithm used in sparse regression, and it has been widely studied by many authors, including [14, 23, 37, 36, 49, 43]. For the case of ℓ_1 norm regularization, the algorithm produces a sequence of iterates by updating in a direction that maximizes the inner product between the current iterate and the negative gradient of the objective function, and at the same time constrains the direction to have a small ℓ_1 norm. When applied to ℓ_1-regularized PageRank optimization problem, we then obtain the following coordinate-wise scheme (recall f is the objective function of ℓ_1-regularized PageRank (2)): for $t \geq 0$,

$$\begin{cases} \text{Choose } i \text{ such that } \left| d_i^{-1} \nabla_i f(x^{(t)}) \right| \text{ is the largest among } [n]; \\ \text{Update } x_i^{(t+1)} = x_i^{(t)} + \eta / d_i. \end{cases} \tag{12}$$

The two main features of this algorithm are: 1) we greedily select the coordinate i at each iteration that maximizes the magnitude of gradient, and 2) we update the current iterate by adding a small step size η to the ith coordinate. This conservative update of the variable counterbalances the greedy selection step, thus making the algorithm more stable.

The stagewise algorithm is known to have implicit regularization effects closely related to ℓ_1 norm regularization [43], and if each component of the ℓ_1-regularized solution has a monotone path, then the sequence of outputs generated by the stagewise algorithm exactly coincides with the ℓ_1 regularization path, as the step size vanishes $\eta \to 0$ [14, 37]. Interestingly, we show that this is indeed our case for ℓ_1-regularized PageRank. The stagewise algorithm is recently advocated

by [43], even when there is no correspondence to the ℓ_1-regularization path, due to its computational efficiency and implicit regularization effect.

The following theorem establishes the monotonicity of the ℓ_1 regularization path. The proof of Theorem 4 is given in Section A.7.

Theorem 4 (**Monotonicity of solution path**). *Let $\widehat{x}(\rho)$ denote the solution for* (2) *indexed by $\rho > 0$. Then, $\widehat{x}(\rho)$ is monotone as a function of ρ, i.e., $\widehat{x}(\rho_0) \leq \widehat{x}(\rho_1)$ whenever $\rho_0 > \rho_1$, where $\leq$ is applied component-wise. Moreover, if $\widehat{x}_i(\rho)$ is positive on node i, the inequality becomes strict.*

Theorem 4 shows that once a node is picked up by ℓ_1-regularized PageRank at some $\rho > 0$, then it will never leave the model thereafter. Furthermore, Theorem 4, combined with known results from [14, 37], guarantees that the sequence of PageRank vectors produced by (12) gives a *provable* approximation to the trajectories of ℓ_1-regularized PageRank as ρ varies, and in the limit as $\eta \to 0$ it *exactly* coincides. The following corollary is thus immediate and we omit the proof.

Corollary 1. *The stagewise algorithm described in* (12) *converges to the ℓ_1-regularized PageRank solution path as the step size goes to* 0*, i.e., $\eta \to 0$.*

Therefore, the stagewise algorithm allows us to explore the entire ℓ_1-regularization path via a single run of simple iterative steps. Another advantage of the stagewise algorithm is that it enjoys the locality property, in that the algorithm only touches the chosen nodes and its neighbors as it progresses. This is obvious from the update step of (12) and the expression of the gradient $\nabla_i f(x)$; if the current iterate $x^{(t)}$ has support set $\mathrm{support}(x^{(t)})$, then the gradient $\nabla f(x^{(t)}) = Qx^{(t)} - \alpha \mathbf{s}$ has nonzero components only at $\mathrm{support}(x^{(t)})$ and its neighbors. This implies that one can compute an approximation to the ℓ_1 regularization path, where nodes around the seed node are part of the path, without touching the whole graph. Thus, when terminated early, the algorithm produces an approximate and partial solution path, with access to only a small number of nodes; and the running time of the entire algorithm depends on the size of the output and not on the size of the full graph.

For the ℓ_1-regularized PageRank, the parameter ρ controls the extent to which the random walk has moved farther from the seed, and so different values of ρ can be used to reveal various scales of local clustering structure around the seed node. Therefore, in the setting of local graph clustering, the stagewise algorithm allows us to provably and efficiently track the evolution of a ℓ_1-PageRank diffusion and better understand the local cluster properties of the graph. This is well-suited for the purpose of exploratory graph analysis, and the idea of using path algorithms for exploring the graph has been also studied in [20]. In addition to the exploratory analysis, the stagewise algorithm can be a competitive algorithm to find the target cluster if the size of the target cluster is small/medium and one needs a fine-scale resolution of the solution path. However, when the size of the target cluster is quite large, using optimization algorithms with a coarse grid of regularization parameter may lead to better computational savings without exploring the entire solution path from scratch. Overall, the stagewise algorithm must be used in a way complementary to the optimization algorithms that directly solve (2). We also refer the readers to [43] for comprehensive study of the stagewise algorithm for general sparse modeling problem.

In practice, the performance of the stagewise algorithm relies on the selection of step size η: when η is large, the stagewise PageRank can fail to approximate the solution path and miss details of the local structure; and when η is small, the algorithm is capable of accurately approximate the solution path, but at the cost of many more iterations. The issue of tuning step size is important for implementing the stagewise algorithm in general, and we refer the reader to [43] for more details. In Section 5.1, we investigate the empirical performance of stagewise PageRank for different values of step size.

5 Empirical evaluation

In this section, we provide an empirical evaluation of our main theoretical results. In particular, we use simulated data to illustrate the behavior of the ℓ_1-regularized PageRank method and the stagewise algorithm, for various parameter settings of the stochastic block model; and we apply our method to an image data set, in order to seek certain segments in the image, given pixels in that segment. To solve Problem (2), we use proximal gradient descent, which is known to enjoy both the locality property and a linear convergence rate [17].

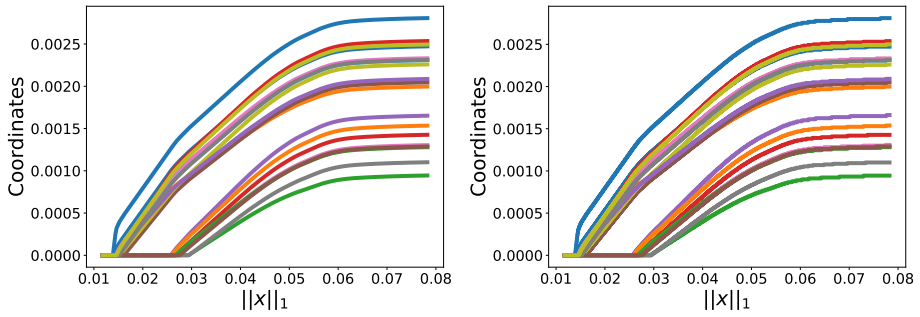
5.1 Simulated data

Here, we run a series of simulations on the graph generated from the stochastic block model. In all cases, the model consists of r clusters, each of which has the same number of nodes and only one of which is the target cluster K. We use the same parameters p and q across different clusters to generate edges within and between clusters. We fix $\alpha = 0.1$.

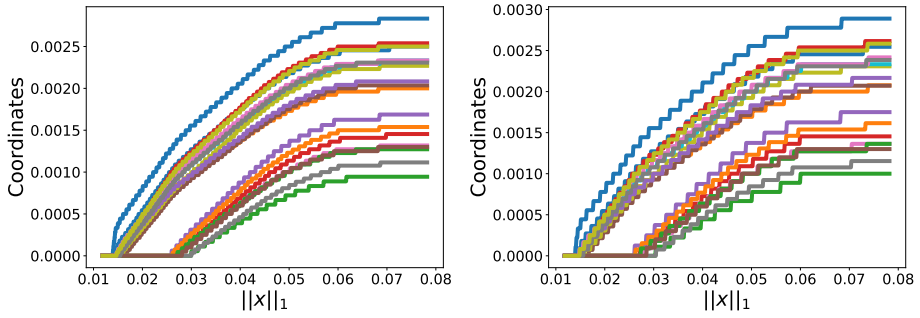
Simulation 1. ℓ_1-regularized solution path and stagewise algorithm. We have seen in Section 4 that the stagewise algorithm generates a provable approximation to the ℓ_1-regularization path. Here, we visually compare the actual solution path to the stagewise algorithm paths for different step sizes. We generate data from the stochastic block model with $p = 0.5, q = 0.002$ and $r = 50$. Each cluster has 20 nodes. Figure 1 shows the ℓ_1-regularization path and stagewise component paths for one particular draw from the stochastic block model. We only show the solution paths for nodes in the target cluster without seed node, among $n = 1000$ nodes. Note that when the step size is small, $\eta = 0.0001$, the stagewise path appears to closely match to the ℓ_1-regularization path; for moderate step size, $\eta = 0.0005$, the stagewise path exhibits some jagged pattern but nevertheless accurately approximates the optimal path; and for relatively large step size, $\eta = 0.001$, while the jagged pattern becomes more evident visually, it is clear that the overall trend still coincides well with that of the ℓ_1 solution path.

Simulation 2. Bounds on the false positives. Next, we study the role of γ in determining the quality of the cluster recovered from noisy graph. In particular, we examine the bound predicted by Theorem 2, $\mathrm{Vol}(\mathrm{FP}) \leq \mathrm{Vol}(K)(\mathcal{O}\left(1/\gamma^2\right) - 1)$, using the simulated data. The graph is generated from the stochastic block model where we fix $k = 100$ and $r = 40$. We vary $p \in \{0.4, 0.6, 0.8, 1\}$, and γ from 0.45 to 0.95 with evenly spaced on the $\frac{1}{\gamma^2}$-scale, to get the total 20 values. The probability q is then determined accordingly. To generate the plots, we solve problem (2) and select the largest ρ that recovers the entire target cluster. Figure 2(a) shows $\mathrm{Vol}(\mathrm{FP})/\mathrm{Vol}(K)$ plotted against $1/\gamma^2$; we see that the error scales approximately linearly, which is consistent with our finding. Shown in Figure 2(b) is the same figure as Figure 2(a), except that each curve is now scaled by p. We see all the curves nearly coincide, which suggests that the result we obtain in Theorem 2 may not be tight. Improving the bound would be an interesting future direction. Finally, Figure 2(c) zooms in Figure 2(a) for large values of γ where $\mathrm{Vol}(\mathrm{FP})/\mathrm{Vol}(K)$ is strictly less than 0.6. In this case, the recovered cluster contains only a small portion of false positives, showing good recovery results.

Simulation 3. Selection of the tuning parameter. Finally, we discuss the practical issue of finding the regularization parameter in (2), or equivalently the number of iterations in the stagewise algorithm. We define the true positive rates and false rates as $\mathrm{TR} := \mathrm{Vol}(\mathrm{TP})/\mathrm{Vol}(K)$ and $\mathrm{FR} := \mathrm{Vol}(\mathrm{FP})/\mathrm{Vol}(\mathrm{support}(\widehat{x}))$, respectively. We will also make use of conductance, which is

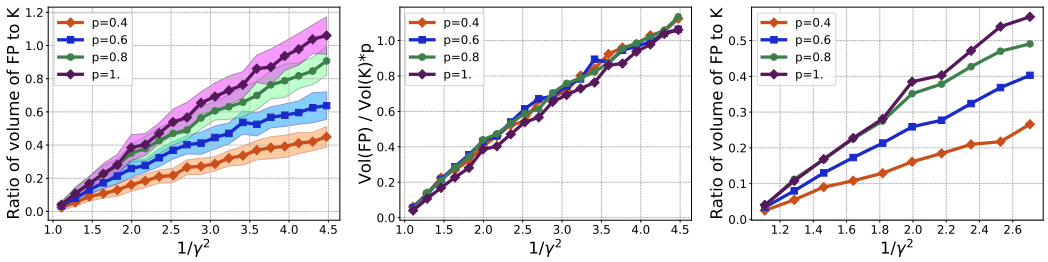


(a) ℓ_1-reg. path (b) Stage. path $\eta = 10^{-4}$



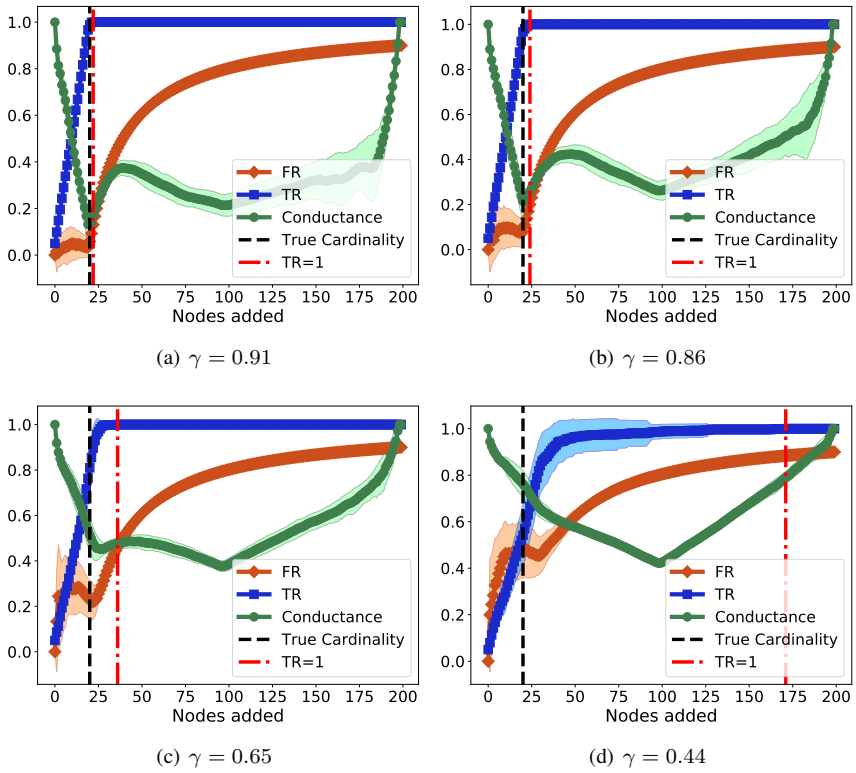
(c) Stage. path $\eta = 5 \cdot 10^{-3}$ (d) Stage. path $\eta = 10^{-3}$

Figure 1: Comparison of the ℓ_1 regularization path and stagewise paths for different step sizes. The profiles are shown only for nodes in $K \backslash S$ among $n = 1000$ nodes. The x-axis is the ℓ_1 norm of the current estimates. For the stagewise algorithm, the results are obtained with 7743, 1530, and 763 iterations, respectively.



(a) Vol(FP)/Vol(K) (b) Vol(FP)/$p \cdot$ Vol(K) (c) Zoomed in for large γ

Figure 2: Volume of false positives for varying γ. Vol(FP)/Vol(K), and its scaled and zoomed-in versions, are plotted against $1/\gamma^2$, for various p. The results are averaged over 30 trials.



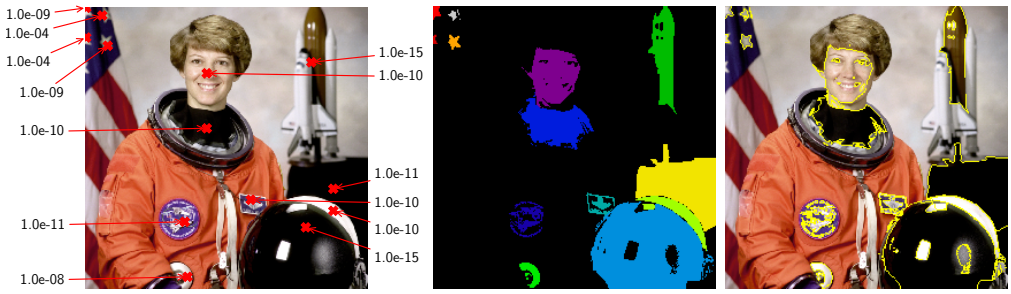
(a) $\gamma = 0.91$ (b) $\gamma = 0.86$

(c) $\gamma = 0.65$ (d) $\gamma = 0.44$

Figure 3: False discovery rates, true positive rates, and conductance for varying γ. The x-axis gives the number of active nodes of the current estimates of the stagewise algorithm, and the vertical lines indicate, respectively, the true cardinality ($k = 20$) and the number of active nodes that reach TR $= 1$.

defined as the ratio $\text{Cut}(S, S^c)/\min(\text{Vol}(S), \text{Vol}(S^c))$, where $\text{Cut}(S, S^c) := \sum_{i \in S, j \in S^c} A_{ij}$. Conductance measures the weight of the edges that are being removed (numerator) divided by the volume of the cluster (denominator). Lower conductance values correspond to better quality clusters. We generate the target cluster using a stochastic block model with $r = 10$ clusters and each cluster has 20 nodes. We set $p = 0.5$ and q is varying in order to generate various γ, as is shown in Figure 3. For each experiment we run stagewise algorithm with $\eta = 0.0001$ and the results are averaged over 30 trials.

Figure 3 illustrates how the false discovery rates and true positive rates and conductance change as the stagewise algorithm explores the ℓ_1-regularization path. For large γ, Figure 3(a), conductance is a good metric for finding the target cluster. This means that we will find the target cluster with low FR and large TR if out of all solutions on the path we choose the one with minimum conductance. As γ gets smaller, the minimum conductance does not relate to the target cluster. However, it is clear from Figures 3(b) and 3(c) that the stagewise algorithm with minimum conductance still finds the target cluster K with good accuracy if the algorithm is terminated early. Finally, in Figure 3(d) we demonstrate a case where γ is small and conductance fails completely to relate to the target cluster. Developing general strategy for selecting the number of iterations, or the regularization parameter, is generally challenging, and we leave this for future work.



(a) Image, seed nodes and ρ parameters (b) Recovered clusters (c) Boundaries of recovered clusters

Figure 4: Recovery of different target clusters. The results are obtained by solving ℓ_1-regularized PageRank problems with the seed nodes and regularization parameter ρ, as depicted in Figure 4(a).

5.2 Image data

In this section, we present the performance of ℓ_1-regularized PageRank on real graphs generated from image data. In particular, given an image, we generate an adjacency matrix using the color and position features of pixels, as described in [38]. Then we run local graph clustering by using certain pixels as seed nodes. For each seed node, we solve a different ℓ_1-regularized problem. In Figure 4 we demonstrate that there exists ρ such that ℓ_1-regularized PageRank finds a good approximation of certain segments in the image. The seed nodes are depicted in Figure 4(a), and the target clusters are self-explanatory by looking at Figure 4(b) and Figure 4(c). We provide details about the seed nodes and the regularization parameter ρ in Figure 4(a). For all clusters we use $\alpha = 0.1$.

6 Discussion

We have examined from a statistical perspective the ℓ_1-regularized PageRank algorithm for the local graph clustering problem. To do so, we reformulated the local graph clustering problem as a statistical recovery problem, where we impose a certain local random model on the graph, and then our task is to recover the target cluster generated from the model. Our results show that the optimal support of the ℓ_1-regularized PageRank vector identifies the target cluster with bounded false positives, and in certain settings exact recovery is also possible. Additionally, we have brought the idea of solution path algorithms from the sparse regression literature to the local graph clustering literature, and we showed that the forward stagewise algorithm gives a provable approximation to the entire ℓ_1 regularization path of this algorithm. This extends the domain of existing local graph clustering algorithm, providing an efficient tool for exploring very large graphs, without the need even to touch the entire graph. This is a primitive that can be highly desirable in modern data analysis, where computational challenges are of great concern.

Acknowledgements. We would like to acknowledge ARO, DARPA, NSF, ONR, and Intel for providing partial support for this work.

References

[1] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

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

[3] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2015.

[4] Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. *Random Structures & Algorithms*, 13(3-4):457–466, 1998.

[5] Arash A Amini, Elizaveta Levina, et al. On semidefinite relaxations for the block model. *The Annals of Statistics*, 46(1):149–179, 2018.

[6] R. Andersen, F. Chung, and K. Lang. Local graph partitioning using pagerank vectors. *FOCS '06 Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 475–486, 2006.

[7] R. Andersen, S. O. Gharan, Y. Peres, and L. Trevisan. Almost optimal local graph clustering using evolving sets. *Journal of the ACM*, 63(2), 2016.

[8] Ery Arias-Castro, Nicolas Verzelen, et al. Community detection in dense random networks. *The Annals of Statistics*, 42(3):940–969, 2014.

[9] Taylor B Arnold and Ryan J Tibshirani. Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics*, 25(1):1–27, 2016.

[10] Richard Arratia and Louis Gordon. Tutorial on large deviations for the binomial distribution. *Bulletin of mathematical biology*, 51(1):125–131, 1989.

[11] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

[12] Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *The Journal of Machine Learning Research*, 17(1):882–938, 2016.

[13] Yudong Chen, Xiaodong Li, Jiaming Xu, et al. Convexified modularity maximization for degree-corrected stochastic block models. *The Annals of Statistics*, 46(4):1573–1602, 2018.

[14] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[15] K. Fountoulakis and J. Gondzio. Performance of First- and Second-Order Methods for L1-Regularized Least Squares Problems. *ArXiv e-prints*, March 2015.

[16] K. Fountoulakis, D. F. Gleich, and M. W. Mahoney. An optimization approach to locally-biased graph algorithms. *Proceedings of the IEEE*, 105(2):256–272, 2017.

[17] K. Fountoulakis, F. Roosta-Khorasani, J. Shun, X. Cheng, and M. W. Mahoney. Variational perspective on local graph clustering. *Mathematical Programming B*, pages 1–21, 2017.

[18] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[19] Chao Gao, Zongming Ma, Anderson Y Zhang, Harrison H Zhou, et al. Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185, 2018.

[20] DF Gleich and Kyle Kloster. Seeded pagerank solution paths. *European Journal of Applied Mathematics*, 27(6):812–845, 2016.

[21] Lennart Gulikers, Marc Lelarge, and Laurent Massoulié. A spectral method for community detection in moderately sparse degree-corrected stochastic block models. *Advances in Applied Probability*, 49(3):686–721, 2017.

[22] Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5(Oct):1391–1415, 2004.

[23] Trevor Hastie, Jonathan Taylor, Robert Tibshirani, Guenther Walther, et al. Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1:1–29, 2007.

[24] Taher H Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526. ACM, 2002.

[25] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[26] K. Kloster and D. F. Gleich. Heat kernel based community detection. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1386–1395, 2014.

[27] I. M. Kloumann and J. M. Kleinberg. Community membership identification from small seed sets. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1366–1375, 2014.

[28] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet of Mathematics*, 6(1):29–123, 2009.

[29] J. Leskovec, K.J. Lang, and M.W. Mahoney. Empirical comparison of algorithms for network community detection. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, pages 631–640, 2010.

[30] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703. ACM, 2014.

[31] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015.

[32] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018.

[33] Mark EJ Newman, Duncan J Watts, and Steven H Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572, 2002.

[34] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

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

[36] Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, pages 1012–1030, 2007.

[37] Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5(Aug):941–973, 2004.

[38] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 22(8):888–905, 2000.

[39] P. Shi, K. He, D. Bindel, and J. Hopcroft. Local Lanczos spectral approximation for community detection. In *Proceedings of ECML-PKDD*, September 2017.

[40] J. Shun, F. Roosta-Khorasani, K. Fountoulakis, and M. W. Mahoney. Parallel local graph clustering. *Proceedings of the VLDB Endowment*, 9(12):1041–1052, 2016.

[41] D. A. Spielman and S. H. Teng. A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM Journal on Scientific Computing*, 42(1): 1–26, 2013.

[42] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[43] Ryan J Tibshirani. A general framework for fast stagewise algorithms. *The Journal of Machine Learning Research*, 16(1):2543–2588, 2015.

[44] N. Veldt, C. Klymko, and D. F. Gleich. Flow-based local graph clustering with better seed set inclusion. *ArXiv e-prints*, November 2018.

[45] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[46] D. Wang, K. Fountoulakis, M. Henzinger, and M. W. Mahoney. Capacity releasing diffusion for speed and locality. *Proceedings of the 34th International Conference on Machine Learning*, 70:3607–2017, 2017.

[47] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich. Local higher-order graph clustering. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 555–564, 2017.

[48] Anderson Y Zhang, Harrison H Zhou, et al. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280, 2016.

[49] Peng Zhao and Bin Yu. Stagewise lasso. *Journal of Machine Learning Research*, 8(Dec): 2701–2726, 2007.

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

[51] Z. A. Zhu, S. Lattanzi, and V. Mirrokni. A local algorithm for finding well-connected clusters. In *Proceedings of the 30th International Conference on Machine Learning*, pages 396–404, 2013.

[52] Hui Zou, Trevor Hastie, Robert Tibshirani, et al. On the âĂIJdegrees of freedomâĂİ of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.

A Proofs of theorems

In this section, we prove all of our theorems. To establish the theorems, we need a few concentration inequalities and intermediate results that shall be used in the proofs. In Section A.1, we give degree concentration inequalities for random graphs, and in Section A.2, we state recovery guarantees of ℓ_1-regularized PageRank on the average graph. Section A.3 gives a few important results on the ℓ_1-regularized PageRank when restricted to the target cluster. Based on these results, we give proofs of our four theorems, respectively, in Section A.4, Section A.5, Section A.6, and Section A.7. All of the proofs of intermediate lemmas are defered to Section B.

A.1 Concentration lemmas

Here, we present several concentration lemmas for degrees of random graphs. The first lemma is consequence of Chernoff's inequality for the sum of independent Bernoulli random variables, and applying union bound (c.f. see Vershynin [45, Theorem 2.3.1]).

Lemma 3 (Adapted from Vershynin [45, Proposition 2.4.1])**.** *Let X_i be the sum of independent Bernoulli random variables, with $\mathbb{E}[X_i] = \mu$ for $i = 1, 2, \ldots, k$. Then, there exists a universal constant $c_0 > 0$ such that if $\mu \geq c_0^{-1}\delta^{-2}\log k$, then with probability at least $1 - 2e^{-c_0\delta^2\mu}$,*

$$\textit{For all } i \in K\text{:} \quad |X_i - \mu| \leq \delta\mu.$$

According to our random model (Definition 2), the degree vector d_K for the target cluster K comprises of random variables which are the sum of independent Bernoulli random variables with common mean $\bar{d}$. Therefore, by applying Lemma 3, it is straightforward to see the following result.

Lemma 4 (Adapted from Vershynin [45, Proposition 2.4.1])**.** *There exists a universal constant $c_0 > 0$ such that if $\bar{d} \geq c_0^{-1}\delta^{-2}\log k$, then with probability at least $1 - 2e^{-c_0\delta^2\bar{d}}$,*

$$\textit{For all } i \in K\text{:} \quad \left|d_i - \bar{d}\right| \leq \delta\bar{d},$$

where $\bar{d}$ is the average degree of the vertices in the cluster K.

The next lemma bounds the number of edges between node $j \in K^c$ and the target cluster K when $q = \mathcal{O}\left(\frac{1}{n}\right)$.

Lemma 5. *Suppose that $q = c/n$ and $k \geq 2(c+3)$ for a positive constant c. Then for n sufficiently large, with probability at least $1 - \mathcal{O}\left(n^{-1}\right)$,*

$$\max_{j \in K^c} \|A_{j,K}\|_1 \leq c + 2.$$

A.2 Exact recovery for expected graph

Here we define the "ground truth" which we obtain by applying ℓ_1-regularized PageRank to the average graph of our random model. Recall the adjacency matrix $\mathbb{E}[A]$ defined in (5), the associated diagonal degree matrix $\mathbb{E}[D]$, and the Laplacian matrix $\mathbb{E}[L]$. Writing $\mathbb{E}[Q] = \alpha\mathbb{E}[D] + \frac{1-\alpha}{2}\mathbb{E}[L]$, we denote the average ℓ_1-regularized PageRank as

$$x^\star = \arg\min_x \left\{ \frac{1}{2} x^\top \mathbb{E}[Q]\, x - \alpha x^\top \mathbf{s} + \rho\alpha \|\mathbb{E}[D]\, x\|_1 \right\}. \tag{13}$$

Compared to (2), we see that the matrices associated with the graph are now replaced by its expected counterpart.

The following lemma shows that there exist a range of ρ values for which the average ℓ_1-regularized PageRank gives an exact recovery for the target cluster, namely $\mathrm{support}(x^\star) = K$.

Lemma 6. *Consider the local random model given in Definition 2 and suppose that Assumption 1 holds. Then, for n sufficiently large, the average ℓ_1-regularized PageRank $x^\star$, defined in* (13)*, identifies K correctly, that is,*

$$\mathrm{support}(x^\star) = K,$$

as long as $\rho \in [\rho^\sharp, \rho^\natural)$, where

$$\rho^\sharp = \frac{q(1-\alpha)}{2\alpha\bar{d}\cdot\mathbb{E}\left[d_{\ell_\sharp}\right] + q(1-\alpha)\left[k\bar{d} + (n-k)\mathbb{E}\left[d_{\ell_\sharp}\right]\right]}, \quad \textit{and} \quad \rho^\natural = \frac{p(1-\alpha)}{\bar{d}\left[(1+\alpha)\bar{d} + (1-\alpha)p\right]},$$

where $\mathbb{E}\left[d_{\ell_\sharp}\right] = \min_{\ell\in K^c}\mathbb{E}[d_\ell]$. Furthermore, $x^\star$ has a closed form expression, $x^\star = u\cdot\mathbf{1}_S + v\cdot\mathbf{1}_K$, where u and v are given by

$$\begin{cases} u = \frac{2\alpha}{(1+\alpha)\bar{d} + (1-\alpha)p}; \\ v = \frac{\frac{1-\alpha}{2}p\cdot u - \rho\alpha\bar{d}}{\alpha\bar{d} + \frac{1-\alpha}{2}q(n-k)}. \end{cases}$$

Lemma 6 shows that $x^\star$ has high probability mass on the seed node with a long tail further away from the seed node. The mass outside the target cluster is being thresholded exactly to zero via ℓ_1-norm regularization, thus identifying the target cluster without any false positives. Of course, in practice, the average graph is not available, and we are instead given an instance of the graph from the random model. In this case, the average ℓ_1-regularized PageRank, $x^\star$, can serve as a ground truth which will allow us to estimate the magnitude of the ℓ_1-regularized PageRank vector by analyzing the error between $x^\star$ and $\widehat{x}$. Lemma 6 is being used in Lemma 9 below.

A.3 Reduced ℓ_1-regularized PageRank

Here, we introduce the following regularized optimization problem:

$$\begin{aligned} \widehat{x}^{\mathrm{R}} &= \arg\min_{x\in\mathbb{R}^n} \left\{ \frac{1}{2} x^\top Q x - \alpha x^\top \mathbf{s} + \rho\alpha\|Dx\|_1 : x_{K^c} = 0 \right\} \\ &= \arg\min_{y\in\mathbb{R}^k} \left\{ \frac{1}{2} y^\top Q_{K,K} y - \alpha y^\top \mathbf{s}_K + \rho\alpha\|D_{K,K} y\|_1 \right\}. \end{aligned} \tag{14}$$

Note that the above reduced problem is defined as the ℓ_1-regularized PageRank problem restricted to the target cluster K. Since our aim is to recover the target cluster K based on the *local* model

(Definition 2), we need to analyze the properties of the solution to this reduced problem. Here, abusing notation, we use $\widehat{x}^{\mathrm{R}}$ to denote the vector either in the original space $\mathbb{R}^n$ or in the reduced space $\mathbb{R}^k$.

We present several lemmas about $\widehat{x}^{\mathrm{R}}$ that will be critical for the proof of theorems. First, we give a guarantee on the recovery of the target cluster for the optimal solution (14).

Lemma 7. *Let $\widehat{x}^{\mathrm{R}}$ be the solution to the reduced problem. Suppose that $(1-\delta)p^2k \geq c_0^{-1}\delta^{-2}\log k$. Then, if ρ satisfies* (7)*, with probability at least $1-6e^{-c_0\delta^2(1-\delta)p^2k}$, we have*

$$\operatorname{support}(\widehat{x}^{\mathrm{R}}) = K.$$

Here c_0 is the same constant appearing in Lemma 3.

For the above result, since by construction $\widehat{x}^{\mathrm{R}}$ is zero outside K, all we need to show is that $\widehat{x}^{\mathrm{R}}_K > 0$ for $\rho \approx \mathcal{O}\left(\frac{\gamma p}{\bar{d}^2}\right)$. Next, we compare the support set of $\widehat{x}^{\mathrm{R}}$ to $\operatorname{support}(\widehat{x})$.

Lemma 8. *For any regularization parameter $\rho \in [0,\infty)$, we have*

$$\operatorname{support}(\widehat{x}^{\mathrm{R}}) \subseteq \operatorname{support}(\widehat{x}).$$

Finally, we have the following estimate on the entries of $\widehat{x}^{\mathrm{R}}$ on $K\backslash S$ (recall u and v are the components of $x^\star$ on S and $K\backslash S$ respectively, see Lemma 6).

Lemma 9. *Suppose that Assumption 1 holds and that $(1-\delta)p^2k \geq c_0^{-1}\delta^{-2}\log k$ for $\delta \leq 0.1$. If*

$$\rho = \left(\frac{1-\alpha}{1+\alpha}\right)^2\left(\frac{1-\delta}{1+\delta}\right)^2\frac{\gamma p}{(1+\delta)\bar{d}^2}, \quad k \geq 5 \ \text{ and } \ \alpha \in [0.1, 0.9],$$

then for n sufficiently large, with probability at least $1-6e^{-c_0\delta^2(1-\delta)p^2k}$, we have

$$\|\widehat{x}^{\mathrm{R}}_{K\backslash S}\|_\infty \leq \underbrace{v}_{=x^\star_{K\backslash S}} + \underbrace{\frac{c_1(1-\alpha)u}{\bar{d}} + 2c_1\delta(v+\rho\alpha)}_{\geq \|\widehat{x}^{\mathrm{R}}_{K\backslash S} - x^\star_{K\backslash S}\|_\infty}.$$

Here c_0 is the same constant appearing in Lemma 3, and c_1 is a positive numerical constant.

A.4 Proof of Theorem 1

The proof of Theorem 1 is now a straightforward combination of Lemma 7 and Lemma 8. By Lemma 7, we know that for $\rho \leq \left(\frac{1-\alpha}{1+\alpha}\right)^2\left(\frac{1-\delta}{1+\delta}\right)^2\frac{\gamma p}{(1+\delta)\bar{d}^2}$, we have $\operatorname{support}(\widehat{x}^{\mathrm{R}}) = K$. Then Lemma 8 gives

$$\operatorname{support}(\widehat{x}^{\mathrm{R}}) = K \subseteq \operatorname{support}(\widehat{x}),$$

proving the result.

A.5 Proof of Theorem 2

To prove Theorem 2, we first need the following lemma for bounding the volume of $\operatorname{support}(\widehat{x})$. This lemma is a stronger version of [15, Theorem 2].

Lemma 10. *For any regularization parameter $\rho > 0$, it holds that*

$$\mathrm{Vol}(\mathrm{support}(\widehat{x})) \leq \frac{1 - d^\top \widehat{x}}{\rho},$$

where $\mathrm{Vol}(\mathrm{support}(\widehat{x})) = \sum_{i \in \mathrm{support}(\widehat{x})} d_i$.

Now, by our choice of ρ (8), then

$$\begin{aligned}\mathrm{Vol}(\mathrm{support}(\widehat{x})) &\leq \left(\frac{1+\alpha}{1-\alpha}\right)^2 \left(\frac{1+\delta}{1-\delta}\right)^2 \frac{(1+\delta)\bar{d}^2}{\gamma p} \\ &= \left(\frac{1+\alpha}{1-\alpha}\right)^2 \left(\frac{1+\delta}{1-\delta}\right)^2 \frac{(1+\delta)k\bar{d}}{\gamma^2}, \end{aligned} \tag{15}$$

where the second step follows from (6). Furthermore, by Theorem 1, $\mathrm{support}(\widehat{x})$ contains the target cluster K, so the errors in $\mathrm{support}(\widehat{x})$ are solely attributed to the presence of false positives. Denoting by FP the set of false positives in $\mathrm{support}(\widehat{x})$, we can write

$$\mathrm{Vol}(\mathrm{support}(\widehat{x})) = \mathrm{Vol}(K) + \mathrm{Vol}(\mathrm{FP}). \tag{16}$$

Since $\mathrm{Vol}(K) = \sum_{i \in K} d_i \geq (1-\delta)k\bar{d}$ by Lemma 4, it follows that

$$\begin{aligned}\mathrm{Vol}(\mathrm{FP}) &= \mathrm{Vol}(\mathrm{support}(\widehat{x})) - \mathrm{Vol}(K) \quad \text{by (16)} \\ &\leq \left(\frac{1+\alpha}{1-\alpha}\right)^2 \left(\frac{1+\delta}{1-\delta}\right)^2 \frac{(1+\delta)k\bar{d}}{\gamma^2} - \mathrm{Vol}(K) \quad \text{by (15)} \\ &\leq \mathrm{Vol}(K) \left[\left(\frac{1+\alpha}{1-\alpha}\right)^2 \left(\frac{1+\delta}{1-\delta}\right)^3 \frac{1}{\gamma^2} - 1 \right], \end{aligned}$$

with probability at least $1 - 2\exp(-c_0\delta^2\bar{d})$. This proves the theorem.

A.6 Proof of Theorem 3

To prove Theorem 3, we first show that with nonvanishing probability, the target cluster has a good seed node that is connected only within K.

Lemma 11. *Suppose that $q = \frac{c}{n}$ for a fixed constant $c > 0$ (Assumption 1). Under the local random model given in Definition 2, if n is sufficiently large, then it holds that*

$$\mathbb{P}\left\{\textit{There exists a node } i \in K \textit{ such that } i \textit{ is not connected to } K^c\right\} \geq 1 - (1 - \exp(-1.5c))^k.$$

Then we select this node as a seed node in the ℓ_1-regularized PageRank.

Now we show that under the conditions of Theorem 3, the optimal solution (14) to the reduced problem, $\widehat{x}^{\mathrm{R}}$, obeys

$$|(Q\widehat{x}^{\mathrm{R}} - \alpha\mathbf{s})_j| = |-\frac{1-\alpha}{2} \cdot A_{j,K}\widehat{x}^{\mathrm{R}}| < \rho\alpha d_j \text{ for } j \in K^c. \tag{17}$$

Notice that the above condition is same as the optimality condition (4) for the full-dimensional problem, for nodes in K^c. Also, by definition, $\widehat{x}^{\mathrm{R}}$ satisfies (4) for nodes in K. Hence, if $\widehat{x}^{\mathrm{R}}$ satisfies (17),

by uniqueness of the solution it follows that $\widehat{x}^{\mathrm{R}} = \widehat{x}$. Since $\mathrm{support}(\widehat{x}^{\mathrm{R}}) \subseteq K$ by construction, this proves that there is no false positive in $\widehat{x}$.

Write $A_{j,K}\widehat{x}^{\mathrm{R}}$ into the sum of two terms,

$$A_{j,K}\widehat{x}^{\mathrm{R}} = A_{j,S}\widehat{x}^{\mathrm{R}}_S + A_{j,K\setminus S}\widehat{x}^{\mathrm{R}}_{K\setminus S}.$$

The first term can be ignored, since by Lemma 11, we choose the seed node to be the one that is solely connected to K. For the second term, we use Hölder's inequality to bound

$$A_{j,K\setminus S}\widehat{x}^{\mathrm{R}}_{K\setminus S} \leq \|A_{j,K\setminus S}\|_1 \|\widehat{x}^{\mathrm{R}}_{K\setminus S}\|_\infty.$$

Applying Lemma 5, Lemma 9, and using the fact that $u \leq \frac{2\alpha}{(1+\alpha)\bar{d}}, v \leq \frac{(1-\alpha)p}{(1+\alpha)\bar{d}^2}$ (which can be deduced from Lemma 6), and $\delta \leq 0.1$ (assumption of theorem), then

$$\begin{aligned}
\left|(Q\widehat{x}^{\mathrm{R}} - \alpha\mathbf{s})_j\right| &\leq \frac{1-\alpha}{2}\|A_{j,K\setminus S}\|_1\|\widehat{x}^{\mathrm{R}}_{K\setminus S}\|_\infty \\
&\leq (0.5c+1)(1-\alpha)\left[\frac{c_1(1-\alpha)u}{\bar{d}} + (1+2c_1\delta)v + 2c_1\delta\cdot\rho\alpha\right] \\
&\leq (0.5c+1)(1-\alpha)\left[\frac{2c_1\alpha(1-\alpha)}{\bar{d}^2(1+\alpha)} + \frac{(1+0.2c_1)(1-\alpha)p}{\bar{d}^2(1+\alpha)} + 0.2c_1\rho\alpha\right],
\end{aligned}$$

where $c_1 > 0$ is a fixed positive constant. Since $\alpha \in [0.1, 0.9]$ by assumption, we can further bound it as

$$\left|(Q\widehat{x}^{\mathrm{R}} - \alpha\mathbf{s})_j\right| \leq (0.5c+1)(0.2c_1+1)\left(\frac{2}{\bar{d}^2} + \rho\alpha\right).$$

We need the above bound to be strictly less than $\rho\alpha d_j$, or

$$(0.5c+1)(0.2c_1+1)\left(\frac{2}{\rho\alpha\bar{d}^2} + 1\right) < d_j.$$

Plugging in expression (10) to the above inequality, and using $\alpha \in [0.1, 0.9]$ and $\delta \leq 0.1$, we obtain

$$\frac{(0.5c+1)(0.2c_1+1)c_2}{\gamma p} \leq (0.5c+1)(0.2c_1+1)\left[\left(\frac{1+\alpha}{1-\alpha}\right)^2\left(\frac{1+\delta}{1-\delta}\right)^2\frac{1+\delta}{\alpha\gamma p} + 1\right] < d_j,$$

for some constant $c_2 > 0$. This proves the theorem.

A.7 Proof of Theorem 4

First, by Rosset and Zhu [36, Proposition 1], we know that the optimal solution path $\widehat{x}(\rho)$ for the ℓ_1-regularized problem (2) is piecewise linear as a function of $\rho > 0$, and in particular this implies that the path is continuous.

Next, we prove that if the support set of $\widehat{x}(\rho)$ remains constant in the interval $[\rho_1, \rho_0]$ ($\rho_0 > \rho_1$), then $\widehat{x}(\rho)$ is strictly decreasing on $[\rho_1, \rho_0]$, i.e., $\widehat{x}(\rho_1) > \widehat{x}(\rho_0)$, where the inequality applies component-wise. Combining with the non-negativity of the solution (Lemma 1) and the continuity of the path $\widehat{x}(\rho)$ then establishes the lemma.

Write $\mathcal{A}(\rho) = \{j : \widehat{x}_j(\rho) > 0\}$ for $\rho > 0$ and choose $\rho_0, \rho_1 > 0$ with $\rho_0 > \rho_1$ such that the support set $\mathcal{A}(\rho) = \mathcal{A}$ does not change for $\rho \in [\rho_1, \rho_0]$. For $j \in \mathcal{A}$, by optimality condition, we have that

$$(Q\cdot\widehat{x}(\rho))_j - \alpha\mathbf{s}_j = -\rho\alpha d_j,$$

and so

$$Q_{\mathcal{A},\mathcal{A}}(\widehat{x}_{\mathcal{A}}(\rho_1) - \widehat{x}_{\mathcal{A}}(\rho_0)) = (\rho_0 - \rho_1)\alpha d_{\mathcal{A}}.$$

Multiplying $Q_{\mathcal{A},\mathcal{A}}^{-1}$ on both sides (note that $Q_{\mathcal{A},\mathcal{A}}$ is positive definite),

$$\widehat{x}_{\mathcal{A}}(\rho_1) - \widehat{x}_{\mathcal{A}}(\rho_0) = \alpha(\rho_0 - \rho_1)Q_{\mathcal{A},\mathcal{A}}^{-1} d_{\mathcal{A}}. \tag{18}$$

Now write

$$Q_{\mathcal{A},\mathcal{A}} = \frac{1+\alpha}{2} \cdot \left(D - \frac{1-\alpha}{1+\alpha} \cdot A\right)_{\mathcal{A},\mathcal{A}} = \frac{1+\alpha}{2} \cdot D_{\mathcal{A},\mathcal{A}}^{1/2} \left(\mathbf{I} - \frac{1-\alpha}{1+\alpha}(D^{-1/2}AD^{-1/2})_{\mathcal{A},\mathcal{A}}\right) D_{\mathcal{A},\mathcal{A}}^{1/2}.$$

Then, for all $v \in \mathbb{R}^{|\mathcal{A}|}$, we have that

$$\begin{aligned} v^\top v \pm v^\top (D^{-1/2}AD^{-1/2})_{\mathcal{A},\mathcal{A}} v &= \sum_{i\in\mathcal{A}} v_i^2 \pm \sum_{(i,j)\in E, i,j\in\mathcal{A}} \frac{2w_{ij}v_iv_j}{\sqrt{d_i d_j}} \\ &> \sum_{(i,j)\in E, i,j\in\mathcal{A}} w_{ij}\left(\frac{v_i}{\sqrt{d_i}} \pm \frac{v_j}{\sqrt{d_j}}\right)^2 \geq 0, \end{aligned}$$

where the second step holds since $d_i > \sum_{j:(i,j)\in E, j\in\mathcal{A}} w_{ij}$. This shows that all the eigenvalues of $(D^{-1/2}AD^{-1/2})_{\mathcal{A},\mathcal{A}}$ have absolute value less than 1. Using the previous fact, then $Q_{\mathcal{A},\mathcal{A}}^{-1}$ can be written as

$$\begin{aligned} Q_{\mathcal{A},\mathcal{A}}^{-1} &= \frac{2}{1+\alpha} D_{\mathcal{A},\mathcal{A}}^{-1/2} \left(I - \frac{1-\alpha}{1+\alpha} \cdot (D^{-1/2}AD^{-1/2})_{\mathcal{A},\mathcal{A}}\right)^{-1} D_{\mathcal{A},\mathcal{A}}^{-1/2} \\ &= \frac{2}{1+\alpha} D_{\mathcal{A},\mathcal{A}}^{-1/2} \left(I + \frac{1-\alpha}{1+\alpha} \cdot (D^{-1/2}AD^{-1/2})_{\mathcal{A},\mathcal{A}} + \left(\frac{1-\alpha}{1+\alpha}\right)^2 \cdot (D^{-1/2}AD^{-1/2})_{\mathcal{A},\mathcal{A}}^2 \right. \\ &\qquad \left. + \cdots\right) D_{\mathcal{A},\mathcal{A}}^{-1/2}. \end{aligned} \tag{19}$$

Now it is easy to see that the right hand side in (18) is positive, i.e., $\alpha(\rho_0 - \rho_1)Q_{\mathcal{A},\mathcal{A}}^{-1} d_{\mathcal{A}} > 0$. This proves the theorem.

B Proofs of lemmas

B.1 Proof of Lemma 1 and Lemma 2

Proof of Lemma 1. By the KKT condition, $\widehat{x}$ must satisfy

$$\nabla_j f(\widehat{x}) = -\rho\alpha\partial\|D\widehat{x}\|_1 \text{ for all } j \in V,$$

or equivalently,

$$(Q\widehat{x})_j - \alpha \mathbf{s}_j = \begin{cases} -\rho\alpha d_j & x_j > 0, \\ \rho\alpha d_j & x_j < 0, \\ [-\rho\alpha d_j, \rho\alpha d_j] & x_j = 0, \end{cases} \tag{20}$$

where $d = \mathrm{diag}(D)$ is the degree vector of the graph. Simple calculation shows that

$$(Q\widehat{x})_j = \frac{(1+\alpha)}{2}\left(d_j\widehat{x}_j - \frac{1-\alpha}{1+\alpha}\cdot\sum_{i:(i,j)\in E}\widehat{x}_i\right)$$
$$= \frac{(1+\alpha)}{2}\left(\sum_{i:(i,j)\in E}\widehat{x}_j - \frac{1-\alpha}{1+\alpha}\cdot\sum_{i:(i,j)\in E}\widehat{x}_i\right). \tag{21}$$

Now assume $j^\star \in V$ is a node such that $\widehat{x}_{j^\star} < 0$ and that $j^\star = \arg\min_{j\in V}\widehat{x}_j$. Then, from KKT condition (20), it must be that $(Q\widehat{x})_{j^\star} > 0$. However, by (21), this is only possible if there is at least one node $i \in V$ with $i \sim j^\star$ such that

$$\widehat{x}_i < \frac{1+\alpha}{1-\alpha}\cdot\widehat{x}_{j^\star}.$$

However, this means that $\widehat{x}_i$ is smaller than $\widehat{x}_{j^\star}$, contradicting to the fact that $j^\star \in V$ is the smallest graph node. This proves the desired result.

□

Proof of Lemma 2. This lemma follows from Lemma 1 and the definition of Q. □

B.2 Proof of Lemma 3 and Lemma 4

Proof of Lemma 3. This lemma is proved in Vershynin [45, Proposition 2.4.1] which we reproduce here for the sake of completeness. By Chernoff's inequality (Vershynin [45, Theorem 2.3.1]), for some constant $c_0 > 0$,

$$\mathbb{P}\left\{|X_i - \mu| \geq \delta\mu\right\} \leq 2e^{-2c_0\delta^2\mu}.$$

Using union bound, then

$$\mathbb{P}\left\{\max_{i\in K}|X_i - \mu| \geq \delta\bar{d}\right\} \leq 2e^{-2c_0\delta^2\mu + \log k}.$$

Since $\log k \leq c_0\delta^2\mu$ by assumption, the result follows. □

Proof of Lemma 4. This lemma follows directly from Lemma 3. □

B.3 Proof of Lemma 5

We have $\|A_{j,K}\|_1 \sim \mathrm{Binom}(k, q)$ for $j \in K^c$, so using tail bound for Binomial distribution [10], for any $u > c$,

$$\mathbb{P}\left\{\|A_{j,K}\|_1 \geq u\right\} \leq \exp\left[-u\log\left(\frac{u}{k\cdot q}\right) - k\left(1-\frac{u}{k}\right)\log\left(\frac{1-u/k}{1-q}\right)\right].$$

Set $u = c + 3$. Then, since $\log(1-x) \geq -x$ for all $0 < x < \frac{1}{2}$, then

$$-k\left(1-\frac{u}{k}\right)\log\left(\frac{1-u/k}{1-q}\right) \leq -k\log(1-u/k) - u\log(1-q) \leq u + \log 2\cdot u \leq 2u = 2(c+3),$$

where the second inequality follows, since $k \geq 2(c+3)$ by assumption, and $q \leq 0.5$ for n large. Then, using union bound,

$$\mathbb{P}\left\{\max_{j \in K_1^c} \|A_{j,K_1}\|_1 \geq c+3\right\} \leq \exp\left[-(c+3)\log\left(\frac{(c+3)n}{c \cdot k}\right) + 2(c+3) + \log(n-k)\right] \leq \mathcal{O}\left(n^{-1}\right),$$

as long as $n \gg k$.

B.4 Proof of Lemma 6

First, the results in Lemma 1 and Lemma 2 hold for any graph, seed vector $\mathbf{s}$, and $\rho > 0$, so the result can also be applied to (13). In particular, the solution $x^\star$ is non-negative, and thus the optimality condition can be expressed as

$$(\mathbb{E}[Q]\, x^\star)_j - \alpha \mathbf{1}_{j \in S} = \begin{cases} -\rho\alpha\mathbb{E}[d_j] & x_j^\star > 0, \\ [-\rho\alpha\mathbb{E}[d_j], 0] & x_j^\star = 0. \end{cases} \tag{22}$$

(Note $\mathbf{s}$ is 1 on the seed node and 0 on the rest.) For $\rho > \frac{1}{d}$ the optimal solution (13) is the zero vector. For $\rho \leq \frac{1}{d}$ the first nonzero nodes appear. From (22), as ρ decreases less than or equal to $\frac{1}{d}$, we can see that the seed node becomes active at first.

Now let $\rho^\natural > 0$ be the regularization parameter for which the next set of nodes enters the active set. Then, for $\rho \in (\rho^\natural, \frac{1}{d}]$, we know that $x_S^\star > 0$ and $x_{S^c}^\star = 0$. Writing $x_S^\star = u\mathbf{1}_S$ for some $u > 0$, we can see that for the seed node $j \in S$,

$$(\mathbb{E}[Q]\, x^\star)_j = \frac{1+\alpha}{2} \cdot \left[\sum_{i:(i,j)\in E} \mathbb{E}[A_{ij}]\, x_j^\star - \frac{1-\alpha}{1+\alpha} \sum_{i:(i,j)\in E} \mathbb{E}[A_{ij}]\, x_i^\star\right] = \frac{1+\alpha}{2} \cdot u \cdot \bar{d}.$$

Substituting in the optimality condition (22), and solving with respect to u,

$$\frac{1+\alpha}{2} \cdot u \cdot \bar{d} - \alpha = -\rho\alpha\bar{d}, \text{ and so } u = \frac{2(\alpha - \rho\alpha\bar{d})}{(1+\alpha)\bar{d}}.$$

It remains to be shown that $x_S^\star = u\mathbf{1}_S$ for the above u satisfies the optimality conditions for $j \in S^c$. To verify this, we use the definition of u and the optimality conditions for $j \in S^c$. We need

$$(\mathbb{E}[Q]\, x^\star)_j \in [-\rho\alpha\mathbb{E}[d_j], 0] \text{ for all } j \in S^c. \tag{23}$$

We have that

$$(\mathbb{E}[Q]\, x^\star)_j = -\frac{1-\alpha}{2} p \cdot u = -\left(\alpha - \rho\alpha\bar{d}\right) \frac{p(1-\alpha)}{\bar{d}(1+\alpha)} \text{ for all } j \in K \backslash S,$$

and

$$(\mathbb{E}[Q]\, x^\star)_j = -\frac{1-\alpha}{2} q \cdot u = -\left(\alpha - \rho\alpha\bar{d}\right) \frac{q(1-\alpha)}{\bar{d}(1+\alpha)} \text{ for all } j \in K^c.$$

Using the above equalities, we get that the optimality conditions are satisfied if and only if

$$\rho > \rho^\natural := \frac{(1-\alpha)p}{\bar{d}\left[(1+\alpha)\bar{d} + (1-\alpha)p\right]}.$$

Next, we prove that, when ρ reaches $\rho^\natural$, the nodes in $K\backslash S$ appear in the active set. To see this, from (22), we can check that the first ρ value allowing nonzero nodes for $K\backslash S$ is given by

$$-\frac{1-\alpha}{2}p\cdot u=-\rho\alpha\bar{d}\text{ , and so }\ \rho=\rho^\natural=\frac{(1-\alpha)p}{\bar{d}\left[(1+\alpha)\bar{d}+(1-\alpha)p\right]}.$$

Now assuming that $x^\star$ takes the form of $u\mathbf{1}_S+v\mathbf{1}_K$, and substituting in the optimality condition (22), we obtain the following equations:

$$\begin{aligned}
&j\in S:\ \underbrace{\frac{1+\alpha}{2}\left[(u+v)\cdot\bar{d}-\frac{1-\alpha}{1+\alpha}\cdot p\cdot v(k-1)\right]}_{=(\mathbb{E}[Q]x^\star)_j}-\alpha=-\rho\alpha\bar{d},\\
&j\in K\backslash S:\ \underbrace{\frac{1+\alpha}{2}\left[v\cdot\bar{d}-\frac{1-\alpha}{1+\alpha}\cdot p\cdot(u+v(k-1))\right]}_{=(\mathbb{E}[Q]x^\star)_j}=-\rho\alpha\bar{d}.
\end{aligned}$$

So, solving with respect to u and v, we obtain

$$\begin{cases}u=\frac{2\alpha}{\left[(1+\alpha)\bar{d}+(1-\alpha)p\right]},\\ v=\frac{\frac{1-\alpha}{2}pu-\rho\alpha\bar{d}}{\alpha\bar{d}+\frac{1-\alpha}{2}q(n-k)}.\end{cases}\tag{24}$$

Also, $x^\star=u\mathbf{1}_S+v\mathbf{1}_K$ with u and v given in (24) satisfies the optimality conditions for K^c if and only if

$$\rho>\rho^\sharp:=\frac{q(1-\alpha)}{2\alpha\bar{d}\cdot\mathbb{E}\left[d_{\ell_\sharp}\right]+q(1-\alpha)\left[k\bar{d}+(n-k)\mathbb{E}\left[d_{\ell_\sharp}\right]\right]},$$

where $\mathbb{E}\left[d_{\ell_\sharp}\right]=\min_{\ell\in K^c}\mathbb{E}\left[d_\ell\right]$. We claim that $\rho^\natural>\rho^\sharp$, in which case we have proved that $x^\star=u\mathbf{1}_S+v\mathbf{1}_K$ satisfies the optimality condition (22) for $\rho\in[\rho^\sharp,\rho^\natural)$, which proves the theorem.

It now remains to check that $\rho^\natural>\rho^\sharp$. With some algebra, we have that

$$\begin{aligned}
\rho^\natural>\rho^\sharp\ &\iff\ 2p\alpha\cdot\bar{d}\cdot\mathbb{E}\left[d_{\ell_\sharp}\right]+pq(1-\alpha)\cdot k\bar{d}+pq(1-\alpha)(n-k)\mathbb{E}\left[d_{\ell_\sharp}\right]\\
&\quad>q(1+\alpha)\bar{d}^2+pq(1-\alpha)\bar{d}\\
&\iff\ 2p\alpha\cdot\bar{d}\cdot\mathbb{E}\left[d_{\ell_\sharp}\right]+pq(1-\alpha)\cdot k\bar{d}+pq(1-\alpha)(n-k)\mathbb{E}\left[d_{\ell_\sharp}\right]\\
&\quad>q(1-\alpha+2\alpha)\bar{d}^2+pq(1-\alpha)\bar{d}\\
&\iff\ 2\alpha\bar{d}(p\mathbb{E}\left[d_{\ell_\sharp}\right]-q\bar{d})+pq(1-\alpha)(k-1)\bar{d}+pq(1-\alpha)(n-k)\mathbb{E}\left[d_{\ell_\sharp}\right]\\
&\quad>q(1-\alpha)\bar{d}^2.
\end{aligned}$$

And, since $\bar{d}=p(k-1)+q(n-k)$, we have

$$q(1-\alpha)\bar{d}^2=q(1-\alpha)\bar{d}\left[p(k-1)+q(n-k)\right],$$

and so

$$\rho^\natural>\rho^\sharp\ \iff\ 2\alpha\bar{d}(p\mathbb{E}\left[d_{\ell_\sharp}\right]-q\bar{d})+(1-\alpha)q(n-k)(p\mathbb{E}\left[d_{\ell_\sharp}\right]-q\bar{d})>0.\tag{25}$$

Under Assumption 1, we know that $p\mathbb{E}\left[d_{\ell_\sharp}\right]=\mathcal{O}\left(1\right)$ while $q\bar{d}=\mathcal{O}\left(\frac{\bar{d}}{n}\right)$, so $p\mathbb{E}\left[d_{\ell_\sharp}\right]-q\bar{d}>0$ for a sufficiently large n. This verifies (25).

B.5 Proof of Lemma 7

Define

$$\check{x}^{\mathrm{R}} = \alpha Q_{K,K}^{-1}(\mathbf{s}_K - \rho d_K).$$

We will prove that $\check{x}^{\mathrm{R}} > 0$ as long as ρ satisfies the condition (7). In particular, this means that $\check{x}^{\mathrm{R}}$ satisfies the optimality condition for the reduced problem (14), from which we can conclude that $\widehat{x}_K^{\mathrm{R}} = \check{x}^{\mathrm{R}} > 0$, proving the result.

Now, by (19), we can write

$$\check{x}^{\mathrm{R}} = \frac{2\alpha}{1+\alpha} D_{K,K}^{-1} \left[\mathbf{I} + \frac{1-\alpha}{1+\alpha} A_{K,K} D_{K,K}^{-1} + \left(\frac{1-\alpha}{1+\alpha}\right)^2 (A_{K,K} D_{K,K}^{-1})^2 + \cdots \right] (\mathbf{s}_K - \rho d_K).$$

Letting $w = D_{K,K}^{-1} \mathbf{s}_K$, it follows that

$$\begin{aligned}
\check{x}^{\mathrm{R}} &= \frac{2\alpha}{1+\alpha} \Big[\sum_{j=0}^{\infty} \left(\frac{1-\alpha}{1+\alpha}\right)^j (D_{K,K}^{-1} A_{K,K})^j w - \rho \sum_{j=0}^{\infty} \left(\frac{1-\alpha}{1+\alpha}\right)^j (D_{K,K}^{-1} A_{K,K})^j \mathbf{1} \Big] \\
&= \frac{2\alpha}{1+\alpha} \Big[w + \frac{1-\alpha}{1+\alpha} D_{K,K}^{-1} A_{K,K} w + \sum_{j=0}^{\infty} \left(\frac{1-\alpha}{1+\alpha}\right)^{j+2} (D_{K,K}^{-1} A_{K,K})^{j+2} w \qquad (26) \\
&\quad - \rho \sum_{j=0}^{\infty} \left(\frac{1-\alpha}{1+\alpha}\right)^j (D_{K,K}^{-1} A_{K,K})^j \mathbf{1} \Big]. \qquad (27)
\end{aligned}$$

Note that $\mathbf{s}_K$ has mass 1 on the seed node, so we have $w = \mathbf{s}_K / d_S$, where d_S is the degree of the seed node. Denoting by FON the first-order neighbors of the seed node in K, i.e., FON $= \{i \in K :$ i is a neighbor of the seed node$\}$, we then have

$$D_{K,K}^{-1} A_{K,K} w = D_{K,K}^{-1} \mathbf{1}_{\mathrm{FON}} / d_S,$$

where $\mathbf{1}_{\mathrm{FON}}$ is the indicator vector for the set FON. Applying the degree concentration Lemma 4, then with probability at least $1 - 2e^{-c_0 \delta^2 \bar{d}}$,

$$D_{K,K}^{-1} A_{K,K} w \geq \frac{1}{(1+\delta)^2 \bar{d}^2} \mathbf{1}_{\mathrm{FON}}. \qquad (28)$$

Next, using Lemma 3, we have that $|\mathrm{FON}| \geq (1-\delta)pk$ with probability at least $1 - 2e^{-c_0 \delta^2 pk}$. Furthermore, for each non-seed node $i \notin$ FON in K, it is connected to nodes in FON with probability p, independently of all other pairs. Then $A_{i,K}^\top \mathbf{1}_{\mathrm{FON}} \mid |\mathrm{FON}| \overset{\mathrm{iid}}{\sim} \mathrm{Binom}(|\mathrm{FON}|, p)$, and thus applying Lemma 3, and using the assumption $(1-\delta)p^2 k \geq c_0^{-1} \delta^{-2} \log k$, we get

$$\begin{aligned}
\mathbb{P}\left\{ A_{i,K}^\top \mathbf{1}_{\mathrm{FON}} \geq (1-\delta)|\mathrm{FON}| p \text{ for all } i \notin \mathrm{FON} \,\middle|\, |\mathrm{FON}| \right\} &\geq 1 - 2e^{-c_0 \delta^2 |\mathrm{FON}| p} \\
&\geq 1 - 2e^{-c_0 \delta^2 (1-\delta) p^2 k}.
\end{aligned}$$

Hence, we can conclude that with probability at least $1 - 4e^{-c_0 \delta^2 (1-\delta) p^2 k}$,

$$A_{i,K}^\top \mathbf{1}_{\mathrm{FON}} \geq (1-\delta)^2 p^2 k \text{ for all } i \notin \mathrm{FON}. \qquad (29)$$

Then,

$$\left(\frac{1-\alpha}{1+\alpha}\right)^2 (D_{K,K}^{-1}A_{K,K})^2 w \geq \left(\frac{1-\alpha}{1+\alpha}\right)^2 (D_{K,K}^{-1}A_{K,K}) \cdot \frac{1}{(1+\delta)^2\bar{d}^2}\mathbf{1}_{\text{FON}}$$
$$\geq \left(\frac{1-\alpha}{1+\alpha}\right)^2 \left(\frac{1-\delta}{1+\delta}\right)^2 \frac{p^2 k}{\bar{d}^2} D_{K,K}^{-1}\mathbf{1}$$
$$\geq \left(\frac{1-\alpha}{1+\alpha}\right)^2 \left(\frac{1-\delta}{1+\delta}\right)^2 \frac{\gamma p}{(1+\delta)\bar{d}^2}\mathbf{1},$$

where the first step applies (28), the second step uses (29), and the third applies Lemma 4 and the fact that $p \cdot k = \gamma\bar{d}$. Returning to (26), we can now see that $\check{x}^{\text{R}} > 0$ as long as ρ is less than $\left(\frac{1-\alpha}{1+\alpha}\right)^2 \left(\frac{1-\delta}{1+\delta}\right)^2 \frac{\gamma p}{(1+\delta)\bar{d}^2}$. This completes the proof of Lemma 7.

B.6 Proof of Lemma 8

Recall that $\widehat{x}$ is the ℓ_1-regularized PageRank on the original graph (2). If $\rho > 1/d_S$, then we can easily check that $\widehat{x} = \widehat{x}^{\text{R}} = 0$, while for $\rho = 1/d_S$, we have $\text{support}(\widehat{x}^{\text{R}}) = \text{support}(\widehat{x}) = S$. To prove that $\text{support}(\widehat{x}^{\text{R}}) \subseteq \text{support}(\widehat{x})$ holds for $\rho < 1/d_S$, we proceed by induction.

Writing $\mathcal{A}_1 = \text{support}(\widehat{x}^{\text{R}})$ and $\mathcal{A}_2 = \text{support}(\widehat{x})$, by the optimality condition, we have

$$\widehat{x}^{\text{R}}_{\mathcal{A}_1} = \alpha Q^{-1}_{\mathcal{A}_1,\mathcal{A}_1}(\mathbf{s}_{\mathcal{A}_1} - \rho d_{\mathcal{A}_1}) \text{ and } \widehat{x}_{\mathcal{A}_2} = \alpha Q^{-1}_{\mathcal{A}_2,\mathcal{A}_2}(\mathbf{s}_{\mathcal{A}_2} - \rho d_{\mathcal{A}_2}).$$

By inductive hypothesis, assume $\mathcal{A}_1 \subseteq \mathcal{A}_2$. According to the expression (19), then we can see that $\widehat{x}_{\mathcal{A}_1} \geq \widehat{x}^{\text{R}}_{\mathcal{A}_1}$, where the inequality applies component-wise. Now let $i \in K$ be a node such that $\widehat{x}^{\text{R}}_i = \widehat{x}_i = 0$. Using the optimality condition, we know that i becomes active for the full-dimensional problem (2) whenever the following condition is satisfied:

$$-\frac{1-\alpha}{2} \sum_{j\sim i, j\in\mathcal{A}_2} w_{ij}\widehat{x}_j = -\rho\alpha d_i. \tag{30}$$

Analogously, the node i becomes active for the reduced problem (14) when the following condition is satisfied:

$$-\frac{1-\alpha}{2} \sum_{j\sim i, j\in\mathcal{A}_1} w_{ij}\widehat{x}^{\text{R}}_j = -\rho\alpha d_i. \tag{31}$$

Comparing the left-hand sides in (30) and (31), it is obvious that under the induction assumption, the left-hand side in (30) is larger than in (31). This implies that $\widehat{x}_i$ becomes active earlier than $\widehat{x}^{\text{R}}_i$, so the induction assumption continues to hold. This completes the proof of the lemma.

B.7 Proof of Lemma 9

Since $\widehat{x}^{\text{R}} = (\widehat{x}^{\text{R}}_S, \widehat{x}^{\text{R}}_{K\setminus S})$ is the solution to the reduced problem (14), fixing $\widehat{x}^{\text{R}}_S$, then $\widehat{x}^{\text{R}}_{K\setminus S}$ is the minimizer of the following optimization problem:

$$\widehat{x}^{\text{R}}_{K\setminus S} = \underset{y\in\mathbb{R}^{k-1}}{\arg\min} \left\{ \frac{1}{2}(\widehat{x}^{\text{R}}_S, y)^\top Q_{K,K}(\widehat{x}^{\text{R}}_S, y) + \rho\alpha\|D\cdot(\widehat{x}^{\text{R}}_S, y)\|_1 \right\}. \tag{32}$$

Due to Lemma 7, the subgradient of $\|D \cdot (\widehat{x}_S^{\mathrm{R}}, y)\|_1$ at $y = \widehat{x}_{K\setminus S}^{\mathrm{R}}$ is given by $d_{K\setminus S}$. Using optimality, it follows that

$$0 = \underbrace{Q_{K\setminus S,K}\widehat{x}^{\mathrm{R}}}_{\text{gradient at } y = \widehat{x}_{K\setminus S}^{\mathrm{R}}} + \rho\alpha d_{K\setminus S}.$$

Next, we can prove that our choice of ρ in Lemma 9 lies in the interval $[\rho^\sharp, \rho^\natural)$.

Lemma 12. *Under the conditions of Lemma 9, we have $\rho \in (\rho^\sharp, \rho^\natural)$.*

Hence, by Lemma 6, the average case ℓ_1-regularized PageRank (13) has support K, so it follows that $x^\star$ is also the solution to the reduced problem

$$x^\star = \min_x \left\{ \frac{1}{2} x^\top \mathbb{E}\left[Q_{K,K}\right] x - \alpha x^\top \mathbf{s}_K + \rho\alpha \|\mathbb{E}\left[D_{K,K}\right] x\|_1 : x_{K^c} = 0 \right\}. \tag{33}$$

(Abusing notation, we use $x^\star$ and $x_K^\star$ interchangeably throughout the proof.) So, using the optimality condition, we obtain

$$0 = \underbrace{\mathbb{E}\left[Q_{K\setminus S,K}\right] x^\star}_{\text{gradient at } y = x_{K\setminus S}^\star} + \rho\alpha \mathbb{E}\left[d_{K\setminus S}\right],$$

where we have $\partial\|\mathbb{E}\left[D\right] \cdot (x_S^\star, y)\|_1 = \mathbb{E}\left[d_{K\setminus S}\right]$ due to Lemma 6 and Lemma 12. Combining the two optimality equations, and adding and subtracting $Q_{K\setminus S,K}x^\star$, it follows that

$$\begin{aligned} 0 &= \mathbb{E}\left[Q_{K\setminus S,K}\right] x^\star - Q_{K\setminus S,K}\widehat{x}^{\mathrm{R}} + \rho\alpha\mathbb{E}\left[d_{K\setminus S}\right] - \rho\alpha d_{K\setminus S} \\ &= (\mathbb{E}\left[Q_{K\setminus S,K}\right] - Q_{K\setminus S,K})x^\star + Q_{K\setminus S,K}(x^\star - \widehat{x}^{\mathrm{R}}) + \rho\alpha(\mathbb{E}\left[d_{K\setminus S}\right] - d_{K\setminus S}). \end{aligned}$$

Writing $Q_{K\setminus S,K}(x^\star - \widehat{x}^{\mathrm{R}}) = Q_{K\setminus S,S}(x_S^\star - \widehat{x}_S^{\mathrm{R}}) + Q_{K\setminus S,K\setminus S}(x_{K\setminus S}^\star - \widehat{x}_{K\setminus S}^{\mathrm{R}})$, and rearranging terms, we get

$$\begin{aligned} Q_{K\setminus S,K\setminus S}(\widehat{x}_{K\setminus S}^{\mathrm{R}} - x_{K\setminus S}^\star) = (\mathbb{E}\left[Q_{K\setminus S,K}\right] - Q_{K\setminus S,K})x^\star + Q_{K\setminus S,S}(x_S^\star - \widehat{x}_S^{\mathrm{R}}) \\ + \rho\alpha(\mathbb{E}\left[d_{K\setminus S}\right] - d_{K\setminus S}). \end{aligned}$$

Taking the ℓ_∞ norm on both sides,

$$\begin{aligned} \|Q_{K\setminus S,K\setminus S}(\widehat{x}_{K\setminus S}^{\mathrm{R}} - x_{K\setminus S}^\star)\|_\infty \le \underbrace{\|(\mathbb{E}\left[Q_{K\setminus S,K}\right] - Q_{K\setminus S,K})x^\star\|_\infty}_{\text{term 1}} + \underbrace{\rho\alpha\|\mathbb{E}[d_{K\setminus S}] - d_{K\setminus S}\|_\infty}_{\text{term 2}} \\ + \underbrace{\|Q_{K\setminus S,S}(\widehat{x}_S^{\mathrm{R}} - x_S^\star)\|_\infty}_{\text{term 3}}. \end{aligned} \tag{34}$$

For term 1, we know that $x^\star = u\mathbf{1}_S + v\mathbf{1}_K$ from Lemma 6, so plugging in to term 1, and thus

$$\begin{aligned} (\mathbb{E}\left[Q_{K\setminus S,K}\right] - Q_{K\setminus S,K})x^\star = -\frac{1-\alpha}{2}(A_{K\setminus S,S} - \mathbb{E}[A_{K\setminus S,S}]) \cdot u \\ + \frac{1+\alpha}{2}(d_{K\setminus S} - \mathbb{E}\left[d_{K\setminus S}\right]) \cdot v - \frac{1-\alpha}{2}(A_{K\setminus S,K} - \mathbb{E}\left[A_{K\setminus S,K}\right])\mathbf{1}_K \cdot v. \end{aligned}$$

Thus,

$$\begin{aligned}
&\text{term 1} = \|(\mathbb{E}\left[Q_{K\setminus S,K}\right] - Q_{K\setminus S,K})x^\star\|_\infty \\
&\le \frac{1-\alpha}{2}\|A_{K\setminus S,S} - \mathbb{E}[A_{K\setminus S,S}]\|_\infty \cdot u \\
&+ \left(\frac{1+\alpha}{2}\|d_{K\setminus S} - \mathbb{E}\left[d_{K\setminus S}\right]\|_\infty + \frac{1-\alpha}{2}\|(A_{K\setminus S,K} - \mathbb{E}\left[A_{K\setminus S,K}\right])\mathbf{1}_K\|_\infty\right)\cdot v \\
&\le \frac{1-\alpha}{2}\|A_{K\setminus S,S} - \mathbb{E}[A_{K\setminus S,S}]\|_\infty \cdot u + \delta\bar{d}\cdot v \le \frac{1-\alpha}{2}u + \delta\bar{d}\cdot v,
\end{aligned}$$

with probability at least $1 - 4e^{-c_0\delta^2 pk}$, where the second step uses the triangle inequality, the third step applies Lemma 3 and Lemma 4, and the last step holds since $|A_{j,S} - p| \le 1$ for all $j \in K\setminus S$. Furthermore, by Lemma 4, we can bound term 2 as

$$\text{term 2} \le \rho\alpha\cdot\delta\bar{d},$$

while for term 3, simple calculation leads to

$$\text{term 3} \le \frac{1-\alpha}{2}\left|\widehat{x}^{\mathrm{R}}_S - x^\star_S\right|.$$

Putting the results together, we have

$$\|Q_{K\setminus S,K\setminus S}(\widehat{x}^{\mathrm{R}}_{K\setminus S} - x^\star_{K\setminus S})\|_\infty \le \frac{1-\alpha}{2}u + \delta\bar{d}\cdot v + \rho\alpha\cdot\delta\bar{d} + \frac{1-\alpha}{2}\left|\widehat{x}^{\mathrm{R}}_S - x^\star_S\right|. \tag{35}$$

Now it remains to upper bound the term $\left|\widehat{x}^{\mathrm{R}}_S - x^\star_S\right|$. In fact, following the same steps as before, we can check that the following holds:

$$\begin{aligned}
\|Q_{S,S}(\widehat{x}^{\mathrm{R}}_S - x^\star_S)\|_\infty \le \underbrace{\|(\mathbb{E}\left[Q_{S,K}\right] - Q_{S,K})x^\star\|_\infty}_{\text{term 4}} + \underbrace{\rho\alpha\|\mathbb{E}[d_S] - d_S\|_\infty}_{\text{term 5}} \\
+ \underbrace{\|Q_{S,K\setminus S}(\widehat{x}^{\mathrm{R}}_{K\setminus S} - x^\star_{K\setminus S})\|_\infty}_{\text{term 6}}.
\end{aligned} \tag{36}$$

Also, we can see that using $x^\star = u\mathbf{1}_S + v\mathbf{1}_K$,

$$\text{term 4} = \|\frac{1+\alpha}{2}(u+v)(d_S - \bar{d}) - \frac{1-\alpha}{2}v((A_{S,K} - \mathbb{E}\left[A_{S,K}\right])\mathbf{1}_K)\|_\infty \le (u+v)\delta\bar{d},$$

where the inequality applies Lemma 3 and Lemma 4. It is also easy to see term $5 \le \rho\alpha\cdot\delta\bar{d}$, and

$$\begin{aligned}
\text{term 6} = \frac{1-\alpha}{2}\|A_{S,K\setminus S}(\widehat{x}^{\mathrm{R}}_{K\setminus S} - x^\star_{K\setminus S})\|_\infty &\le \frac{1-\alpha}{2}\|A_{S,K\setminus S}\|_1\|\widehat{x}^{\mathrm{R}}_{K\setminus S} - x^\star_{K\setminus S}\|_\infty \\
&\le \frac{1-\alpha}{2}(1+\delta)p(k-1)\|\widehat{x}^{\mathrm{R}}_{K\setminus S} - x^\star_{K\setminus S}\|_\infty,
\end{aligned}$$

where the second step uses Hölder's inequality and the next step applies Lemma 3. Furthermore, $Q_{S,S} = \frac{1+\alpha}{2}d_S$ by definition, which is bounded below by $\frac{1+\alpha}{2}(1-\delta)\bar{d}$. Thus, from (36), we get

$$\|\widehat{x}^{\mathrm{R}}_S - x^\star_S\|_\infty \le \frac{\delta\bar{d}(u+v) + \rho\alpha\cdot\delta\bar{d} + \frac{1-\alpha}{2}(1+\delta)p(k-1)\|\widehat{x}^{\mathrm{R}}_{K\setminus S} - x^\star_{K\setminus S}\|_\infty}{\frac{1+\alpha}{2}(1-\delta)\bar{d}}.$$

Finally, return to (35) and substitute the above bound in place of $\|\widehat{x}^{\mathrm{R}}_S - x^\star_S\|_\infty$, then

$$\|Q_{K\setminus S,K\setminus S}(\widehat{x}^{\mathrm{R}}_{K\setminus S} - x^\star_{K\setminus S})\|_\infty \le \frac{1-\alpha}{2}u + \delta\bar{d}\cdot v + \rho\alpha\cdot\delta\bar{d} + \left(\frac{1-\alpha}{1+\alpha}\right)\frac{\delta\bar{d}(u+v) + \rho\alpha\cdot\delta\bar{d} + \frac{1-\alpha}{2}(1+\delta)p(k-1)\|\widehat{x}^{\mathrm{R}}_{K\setminus S} - x^\star_{K\setminus S}\|_\infty}{(1-\delta)\bar{d}}. \quad (37)$$

The following lemma concerns the local strong convexity of the matrix $Q_{K\setminus S,K\setminus S}$ in ℓ_∞ norm.

Lemma 13. *Under the conditions of Lemma 9, with probability at least* $1 - 2e^{-c_0\delta^2\bar{d}}$,

$$\|Q_{K\setminus S,K\setminus S}y\|_\infty \ge \alpha(1-\delta)\bar{d}\|y\|_\infty \text{ for all } y \in \mathbb{R}^{k-1}.$$

Applying Lemma 13 to the left-hand side of (37), and rearranging terms and simplifying, we have that

$$\left[\alpha(1-\delta)\bar{d} - \frac{\gamma}{2}\left(\frac{1-\alpha}{1+\alpha}\right)\left(\frac{1+\delta}{1-\delta}\right)\right]\|\widehat{x}^{\mathrm{R}}_{K\setminus S} - x^\star_{K\setminus S}\|_\infty \le \left(\frac{1-\alpha}{2} + \frac{1-\alpha}{1+\alpha}\frac{\delta}{1-\delta}\right)u + \left(\delta\bar{d} + \frac{1-\alpha}{1+\alpha}\frac{\delta}{1-\delta}\right)v + \rho\alpha\left(\delta\bar{d} + \frac{1-\alpha}{1+\alpha}\frac{\delta}{1-\delta}\right),$$

where we use $\alpha < 1$ and $p(k-1) = \gamma\bar{d}$. By assumption of the lemma, we have $\delta \le 0.1$ and also $\alpha \in [0.1, 0.9]$ while $\gamma \in (0,1)$ by definition (6), so

$$\left[\alpha(1-\delta)\bar{d} - \frac{\gamma}{2}\left(\frac{1-\alpha}{1+\alpha}\right)\left(\frac{1+\delta}{1-\delta}\right)\right] \ge c_1\bar{d} \quad \text{and} \quad \frac{1-\alpha}{1+\alpha}\frac{\delta}{1-\delta} \le \delta\bar{d},$$

for some numerical constant $c_1 > 0$. As a result,

$$c_1\bar{d}\|\widehat{x}^{\mathrm{R}}_{K\setminus S} - x^\star_{K\setminus S}\|_\infty \le (1-\alpha)u + 2\delta\bar{d}(v + \rho\alpha).$$

Dividing both sides by $c_1\bar{d}$,

$$\|\widehat{x}^{\mathrm{R}}_{K\setminus S} - x^\star_{K\setminus S}\|_\infty \le \frac{c_1^{-1}(1-\alpha)u}{\bar{d}} + 2c_1^{-1}\delta(v + \rho\alpha).$$

Using the triangle inequality, Lemma 9 now follows.

B.8 Proof of Lemma 10

From the optimality condition (4), we multiply $\mathbf{1} \in \mathbb{R}^n$ on both sides. Then, since $L\mathbf{1} = 0$, we get

$$\alpha d^\top\widehat{x} - \alpha \le -\rho\alpha\sum_{j:\widehat{x}_j>0} d_j = -\rho\alpha\mathrm{Vol}(\mathrm{support}(\widehat{x})).$$

Dividing by $-\rho\alpha$,

$$\mathrm{Vol}(\mathrm{support}(\widehat{x})) \le \frac{1 - d^\top\widehat{x}}{\rho}.$$

B.9 Proof of Lemma 11

This lemma follows from simple probability calculation. We have

$$\begin{aligned}\mathbb{P}\left\{\cup_{i\in K}(i \text{ is not connected to } K^c)\right\} &= 1-\mathbb{P}\left\{\cap_{i\in K}(i \text{ is connected to } K^c)\right\}\\ &= 1-\left(\mathbb{P}\left\{i \text{ is connected to } K^c\right\}\right)^k\\ &= 1-\left(1-\mathbb{P}\left\{i \text{ is not connected to } K^c\right\}\right)^k\\ &= 1-\left(1-(1-q)^{n-k}\right)^k\\ &\geq 1-\left(1-(1-q)^n\right)^k.\end{aligned}$$

Here, the second and fourth equality follow since each pair of nodes has an edge independently of all other pairs. Now let n is sufficiently large so that $q \leq 0.5$. Then, since $(1-q)^n \geq \exp(-1.5c)$, we have that

$$\mathbb{P}\left\{\cup_{i\in K}(i \text{ is not connected to } K^c)\right\} \geq 1-\left(1-\exp(-1.5c)\right)^k,$$

as desired.

B.10 Proof of Lemma 12

Letting $\bar{c} = \left(\frac{1-\alpha}{1+\alpha}\right)\left(\frac{1-\delta}{1+\delta}\right)^2\frac{\gamma}{1+\delta}$, we can write $\rho = \frac{\bar{c}(1-\alpha)p}{(1+\alpha)\bar{d}^2}$. Following the same steps as (25) (proof of Lemma 6), we can see

$$\begin{aligned}\rho > \rho^\sharp \quad &\Longleftrightarrow \quad 2\bar{c}p\alpha\cdot\bar{d}\cdot\mathbb{E}\left[d_{\ell_\sharp}\right] + \bar{c}pq(1-\alpha)\cdot k\bar{d} + \bar{c}pq(1-\alpha)(n-k)\mathbb{E}\left[d_{\ell_\sharp}\right] > q(1+\alpha)\bar{d}^2\\ &\Longleftrightarrow \quad 2\alpha\bar{d}(\bar{c}p\mathbb{E}\left[d_{\ell_\sharp}\right]-q\bar{d}) + (1-\alpha)q(n-k)(\bar{c}p\mathbb{E}\left[d_{\ell_\sharp}\right]-q\bar{d}) \qquad (38)\\ &\qquad\qquad - pq(1-\alpha)(k-1)\bar{d}(1-\bar{c}) > 0.\end{aligned}$$

Since $\delta \leq 0.1$ and $\alpha \in [0.1, 0.9]$, we know that $\bar{c} \geq c'\gamma$ for some constant $c' > 0$. Also, since $q = \frac{c}{n}$ while $p = k = \mathcal{O}(1)$ by Assumption 1, we have $\gamma = \mathcal{O}(1)$, and thus

$$\begin{cases}2\alpha\bar{d}(\bar{c}p\mathbb{E}\left[d_{\ell_\sharp}\right]-q\bar{d}) + (1-\alpha)q(n-k)(\bar{c}p\mathbb{E}\left[d_{\ell_\sharp}\right]-q\bar{d}) = \mathcal{O}\left(\bar{d}\right);\\ pq(1-\alpha)(k-1)\bar{d}(1-\bar{c}) = \mathcal{O}\left(\frac{k\bar{d}}{n}\right).\end{cases}$$

Therefore, if n is sufficiently large, the first two terms on the right-hand side of (38) are positive and much larger than the last term. This proves that $\rho > \rho^\sharp$.

To see $\rho < \rho^\natural$, it suffices to check

$$\left(\frac{1-\alpha}{1+\alpha}\right)^2\frac{p}{\bar{d}^2} < \frac{p(1-\alpha)}{\bar{d}\left[(1+\alpha)\bar{d}+(1-\alpha)p\right]} = \rho^\natural.$$

Rearranging the terms, equivalently we need $2\alpha(1+\alpha)\bar{d} > (1-\alpha)^2 p$. Since $\bar{d} \geq p\cdot(k-1)$, we then have $2\alpha(1+\alpha)\cdot(k-1) > (1-\alpha)^2$ which, due to the assumption $\alpha \in [0.1, 0.9]$, holds for any $k \geq 5$.

B.11 Proof of Lemma 13

Denoting by $\mathbf{1}$ the vector whose entries are all ones, we have

$$\begin{aligned} Q_{K\setminus S,K\setminus S} &= \alpha D_{K\setminus S,K\setminus S} + \frac{1-\alpha}{2}(D_{K\setminus S,K\setminus S} - A_{K\setminus S,K\setminus S}) \\ &= \alpha D_{K\setminus S,K\setminus S} + \frac{1-\alpha}{2}(\text{diag}\,(A\mathbf{1})_{K\setminus S,K\setminus S} - A_{K\setminus S,K\setminus S}) \\ &= \alpha D_{K\setminus S,K\setminus S} + \frac{1-\alpha}{2}\text{diag}\left(A_{K\setminus S,S\cup K^c}\mathbf{1}_{S\cup K^c}\right) + \frac{1-\alpha}{2}L_{K\setminus S,K\setminus S}, \end{aligned}$$

where $L_{K\setminus S,K\setminus S}$ is the graph Laplacian of the sub-graph induced by $A_{K\setminus S,K\setminus S}$.

Now assume $\|y\|_\infty = 1$. Without loss of generality, we will assume $y_1 = 1$ (the same argument holds in the case of $y_1 = -1$). Then, by definition of graph Laplacian, it is easy to see $(L_{K\setminus S,K\setminus S} \cdot y)_1 \geq 0$, and so

$$(Q_{K\setminus S,K\setminus S} \cdot y)_1 \geq \alpha d_1.$$

Hence, applying degree concentration to d_1 (Lemma 4),

$$\|Q_{K\setminus S,K\setminus S} \cdot y\|_\infty \geq (Q_{K\setminus S,K\setminus S} \cdot y)_1 \geq \alpha d_1 \geq (1-\delta)\alpha\bar{d},$$

with probability at least $1 - 2e^{-c_0\delta^2\bar{d}}$, proving the result.