Journal of Machine Learning Research 22 (2021) 1-54 Submitted 1/20; Revised 4/21; Published 6/21

Statistical guarantees for local graph clustering

Wooseok Ha[*] HAYWSE@BERKELEY.EDU
Department of Statistics,
University of California at Berkeley,
Berkeley, CA, USA.

Kimon Fountoulakis[*] KFOUNTOU@UWATERLOO.CA
School of Computer Science,
University of Waterloo,
Waterloo, ON, Canada.

Michael W. Mahoney MMAHONEY@STAT.BERKELEY.EDU
ICSI and Department of Statistics,
University of California at Berkeley,
Berkeley, CA, USA.

Editor: Vahab Mirrokni

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 (Fountoulakis et al., 2019) 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 empirically that ℓ_1-regularized PageRank has a state-of-the-art performance on many real graphs, demonstrating the superiority of the method. From a computational perspective, we show that the solution path of ℓ_1-regularized PageRank is monotonic. This allows for the application of the forward stagewise algorithm, which approximates the entire solution path in running time that does not depend on the size of the whole graph. Finally, we show that ℓ_1-regularized PageRank and approximate personalized PageRank (APPR) (Andersen et al., 2006), another very popular method for local graph clustering, are equivalent in the sense that we can lower and upper bound the output of one with the output of the other. Based on this relation, we establish for APPR similar results to those we establish for ℓ_1-regularized PageRank.

Keywords: clustering, local graph clustering, PageRank, seed expansion, ℓ_1-regularization.

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*:

[*]. Equal contribution

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 is 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* (Fountoulakis et al., 2019), 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., Andersen et al. (2006); Zhu et al. (2013), 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 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. The model we consider is very general; it concerns the target cluster and its adjacent nodes; and it encompasses the stochastic block model (Holland et al., 1983; Abbe, 2017) and the planted clustering model (Alon et al., 1998; Arias-Castro and Verzelen, 2014) 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. In particular, the cluster is recovered through the support set of the ℓ_1-regularized PageRank vector and we give rigorous bounds on the false positives and false negatives of the recovered cluster. Furthermore, observe that our statistical perspective is more aligned with statistical guarantees for the sparse regression problem (and the lasso problem (Tibshirani, 1996)), 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. Empirically we demonstrate the ability of the method to recover the target cluster in a range of real-world data graphs. Finally, we establish an equivalence between ℓ_1-regularized PageRank and the very popular local graph clustering algorithm from Andersen et al. (2006); Zhu et al. (2013). This allows us to prove similar average case guarantees for the algorithm of Andersen et al. (2006); Zhu et al. (2013) as well.

1.1 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—for example, personalized PageRank and flow-based methods; and work in statistics on stochastic graph models. We discuss each in turn.

Local graph clustering The origins of local graph clustering are with the work of Spielman and Teng (2013). Subsequent to their original results, there has been a great deal of follow-up work on local graph clustering procedures, including with random walks (Andersen et al., 2006; Zhu et al., 2013), local Lanczos spectral approximations (Shi et al., 2017), evolving sets (Andersen et al., 2016), seed expansion methods (Kloumann and Kleinberg, 2014), optimization-based approaches (Fountoulakis et al., 2019, 2017), and local flow methods (Orecchia and Zhu, 2014; Wang et al., 2017; Veldt et al., 2019; Fountoulakis et al., 2020b,a). There also exist local higher-order clustering (Yin et al., 2017), linear algebra approaches (Shi et al., 2017), spectral methods based on Heat Kernel PageRank (Kloster and Gleich, 2014), and parallel local spectral approaches (Shun et al., 2016). 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 (Spielman and Teng, 2013; Andersen et al., 2006; Fountoulakis et al., 2019; Orecchia and Zhu, 2014; Veldt et al., 2019; Wang et al., 2017; Fountoulakis et al., 2020b). More recently, Green et al. (2019) studied local graph clustering in a traditional statistical learning setup where they aim to identify density clusters around a given seed node.

Stochastic block model There are 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 (Abbe and Sandon, 2015; Abbe et al., 2015; Zhang and Zhou, 2016; Massoulié, 2014; Mossel et al., 2018; Newman et al., 2002; Mossel et al., 2015; Rohe et al., 2011; Amini and Levina, 2018; Abbe, 2017). Recent work has also generalized the SBM to a degree-corrected block model, to capture degree heterogeneity of the network (Chen et al., 2018; Gulikers et al., 2017; Zhao et al., 2012; Gao et al., 2018). 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 (Abbe, 2017).

We should emphasize that the traditional graph partitioning problem is quite different than the local graph clustering problem that we consider in this paper. Among other things, the former partitions all the vertices of a graph into different clusters, while for the latter problem our objective is to find a single cluster given a seed node in the cluster; the former takes as input a graph, while the latter takes as input a graph and a seed set of nodes; and the former runs in time depending on the size of the graph, while the latter runs in time depending on the size of the output, but is otherwise independent of the size of the graph.

1.2 Notation

We write $[n] = \{1, \dots, n\}$ 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., $A_{ij} = w_{ij}$ if $(i, j) \in E$, and 0 otherwise. 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 defined as $\mathrm{Vol}(B) = \sum_{i\in B} d_i$ for $B \subseteq V$. We denote by $L = D - A$ the graph Laplacian; and $Q \coloneqq \alpha D + \frac{1-\alpha}{2} L$. If $v \in \mathbb{R}^n$ is a vector defined on V, we denote by $\mathrm{support}(v)$ the support set of v, i.e., $\mathrm{support}(v) := \{i \in V \mid v_i \neq 0\}$.

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. Local graph clustering from a variational point of view

In this section, we briefly review and motivate ℓ_1-regularized PageRank (Fountoulakis et al., 2019), an optimization formulation of local graph clustering to find a local cluster around a given seed node. We then study some properties of the output vector of ℓ_1-regularized PageRank that will provide further intuition on the method.

2.1 Background on ℓ_1-regularized PageRank

PageRank (Page et al., 1999; Brin and Page, 1998) is a popular approach for ranking the importance of the nodes given 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 on the graph, i.e.,

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

where $W = (I + AD^{-1})/2$ is the lazy random walk operator 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 $\{1, 2, \ldots, n\}$.

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. For example, when the node i is given as the seed node, we consider the input distribution $\mathbf{s}$ to be the discrete Dirac measure such that $\mathbf{s}_i = 1$ and zero elsewhere. This "personalized" PageRank (Haveliwala, 2002) 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 diffusion process in (1) defining personalized PageRank performs a lazy random walk with probability $1-\alpha$ and "teleports" a random walker back to the original seed node with probability α.

From the definition itself, the personalized PageRank vector can be obtained by solving the linear system (1). Unfortunately, this step 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 Andersen et al. (2006), 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 (see Section 4 for more details on the APPR algorithm).

Gleich and Mahoney (2014) show that APPR implicitly solves a constrained ℓ_1-regularized ℓ_2 objective min-cut problem, drawing connection between APPR and ℓ_1-regularized optimization. Motivated by this, Fountoulakis et al. (2019) 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 all nodes that are dissimilar to the seed node, thus 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 (Fountoulakis et al., 2019) showed that proximal gradient descent (ISTA) can solve the ℓ_1-regularized PageRank minimization problem with access to only a small

portion of the entire graph, i.e., without even touching the entire graph, therefore allowing the method to easily scale to very large-scale graphs (Shun et al., 2016). They also showed that the optimality conditions characterizing ℓ_1-regularized PageRank implies the termination criterion of APPR, and in particular, the worst-case performance guarantees of both methods are identical.

In this paper, we investigate the statistical performance of ℓ_1-regularized PageRank by reformulating the local graph clustering into the problem of sparse support recovery. Here we give a more precise definition of the ℓ_1-regularized PageRank optimization problem from Fountoulakis et al. (2019) 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 vector 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.

Note that our definition of ℓ_1-regularized PageRank is consistent with the original formulation (Fountoulakis et al., 2019, Equation (8)) by change of variables $D^{1/2}x = q$. To better understand the objective function of (2), when the regularization parameter ρ is set to 0, one can easily check that the re-scaled version of the output solution $D\widehat{x}$ recovers the original PageRank solution, that is, $D\widehat{x} = p^{\text{PR}}$ satisfies the stationary equation (1). For the stationary personalized PageRank vector p^{PR}, mass is concentrated around the seed node, meaning that after ordering, it has long tail for nodes far away from the seed node. Importantly, we can then efficiently cut this tale using ℓ_1 norm regularization, without even having to compute the long tail.

2.2 Properties

Here, we state some properties of the ℓ_1-regularized PageRank vector that will be useful for our analysis, as well as for gaining insight into the method. The proof of these lemmas can be found in Appendix C.

First, the following lemma guarantees that the ℓ_1-regularized PageRank vector is non-negative. This result should be natural, because ℓ_1-regularized PageRank computes the importance scores of the nodes relative to the seed node, which cannot be negative.

Lemma 2 (Non-negativity of ℓ_1-reg. PageRank.) *Let $\widehat{x}$ be the optimal output solution given in Definition 1. Then $\widehat{x}$ is non-negative, i.e., $\widehat{x}_j \geq 0$ for all $j \in V$.*

The next lemma guarantees that the gradient of f at the optimal solution $\widehat{x}$ must be non-positive. Since the ℓ_1-regularized PageRank problem is strongly convex (the minimum eigenvalue of Q is > 0), by KKT condition, this characterization is both necessary and sufficient for $\widehat{x}$ to be the unique solution. We frequently use this lemma in the proof of our subsequent results.

Lemma 3 (Optimality condition for ℓ_1-reg. PR minimization.) *Let* $\mathrm{support}(\widehat{x}) := \{i \in V \mid \widehat{x}_i \neq 0\}$ *be the support set of the optimal solution. Then*

$$\nabla_i f(\widehat{x}) = (Q\widehat{x})_i - \alpha \mathbf{s}_i = \begin{cases} -\rho\alpha d_i, & \text{if } i \in \mathrm{support}(\widehat{x}), \\ \in [-\rho\alpha d_i, 0], & \text{if } i \notin \mathrm{support}(\widehat{x}) \text{ and } i \text{ is a neighbor of nonzero node}, \\ 0, & \text{otherwise.} \end{cases}$$

Finally, the next result shows that the solution path of the ℓ_1-regularized PageRank problem is monotone, meaning that when the output of ℓ_1-regularized PageRank is parametrized as a function of $\rho > 0$, the trajectory $\widehat{x}(\rho)$ changes in a monotonic manner as ρ varies.

Lemma 4 (Monotonicity of ℓ_1-reg. PageRank.) *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.*

Lemma 4 shows that once a node enter the model at some number $\rho > 0$, then it will never leave the model thereafter. This result is intuitive since the importance score of the nodes will increase as the amount of regularization or shrinkage decreases. In Section 5, we will use this monotonic property in a crucial way to develop an iterative algorithm that approximates the entire solution path.

3. Statistical guarantees under random model

In this section, we study the recovery guarantees for ℓ_1-regularized PageRank under a random graph model. We begin by introducing a random model that we consider for generating a target cluster. In particular we will assume the graph is generated according to the following model.

Definition 5 (Local random model.) *Given a graph* $G = (V, E)$ *that has* n *vertices, let* $K \subseteq V$ *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*.*

According to Definition 5, 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 population version of the graph, which is the graph induced by the expected adjacency matrix $\mathbb{E}[A]$, where the expectation is taken with respect to the distribution defined by our random model. That is, the population graph is an undirected graph $\overline{G} = (V, E)$ whose adjacency matrix is $\mathbb{E}[A]$, where

$$\mathbb{E}[A_{ij}] = \begin{cases} p & \text{if } i \in K \text{ and } j \in K, \\ q & \text{if } i \in K \text{ and } j \in K^c, \\ \text{Any value} & \text{if } i \in K^c \text{ and } j \in K^c. \end{cases} \tag{3}$$

The expected degree matrix is similarly denoted by $\mathbb{E}[D]$ and the expected graph Laplacian is defined as $\mathbb{E}[L] = \mathbb{E}[D] - \mathbb{E}[A]$. The model in Definition 5 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 random model is fairly general, and it covers several popular random graph models appearing in the literature, including the stochastic block model (SBM) (Holland et al., 1983; Abbe, 2017) and the planted clustering model (Alon et al., 1998; Arias-Castro and Verzelen, 2014; Chen and Xu, 2016). 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 piece of 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 expected degree, $\mathbb{E}[d_i] = p(k-1) + q(n-k)$ for all $i \in K$, which we denote by $\bar{d}$. For the nodes ℓ outside K, we write $\mathbb{E}[d_\ell]$ to denote its expected degree, where the expectation is taken with respect to a distribution that generates the graph in K^c. Conductance measures the weight of the edges that are being removed over the volume of the cluster—formally it is defined as the ratio $\mathrm{Cut}(S, S^c)/\min(\mathrm{Vol}(S), \mathrm{Vol}(S^c))$, where $\mathrm{Cut}(S, S^c) := \sum_{i \in S, j \in S^c} A_{ij}$. From Definition 5, the conductance of the target cluster of the population graph $\overline{G}$ is given by

$$\overline{\mathrm{Cond}} = 1 - \gamma, \quad \text{where } \gamma := \frac{p \cdot (k-1)}{\bar{d}} \in (0, 1). \tag{4}$$

Here γ can be viewed as the ratio of the random walker staying inside K under the population graph. (Note $\bar{d}$ is the expected degree of the target cluster and $p \cdot (k-1)$ is the expected degree of the target cluster when restricted to the subgraph within K.)

As in the worst-case analysis of local graph clustering (Andersen et al., 2006; Zhu et al., 2013), conductance $\overline{\mathrm{Cond}}$, or equivalently the number γ, will play a crucial role in determining the behavior of ℓ_1-regularized PageRank under the local graph model. In particular, we see that in the extreme scenario where $\gamma = 1$, we have $q = 0$ indicating perfect separability of the target cluster from the rest, while for $\gamma = 0$, we have $p = 0$ meaning there is no signal to ever recover. With this definition, we can also write $p(k-1) = \gamma\bar{d}$ and $q(n-k) = (1-\gamma)\bar{d}$.

3.1 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 as in Definition 5, 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. Specifically, for $\delta > 0$, define

$$\rho(\delta) := \left(\frac{1-\alpha}{1+\alpha}\right)^2 \left(\frac{1-\delta}{1+\delta}\right)^2 \frac{\gamma p}{(1+\delta)\bar{d}^2} = \mathcal{O}\left(\frac{\gamma p}{\bar{d}^2}\right), \tag{5}$$

where α is the teleportation constant and where $p, \gamma, \bar{d}$ are the parameters of the random model defined in (4). Then, if we solve the convex problem (2) with $\rho \leq \rho(\delta)$, the optimal solution fully recovers the target cluster K, as long as the seed node is initialized inside K.

Theorem 6 (Full recovery.) *Suppose that* $p^2k \geq \mathcal{O}\left(\delta^{-2}\log k\right)$. *If the regularization parameter satisfies* $\rho \leq \rho(\delta)$ *where* $\rho(\delta)$ *is defined in* (5)*, then the solution to Problem* (2) *fully recovers the cluster* K*, i.e.,*

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

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

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 $B \subseteq V$ is given by $\operatorname{Vol}(B) = \sum_{i\in B} d_i$.

Theorem 7 (Bounds on false positives.) *Suppose the same conditions as Theorem 6. If the regularization parameter satisfies* $\rho \geq \rho(\delta)$*, then the solution to Problem* (2) *satisfies the bound*

$$\operatorname{Vol}(FP) \leq \operatorname{Vol}(K)\big[\underbrace{\left(\frac{1+\alpha}{1-\alpha}\right)^2\left(\frac{1+\delta}{1-\delta}\right)^3\frac{1}{\gamma^2}-1}_{=\mathcal{O}\left(\frac{1}{\gamma^2}\right)-1}\big], \tag{6}$$

with probability at least $1 - 6\exp(-\mathcal{O}\left(\delta^2 p^2 k\right))$,[2] *where* $FP = \{i \in \operatorname{support}(\widehat{x}) : i \in K^c\}$ *is the collection of false positive nodes.*

The proof of Theorem 6 and Theorem 7 is given in Appendix B.1 and Appendix B.2, respectively.

To give a brief sketch of the proof, we first consider the following reduced version of the ℓ_1-regularized PageRank problem, (see Appendix A.3 for details on this)

$$\min_{x\in\mathbb{R}^n}\left\{\frac{1}{2}x^\top Qx - \alpha x^\top \mathbf{s} + \rho\alpha\|Dx\|_1 : x_{K^c} = 0\right\}. \tag{7}$$

The reduced problem (7) is introduced because it allows us to analyze the properties of the solution while restricting the matrices Q and D to the nodes in K only. Note that when there are no false positives, the optimal solution (2) coincides with the solution to Problem (7). More generally, we can prove that the support set of $\widehat{x}$ is always bigger than the support set of the solution to (7). Furthermore, we can show that when $\rho \leq \rho(\delta)$, the solution to (7) is strictly positive over K. Putting these together then establishes Theorem 6. On the other hand, using the optimality condition Lemma 3, we can obtain an upper bound on the volume of the optimal solution $\widehat{x}$. Since, by Theorem 6, $\widehat{x}$ entirely recovers the target cluster K, the errors in the support set of $\widehat{x}$ are solely due to the presence of false positives. Combining the two results, we can get an upper bound on the false positives, therefore proving Theorem 7.

The results of Theorem 6 and Theorem 7 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

1. More precisely, we assume $(1-\delta)p^2k \geq c_0^{-1}\delta^{-2}\log k$ for a fixed constant $c_0 > 0$. Then with probability at least $1 - 6e^{-c_0\delta^2(1-\delta)p^2k}$, the statement in the theorem holds.
2. The same probability bound holds as in Theorem 6.

find small- and meso-scale clusters in massive graphs (Leskovec et al., 2009, 2010). In addition, Theorem 6 indicates that if γ is small, 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 (equivalently a smaller value of ρ 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 7 shows that the volume of false positives grows quadratically as $1/\gamma$, so we need γ to be well-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 recovered cluster to keep high mass inside K.

We remark several other comments regarding the results. First, we suspect the current bound we obtain in (6) may not be tight with respect to α and other constants, and especially the factor $(\frac{1+\alpha}{1-\alpha})^2$ may be an artifact of our proof. Studying the lower bound on the performance of the method, as well as obtaining an improved bound on false positives, is therefore an interesting future direction to pursue. Furthermore, on the basis of 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.

3.2 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 $p = \mathcal{O}(1), k = \mathcal{O}(1)$, 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 $q = \frac{c}{n}$ for a fixed numerical constant $c > 0$.*

As we noted previously, 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 (Leskovec et al., 2009, 2010). 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 statement of exact recovery guarantees for ℓ_1-regularized PageRank when applied to the noisy graph generated from Definition 5. In particular, the fact that $q = \mathcal{O}\left(n^{-1}\right)$ 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.

Theorem 8 (No false positives.) *Suppose the same conditions as Theorem 6. Assume that Assumption 1 holds and that the size of the target cluster k is $\geq 2(c+3)$. Fix $\delta \leq 0.1$. If the regularization parameter satisfies $\rho \geq \rho(\delta)$, then there is a good starting node in K such that the solution to*

Problem (2) *with that node as a seed node and with teleportation parameter* $\alpha \in [0.1, 0.9]$ *satisfies*

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

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] *as long as*

$$\frac{C(0.5c+1)}{\gamma p} = \mathcal{O}\left(\frac{1}{\gamma p}\right) < d_j, \tag{8}$$

for all node $j \in K^c$ *adjacent to* K*, where* $C > 0$ *is a universal constant.*

The proof is given in Appendix B.3.

The proof of Theorem 8 proceeds by showing that under the conditions of the theorem, the solution to the reduced problem (7) obeys the optimality condition for the full-dimensional ℓ_1-regularized PageRank problem (Lemma 3). To do so, we use concentration inequalities to bound the difference between the solution to (7) and the ℓ_1-regularized PageRank vector on the population graph $\overline{G} = (V, E)$ (see Appendix A.2). Once we establish the equivalence of the solutions for both problems, the result follows since by construction the support set of the solution to (7) is contained in the target cluster.

In the statement of Theorem 8, we required the conditions $\alpha \in [0.1, 0.9]$ and $\delta \leq 0.1$ mainly to avoid overly complicated constants; while this simplifies the presentation of the theorem, it is not difficult to show that a similar result holds more generally. Importantly, when combined with Theorem 6 (full recovery of the cluster), our result Theorem 8 immediately establishes that ℓ_1-regularized PageRank recovers the target cluster exactly, even when the target cluster is constant-sized. We state this result informally in the following, which requires no proof.

Corollary 9 (Exact recovery; informal statement.) *Under the same assumptions as Theorem 6 and Theorem 8, there is a good starting node in* K *such that* ℓ_1*-regularized PageRank parameterized with that node as a seed node satisfies*

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

with nonvanishing probability.

It should be noted that a sort of condition like (8) 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 6 and Theorem 7, condition (8) rules out some specific models to which Theorem 8 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 $\mathcal{O}(1/\gamma p) = \mathcal{O}(1)$. In practice, condition (8) 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.

3. More precisely, we 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.

3.3 Comparison with existing results

The local graph clustering problem has been relatively well-studied in the area of theoretical computer science, and the existing works largely focus on the worst-case guarantees. We now compare our results through the random graph model with the current known state-of-the-art worst-case results, given by Zhu et al. (2013).

First, the main result Theorem 1 of Zhu et al. (2013), when applied to our population graph $\overline{G}$, implies that $\mathrm{Vol}(\mathrm{FP}), \mathrm{Vol}(\mathrm{FN}) \leq \mathrm{Vol}(K) \cdot \mathcal{O}\left((1-\gamma)\log k\right)$, as long as $\mathrm{Gap} = \mathcal{O}\left(1/((1-\gamma)\log k)\right) \geq \mathcal{O}(1)$. When $\gamma = \mathcal{O}(1) \in (0,1)$, our Theorem 6 states that if $pk^2 \geq \mathcal{O}(\log k)$, the output of the ℓ_1-regularized PageRank model does not contain any false negative. This cannot be deduced from Zhu et al. (2013). In addition, our general bound on false positive, i.e., $\mathrm{Vol}(\mathrm{FP}) \leq \mathrm{Vol}(K) \cdot \left(\mathcal{O}\left(1/\gamma^2\right) - 1\right)$ in Theorem 7, is better than the worst-case bound of Zhu et al. (2013) in the regime of large γ, which is typically the case for many interesting scenarios. For instance, when the expected target conductance $\overline{\mathrm{Cond}} = 1 - \gamma$ is small and fixed, the bound of the worst-case result degrades as the size of the target cluster k increases, whereas our result is improved by increasing the probability bound. In the regime of $p = \mathcal{O}(1), q = \mathcal{O}(1/n)$, and $k = \mathcal{O}(1)$ (hence $\gamma = \mathcal{O}(1)$), our Theorem 8 shows that the output even contains no false positive. In this particular case, the strong separability ($p = \mathcal{O}(1), q = \mathcal{O}(1/n)$) corresponds to a constant signal-to-noise ratio, since even for $q = \mathcal{O}(1/n)$ there are still a constant amount of edges outgoing from the target cluster, while the internal edges inside the target cluster is also constant. Although in practice the exact recovery of the target cluster may be a strong requirement, nevertheless, for real world clusters with high signal-to-noise ratio, ℓ_1-regularized PageRank can still reconstruct the ground truth clusters more or less exactly (see, for instance, Section 6.2).

Finally we briefly give a comparison of our theoretical results with the information theoretic results of Chen and Xu (2016) in the special case of planted clustering model. To ease and simplify the comparison, we only consider the case where $p = O(1)$ and $q = O(\log n/n)$. In this case, Chen and Xu (2016, Theorem 2.5) implies that a solution to a SDP achieves exact recovery as long as $k \geq O(\log n)$, whereas our Theorem 6 and 7 suggest that in the same regime ℓ_1-regularized PageRank fully recovers the target cluster while picking up a constant proportion of false positives. Importantly, ℓ_1-regularized PageRank is essentially a local method, while Chen and Xu (2016)'s SDP is a global method that explores the entire graph.

4. Equivalence between ℓ_1-regularized PageRank and APPR

In this section, we illuminate a deep connection between approximate personalized PageRank (APPR) and ℓ_1-regularized PageRank, and based on this result, we provide novel statistical recoverability guarantees for APPR when the graph is generated from the random graph model.

Approximate personalized PageRank (APPR), first studied in Andersen et al. (2006) and followed up by a number of subsequent works, is at the heart of local graph clustering whose central idea is to approximate the original PageRank vector without ever touching the entire graph. Gleich and Mahoney (2014) show that APPR implicitly corresponds to adding a ℓ_1 norm regularization term to the ℓ_2 norm formulation of a min-cut linear program, explaining why APPR gives rise to very sparse solutions. Fountoulakis et al. (2019) further observed that the APPR algorithm is equivalent to the iterative coordinate descent solver applied to the PageRank linear system (1) with a specifically designed termination criterion. In our notation, we can write the algorithm in a simple and compact way:

- Given a parameter $\rho > 0$, initialize $x^{(0)} = 0$;
- For each $k \geq 0$, while $\|D^{-1}\nabla f(x^{(k)})\|_\infty \geq \rho\alpha$, iterate the steps[4]

$$\begin{cases} \text{Choose an } i \in [n] \text{ such that } \nabla_i f(x^{(k)}) \leq -\rho\alpha d_i; \\ \text{Update } x_i^{(k+1)} = x_i^{(k)} - d_i^{-1}\nabla_i f(x^{(k)}). \end{cases} \tag{9}$$

- Output $x^{(k^\star)} =: x^{\text{APPR}}$.

The output of the APPR algorithm is the personalized score vector x^{APPR} that efficiently approximates the original PageRank. Note that the updating formula (9) requires one evaluation of the gradient which can be done by accessing only neighboring nodes of the selected entry i. As a result the APPR algorithm can solve the PageRank equation (1) extremely efficient; and in particular the running time of the algorithm depends on the size of the output rather than the size of the whole graph.

APPR was developed purely from an algorithmic perspective and the output of the algorithm depends on which coordinate i is chosen at every iteration (see the updating step (9)). This property, while allowing the algorithm to enjoy the *locality* property and resulting in the output vector that is highly sparse, can lead to different results of local clustering in principle (albeit the results may look more or less similar). On the other hand, ℓ_1-regularized PageRank eliminates this issue by decoupling the locality/sparsity of the local clustering output vector from the algorithmic issue of running in time independent of the size of the graph. In particular, any convex optimization algorithm that exploits the locality of the problem, such as proximal gradient descent (ISTA), can be employed to minimize the ℓ_1-regularized objective function (2) while touching only the neighbors of the current non-zero nodes. Therefore the formulation via ℓ_1-regularized PageRank achieves two objectives of the graph processing that are desirable for local graph clustering, i.e., both the locality/sparsity of the solution and the "strongly local" running time of the algorithm.

It is shown in Fountoulakis et al. (2019) that ℓ_1-regularized PageRank can be viewed as a variational formulation of APPR, in the sense that the optimality condition for the ℓ_1-regularized objective function (Lemma 3) implies the termination criterion of APPR, i.e., $\|D^{-1}\nabla f(x^{(k)})\|_\infty < \rho\alpha$. In this work, we strengthen the connection between the output of APPR and the output of ℓ_1-regularized PageRank. In particular, we show that with appropriate choices of the regularization parameters, both approaches become equivalent in terms of cluster recovery.

Theorem 10 (Equivalence between ℓ_1-reg. PR and APPR.) *Let $\widehat{x}(\rho)$ be the ℓ_1-regularized PageRank vector* (2)*, let $x^{APPR}(\rho)$ be the output of APPR* (9) *at the regularization parameter ρ. Then for any $\rho_0 > 0$,*

$$\text{support}(\widehat{x}(\rho_0)) \subseteq \text{support}(x^{APPR}(\rho_0)) \subseteq \text{support}(\widehat{x}((1-\alpha)\cdot\rho_0/2)).$$

That is, for any parameter $\rho = \rho_0 > 0$, the support set of the output of APPR is, respectively, a superset and a subset of that of ℓ_1-regularized PageRank at $\rho = \rho_0$ and $\rho = (1-\alpha)\cdot\rho_0/2$.

4. It can be shown that if APPR is initialized at $x^{(0)} = 0$ then $\nabla f(x^{(k)}) \leq 0$ for all $k \geq 0$. In this case, the termination criterion of APPR reads as $\nabla_i f(x^{(k)}) > -\rho\alpha d_i$ for all $i \in [n]$.

The proof is given in Appendix B.4.

Theorem 10 establishes rather a stronger equivalence between the ℓ_1-regularized PageRank and the APPR approaches than was established in the prior work of Fountoulakis et al. (2019). Importantly, using this result, various statistical guarantees for the output cluster of APPR directly follow from the results we establish for the output of ℓ_1-regularized PageRank. While analyzing APPR's output in the random graph setting involves technical challenges due to its algorithmic nature, the fact that the output cluster of ℓ_1-regularized PageRank is given by the optimization solution allows us to analyze statistical properties more easily. In the previous section (Section 3), we stuided the recovery guarantees for ℓ_1-regularized PageRank under a random graph model. In the next section we will show how this guarantee can be transferred to the output cluster of APPR using Theorem 10.

4.1 Approximate personalized PageRank on random graph

One advantage of variational formulation of local graph clustering, via ℓ_1-regularized convex program (2), is that it allows tractable analysis of the method in random graphs. Given the connection between ℓ_1-regularized PageRank and APPR established in Theorem 10, this further allows us to obtain the statistical guarantees of the APPR algorithm under the random graph model. We formally state this result in the following theorem, which is a simple consequence of Theorem 10, in combination with those obtained in Section 3.1 and 3.2. (We write $\mathrm{FP}(x^{\mathrm{APPR}}) = \{i \in \mathrm{support}(x^{\mathrm{APPR}}) : i \in K^c\}$ to denote the collection of false positive nodes in the APPR's output.)

Theorem 11 (Recovery guarantees for APPR.) *Consider the APPR algorithm given in* (9) *with parameter $\rho = \rho(\delta)$. Under the same conditions as Theorem 6, the output vector $x^{APPR}(\rho(\delta))$ satisfies*

$$K \subseteq \mathrm{support}(x^{APPR}) \quad \textit{and} \quad \mathrm{Vol}(\mathit{FP}(x^{APPR})) \leq \mathrm{Vol}(K) \left[\mathcal{O}\left(\frac{1}{\gamma^2}\right) - 1 \right],$$

with nonvanishing probability. Furthermore, under the same conditions as Theorem 8, there is a good starting node in K such that the APPR's output vector parameterized with that node as a seed node satisfies

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

with nonvanishing probability, as long as

$$\mathcal{O}\left(\frac{1}{\gamma p}\right) < d_j,$$

for all node $j \in K^c$ adjacent to K.

The proof is given in Appendix B.5.

5. Stagewise PageRank and solution paths

The ℓ_1-regularized PageRank problem (2) is convex and there are numerous ways to solve it using convex optimization techniques. In Fountoulakis et al. (2019), the authors apply proximal gradient descent (ISTA) and show that the algorithm enjoys the *locality* property, meaning that the algorithm touches at most the optimal support set and its neighbors, while inheriting the fast convergence property of the proximal gradient descent. Optimization algorithms typically solve the problem

at a single value of ρ, or a sequence of multiple values. On the other hand, "path algorithms" are designed to solve the problem for all values of a regularization parameter $\rho \in (0, \infty]$, or a subset of it when terminated early. This is more desirable when we need solutions over a list of parameter values.

The idea of designing path algorithms has already gained much attention in the sparse regression literature, first pioneered by Efron et al. (2004) and further developed by Zou et al. (2007); Hastie et al. (2004); Arnold and Tibshirani (2016). This has rendered the exploration of full regression coefficient paths relatively inexpensive. Unlike regression setting, however, this type of path algorithm has been less studied in the setting of local graph clustering (Gleich and Kloster, 2016). Motivated by the path algorithms developed in the sparse regression setting, we consider the following coordinate-wise algorithm for local graph clustering:

- Initialize $x^{(0)} = 0$;
- For each $k \geq 0$, iterate the steps

$$\begin{cases} \text{Choose an } i \in [n] \text{ such that } d_i^{-1} \nabla_i f(x^{(k)}) \leq 0 \text{ is the smallest among } [n]; \\ \text{Update } x_i^{(k+1)} = x_i^{(k)} + d_i^{-1} \eta. \end{cases} \tag{10}$$

- Output a sequence $\{x^{(0)}, x^{(1)}, \ldots\}$.

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 above algorithm is called the "stagewise" algorithm, and in fact it has been widely studied by many authors (Efron et al., 2004; Hastie et al., 2007; Rosset et al., 2004; Rosset and Zhu, 2007; Zhao and Yu, 2007; Tibshirani, 2015). The stagewise algorithm is known to have implicit regularization effect closely related to ℓ_1 norm regularization (Tibshirani, 2015), and in particular, if each component of the ℓ_1-regularized solution has a monotone path, then the stagewise path exactly coincides with that of ℓ_1 regularization, for $\eta \to 0$ (Efron et al., 2004; Rosset et al., 2004). Recalling our earlier result on the monotone path property of ℓ_1-regularized PageRank (Lemma 4), the following result is then immediate.

Corollary 12 *The output sequence of stagewise algorithm described in* (10) *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 *provably* 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 (10) and the expression of the gradient $\nabla_i f(x)$. Thus, when terminated early, the algorithm produces an approximate and partial solution path without making access to the entire graph.

For the ℓ_1-regularized PageRank method, the parameter ρ controls the extent to which the random walk has moved farther from the seed, so different values of ρ 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-regularized

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 Gleich and Kloster (2016). 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 and/or 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 complementary way to the optimization algorithms that directly solve (2). We also refer the readers to Tibshirani (2015) for comprehensive study of the stagewise algorithm for general sparse modeling problem.

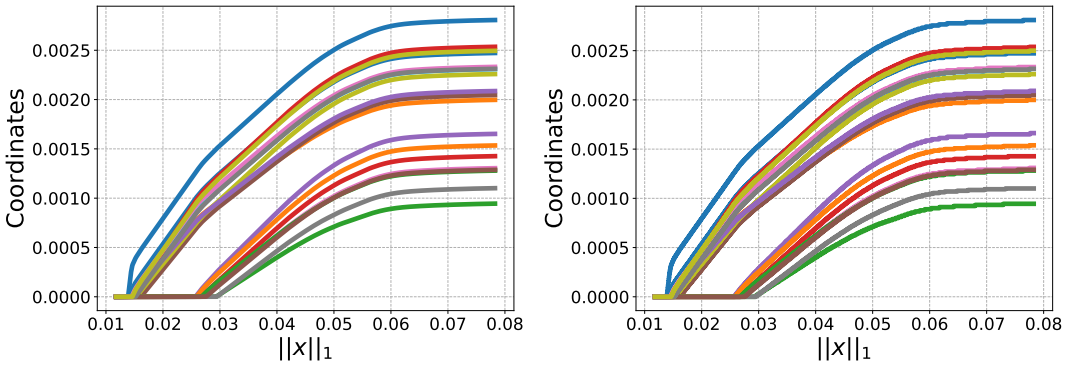
In Figure 1, we compare the actual solution path to the stagewise algorithm paths for different step sizes. We generate data from the stochastic block model with $p = 0.5, q = 0.002$ and $r = 50$. Each cluster has 20 nodes. Figure 1 shows the ℓ_1-regularization path and stagewise component paths for one particular draw from the stochastic block model. Here we only show the solution paths for nodes in the target cluster without seed node, among $n = 1000$ nodes (each color line corresponds to different nodes in the target cluster). 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.

6. Numerical evaluation

In this section, we provide a detailed numerical evaluation, to illustrate the performance of ℓ_1-regularized PageRank on synthetic and real data. We have conducted a comprehensive experiments to demonstrate the state-of-the-art performance of the method across a wide range of real graphs, which has not been investigated at this level of extent in the prior works. To measure the quality of the recovered cluster, we define Precision and Recall as $\mathrm{Vol}(\mathrm{TP})/\mathrm{Vol}(\mathrm{support}(\widehat{x}))$ and $\mathrm{Vol}(\mathrm{TP})/\mathrm{Vol}(K)$, respectively. The F1score is the harmonic mean of precision and recall, $\left(2/\left(\mathrm{Precision}^{-1} + \mathrm{Recall}^{-1}\right)\right)$. We will also make use of conductance, where recall $\mathrm{Cond} = \mathrm{Cut}(S, S^c)/\min\left(\mathrm{Vol}(S), \mathrm{Vol}(S^c)\right)$. The lower the conductance value is, the better the quality of the cluster is. For our experiments, we solve problem (2) using a proximal coordinate descent algorithm, which enjoys both the "strongly local" running time property (running time depends on the size of cluster rather than the entire graph) as well as linear convergence (Fountoulakis et al., 2019). If we further need to explore the entire solution path of (2), the stagewise algorithm with small step size is used.

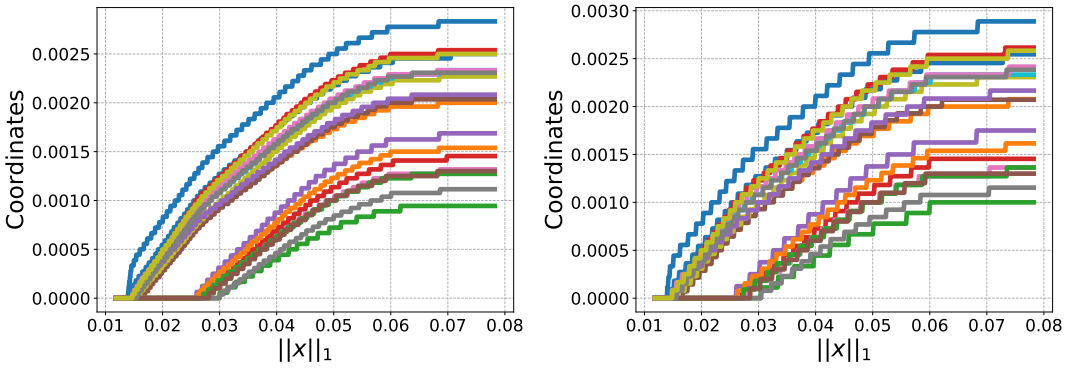
6.1 Simulated data

In Figure 2 we demonstrate the performance of the stagewise algorithm (10) with $\eta = 10^{-4}$ as γ increases. For the experiments in Figure 2, we fix the teleportation parameter $\alpha = 0.1$. We generate graphs from the stochastic block model which consists of 10 clusters, each of which has 20 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. Here we set $p = 0.5$ and q is varying in order to generate various γ as is shown in Figure 2. The results are averaged over 30 trials. For this



(a) ℓ_1-reg. path

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



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

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

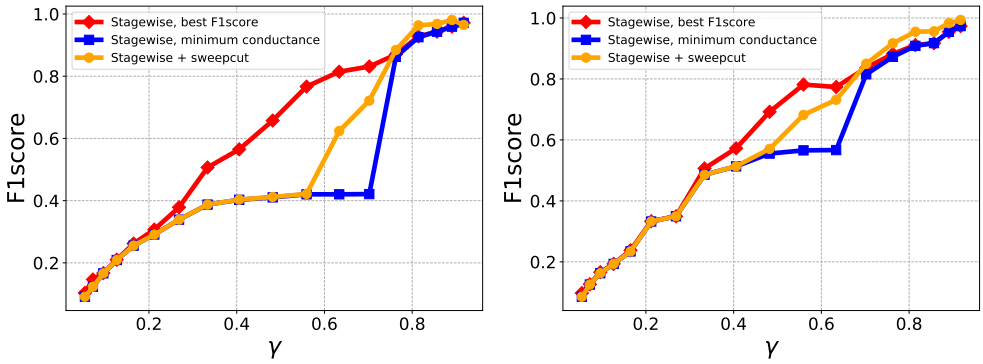
Figure 1: Comparison of ℓ_1 regularization path and stagewise paths for different step sizes η. The profiles are shown only for nodes in $K \backslash S$ among $n = 1000$ nodes. Each color in the plot corresponds to different nodes. The x-axis is the ℓ_1 norm of the current estimates. For stagewise, the results are obtained with $7706, 1527$, and 764 iterations respectively.

experiment we could use ℓ_1-regularized PageRank or APPR, but all these methods are similar with stagewise having the benefit of not needing to choose a regularization parameter. In particular, in Section 5 we show that for small enough step size η, the stagewise algorithm explores the piecewise constant solution path of the ℓ_1-regularized PageRank problem (see Corollory 12). Therefore, Figure 2 reflects the performance of ℓ_1-regularized PageRank as well. Moreover, in Theorem 10 we show an equivalence result between APPR and ℓ_1-regularized PageRank, and later on in Figure 4 we illustrate this equivalence in practice on real data as well. Therefore, Figure 2 reflects the approximate performance of APPR as well.

There are two subfigures in Figure 2. In Subfigure 2(a), we illustrate the performance when the stagewise algorithm touches at most $3/4$ of the overall nodes in the graph. After this threshold, the algorithm is terminated. In Subfigure 2(b), we illustrate the performance when the stagewise algorithm touches at most $1/2$ of the nodes. We introduce this threshold for two reasons. First, in local graph clustering, it does not make sense to explore more than half of the graph. One could do this of course, but if one is willing to touch more than half of the graph, they could as well have utilized non-local algorithms. Second, we do this to illustrate the importance of not letting the algorithm to explore a huge part of the graph if the stagewise algorithm (or ℓ_1-regularized PageRank) is combined with metrics such as conductance. If one does this, then they take the risk of finding a larger or a smaller cluster that has smaller conductance than the target cluster but is very different than the target cluster, thus resulting in small F1score. In each subfigure, we illustrate the performance of the stagewise algorithm in three cases. The first (red line with rhombs, "Stagewise, best F1score") is a best-case scenario where we select the solution with the best F1score out of all the solutions that are produced by the stagewise algorithm (10). The second scenario (blue line with squares, "Stagewise, minimum conductance") shows a more realistic case where we select the solution with minimum conductance out of all the solutions that are produced by the stagewise algorithm. The third scenario (orange line with circles, "Stagewise + sweep-cut") also shows a realistic case where at each iteration of the stagewise algorithm we perform sweep-cut to find a cluster of small conductance and out of all iterations we return the cluster with the smallest conductance. The sweep cut rounding procedure is a common technique to post-process the output of local graph clustering methods—for details, we refer the reader to one of these papers (Andersen et al., 2006; Zhu et al., 2013).

In both subfigures, the F1score for the "Stagewise, best F1score" scenario scales linearly as a function of γ. On the other hand, in Subfigure 2(a) the scenario where we select the solution with minimum conductance scales sub-linearly as a function of γ until a phase-transition around $\gamma \approx 0.75$ after which, the scenario of minimum conductance matches the best-case scenario. In Subfigure 2(a) we get similar results for the "Stagewise + sweep-cut" scenario, although the gap with the best-case scenario becomes smaller. In Subfigure 2(b) we see that the gap between "Stagewise, best F1score" scenario and the other two scenaria is much smaller. As we discussed above, the gap closes because in Subfigure 2(b) we do not allow the stagewise algorithm to explore more than half of the nodes in the graph, and the method avoids returning clusters of small conductance but low F1score.

In Figure 3 we demonstrate a much more detailed view of the performance of the stagewise algorithm for four representative γ's. More precisely, we show that for large and medium values of γ there exists a solution in the solution path of the ℓ_1-regularized PageRank, which is found by the stagewise algorithm, and it recovers the target cluster. While for small $\gamma < 0.5$, there is no solution in the solution path of the ℓ_1-regularized PageRank that recovers the target cluster with high



(a) F1score against γ. Stagewise terminated when $3/4$ of the overall nodes are nonzero.

(b) F1score against γ. Stagewise terminated when $1/2$ of the overall nodes are nonzero.

Figure 2: In this figure we demonstrate performance (F1score) of the stagewise algorithm (10) as γ increases. We demonstrate three scenaria. In the first scenario (red line with rhombs, "Stagewise, best F1score") we select the solution with the best F1score out of all the solutions that are produced by the stagewise algorithm. In the second scenario (blue line with squares, "Stagewise, minimum conductance") we select the solution with minimum conductance out of all the solutions that are produced by the stagewise algorithm. In the third scenario (orange line with circles, "Stagewise + sweep-cut") at each iteration of the stagewise algorithm we perform sweep-cut to find a cluster of small conductance and out of all iterations we return the cluster with the smallest conductance.

accuracy. For Figure 3 we use the same experiment setting as in Figure 2. For large γ, Figure 3(a), we observe that when the stagewise algorithm recovers about 20 nodes, then these nodes correspond to very high precision and recall, and as the number of nodes in the solution increases then precision decreases. We also observe that conductance of the recovered cluster is a good metric for finding the target cluster. Meaning that we will find the target cluster with high precision and recall 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 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 of solution fails to relate to the target cluster and there is no output of the stagewise algorithm that recovers the target cluster accurately.

6.2 Real data experiments

Next, we test the performance of local graph clustering methods using biology networks and social networks. All the graphs that we consider are unweighted and undirected. For a summary of the datasets see Table 1. All datasets that are used come with *suggested* ground-truth clusters for social networks and with a gold standard for biology networks. However, we filter the given ground-truth clusters by measuring their conductance values. In particular, we keep only the ground-truth clusters that have conductance value less than or equal to 0.6. This means we keep the clusters that the number of edges that cross the cluster is 60% of the volume of the cluster. This way we obtain a wide range of clusters from "good" (small conductance) to noisy (large conductance, i.e., close to 0.6).

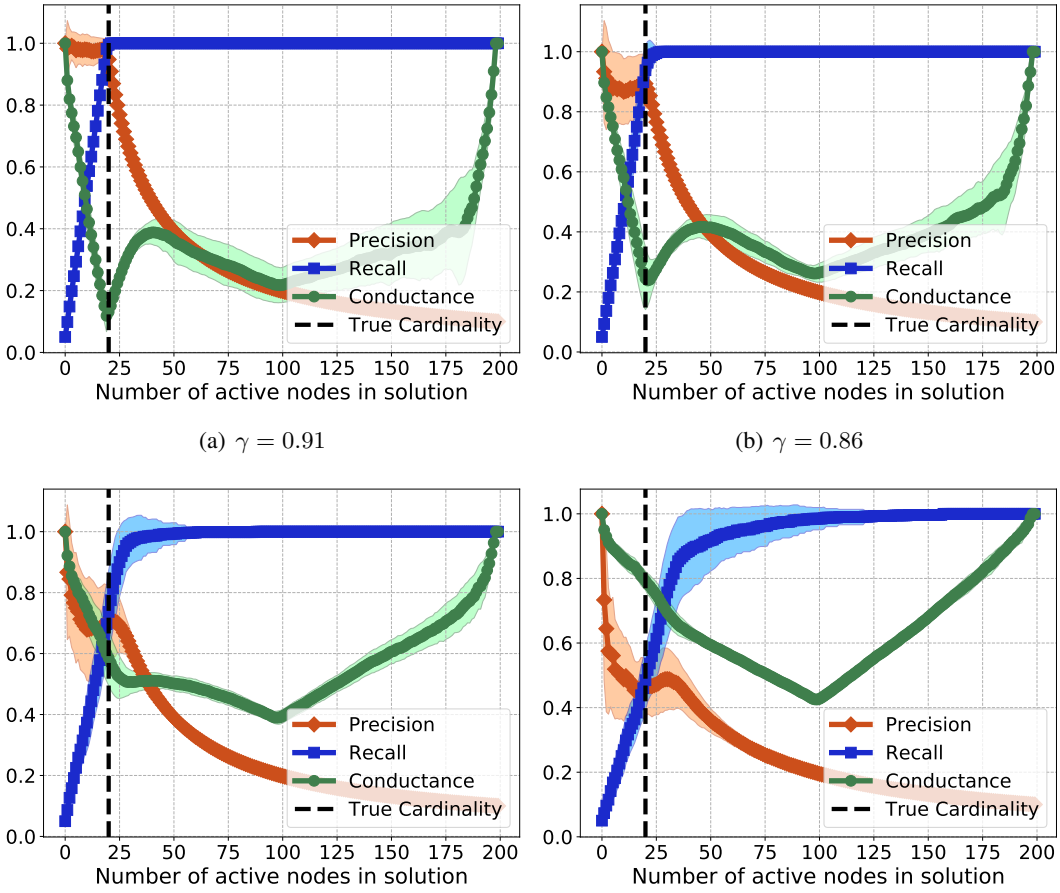
Table 1: Summary of datasets

dataset	number of nodes	number of edges	description
Sfld	232	15570	pairwise similarities of blasted sequences of proteins
PPI-mips	1096	13221	protein-protein interaction network
FB-Johns55	5157	186572	Facebook social network for Johns Hopkins University
Colgate88	3482	155043	Facebook social network for Colgate University
Orkut	3072441	117185083	Large-scale on-line social network

6.2.1 DATASETS

Sfld. This dataset contains pairwise similarities of blasted sequences of 232 proteins belonging to the amidohydrolase superfamily (Brown et al., 2006). There are 232 nodes and 31140 edges in this graph. A gold standard is provided describing families within the given superfamily. According to the gold standard the amidrohydrolase superfamily contains 29 families/clusters. However, after filtering the families we find that only two have conductance value less than 0.6. (This should not be surprising, given that the graph is so dense.) In Table 2 we present properties of the two clusters that correspond to urease.0 and AMP amidrohydrolase superfamilies.

PPI-mips. This dataset is a protein-protein interaction graph of mammalian species (Pagel et al., 2004). There are 1096 nodes and 26442 edges in this graph. In Table 2 we provide all *suggested* ground-truth clusters that have conductance less than 0.6.



(a) $\gamma = 0.91$

(b) $\gamma = 0.86$

(c) $\gamma = 0.65$

(d) $\gamma = 0.44$

Figure 3: In Figures 3(a), 3(b) and 3(c) we illustrate that for large and medium values of γ, the stagewise algorithm recovers the target cluster. While for small γ in Figure 3(d), the stagewise algorithm k does not recover the target cluster with high accuracy. The x-axis gives the number of nonzero nodes in the solution of the stagewise. The vertical line indicates the cardinality of the target cluster ($k = 20$).

Table 2: "Ground truth" clusters for the Sfld and PPI-mips datasets. Full details about the datasets can be found in the original paper (Pagel et al., 2004). In this table we report the volume, the number of nodes and the conductance of each ground truth cluster.

dataset	feature	feature (short name)	volume	nodes	conductance
Sfld	urease.0	urease	16209	100	0.42
	AMP	AMP	1721	28	0.56
PPI-mips	Actin-associated-proteins	Actin	870	24	0.36
	Anaphase-promoting-complex	Anaphase	165	11	0.33
	Cdc28p-complexes	Cdc28p	135	10	0.33
	Coat-complexes	Coat	676	19	0.49
	cytoplasmic-ribosomal-large-subunit	ct-large	9720	81	0.33
	cytoplasmic-ribosomal-small-subunit	ct-small	4788	57	0.33
	F0-F1-ATP-synthase	F0-F1-ATP	315	15	0.33
	H+-transporting-ATPase-vacuolar	H+	315	15	0.33
	mitochondrial-ribosomal-large-subunit	mc-large	1488	32	0.33
	mitochondrial-ribosomal-small-subunit	mc-small	273	14	0.33
	Mitochondrial-translation-complexes	mc-complex	281	12	0.53
	mRNA-splicing	mRNA	1861	33	0.43
	Nuclear-pore-complex	Nuclear	614	18	0.50
	RNA-polymerase-II-holoenzyme	RNA	1487	29	0.45
	Spindle-pole-body	Spindle	981	22	0.52
	TRAPP-complex	TRAPP	135	10	0.33
	tRNA-splicing	tRNA	165	11	0.33
	19-22S-regulator	19-22S	459	18	0.33
	20S-proteasome	20S	315	15	0.33

FB-Johns55. This graph is a Facebook anonymized dataset on a particular day in September 2005 for a student social network at John Hopkins university. The graph is unweighted and it represents "friendship" ties. The data form a subset of the Facebook100 data-set from Traud et al. (2011, 2012). This graph has 5157 nodes and 186572 edges. This dataset comes along with 6 features, i.e., second major, high school, gender, dorm, major index and year. We construct "ground truth" clusters by using the features for each node. In particular, we consider nodes with the same value of a feature to be a cluster, e.g., students of year 2009. As analyzed in the original paper that introduced these datasets (Traud et al., 2012), for FB-Johns55 the year and major index features give non-trivial "assortativity coefficients", which is a "local measure of homophily". This agrees with the ground truth clusters we find after applying our filtering technique. The clusters per graph are shown in Table 3.

Colgate88. This graph is constructed similarly to the FB-Johns55 graph but for Colgate University. We apply the same filtering techniques as for FB-Johns55 and we present the filtered clusters in Table 3. There are 3482 nodes and 155043 edges in this graph.

Orkut. Orkut is a free on-line social network where users form friendship each other. Orkut also allows users form a group which other members can then join. This dataset has 3072441 nodes and 117185083 edges. It can be downloaded from Leskovec and Krevl (2014). This dataset comes with

Table 3: "Ground truth" clusters for the FB-Johns55 and Colgate88 datasets.

dataset	feature	volume	nodes	conductance
FB-Johns55	year 2006	81893	845	0.54
	year 2007	89021	842	0.49
	year 2008	82934	926	0.39
	year 2009	33059	910	0.21
	major index 217	10697	201	0.26
	second major 0	178034	2844	0.51
	dorm 0	137166	2121	0.52
	gender 1	181656	2144	0.46
	gender 2	173524	2598	0.48
Colgate88	year 2004	14888	230	0.54
	year 2005	50643	501	0.50
	year 2006	62065	557	0.48
	year 2007	68382	589	0.41
	year 2008	62430	641	0.29
	year 2009	35379	641	0.11
	secondMajor 0	175239	2107	0.54
	dorm 0	100414	1157	0.53
	gender 1	162759	1695	0.48
	gender 2	123724	1485	0.55

5000 ground truth communities, which we filter by maintaining the clusters with conductance less than or equal to 0.6. Out of the 5000 communities only 282 pass our filtering test.

The above real graphs may not be generated from the random model that we consider in Section 3, nonetheless, we have evaluated the method to illustrate the scenarios that are not just idealized. Moreover, the Facebook social networks and the biology networks are expected to exhibit homophily structure for some ground truth clusters that are highly inter-connected relative to the rest of the graph. These block-structured networks belong to a special class of networks and we believe that random graph models, such as stochastic block model, can closely approximate it.

6.2.2 METHODS

We compare the performance of ℓ_1-regularized PageRank with state-of-the-art local graph clustering algorithms. There are two categories of local graph clustering methods: local spectral (Andersen et al., 2006) and it's follow-up work (Zhu et al., 2013); and local flow-based methods (Lang and Rao, 2004; Andersen and Lang, 2008; Orecchia and Zhu, 2014; Veldt et al., 2016).

Approximate personalized PageRank (APPR): In Andersen et al. (2006) and Zhu et al. (2013) an approximate personalized PageRank (APPR) algorithm is proposed, where the personalized PageRank linear system is solved approximately using a local diffusion process. Both papers study nearly identical algorithms, with the latter paper giving better theoretical guarantees based on a stronger assumption. In particular, the authors in Zhu et al. (2013) assume that the internal connectivity of the

target cluster (minimum conductance of the induced subgraph for the target cluster) is quadratically stronger than the conductance of the target cluster.

SimpleLocal (SL): Out of the three flow-based methods, FlowImprove, LocalFlowImprove and SimpleLocal in Andersen and Lang (2008); Orecchia and Zhu (2014); Veldt et al. (2016), respectively, SimpleLocal Veldt et al. (2016) simplifies and generalizes the methods proposed in Andersen and Lang (2008); Orecchia and Zhu (2014), while having similar theoretical and practical guarantees in terms of quality of the output. Depending on the parameter tuning of SimpleLocal, one can obtain FlowImprove and one of the two variants of LocalFlowImprove. In particular, there are two variants of LocalFlowImprove in Orecchia and Zhu (2014). An "exact" and an "inexact" variant. The "exact" variant is equivalent to SimpleLocal. The only thing that differs is how one solves the sub-problems at each iteration, see Section 8 in Fountoulakis et al. (2020a) for details. In the second variant, each sub-problem is solved using Dinic's algorithm with early termination. This version has slightly worse guarantees in terms of conductance than SimpleLocal, although in practice we expect both methods to perform similarly. We expect that in practice the differences between the methods will be merely computational. Also, even the computational differences are expected to be minor since all methods have strongly-local running time. Since there exists no implementation of the inexact variant of the LocalFlowImprove algorithm we choose not to implement it.

Therefore, we will only use SimpleLocal in our experiments, and we will use different parameter tuning such that we obtain performance for a wide-range of methods. Moreover, SimpleLocal requires initialization with a set of seed nodes that ideally has some overlap with the target cluster. Therefore, in this paper, we will use SimpleLocal with three different initialization techniques. Below we comment on why we choose the following inputs to SimpleLocal.

1. ℓ_1-reg. PR-SL: SimpleLocal using the output of ℓ_1-reg. PR as input.

2. BFS-SL: We will initialize SimpleLocal using the output of a breadth-first-search-type (BFS) algorithm starting from a given seed node. The algorithm that is used for initialization of SimpleLocal is shown in Algorithm 1. Details about how we use this algorithm are given in the next subsection. Briefly, Algorithm 1 is used as a seed node expansion technique, which has also been used in Veldt et al. (2016).

Note that all these flow-based methods are "improve" methods, i.e., they are not stand-alone, and they require initial input from some other stand-alone method, such as APPR or ℓ_1-reg. PR. This is both a theoretical and a practical argument. It is a theoretical argument because in the theoretical analysis of SimpleLocal (see Theorem 3 in Orecchia and Zhu (2014)), the input to SimpleLocal is required to have sufficient overlap with the target cluster. Otherwise, currently there is no guarantee that flow-improve methods output a reasonable approximation to the target cluster in terms of its conductance value (Theorem 1 in Orecchia and Zhu (2014)) or in terms of false and true positives (Theorem 3 in Orecchia and Zhu (2014)). Theorem 3 in Orecchia and Zhu (2014) suggests that sufficient overlap can be achieved either by using a local spectral method such as APPR or ℓ_1-reg. PR. In this paper, we use only the latter. This is because, as shown in Fountoulakis et al. (2019), theoretically and empirically APPR is very similar to ℓ_1-reg. PR. We also show extensive experiments in this section about the similarity of APPR and ℓ_1-reg. PR. In practice one could use heuristics such as a BFS-type algorithm to slightly expand a seed node within a target cluster and then provide the output of the BFS-type as input to SimpleLocal, i.e., see also Veldt et al. (2016). In Algorithm 1 we provide a pseudo-code for the BFS-type algorithm that we use in this

paper. Basically, the only difference with a standard BFS algorithm is that we explore neighbors in batches, which allows us to terminate the BFS algorithm after a given number of steps.

6.2.3 PARAMETER TUNING

APPR and ℓ_1-reg. PR have two parameters, the teleportation parameter α and a tolerance/ℓ_1-regularization parameter ρ. The teleportation parameter should be set according to the reciprocal of the mixing time of a random walk within the target cluster, which is equal to the smallest nonzero eigenvalue of the normalized Laplacian for the subgraph that corresponds to the target cluster. See Zhu et al. (2013) for details. Let us denote the eigenvalue with λ. In our case the target cluster is a given ground truth cluster. We use this information to set parameter α. In particular, for each node in the target cluster we run APPR and ℓ_1-reg. PR 4 times where α is set based on a range of values in $[\lambda/2, 2\lambda]$ with a step of $(2\lambda - \lambda/2)/4$. The regularization parameter of ℓ_1-reg. PR has nearly identical purpose as the tolerance parameter of APPR. Both parameters are set to be proportional to inverse of the volume of the target cluster, as suggested by theoretical arguments in this paper as well as in Andersen et al. (2006); Zhu et al. (2013); Fountoulakis et al. (2019). For each parameter setting we use sweep cut to round the output of APPR and ℓ_1-reg. PR to find a cluster of low conductance. The sweep cut rounding procedure is a common technique to post-process the output of local graph clustering methods—for details, we refer the reader to Andersen et al. (2006); Zhu et al. (2013); Fountoulakis et al. (2019). We use a proximal coordinate descent algorithm (Fountoulakis et al., 2019) to solve the ℓ_1-reg. PR problem. Over all parameter settings, we return the cluster with the lowest conductance value as an output of APPR and ℓ_1-reg. PR.

SimpleLocal has only one parameter denoted by δ. The δ parameter controls how localized the output is going to be around the input seed nodes (Veldt et al., 2016), it also controls the quality of the output in terms of its conductance value (Theorem 1 in Orecchia and Zhu (2014)) or in terms of false and true positives (Theorem 3 in Orecchia and Zhu (2014)). The parameter has to satisfy $\delta \geq 0$, and according to Orecchia and Zhu (2014) it should be set such that $\text{Vol}(R)/\text{Vol}(V - R) + \delta = 1/(3/\sigma + 3)$, where R is the set of seed nodes that is given as input to SimpleLocal, and σ is a parameter that satisfies $\sigma \in [\text{Vol}(R)/\text{Vol}(V - R), 1]$. It is suggested in Orecchia and Zhu (2014) to set $\sigma = \mathcal{O}(\text{Vol}(R \cap K)/\text{Vol}(K))$, where K denotes the target cluster. Therefore, in our experiments we set

$$\sigma = \min \left[\max \left(\frac{\text{Vol}(R)}{\text{Vol}(V - R)}, \frac{\text{Vol}(R \cap K)}{\text{Vol}(K)} \right), 1 \right].$$

When we use ℓ_1-reg. PR as input to SimpleLocal, this parameter setting also provides the theoretical guarantees which are claimed in Orecchia and Zhu (2014) (assuming that all assumptions about the target cluster are satisfied).

We set parameter *steps* $= n$ in Algorithm 1 and we terminate Algorithm 1 when the current set of seed nodes, denoted by *seeds*, satisfies $\text{Vol}(\textit{seeds} \cap K)/\text{Vol}(K) \geq 0.75$, i.e., the output of Algorithm 1 has to have at least 75% overlap with the target cluster, or when $\text{Vol}(\textit{seeds})/\text{Vol}(G) \geq 0.25$, i.e., the volume of the output is larger than 25% the volume of the graph. This setting is also motivated by Theorem 3 in Orecchia and Zhu (2014), which mentions that the input to SimpleLocal has 75% overlap with the target cluster. However, since this relies on the assumption that the input set of nodes is the output of APPR or ℓ_1-reg. PR, which are themselves clustering algorithms, while

Algorithm 1 is not, then we also add the second termination criterion that the output cannot have volume more than 25% the volume of the graph.

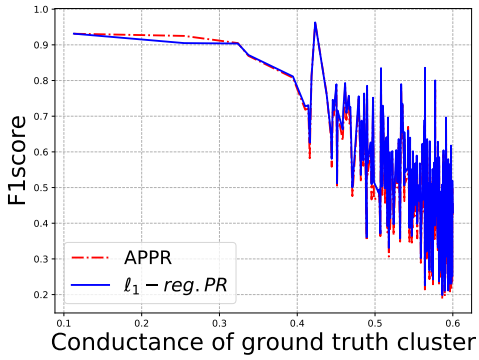
Input : *seeds* - set of seed nodes that we want to expand
Output : *seeds* - updated/expanded set of seeds
Parameters: *steps* - number of steps of the algorithm
Create queue Q;
Set all nodes in *seeds* as visited;
Set $step = 1$;
while $step \leq steps$ **do**
  $k = 1$ and $l =$ size of Q;
  while $k \leq l$ **do**
    remove node u from head of Q;
    mark and enqueue all (unvisited) neighbours of u;
    Add newly visited nodes in set *seeds*;
    increase k;
  end
  increase *step*;
end

Algorithm 1: A modified BFS algorithm for seed set expansion

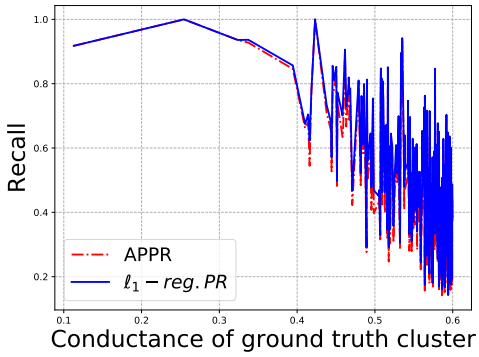
6.2.4 RESULTS

APPR and ℓ_1-reg. PR are similar. We start our empirical observations by comparing APPR and ℓ_1-reg. PR. We already proved in Theorem 10 that these methods are similar. However, we would like to point out some additional details, which will help us justify simplification of the experiments for the remainder of this section. In Figure 4, we make a much more extended empirical comparison between the two local spectral methods, where we compare their precision, recall and F1score. We use the Orkut dataset, and we compare the two methods using all 282 ground truth clusters of this dataset. In this figure, we present average results over all nodes for each given ground truth cluster. We observe that APPR and ℓ_1-reg. PR produce output with nearly identical precision, recall and F1score. We attribute any minor differences between the two methods to the minor differences between the optimality conditions of ℓ_1-reg. PR and APPR. Moreover, APPR and proximal coordinate descent for ℓ_1-reg. PR have similar running time in practice for producing the results in Figure 4. This is justified by worst-case analysis in Fountoulakis et al. (2019) as well as our average-case analysis in Theorem 11. In particular, each method required 75 hours to produce Figure 4, which required 304960 calls to APPR and ℓ_1-reg. PR. Due to the similarity of the two approaches, we will only use ℓ_1-reg. PR in the subsequent experiments. Note that in this experiment we do not show performance of flow-based methods because based on our empirical observations on small datasets like FB-Johns55 and Colgate88 it would have taken close to a month to obtain similar plots like in Figure 4. (Improving this situation for local flow improve methods is an important direction for future work.) For example, see the running times of flow-based methods for FB-Johns55 and Colgate88 in Table 7.

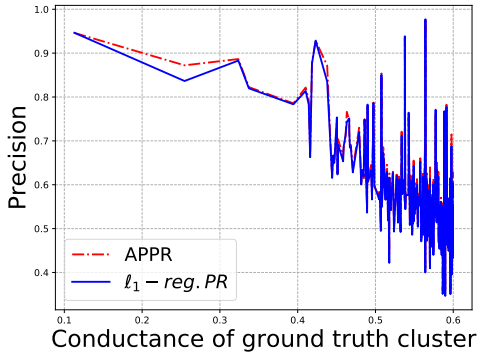
APPR and ℓ_1-reg. PR work well for low conductance target clusters. Here, we comment on the performance of APPR and ℓ_1-reg. PR in Figure 4 on the Orkut dataset. We observe that performance of both methods decreases as the conductance of the target cluster increases, i.e., as the cluster



(a) F1score vs conductance



(b) Recall vs conductance



(c) Precision vs conductance

Figure 4: We demonstrate the similarity of APPR and ℓ_1-reg. PR by illustrating F1score, recall and precision vs conductance of ground truth clusters plots for the Orkut dataset (282 ground truth clusters). This plot demonstrates two properties of the algorithms. First, it demonstrates performance of ℓ_1-reg. PR and APPR as conductance of the target cluster increases (horizontal-axis). Observe that as conductance becomes larger then overall performances decreases. Second, it demonstrates that the methods ℓ_1-reg. PR and APPR get nearly identical results, which is also justified by theoretical arguments. See the main text for details.

quality gets worse. This is an expected outcome that is predicted by the average-case analysis of this paper. However, it is important to mention that, for target clusters with conductance close to 0.4, the F1score of ℓ_1-reg. PR is around 0.8. This is surprisingly good performance since a cluster with conductance 0.4 means that roughly half of the edges of the target cluster cross the cluster boundary (which implies that such target clusters are actually quite noisy and not of particularly high quality).

Results for biology datasets and Sfld and PPI-mips. The results for the biology datasets are shown in Table 4. We present average results over all nodes for each given ground truth cluster. We denote with bold numbers the performance number of a method when it has the *largest F1score* among all methods.

We note the consistent state-of-the-art performance of ℓ_1-reg. PR for all clusters in Table 4. For some ground truth clusters, ℓ_1-regularized PageRank perfectly recovers the target clusters, which is mainly attributed to the fact that the ground truth clusters have strong separability property (see also Corollary 9 of the main text). In most experiments, SimpleLocal did not improve the performance of the input of ℓ_1-reg. PR, but also it did not make it worse. In some cases, like the Spindle ground truth cluster in the PPI-mips dataset, SimpleLocal decreased the performance of ℓ_1-reg. PR in terms of the F1score. This is because SimpleLocal found clusters that have smaller conductance, but that do not correspond to clusters with the highest F1score. This is a known issue that has been mentioned in Fountoulakis et al. (2017). BFS-SL has the worst performance among all methods in most experiments. In fact, BFS-SL performs well only for the usrease ground truth cluster in the Sfld dataset. The performance of BFS-SL is especially poor for all ground truth clusters in the PPI-mips dataset. It is important to mention that we did experiment with different parameter tuning for both BFS-type Algorithm 1 and SimpleLocal for the BFS-SL method, but the performance was poor for all settings of parameters that we tried. We attribute the poor performance of BFS-SL in the BFS-type algorithm, which provides the input to SimpleLocal. In particular, the BFS-type Algorithm 1 is not related to clustering in a general sense, and this translates to poor quality input to SimpleLocal. As is mentioned in the theoretical analysis in Orecchia and Zhu (2014); Veldt et al. (2016), SimpleLocal requires as input the output of a local spectral method such as ℓ_1-reg. PR in order to perform well, which is also verified by the results in Table 4.

The corresponding running times for each method are shown in Table 5. Each numeric value in this table demonstrates the total running time to run a method for a given ground truth cluster. The running time is the sum of the running times for each node in the ground truth cluster. Note that ℓ_1-reg. PR is the fastest method.

Results social networks FB-Johns55 and Colgate88. The results for the social network graphs are shown in Table 6. There are a lot of interesting observation for this set of experiments. First, ℓ_1-reg. PR outperforms BFS-LS, with the exception of the ground truth clusters of major index 217 in FB-Johns55, where BFS-LS has a 0.03 larger F1score, and the ground truth cluster of year 2009 in Colgate88, where BFS-SL has the same F1score as ℓ_1-reg. PR. We observed two reasons that BFS-SL has worse performance in most experiments. The first reason is that BFS-SL outputs a cluster that has smaller conductance than the output cluster of ℓ_1-reg. PR, but better conductance is often not related to the ground truth cluster, especially in cases where the ground truth cluster itself has large conductance. We attribute this behavior to SL because as an algorithm it attempts to find a cluster with small conductance. The second reason is that the input to SL from BFS-type Algorithm 1 is not a good approximation to the ground truth cluster, which is a required property of SL such that it performs well.

Table 4: Results for biology datasets Sfld and PPI-mips. In this table, we present average results of F1score over all nodes for each given ground truth cluster. We denote with bold numbers the performance number of a method when it has the largest score among all methods.

dataset	feature	ℓ_1-reg. PR	BFS-SL	ℓ_1-reg. PR-SL
Sfld	urease	**0.75**	0.42	0.38
	AMP	**0.86**	**0.86**	**0.86**
PPI-mips	Actin	**0.98**	0.04	**0.98**
	Anaphase	**1.00**	0.09	**1.00**
	Cdc28p	**1.00**	0.10	**1.00**
	Coat	**0.85**	0.04	**0.85**
	ct-large	**1.00**	0.02	**1.00**
	ct-small	**1.00**	0.05	**1.00**
	F0-F1-ATP	**1.00**	0.06	**1.00**
	H+	**1.00**	0.06	**1.00**
	mc-large	**1.00**	0.03	**1.00**
	mc-small	**1.00**	0.07	**1.00**
	mc-complex	0.78	0.07	**0.79**
	mRNA	**0.93**	0.03	**0.93**
	Nuclear	**0.85**	0.05	**0.85**
	RNA	**0.87**	0.03	0.80
	Spindle	**0.85**	0.03	0.82
	TRAPP	**1.00**	0.10	**1.00**
	tRNA	**1.00**	0.09	**1.00**
	19-22S	**1.00**	0.05	**1.00**
	20S	**1.00**	0.06	**1.00**

The second set of observations is about ℓ_1-reg. PR-SL. For most ground truth clusters, we observe that ℓ_1-reg. PR-SL performs worse than or on par with ℓ_1-reg. PR, with the exception of ground truth clusters year 2009 and major index 217 in FB-Johns55 and ground truth clusters of years 2008 and 2009 in Colgate88. When ℓ_1-reg. PR-SL makes the input of ℓ_1-reg. PR worse, it is clearly because the former finds a cluster with better conductance value which does not relate to the ground truth cluster. We observe this behavior very often when the target cluster does not have small conductance value, and this is also confirmed through our simulation study (see Figure 3(d) in Section 6.1). When ℓ_1-reg. PR-SL performs better it is because the ground truth cluster has small conductance but not small enough such that ℓ_1-reg. PR performs well by itself. In particular, ℓ_1-reg. PR leaks more mass outside of the target cluster than it should, and this results in small precision. This is a well-known problem that has been also observed in Fountoulakis et al. (2017), which can be fixed by SL.

The corresponding running times for each method are shown in Table 7. The running time is the sum of the running times for each node in the ground truth cluster. Note that ℓ_1-reg. PR is the fastest method.

Table 5: Running times for biology datasets Sfld and PPI-mips. The numeric entries of the table show the total time in seconds over all nodes of a given ground truth cluster. The running times of ℓ_1-reg. PR where less than 0.1 for most experiments, and they have been rounded up to 0.1.

dataset	feature	ℓ_1-reg. PR	BFS-SL	ℓ_1-reg. PR-SL
Sfld	urease	17.2	29.8	151.2
	AMP	0.2	2.5	1.6
PPI-mips	Actin	0.1	1.8	0.6
	Anaphase	0.1	0.8	0.1
	Cdc28p	0.1	0.7	0.1
	Coat	0.1	1.5	0.3
	ct-large	3.6	8.5	16.5
	ct-small	1.3	10.0	6.6
	F0-F1-ATP	0.1	1.0	0.1
	H+	0.1	1.1	0.1
	mc-large	0.3	3.3	2.0
	mc-small	0.1	0.9	0.1
	mc-complex	0.1	0.8	0.2
	mRNA	0.3	3.5	2.4
	Nuclear	0.1	1.3	0.3
	RNA	0.2	1.5	2.7
	Spindle	0.1	2.3	0.7
	TRAPP	0.1	0.7	0.1
	tRNA	0.1	0.7	0.1
	19-22S	0.1	1.3	0.2
	20S	0.1	1.0	0.1

Memory scalability for small- and large-scale graphs. One of the main motivations of the ℓ_1-regularized PageRank model is the low memory requirements for each seed node. In Table 8 we illustrate performance in terms of memory requirements for solving the ℓ_1-regularized PageRank using proximal coordinate descent. In particular, we illustrate how memory requirement increases as ρ decreases. As is expected, for large values of ρ the output solution is very sparse and the algorithms have very low memory requirements. As ρ gets smaller then the solutions become denser and memory requirements increase. We observe similar performance for FB-Johns55 (small dataset) and Orkut (large dataset).

7. Conclusion

In this paper, we have examined the ℓ_1-regularized PageRank optimization problem for local graph clustering. In a local graph clustering problem, the objective is to find a single target cluster in a large graph, given a seed node in the cluster, and to do so in a running time that does not depend on the size of the entire graph, but instead that depends on the size of the output cluster. Algorith-

Table 6: Results for Facebook datasets FB-Johns55 and Colgate88. In this table we present average results of F1score over all nodes for each given ground truth cluster. We denote with bold numbers the performance number of a method when it has the largest score among all methods.

dataset	feature	ℓ_1-reg. PR	BFS-SL	ℓ_1-reg. PR-SL
FB-Johns55	year 2006	**0.32**	0.13	0.23
	year 2007	**0.43**	0.17	0.31
	year 2008	**0.50**	0.34	0.36
	year 2009	0.84	0.78	**0.89**
	major index 217	0.85	0.88	**0.88**
	second major 0	**0.41**	0.20	0.20
	dorm 0	**0.46**	0.13	0.08
	gender 1	**0.42**	0.21	0.21
	gender 2	**0.46**	0.19	0.18
Colgate88	year 2004	0.42	0.29	**0.44**
	year 2005	**0.44**	0.15	0.43
	year 2006	**0.46**	0.27	0.39
	year 2007	**0.54**	0.25	0.46
	year 2008	0.75	0.56	**0.88**
	year 2009	0.96	0.96	**0.98**
	second major 0	**0.49**	0.24	0.25
	dorm 0	**0.46**	0.04	0.26
	gender 1	**0.45**	0.21	0.25
	gender 2	**0.34**	0.20	0.26

mic results (i.e., running-time bounds to achieve a given cluster quality) for local graph clustering abound, but our results are the first statistical results (i.e., where one is interested in recovering a cluster under an hypothesized model). Under our local random model, we show that the optimal support of ℓ_1-regularized PageRank identifies the target cluster with bounded false positives, and in certain settings exact recovery is also possible. We further establish a strong connection between ℓ_1-regularized PageRank and approximate personalized PageRank (APPR), based on which we obtain similar statistical guarantees on APPR under the random model. 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. Finally we demonstrate the state-of-the-art performance of ℓ_1-regularized PageRank on both simulated and real data graphs.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (NSF grant IIS-9988642) and the Multidisciplinary Research Program of the Department of Defense

Table 7: Running times for Facebook datasets FB-Johns55 and Colgate88. The numeric entries of the table show the total time in seconds over all nodes of a given ground truth cluster.

dataset	feature	ℓ_1-reg. PR	BFS-SL	ℓ_1-reg. PR-SL
FB-Johns55	year 2006	307	28126	56022
	year 2007	453	28083	61462
	year 2008	1125	25855	62668
	year 2009	852	15736	15090
	major index 217	80	728	768
	second major 0	8866	91500	308041
	dorm 0	10865	70084	260485
	gender 1	1533	69265	205031
	gender 2	13080	83836	301843
Colgate88	year 2004	11	1482	659
	year 2005	70	5800	7679
	year 2006	99	5471	11398
	year 2007	154	6198	11641
	year 2008	287	4756	6900
	year 2009	530	2882	2507
	second major 0	2876	24525	71937
	dorm 0	449	15184	36992
	gender 1	1533	20347	53016
	gender 2	430	16854	39537

(MURI N00014-00-1-0637). This material is also based in part on research sponsored by DARPA and the Air Force Research Laboratory under agreement number FA8750-17-2-0122. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA and the Air Force Research Laboratory or the U.S. Government. We would also like to acknowledge ARO, NSF, ONR, and Berkeley Institute for Data Science for providing partial support of this work.

Table 8: Memory scalability for proximal coordinate descent for solving the ℓ_1-regularized PageRank as ρ decreases. We illustrate averaged results over 100 trials, i.e., seed nodes. The numbers in the first row of each dataset are Megabytes. Note that to compute the memory requirements for storing the adjacency matrix we did not store the edge weights since the graphs are unweighted.

dataset	statistics	ρ =1.0e-1	1.0e-2	1.0e-3	1.0e-4	1.0e-5	1.0e-6	1.0e-7
Orkut	Average memory	0.11	0.20	0.27	1.00	3.51	22.8	100
	$\frac{\text{Average memory}}{\text{mem. for adjacency}}$	1.2e-4	2.1e-4	2.9e-4	1.0e-3	3.6e-3	2.4e-2	1.0e-1
	Average nodes touched	72	72	131	1474	10788	83139	443857
	Average nodes in solution	1	1	4	33	266	2401	22159
	Number of iterations	1	2	10	104	1050	11370	132153
FB-Johns55	Average memory	0.09	0.17	0.20	0.75	1.44	2.41	2.65
	$\frac{\text{Average memory}}{\text{mem. for adjacency}}$	6.3e-2	1.1e-1	1.3e-1	4.9e-1	9.5e-1	1.5e+0	1.7e+0
	Average nodes touched	66	68	114	959	3113	5079	5157
	Average nodes in solution	1	1	3	36	367	4201	5157
	Number of iterations	1	3	10	107	1746	41820	539493

Appendix A. Some auxiliary lemmas

To establish the theorems, we need a few concentration inequalities and intermediate results that shall be used in the proof. 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 population graph. Section A.3 gives a few important results on the ℓ_1-regularized PageRank when restricted to the target cluster.

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 (2018, Theorem 2.3.1)).

Lemma 13 (Adapted from Vershynin (2018, 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 $1 - 2e^{-c_0\delta^2\mu}$,*

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

According to our random model (Definition 5), 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 13, it is straightforward to see the following result.

Lemma 14 (Adapted from Vershynin (2018, 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 $1 - 2e^{-c_0\delta^2\bar{d}}$,*

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

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 15 *Suppose that $q = c/n$ and that $k \geq 2(c+3)$ 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 of target cluster in population graph

Next, we define the "ground truth" PageRank vector, which we obtain by applying ℓ_1-regularized PageRank to the population graph $\overline{G}$ of our random model. Recall the adjacency matrix $\mathbb{E}[A]$ defined in (3), 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 population version of ℓ_1-regularized PageRank as

$$x^\star = \arg\min_x \left\{ \frac{1}{2} x^\top \mathbb{E}[Q]\, x - \alpha x^\top \mathbf{s} + \rho\alpha \|\mathbb{E}[D]\, x\|_1 \right\}. \tag{11}$$

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

The following lemma shows that there exist a range of ρ values for which the solution to the population version of the ℓ_1-regularized PageRank minimization problem (11) gives an exact recovery for the target cluster.

Lemma 16 (Exact recovery for population PageRank vector.) *Consider the local random model given in Definition 5 and suppose that Assumption 1 holds. Then, for n sufficiently large, the "ground truth" PageRank vector $x^\star$, defined in* (11)*, identifies K correctly, i.e.,*

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

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 $\mathbb{E}\left[d_{\ell_\sharp}\right] = \min_{\ell\in K^c}\mathbb{E}\left[d_\ell\right]$. 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

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

From the closed form expression, we can see that $x^\star$ has high probability mass on the single 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 population graph is unknown, and we are instead given an instance of the graph from the random model. In this case, the ground truth vector $x^\star$ allows us to estimate the magnitude of the ℓ_1-regularized PageRank vector $\widehat{x}$ by analyzing the error between $x^\star$ and $\widehat{x}$. For more details, we refer the reader to the proof of Lemma 19.

A.3 ℓ_1-regularized PageRank restricted to target cluster

Consider the ℓ_1-regularized PageRank problem restricted to the target cluster K.

$$\begin{aligned} \widehat{x}^{\mathrm{R}} &= \operatorname*{arg\,min}_{x\in\mathbb{R}^n}\left\{\frac{1}{2}x^\top Qx - \alpha x^\top\mathbf{s} + \rho\alpha\|Dx\|_1 : x_{K^c} = 0\right\} \\ &= \operatorname*{arg\,min}_{y\in\mathbb{R}^k}\left\{\frac{1}{2}y^\top Q_{K,K}y - \alpha y^\top\mathbf{s}_K + \rho\alpha\|D_{K,K}y\|_1\right\}. \end{aligned} \tag{12}$$

Since our aim is to recover the target cluster K based on the *local* model (Definition 5), it is natural 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 (12).

Lemma 17 *Let $\widehat{x}^{\mathrm{R}}$ be the solution to the reduced problem* (12)*. Suppose that $(1-\delta)p^2k \geq c_0^{-1}\delta^{-2}\log k$. Then, if the regularization parameter satisfies $\rho \leq \rho(\delta)$, we have*

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

with probability at least $1-6e^{-c_0\delta^2(1-\delta)p^2k}$. Here c_0 is the same constant appearing in Lemma 13.

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

Lemma 18 *For any regularization parameter $\rho \in [0,\infty)$, we have*

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

Finally we have the following estimate on the maximum coordinate of $\widehat{x}^{\mathrm{R}}$ on $K\backslash S$ (recall u and v are respectively the components of $x^\star$ on S and $K\backslash S$; see Lemma 16).

Lemma 19 *Let $\delta \leq 0.1$ and suppose that $(1-\delta)p^2k \geq c_0^{-1}\delta^{-2}\log k$. Assume that Assumption 1 holds and that the size of the target cluster $k \geq 5$. If the regularization parameter satisfies $\rho \geq \rho(\delta)$, then for n sufficiently large, the following bound holds*

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

with probability at least $1-6e^{-c_0\delta^2(1-\delta)p^2k}$. Here c_0 is the same constant appearing in Lemma 13 and c_1 is a positive numerical constant.

Appendix B. Proofs of theorems

In this section we prove all of our theorems. Based on the lemmas presented above we give proofs of our five theorems, respectively, in Section B.1, B.2, B.3, B.4, and B.5.

B.1 Proof of Theorem 6

The proof of Theorem 6 is a straightforward combination of Lemma 17 and 18. Specifically, by Lemma 17, we know that for $\rho \leq \rho(\delta)$, we have that $\mathrm{support}(\widehat{x}^{\mathrm{R}}) = K$. Then applying Lemma 18 we have

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

thus proving the result.

B.2 Proof of Theorem 7

To prove Theorem 7, we first need the following lemma for bounding the volume of $\mathrm{support}(\widehat{x})$. This lemma is a stronger version of Fountoulakis and Gondzio (2015, Theorem 2).

Lemma 20 *For any regularization parameter $\rho > 0$, it holds that*

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

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

Now by our choice of $\rho \geq \rho(\delta)$ (5) we have

$$\begin{aligned}\mathrm{Vol}(\mathrm{support}(\widehat{x})) \leq \frac{1}{\rho} \leq \frac{1}{\rho(\delta)} &= \left(\frac{1+\alpha}{1-\alpha}\right)^2 \left(\frac{1+\delta}{1-\delta}\right)^2 \frac{(1+\delta)\bar{d}^2}{\gamma p} \\ &= \left(\frac{1+\alpha}{1-\alpha}\right)^2 \left(\frac{1+\delta}{1-\delta}\right)^2 \frac{(1+\delta)k\bar{d}}{\gamma^2}, \end{aligned} \tag{13}$$

where the second step follows from (4). Furthermore, by Theorem 6, $\mathrm{support}(\widehat{x})$ 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

$$\mathrm{Vol}(\mathrm{support}(\widehat{x})) = \mathrm{Vol}(K) + \mathrm{Vol}(\mathrm{FP}). \tag{14}$$

Since $\mathrm{Vol}(K) = \sum_{i \in K} d_i \geq (1-\delta)k\bar{d}$ by Lemma 14, it follows that

$$\begin{aligned}\mathrm{Vol}(\mathrm{FP}) &= \mathrm{Vol}(\mathrm{support}(\widehat{x})) - \mathrm{Vol}(K) \quad \text{by (14)} \\ &\leq \left(\frac{1+\alpha}{1-\alpha}\right)^2 \left(\frac{1+\delta}{1-\delta}\right)^2 \frac{(1+\delta)k\bar{d}}{\gamma^2} - \mathrm{Vol}(K) \quad \text{by (13)} \\ &\leq \mathrm{Vol}(K) \left[\left(\frac{1+\alpha}{1-\alpha}\right)^2 \left(\frac{1+\delta}{1-\delta}\right)^3 \frac{1}{\gamma^2} - 1 \right], \end{aligned}$$

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

B.3 Proof of Theorem 8

Under Assumption 1, with nonvanishing probability we can find a good seed node in the target cluster that is connected solely to K.

Lemma 21 *Suppose that Assumption 1 holds, i.e., $q = \frac{c}{n}$ for a fixed constant $c > 0$. Under the local random model given in Definition 5, if n is sufficiently large, then*

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

We select this node as a seed node in the ℓ_1-regularized PageRank problem.

Next we show that under the conditions of Theorem 8, the optimal solution to the reduced problem, $\widehat{x}^{\mathrm{R}} = \widehat{x}^{\mathrm{R}}(\rho(\delta))$ in (12), obeys

$$|(Q\widehat{x}^{\mathrm{R}} - \alpha \mathbf{s})_j| = |-\frac{1-\alpha}{2} \cdot A_{j,K}\widehat{x}^{\mathrm{R}}| < \rho(\delta)\alpha d_j \text{ for } j \in K^c. \tag{15}$$

Note that the above inequality is simply the optimality condition for the full-dimensional problem on K^c (see Lemma 3). Since, by the optimality conditions for the reduced problem, $\widehat{x}^{\mathrm{R}}$ also satisfies the

full-dimensional optimality condition on K. By uniqueness of the solution this means that $\widehat{x}^{\mathrm{R}} = \widehat{x}$. Also, since $\mathrm{support}(\widehat{x}^{\mathrm{R}}) \subseteq K$ by construction, this proves that there is no false positive in $\widehat{x}$.

Now 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 21 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

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

Applying Lemma 15, 19, and using the fact that $u \le \frac{2\alpha}{(1+\alpha)\bar{d}}, v \le \frac{(1-\alpha)p}{(1+\alpha)\bar{d}^2}$ (which can be deduced from Lemma 16), and that $\delta \le 0.1$ while $\alpha \in [0.1, 0.9]$ (by assumptions of the theorem), then

$$\begin{aligned}\left|(Q\widehat{x}^{\mathrm{R}} - \alpha\mathbf{s})_j\right| \le \frac{1-\alpha}{2}\|A_{j,K\setminus S}\|_1\|\widehat{x}^{\mathrm{R}}_{K\setminus S}\|_\infty &\le (0.5c+1)\left[\frac{c_1(1-\alpha)u}{\bar{d}} + (1+2c_1\delta)v + 2c_1\delta\cdot\rho(\delta)\alpha\right] \\ &\le (0.5c+1)C_1\left(\frac{2}{\bar{d}^2} + \rho(\delta)\alpha\right),\end{aligned}$$

where $C_1 > 0$ is a positive constant. Then, to prove (15), it suffices to show

$$(0.5c+1)C_1\left[\frac{2}{\rho(\delta)\alpha\bar{d}^2} + 1\right] < d_j.$$

Plugging in the definition of $\rho(\delta)$ (5) to the above, and using $\alpha \in [0.1, 0.9]$ and $\delta \le 0.1$, we obtain

$$\begin{aligned}(0.5c+1)C_1\left[\frac{2}{\rho(\delta)\alpha\bar{d}^2} + 1\right] &= (0.5c+1)C_1\left[\left(\frac{1+\alpha}{1-\alpha}\right)^2\left(\frac{1+\delta}{1-\delta}\right)^2\frac{1+\delta}{\alpha\gamma p} + 1\right] \\ &\le \frac{(0.5c+1)C_2}{\gamma p},\end{aligned}$$

for some constant $C_2 > 0$. By the assumption (8) this means that $\rho(\delta)^{-1}\alpha^{-1}|(Q\widehat{x}^{\mathrm{R}} - \alpha\mathbf{s})_j| < d_j$, and thus we have proved the claim (15).

B.4 Proof of Theorem 10

Fix $\rho_0 > 0$. First, we prove the first part of the theorem, i.e., $\mathrm{support}(\widehat{x}(\rho_0)) \subseteq \mathrm{support}(x^{\mathrm{APPR}}(\rho_0))$. When $\rho_0 > 1/d_S$, this relation holds trivially because both $\widehat{x}(\rho_0)$ and $x^{\mathrm{APPR}}(\rho_0)$ do not contain any nodes in the support set. To prove that the relation holds for $\rho_0 \le 1/d_S$, we will proceed by induction. Specifically, we will prove that

$$\begin{cases}\widehat{x}_i(\rho_0) < x_i^{\mathrm{APPR}}(\rho_0), & i \in \mathrm{support}(x_i^{\mathrm{APPR}}(\rho_0)), \\ \widehat{x}_i(\rho_0) = x_i^{\mathrm{APPR}}(\rho_0) = 0, & i \notin \mathrm{support}(x_i^{\mathrm{APPR}}(\rho_0)),\end{cases} \tag{16}$$

for all $\rho_0 \in (0, 1/d_S]$. Assuming that this holds, it is straightforward to follow that

$$\mathrm{support}(\widehat{x}(\rho_0)) \subseteq \mathrm{support}(x^{\mathrm{APPR}}(\rho_0)),$$

thus proving the claim.

Now we turn to proving (16). When $\rho_0 = 1/d_S$, we can see from the algorithm (9) that $x^{\text{APPR}}(\rho_0)$ contains the seed node as the only element in the support set, while by Lemma 3, we have $\widehat{x} = 0$, which proves (16). Next, assuming that it holds at some $\rho_0 \in (0, 1/d_S]$, let $\widetilde{\rho}_0 > 0$ be chosen such that $\widetilde{\rho}_0 < \rho_0$ and such that there appears a node $i \in V$ that violates the condition (16) for the first time. If $i \in \text{support}(x^{\text{APPR}}(\widetilde{\rho}_0))$, since the path $\widehat{x}(\rho)$ is continuous by the proof of Lemma 4, this means that i is the first node such that $\widehat{x}_i(\widetilde{\rho}_0) = x_i^{\text{APPR}}(\widetilde{\rho}_0) > 0$ (there may exist multiple nodes that simultaneously violate (16) at $\rho = \widetilde{\rho}_0$ but the same statement still applies). In this case, by induction, we know that $\widehat{x}_j(\widetilde{\rho}_0) \le x_j^{\text{APPR}}(\widetilde{\rho}_0)$ for $j \neq i$. Then

$$\begin{aligned}(D^{-1}\nabla f(\widehat{x}(\widetilde{\rho}_0)))_i &= \frac{1+\alpha}{2}\widehat{x}_i(\widetilde{\rho}_0) - \frac{1-\alpha}{2}(D^{-1}A\widehat{x}(\widetilde{\rho}_0))_i - \alpha \mathbf{s}_i \\ &\ge \frac{1+\alpha}{2}x_i^{\text{APPR}}(\widetilde{\rho}_0) - \frac{1-\alpha}{2}(D^{-1}Ax^{\text{APPR}}(\widetilde{\rho}_0))_i - \alpha \mathbf{s}_i \\ &= (D^{-1}\nabla f(x^{\text{APPR}}(\widetilde{\rho}_0)))_i,\end{aligned}$$

where the inequality holds since $(D^{-1}Ax)_i = \sum_{j\sim i} w_{ji}x_j/d_i$ for any $x \in \mathbb{R}^{|V|}$. By the optimality condition of (2) (Lemma 3), the left-hand side must be $(D^{-1}\nabla f(\widehat{x}(\widetilde{\rho}_0)))_i = -\widetilde{\rho}_0\alpha$ which gives

$$(D^{-1}\nabla f(x^{\text{APPR}}(\widetilde{\rho}_0)))_i \le -\widetilde{\rho}_0\alpha.$$

However, this is a contradiction to the termination criterion of the APPR algorithm (9), and in particular, we must have $i \notin \text{support}(x^{\text{APPR}}(\widetilde{\rho}_0))$ and $\widehat{x}_i(\widetilde{\rho}_0) > 0$ by the condition (16). Then, by Lemma 3, this implies that $(D^{-1}\nabla f(\widehat{x}(\widetilde{\rho}_0)))_i = -\widetilde{\rho}_0\alpha$, and following the same steps as above we can see that $-\widetilde{\rho}_0\alpha > (D^{-1}\nabla f(x^{\text{APPR}}(\widetilde{\rho}_0)))_i$. However, this again contradicts the termination criterion of APPR, thus proving the first part of the theorem.

Next, we prove the second part of the theorem, i.e., $\text{support}(x^{\text{APPR}}(\rho_0)) \subseteq \text{support}(\widehat{x}(\rho_1))$ for $\rho_1 = (1-\alpha)\cdot\rho_0/2$. We first claim that for every node $i \in \text{support}(x^{\text{APPR}}(\rho_0))$, we must have $-\rho\alpha d_i < \nabla_i f(x^{\text{APPR}}(\rho_0)) \le -\frac{1-\alpha}{2}\rho\alpha d_i$. The first inequality is obvious from the termination criterion of APPR. To see why the second inequality holds, if a node $i \in V$ is selected at iteration k of the APPR algorithm, then by the update step (9), we have $x_i^{(k+1)} = x_i^{(k)} - d_i^{-1}\nabla_i f(x^{(k)})$. Then, the gradient of $f(x^{(k)})$ on node i is updated as

$$\begin{aligned}\nabla_i f(x^{(k+1)}) &= \left[\left(\frac{1+\alpha}{2}D - \frac{1-\alpha}{2}A\right)x^{(k+1)}\right]_i - \alpha\mathbf{s}_i \\ &= \frac{1+\alpha}{2}d_i x_i^{(k)} - \frac{1+\alpha}{2}\nabla_i f(x^{(k)}) - \frac{1-\alpha}{2}\sum_{j\sim i} w_{ji}x_j^{(k)} - \alpha\mathbf{s}_i \\ &= \nabla_i f(x^{(k)}) - \frac{1+\alpha}{2}\nabla_i f(x^{(k)}) = \frac{1-\alpha}{2}\nabla_i f(x^{(k)}).\end{aligned} \tag{17}$$

By the termination criterion of APPR, we know that $\nabla_i f(x^{(k)}) \le -\rho_0\alpha d_i$ for all $i \in V$. This shows that the right-hand side of the above equation is $\le -\rho_0\alpha d_i$, and in particular, taking $k \to \infty$, we obtain $\nabla_i f(x^{\text{APPR}}(\rho_0)) \le -\frac{1-\alpha}{2}\rho_0\alpha d_i$, which proves the claim.

Now consider the following strategy of continuously adding small mass to the vector $x^{\text{APPR}}(\rho_0)$:

1. Start with $x^{\text{APPR}}(\rho_0)$. Find the node i whose gradient is smallest, i.e.,

$$j_1 \in \arg\min_{i\in V} \nabla_i f(x^{\text{APPR}}(\rho_0)).$$

2. Add mass to the node j_1 until some other node $j_2 \in V$ has the gradient as small as j_1. This event always happens because by the work above (17) we can see that the gradient on node j_1 is increasing while the gradients on the other nodes are either decreasing (if it is a neighboring node of j_1) or stay the same.

3. Add mass to the nodes (j_1, j_2) such that gradients on j_1 and j_2 are increasing at the same rate,[5] and until some other node $j_3 \in V$ has the gradient as small as j_1 and j_2.

4. Continue this step until the gradient on the node j_1 becomes $-\frac{1-\alpha}{2}\rho_0\alpha d_i$.

Writing $\widetilde{x}^{\text{APPR}}$ to denote the output vector of the above procedure, it is easy to see that $\widetilde{x}^{\text{APPR}}$ indeed satisfies the optimality condition of the ℓ_1-regularized PageRank minimization problem, given in Lemma 3, with $\rho = \rho_1$. This in turn shows that $\text{support}(x^{\text{APPR}}) \subseteq \text{support}(\widehat{x}(\rho_1))$, since by the uniqueness of the optimal solution we have $\widetilde{x}^{\text{APPR}} = \widehat{x}(\rho_1)$. This completes the proof of Theorem 10.

B.5 Proof of Theorem 11

First, by Theorem 10, we know that

$$\text{support}(\widehat{x}(\rho(\delta))) \subseteq \text{support}(x^{\text{APPR}}(\rho(\delta))) \subseteq \text{support}(\widehat{x}((1-\alpha)\cdot\rho(\delta)/2)).$$

Since $\widehat{x}(\rho(\delta))$ contains the target cluster by Theorem 6, therefore together with the relation above we obtain $K \subseteq \text{support}(x^{\text{APPR}})$. To see that the false positives in x^{APPR} are also bounded, we can follow the same step as the proof of Theorem 7 (see Section B.2) to show that

$$\text{Vol}(\text{FP}(\widehat{x}((1-\alpha)\cdot\rho(\delta)/2))) \leq \text{Vol}(K)\left[\left(\frac{1+\alpha}{1-\alpha}\right)^2\left(\frac{1+\delta}{1-\delta}\right)^3\frac{2}{(1-\alpha)\gamma^2} - 1\right].$$

From the relation between $\text{support}(x^{\text{APPR}}(\rho(\delta)))$ and $\text{support}(\widehat{x}((1-\alpha)\cdot\rho(\delta)/2))$ it directly follows that

$$\text{Vol}(\text{FP}(x^{\text{APPR}})) \leq \text{Vol}(K)\left[\left(\frac{1+\alpha}{1-\alpha}\right)^2\left(\frac{1+\delta}{1-\delta}\right)^3\frac{2}{(1-\alpha)\gamma^2} - 1\right].$$

Finally to prove the exact recovery case, it suffices to show that $\text{support}(\widehat{x}((1-\alpha)\cdot\rho(\delta)/2))$ contains no false positives. In order for this all we need to do is to replace $\rho(\delta)$ by $\rho((1-\alpha)/2\cdot\delta)$ in the proof of Theorem 8. Then the same proof steps go through and we finally get the condition

$$\frac{2C(0.5c+1)}{(1-\alpha)\gamma p} = \mathcal{O}\left(\frac{1}{\gamma p}\right) < d_j,$$

in place of (8). This completes the proof of the theorem.

Appendix C. Proofs of lemmas

We prove all of our lemmas in this section.

5. This is always possible since for any positive $\epsilon \in \mathbb{R}^{|V|}$, we know that $\sum_{i\in V}(\nabla_i f(x+\epsilon) - \nabla_i f(x)) > 0$. So, in order to increase the gradients at the same rate, we need to find a direction ϵ such that the summands $\nabla_j f(x+\epsilon) - \nabla_j f(x)$ are same for nodes j that are selected.

C.1 Proof of Lemma 2

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,

$$(Q\widehat{x})_j - \alpha \mathbf{s}_j = \begin{cases} -\rho\alpha d_j, & x_j > 0, \\ \rho\alpha d_j, & x_j < 0, \\ [-\rho\alpha d_j, \rho\alpha d_j], & x_j = 0, \end{cases} \tag{18}$$

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

$$\begin{aligned}(Q\widehat{x})_j &= \frac{(1+\alpha)}{2}\left(d_j\widehat{x}_j - \frac{1-\alpha}{1+\alpha}\cdot\sum_{i:(i,j)\in E}\widehat{x}_i\right) \\ &= \frac{(1+\alpha)}{2}\left(\sum_{i:(i,j)\in E}\widehat{x}_j - \frac{1-\alpha}{1+\alpha}\cdot\sum_{i:(i,j)\in E}\widehat{x}_i\right). \end{aligned} \tag{19}$$

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 the condition (18), it must be that $(Q\widehat{x})_{j^\star} > 0$. However by (19) 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 node in the graph. This proves the desired result.

C.2 Proof of Lemma 3

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

C.3 Proof of Lemma 4

First, by Rosset and Zhu (2007, Proposition 1), we know that the optimal solution path $\widehat{x}(\rho)$ for the ℓ_1-regularized minimization (2) is piecewise linear as a function of $\rho > 0$. 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. This, together with the non-negativity of the solution (Lemma 2) 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 Lemma 3, 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),

$$\widehat{x}_{\mathcal{A}}(\rho_1) - \widehat{x}_{\mathcal{A}}(\rho_0) = \alpha(\rho_0 - \rho_1) Q_{\mathcal{A},\mathcal{A}}^{-1} d_{\mathcal{A}}. \tag{20}$$

Now write

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

For all $z \in \mathbb{R}^{|\mathcal{A}|}$, we have that

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

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 $(D^{-1/2} A D^{-1/2})_{\mathcal{A},\mathcal{A}}$ have absolute value less than 1. Using the previous fact, $Q_{\mathcal{A},\mathcal{A}}^{-1}$ can be written as

$$\begin{aligned} Q_{\mathcal{A},\mathcal{A}}^{-1} &= 2(1+\alpha)^{-1} \cdot D_{\mathcal{A},\mathcal{A}}^{-1/2} \left[I - \frac{1-\alpha}{1+\alpha} \cdot (D^{-1/2} A D^{-1/2})_{\mathcal{A},\mathcal{A}} \right]^{-1} D_{\mathcal{A},\mathcal{A}}^{-1/2} \\ &= 2(1+\alpha)^{-1} \cdot D_{\mathcal{A},\mathcal{A}}^{-1/2} \left[I + \frac{1-\alpha}{1+\alpha} \cdot (D^{-1/2} A D^{-1/2})_{\mathcal{A},\mathcal{A}} + \left(\frac{1-\alpha}{1+\alpha} \right)^2 \cdot (D^{-1/2} A D^{-1/2})_{\mathcal{A},\mathcal{A}}^2 \right. \\ &\qquad \left. + \cdots \right] D_{\mathcal{A},\mathcal{A}}^{-1/2}. \end{aligned} \tag{21}$$

Now it is easy to see that the right hand side in (20) is positive, i.e., $\alpha(\rho_0 - \rho_1) Q_{\mathcal{A},\mathcal{A}}^{-1} d_{\mathcal{A}} > 0$. This completes the proof of Lemma 4.

C.4 Proof of Lemma 13

This lemma is proved in Vershynin (2018, Proposition 2.4.1) which we reproduce here for completeness. By Chernoff's inequality ((Vershynin, 2018, Theorem 2.3.1)), for some constant $c_0 > 0$,

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

Using union bound,

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

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

C.5 Proof of Lemma 14

This lemma follows directly from Lemma 13.

C.6 Proof of Lemma 15

We have $\|A_{j,K}\|_1 \sim \text{Binom}(k, q)$ for $j \in K^c$, so using tail bound for Binomial distribution ((Arratia and Gordon, 1989)), for any $t > c$,

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

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

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

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

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

as long as $n \gg k$.

C.7 Proof of Lemma 16

First the results in Lemma 2 and Lemma 3 hold for any graph, seed vector $\mathbf{s}$, and $\rho > 0$, so the result can also be applied to (11). In particular, the solution $x^\star$ is non-negative, and thus the optimality condition can be expressed as

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

(Note $\mathbf{s}$ is 1 on the seed node and 0 on the rest.) For $\rho > \frac{1}{d}$ the optimal solution (11) 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 $j \in S$,

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

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

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

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

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

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:

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

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

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

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

Also, 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 $p\mathbb{E}\left[d_{\ell_\sharp}\right] = \mathcal{O}(1)$ 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).

C.8 Proof of Lemma 17

Define

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

We will prove that $\check{x}^{\mathrm{R}} = \check{x}^{\mathrm{R}}(\rho) > 0$ for any $\rho \leq \rho(\delta)$ where $\rho(\delta)$ is defined in (5). In particular, this means that $\check{x}^{\mathrm{R}}$ satisfies the optimality condition for the reduced problem (12), implying that $\widehat{x}_K^{\mathrm{R}} = \check{x}^{\mathrm{R}} > 0$ and thus proving the result.

First, using (21), we can write

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

Let $w = D_{K,K}^{-1}\mathbf{s}_K$. Then

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

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 14, then with probability at least $1 - 2e^{-c_0\delta^2\bar{d}}$,

$$D_{K,K}^{-1}A_{K,K}w \geq \frac{1}{(1+\delta)^2\bar{d}^2}\mathbf{1}_{\mathrm{FON}}. \tag{27}$$

Next, using Lemma 13, we have that $|\text{FON}| \geq (1-\delta)pk$ with probability at least $1-2e^{-c_0\delta^2 pk}$. Furthermore, for each non-seed node $i \notin \text{FON}$ in K, it is connected to nodes in FON with probability p, independently of all other pairs. Then $A_{i,K}^\top \mathbf{1}_{\text{FON}} \mid |\text{FON}| \overset{\text{iid}}{\sim} \text{Binom}(|\text{FON}|, p)$, and thus applying Lemma 13, and using the assumption $(1-\delta)p^2k \geq c_0^{-1}\delta^{-2}\log k$, we get

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

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

$$A_{i,K}^\top \mathbf{1}_{\text{FON}} \geq (1-\delta)^2 p^2 k \text{ for all } i \notin \text{FON}. \tag{28}$$

Then,

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

where the first step applies (27), the second step uses (28), and the third applies Lemma 14 and the fact that $p \cdot k = \gamma\bar{d}$. Returning to (26), we can now see that $\breve{x}^{\text{R}} > 0$ as long as ρ is less than $\rho(\delta) = \left(\frac{1-\alpha}{1+\alpha}\right)^2 \left(\frac{1-\delta}{1+\delta}\right)^2 \frac{\gamma p}{(1+\delta)\bar{d}^2}$. This completes the proof of Lemma 17.

C.9 Proof of Lemma 18

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 (21), 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:

$$-\frac{1-\alpha}{2}\sum_{j\sim i, j\in\mathcal{A}_2} w_{ij}\widehat{x}_j = -\rho\alpha d_i. \tag{29}$$

Analogously, the node i becomes active for the reduced problem (12) when the following condition is satisfied:

$$-\frac{1-\alpha}{2}\sum_{j\sim i, j\in\mathcal{A}_1} w_{ij}\widehat{x}^{\text{R}}_j = -\rho\alpha d_i. \tag{30}$$

Comparing the left-hand sides in (29) and (30), it is obvious that under the induction assumption, the left-hand side in (29) is larger than in (30). 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.

C.10 Proof of Lemma 19

Since $\widehat{x}^{\mathrm{R}} = (\widehat{x}^{\mathrm{R}}_S, \widehat{x}^{\mathrm{R}}_{K\setminus S})$ is the solution to the reduced problem (12), fixing $\widehat{x}^{\mathrm{R}}_S$, then $\widehat{x}^{\mathrm{R}}_{K\setminus S}$ is the minimizer of the following optimization problem:

$$\widehat{x}^{\mathrm{R}}_{K\setminus S} = \operatorname*{arg\,min}_{y\in\mathbb{R}^{k-1}} \left\{ \frac{1}{2}(\widehat{x}^{\mathrm{R}}_S, y)^\top Q_{K,K}(\widehat{x}^{\mathrm{R}}_S, y) + \rho\alpha\|D\cdot(\widehat{x}^{\mathrm{R}}_S, y)\|_1 \right\}. \tag{31}$$

Due to Lemma 17, the subgradient of $\|D\cdot(\widehat{x}^{\mathrm{R}}_S, y)\|_1$ at $y = \widehat{x}^{\mathrm{R}}_{K\setminus S}$ is given by $d_{K\setminus S}$. Using optimality, it follows that

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

Next, we can prove that our choice of $\rho = \rho(\delta)$ lies in the interval $[\rho^\sharp, \rho^\natural)$.

Lemma 22 *Under the conditions of Lemma 19, we have $\rho = \rho(\delta) \in (\rho^\sharp, \rho^\natural)$.*

Hence, by Lemma 16, the population version of ℓ_1-regularized PageRank (11) has support K, so it follows that $x^\star$ is also the solution to the reduced problem

$$x^\star = \min_x \left\{ \frac{1}{2}x^\top \mathbb{E}\left[Q_{K,K}\right] x - \alpha x^\top \mathbf{s}_K + \rho\alpha\|\mathbb{E}\left[D_{K,K}\right] x\|_1 : x_{K^c} = 0 \right\}. \tag{32}$$

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

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

where we have $\partial\|\mathbb{E}\left[D\right]\cdot(x^\star_S, y)\|_1 = \mathbb{E}\left[d_{K\setminus S}\right]$ due to Lemma 16 and Lemma 22. 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 $Q_{K\setminus S,K}(x^\star - \widehat{x}^{\mathrm{R}}) = Q_{K\setminus S,S}(x^\star_S - \widehat{x}^{\mathrm{R}}_S) + Q_{K\setminus S,K\setminus S}(x^\star_{K\setminus S} - \widehat{x}^{\mathrm{R}}_{K\setminus S})$, and rearranging terms, we get

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

Taking the ℓ_∞ norm on both sides,

$$\begin{aligned} \|Q_{K\setminus S,K\setminus S}(\widehat{x}^{\mathrm{R}}_{K\setminus S} - x^\star_{K\setminus S})\|_\infty \le \underbrace{\|(\mathbb{E}\left[Q_{K\setminus S,K}\right] - Q_{K\setminus S,K})x^\star\|_\infty}_{\text{term 1}} + \underbrace{\rho\alpha\|\mathbb{E}[d_{K\setminus S}] - d_{K\setminus S}\|_\infty}_{\text{term 2}} \\ + \underbrace{\|Q_{K\setminus S,S}(\widehat{x}^{\mathrm{R}}_S - x^\star_S)\|_\infty}_{\text{term 3}}. \end{aligned} \tag{33}$$

For term 1 we know that $x^\star = u\mathbf{1}_S + v\mathbf{1}_K$ from Lemma 16, so plugging in to term 1,

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

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 13 and Lemma 14, and the last step holds since $|A_{j,S} - p| \le 1$ for all $j \in K\setminus S$. Furthermore, by Lemma 14, 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

$$\|Q_{K\setminus S,K\setminus S}(\widehat{x}^{\mathrm{R}}_{K\setminus S} - x^\star_{K\setminus S})\|_\infty \le \frac{1-\alpha}{2}u + \delta\bar{d} \cdot v + \rho\alpha \cdot \delta\bar{d} + \frac{1-\alpha}{2}\left|\widehat{x}^{\mathrm{R}}_S - x^\star_S\right|. \tag{34}$$

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

$$\|Q_{S,S}(\widehat{x}^{\mathrm{R}}_S - x^\star_S)\|_\infty \le \underbrace{\|(\mathbb{E}\left[Q_{S,K}\right] - Q_{S,K})x^\star\|_\infty}_{\text{term 4}} + \underbrace{\rho\alpha\|\mathbb{E}[d_S] - d_S\|_\infty}_{\text{term 5}} + \underbrace{\|Q_{S,K\setminus S}(\widehat{x}^{\mathrm{R}}_{K\setminus S} - x^\star_{K\setminus S})\|_\infty}_{\text{term 6}}. \tag{35}$$

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 13 and Lemma 14. It is also easy to see that term 5 $\le \rho\alpha \cdot \delta\bar{d}$, while

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

where the second step uses Hölder's inequality and the next step applies Lemma 13. 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 (35), we get

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

Finally, return to (34) and substitute the above bound in place of $\|\widehat{x}^{\mathrm{R}}_S - x^\star_S\|_\infty$, then

$$\begin{aligned}\|Q_{K\setminus S,K\setminus S}(\widehat{x}^{\mathrm{R}}_{K\setminus S} - x^\star_{K\setminus S})\|_\infty &\leq \frac{1-\alpha}{2}u + \delta\bar{d}\cdot v + \rho\alpha\cdot\delta\bar{d} \\ &\quad + \left(\frac{1-\alpha}{1+\alpha}\right)\frac{\delta\bar{d}(u+v) + \rho\alpha\cdot\delta\bar{d} + \frac{1-\alpha}{2}(1+\delta)p(k-1)\|\widehat{x}^{\mathrm{R}}_{K\setminus S} - x^\star_{K\setminus S}\|_\infty}{(1-\delta)\bar{d}}. \quad (36)\end{aligned}$$

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

Lemma 23 *Under the conditions of Lemma 19, with probability at least* $1 - 2e^{-c_0\delta^2\bar{d}}$,

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

Applying Lemma 23 to the left-hand side of (36), and rearranging terms and simplifying, we have that

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

where we use $\alpha < 1$ and $p(k-1) = \gamma\bar{d}$. By assumption of the lemma, we have that $\delta \leq 0.1$ and that $\alpha \in [0.1, 0.9]$ while $\gamma \in (0,1)$ by definition (4), so

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

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

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

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

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

Using the triangle inequality, Lemma 19 now follows.

C.11 Proof of Lemma 20

By Lemma 3 we know that

$$(Q\widehat{x} - \alpha\mathbf{s})_i \begin{cases} = -\rho\alpha d_i, & \widehat{x}_i > 0, \\ \leq 0, & \widehat{x}_i = 0. \end{cases}$$

Summing over i's, we have

$$\begin{aligned} \sum_i (Q\widehat{x} - \alpha\mathbf{s})_i &= \alpha \sum_i d_i \widehat{x}_i - \alpha \\ &\leq -\rho\alpha \sum_{j:\widehat{x}_j > 0} d_j = -\rho\alpha \cdot \mathrm{Vol}(\mathrm{support}(\widehat{x})), \end{aligned}$$

where the first equality holds since $\mathbf{1}^\top Q\widehat{x} = \mathbf{1}^\top(\alpha D + (1-\alpha)/2 \cdot L)\widehat{x} = \alpha \sum_i d_i \widehat{x}_i$. Dividing both sides by $-\rho\alpha$, we get

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

C.12 Proof of Lemma 21

This result follows from simple probability calculation. We have

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

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

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

C.13 Proof of Lemma 22

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 $\rho = \rho(\delta) = \frac{\bar{c}(1-\alpha)p}{(1+\alpha)\bar{d}^2}$. Following the same steps as (25) (proof of Lemma 16), we can see

$$\begin{aligned} \rho > \rho^\sharp \iff & 2\bar{c}p\alpha \cdot \bar{d} \cdot \mathbb{E}\left[d_{\ell_\sharp}\right] + \bar{c}pq(1-\alpha) \cdot k\bar{d} + \bar{c}pq(1-\alpha)(n-k)\mathbb{E}\left[d_{\ell_\sharp}\right] > q(1+\alpha)\bar{d}^2 \\ \iff & 2\alpha\bar{d}(\bar{c}p\mathbb{E}\left[d_{\ell_\sharp}\right] - q\bar{d}) + (1-\alpha)q(n-k)(\bar{c}p\mathbb{E}\left[d_{\ell_\sharp}\right] - q\bar{d}) \\ & \qquad - pq(1-\alpha)(k-1)\bar{d}(1-\bar{c}) > 0. \end{aligned} \tag{37}$$

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 that $\gamma = \mathcal{O}(1)$, 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 (37) 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, we equivalently need to show $2\alpha(1+\alpha)\bar{d} > (1-\alpha)^2 p$. 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 $k \geq 5$.

C.14 Proof of Lemma 23

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 14),

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

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

References

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

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.

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.

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.

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

Reid Andersen and Kevin J Lang. An algorithm for improving graph partitions. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 651–660, 2008.

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

Reid Andersen, Shayan Oveis Gharan, Yuval Peres, and Luca Trevisan. Almost optimal local graph clustering using evolving sets. *Journal of the ACM (JACM)*, 63(2):15, 2016.

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

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.

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

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.

Shoshana D Brown, John A Gerlt, Jennifer L Seffernick, and Patricia C Babbitt. A gold standard set of mechanistically diverse enzyme superfamilies. *Genome biology*, 7(1):R8, 2006.

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.

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

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Kimon Fountoulakis and Jacek Gondzio. Performance of first-and second-order methods for big data optimization. *arXiv preprint arXiv:1503.03520*, 2015.

Kimon Fountoulakis, David F Gleich, and Michael W Mahoney. An optimization approach to locally-biased graph algorithms. *Proceedings of the IEEE*, 105(2):256–272, 2017.

Kimon Fountoulakis, Farbod Roosta-Khorasani, Julian Shun, Xiang Cheng, and Michael W Mahoney. Variational perspective on local graph clustering. *Mathematical Programming*, 174(1-2): 553–573, 2019.

Kimon Fountoulakis, Meng Liu, David F Gleich, and MW Michael. Flow-based algorithms for improving clusters: A unifying framework, software, and performance. *arXiv preprint arXiv:2004.09608*, 2020a.

Kimon Fountoulakis, Di Wang, and Shenghao Yang. p-norm flow diffusion for local graph clustering. In *International Conference on Machine Learning*, pages 3222–3232. PMLR, 2020b.

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

David Gleich and Michael Mahoney. Anti-differentiating approximation algorithms: A case study with min-cuts, spectral, and flow. In *International Conference on Machine Learning*, pages 1018–1025. PMLR, 2014.

David F Gleich and Kyle Kloster. Seeded PageRank solution paths. *European Journal of Applied Mathematics*, 27(6):812–845, 2016.

Alden Green, Sivaraman Balakrishnan, and Ryan J Tibshirani. Local spectral clustering of density upper level sets. *arXiv preprint arXiv:1911.09714*, 2019.

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.

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.

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

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

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

Kyle Kloster and David F Gleich. Heat kernel based community detection. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1386–1395. ACM, 2014.

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

Kevin Lang and Satish Rao. A flow-based method for improving the expansion or conductance of graph cuts. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 325–337. Springer, 2004.

Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

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

Jure Leskovec, Kevin J Lang, and Michael Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on World wide web*, pages 631–640. ACM, 2010.

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.

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.

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

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.

Lorenzo Orecchia and Zeyuan Allen Zhu. Flow-based algorithms for local graph clustering. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1267–1286. SIAM, 2014.

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

Philipp Pagel, Stefan Kovac, Matthias Oesterheld, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Pekka Mark, Volker Stümpflen, Hans-Werner Mewes, et al. The MIPS mammalian protein–protein interaction database. *Bioinformatics*, 21(6):832–834, 2004.

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

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

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.

Pan Shi, Kun He, David Bindel, and John E Hopcroft. Local Lanczos spectral approximation for community detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 651–667. Springer, 2017.

Julian Shun, Farbod Roosta-Khorasani, Kimon Fountoulakis, and Michael W Mahoney. Parallel local graph clustering. *Proceedings of the VLDB Endowment*, 9(12):1041–1052, 2016.

Daniel A Spielman and Shang-Hua 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.

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

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

Amanda L Traud, Eric D Kelsic, Peter J Mucha, and Mason A Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, 53(3):526–543, 2011.

Amanda L Traud, Peter J Mucha, and Mason A Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.

Nate Veldt, David Gleich, and Michael Mahoney. A simple and strongly-local flow-based method for cut improvement. In *International Conference on Machine Learning*, pages 1938–1947, 2016.

Nate Veldt, Christine Klymko, and David F Gleich. Flow-based local graph clustering with better seed set inclusion. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 378–386. SIAM, 2019.

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

Di Wang, Kimon Fountoulakis, Monika Henzinger, Michael W Mahoney, and Satish Rao. Capacity releasing diffusion for speed and locality. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3598–3607. JMLR. org, 2017.

Hao Yin, Austin R Benson, Jure Leskovec, and David F Gleich. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 555–564. ACM, 2017.

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

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

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

Zeyuan Allen Zhu, Silvio Lattanzi, and Vahab S Mirrokni. A local algorithm for finding well-connected clusters. In *Proceedings of the 30th International Conference on Machine Learning*, pages 396–404, 2013.

Hui Zou, Trevor Hastie, and Robert Tibshirani. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.