Statistical guarantees for local graph clustering

Wooseok Ha
UC Berkeley

Kimon Fountoulakis
University of Waterloo

Michael W. Mahoney
ICSI and UC Berkeley

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 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.

1 Introduction

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

In this paper, we consider the so-called ℓ_1*-regularized PageRank algorithm* [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. This model 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.

Literature review 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, in-

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

cluding with random walks [Andersen et al., 2006], 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 [Wang et al., 2017].

There are also numerous papers in statistics on partitioning random graphs. Arguably, the stochastic block model (SBM) is the most commonly employed random model for graph partitioning [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]. 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].

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

2 Background on ℓ_1-regularized PageRank

PageRank [Page et al., 1999, Brin and Page, 1998] is a popular approach for ranking the nodes of a graph. It is defined as the stationary distribution of a Markov chain, which is encoded by a convex combination of the input distribution $\mathbf{s} \in \mathbb{R}^n$ and the (lazy) random walk operator W:

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

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

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. 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. Motivated by this, Fountoulakis et al. [2019] recently proposed the *ℓ_1-regularized PageRank optimization problem*. Unlike APPR, the solution method for the ℓ_1-regularized PageRank optimization problem is purely optimization-based. It uses an ℓ_1 norm regularization to set automatically to be zero nodes dissimilar to the seed node, thereby resulting in a highly sparse output. In this manner, ℓ_1-regularized PageRank can estimate the personalized ranking, while maintaining the most relevant nodes at the same time. Prior work [Fountoulakis et al., 2019] also showed that proximal gradient descent (ISTA) can solve the ℓ_1-regularized PageRank problem, with access to only a small portion of the entire graph, i.e., without even touching the entire graph, thereby allowing the method to easily scale to very large-scale graphs.

In this paper, we investigate the statistical performance of ℓ_1-regularized PageRank by reformulating the local graph clustering into the problem of sparse recovery. Here is a more precise definition of the ℓ_1-regularized PageRank optimization problem from [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 [Fountoulakis et al., 2019] on the graph is defined as

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

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

To see the intuition behind (2), observe that if we set $\rho = 0$ and $\widehat{p}^{\mathrm{PR}} = D\widehat{x}$, then we can see that it recovers the original PageRank solution of (1). In other words, the optimization

problem in (2) adds an additional ℓ_1 norm regularization to the quadratic objective of the linear system (1) and change variables to $x = D^{-1}p^{\text{PR}}$. The fact that the output cluster of ℓ_1-regularized PageRank is given by the optimization solution allows us to analyze statistical properties under random graph model more easily.

Properties. Here, we state some properties of ℓ_1-regularized PageRank that will be useful for our analysis. First, the following lemma guarantees that the ℓ_1-regularized PageRank vector is non-negative.

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

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

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

$$\nabla_i f(\widehat{x}) = (Q\widehat{x})_i - \alpha \mathbf{s}_i$$

$$\begin{cases} = -\rho\alpha d_i & \widehat{x}_i > 0, \\ \in [-\rho\alpha d_i, 0] & \widehat{x}_i = 0,\ i \textit{ is a neighbor of nonzero node}, \\ = 0 & \textit{otherwise.} \end{cases}$$

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

3 Statistical guarantees under random model

In this section, we introduce a random model that we consider for generating a target cluster, and then we provide recovery guarantees for ℓ_1-regularized PageRank. The details of the proofs are referred to our longer version of the present work [Ha et al., 2019].

3.1 Random graph model

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

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

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

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

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 2 allows us to formulate the problem of local graph clustering as the recovery of a target cluster. Since we are interested in recovering a single target cluster, it is natural to make assumptions only for nodes in the target cluster and nodes adjacent to the target cluster, and to leave the interactions between other nodes unspecified.

The employed random model is also fairly general and 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 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 $i \in K$, 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 any distribution. Conductance measures the weight of the edges that are being removed over the volume of the cluster— formally it is defined as the ratio $\text{Cut}(S, S^c)/\min(\text{Vol}(S), \text{Vol}(S^c))$, where $\text{Cut}(S, S^c) := \sum_{i\in S, j\in S^c} A_{ij}$. From Definition 2, the conductance of the target cluster of the expected graph $\bar{G}$ is given by

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

Here γ can be viewed as the ratio of the random walker

staying inside K under the expected 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 $\overline{\text{Cond}}$, or equivalently the number γ, will play a crucial in determining the behavior of ℓ_1-regularized PageRank under the local graph model. In particular, note 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 recover. With this definition, we can also write $p(k-1) = \gamma\bar{d}$ and $q(n-k) = (1-\gamma)\bar{d}$.

3.2 Recovery of target cluster with bounded false positives

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

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

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

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

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

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

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

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

$$\rho = \left(\frac{1-\alpha}{1+\alpha}\right)^2 \left(\frac{1-\delta}{1+\delta}\right)^2 \frac{\gamma p}{(1+\delta)\bar{d}^2}, \tag{6}$$

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

$$\text{Vol}(FP) \leq \text{Vol}(K)\big[\underbrace{\left(\frac{1+\alpha}{1-\alpha}\right)^2 \left(\frac{1+\delta}{1-\delta}\right)^3 \frac{1}{\gamma^2} - 1}_{=\mathcal{O}\left(\frac{1}{\gamma^2}\right)-1}\big], \tag{7}$$

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

The above results, Theorem 1 and Theorem 2, show several regimes where ℓ_1-regularized PageRank can fully recover the target cluster with nonvanishing probability. In particular, when $p = \mathcal{O}(1)$, the size of the target cluster, k, is required to be larger than $\mathcal{O}(\log k)$, which includes the constant size $k = \mathcal{O}(1)$. This is often the regime of interest for local graph clustering, where the goal is to find small- and meso-scale clusters in massive graphs [Leskovec et al., 2009, 2010]. In addition, Theorem 1 indicates that if γ is small, then we need to set ρ to be small to recover the entire cluster. Intuitively, more mass will leak out to K^c for small γ, so we need to run more steps of random walk (ρ smaller in our optimization framework) to find the right cluster. However, this means that the ℓ_1-regularized PageRank vector will also pick up many nonzero nodes in K^c, resulting in many false positives in the support set. Indeed, Theorem 2 shows that the volume of false positives grows quadratically as $1/\gamma$, so we need γ to be bounded to get a meaningful recovery from local clustering. In the case of $p = \mathcal{O}(1), k = \mathcal{O}(1)$, this amounts to requiring that $q = \mathcal{O}\left(\frac{1}{n}\right)$ in order for the target cluster to keep high mass inside K.

Several other comments are worth making regarding these results. First, the current bound we obtain in (7) may not be tight with respect to α and other constants, and the factor $(\frac{1+\alpha}{1-\alpha})^2$ may be an artifact of our proof. 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, based on our empirical results, ℓ_1-regularized PageRank performs well across a broad range of α values, and we have not seen much difference in terms of performance among different α's. The role of α in ℓ_1-regularized PageRank is closely tied to the regularization parameter ρ, and we leave the question of selecting optimal α for future work.

3.3 Exact recovery of target cluster with no false positives

Next, we study the scenarios under which ℓ_1-regularized PageRank can exhibit a stronger recovery guarantee. Specifically, under some additional conditions, we show that the support set of the optimal solution (2) identifies

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

[2] The same probability bound as Theorem 1.

the target cluster exactly, without making any false positives. For this stronger exact recovery result, we require the following assumption about the parameters of the model.

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

As we noted above, the setting $k = \mathcal{O}(1)$ is often the case of interest for local graph clustering, where we would like to identify small- and medium-scale structure in large graphs [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 statements of exact recovery guarantees for ℓ_1-regularized PageRank when applied to the noisy graph generated from Definition 2. In particular, the fact that $q = \mathcal{O}\left(n^{-1}\right)$ 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 3 (No false positives). *Suppose the same conditions as Theorem 1, and assume also that Assumptions 1 holds. If $k \geq 2(c+3)$, $\alpha \in [0.1, 0.9]$, $\delta \geq 0.1$, and*

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

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

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

as long as

$$\frac{C(0.5c+1)}{\gamma p} = \mathcal{O}\left(\frac{1}{\gamma p}\right) < d_j, \tag{9}$$

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

We require $\alpha \in [0.1, 0.9]$ and $\delta \leq 0.1$ in the condition (8) to avoid overly complicated constants; while this simplifies the statements of the theorem, it is not difficult to show that a similar result holds more generally. While Theorem 3 guarantees no false positives in the solution of ℓ_1-regularized PageRank, when combined with Theorem 1, it immediately establishes that ℓ_1-regularized PageRank recovers the target cluster exactly, even when the target cluster is constant-sized.

Corollary 1 (Exact recovery). *Under the same assumptions as Theorem 1 and Theorem 3, there is a good starting node in K such that ℓ_1-regularized PageRank parameterized with that node as a seed node satisfies*

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

with nonvanishing probability.

Some sort of condition like (9) about the realized degree seems necessary in order that the ℓ_1-regularized PageRank has no false positives. The optimization program (2) assigns less weights to low degree nodes in the ℓ_1 penalty, so any nodes adjacent to K will become active unless the ℓ_1-regularized PageRank penalizes them with nontrivial weights. Unlike Theorem 1 and Theorem 2, condition (9) rules out some specific models to which Theorem 3 can be applied. For example, planted clustering model with $p = \mathcal{O}(1)$ and $q = \mathcal{O}(1/n)$ does not satisfy this condition because the degrees in K^c do not concentrate. For the stochastic block model, this condition is still satisfied if nodes adjacent to the target cluster belong to the clusters with degree larger than $\mathcal{O}(1/\gamma p)$. In practice, condition (9) 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.

3.4 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 random graph model with the current known state-of-the-art worst-case results, given by [Zhu et al., 2013]. First, [Zhu et al., 2013, Theorem 1], when applied to our "expected" graph, implies that $\text{Vol(FP)}, \text{Vol(FN)} \leq \text{Vol}(K) \cdot \mathcal{O}\left((1-\gamma)\log k\right)$, as long as $\text{Gap} = \mathcal{O}\left(1/((1-\gamma)\log k)\right) \geq \mathcal{O}(1)$. When $\gamma = \mathcal{O}(1) \in (0,1)$, our Theorem 1 states that if $pk^2 \geq \mathcal{O}(\log k)$, the output of the algorithm does not contain any false negative, which cannot be deduced from Zhu et al. [2013]. In addition, our general bound on false positive, i.e., $\text{Vol(FP)} \leq \text{Vol}(K) \cdot \mathcal{O}\left(1/\gamma^2 - 1\right)$ in Theorem 2, is better than the worst-case bound of Zhu et al. [2013]'s result in the regime of large γ which is of many practical interest; for instance, when the expected target conductance $\overline{\text{Cond}} = 1 - \gamma$ 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

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

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 3 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 exactly (see, for instance, Section 4.2).

4 Numerical experiments

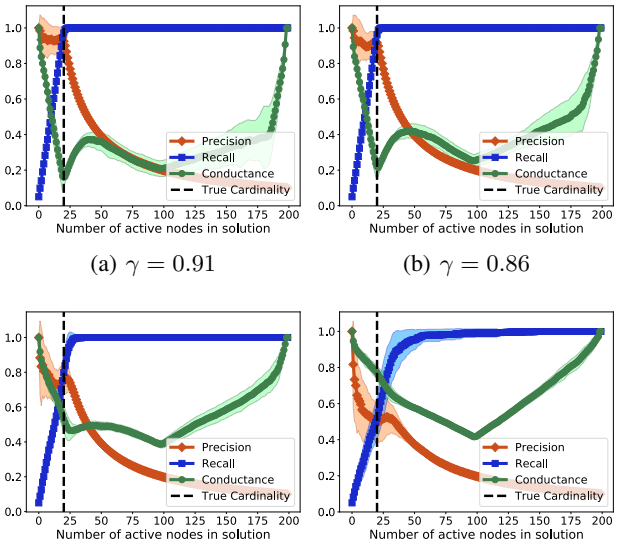
We now illustrate the performance of ℓ_1-regularized PageRank on synthetic and real data. To measure the quality of the recovered cluster, we define Precision and Recall as $\mathrm{Vol}(\mathrm{TP})/\mathrm{Vol}(\mathrm{support}(\widehat{x}))$ and $\mathrm{Vol}(\mathrm{TP})/\mathrm{Vol}(K)$ respectively. The F1 score 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. For our experiments, we solve problem (2) using a proximal coordinate descent algorithm, which enjoys the locality property (running time depends on the size of cluster rather than the entire graph) and linear convergence [Fountoulakis et al., 2019].

4.1 Simulated data

First we run a series of simulations to examine the ability of ℓ_1-reg. PR to recover the target cluster. More precisely, we show that for large and medium values of γ there exists a parameter ρ such that the support of the ℓ_1-reg. PR solution recovers the target cluster. While for small $\gamma < 0.5$, ℓ_1-reg. PR does not recover the target cluster with high accuracy.

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 number of nodes and only one of which is the target cluster K. We use the same parameters p and q across different clusters to generate edges within and between clusters. Here we set $p = 0.5$ and q is varying in order to generate various γ as is shown in Figure 1. We use four settings of γ, one that is favorable, one that is not, and two in-between those two cases. For each experiment, we solve (2) over a range of ρ's (i.e., to obtain the ℓ_1-regularized PageRank solution path) and the results are averaged over 30 trials.

For large γ, Figure 1(a), we observe that when ℓ_1-reg. PR 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



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

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

Figure 1: In Figures 1(a), 1(b) and 1(c) we illustrate that for large and medium values of γ, ℓ_1-reg. PR recovers the target cluster. While for small γ in Figure 1(d), ℓ_1-reg. PR does not recover the target cluster with high accuracy. The x-axis gives the number of nonzero nodes in the solution of ℓ_1-reg. PR as the parameter ρ decreases. The vertical line indicates the cardinality of the target cluster ($k = 20$).

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 1(b) and 1(c) that ℓ_1-regularized PageRank with minimum conductance still finds the target cluster K with good accuracy if the algorithm is terminated early. Finally, in Figure 1(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 ℓ_1-reg. PR that recovers the target cluster accurately.

4.2 Real data

In this section we apply ℓ_1-reg. PR to biology networks and social networks. All the real graphs that are used come with a number of ground truth clusters and we compare the performance of ℓ_1-reg. PR with state-of-the-art local graph clustering algorithms to recover the clusters. For details on the experimental analysis, including the description of datasets, parameter tuning, additional results and running times, we refer the reader to [Ha et al., 2019, Section 5].

4.2.1 Datasets

We apply local graph clustering to five different biology and social networks. The dataset *Sfld* contains pairwise similarities of blasted sequences of 232 proteins belonging to the amidohydrolase superfamily [Brown et al., 2006]. There are 232 nodes and 15570 edges in this graph. *PPI-mips* is a protein-protein interaction graph of mammalian species [Pagel et al., 2004]. There are 1096 nodes and 26442 edges in this graph. *FB-Johns55* and *Colgate88* are Facebook anonymized datasets on a particular day in September 2005 for a student social network at John Hopkins university and Colgate university [Traud et al., 2011, 2012]. These graphs have 5157 nodes and 186572 edges and 3482 nodes and 155043 edges respectively. *Orkut* is a free on-line social network where users form friendship each other. This dataset has 3072441 nodes and 117185083 edges [Leskovec and Krevl, 2014].

4.2.2 Baseline methods

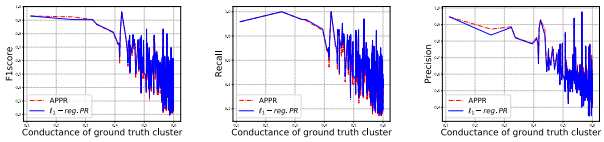
We consider various baseline methods to perform and compare local graph clustering on the real data graphs.

APPR: [Andersen et al., 2006] proposes an Approximate Personalized PageRank algorithm, where the personalized PageRank linear system (1) is solved approximately using a local diffusion process. Strong theoretical guarantees of the algorithm based on conductance measure are presented in [Andersen et al., 2006, Zhu et al., 2013].

SimpleLocal (SL): For local flow-based methods, we consider the one proposed in [Veldt et al., 2016] (SimpleLocal) which simplifies and generalizes the methods in [Lang and Rao, 2004, Andersen and Lang, 2008, Orecchia and Zhu, 2014], while having similar theoretical and practical guarantees in terms of quality of the output. Since flow-based methods require initial input from some other stand-alone method, we consider two initialization techniques: ℓ_1-reg. PR-SL is SL using the output of ℓ_1-regularized PageRank as input; BFS-SL initializes SL 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 SL is shown in Algorithm 1 in [Ha et al., 2019], which has also been used in [Veldt et al., 2016].

4.2.3 Experimental results

APPR and ℓ_1-reg. PR are similar. In [Fountoulakis et al., 2019], it is shown that ℓ_1-regularized PageRank can be viewed as a variational version of APPR (See also [Ha et al., 2019, Theorem 5] for a concrete result). Here we present a comprehensive empirical evidence by comparing APPR and ℓ_1-regularized PageRank on the Orkut dataset (282 ground truth clusters). Specifically we compare their precision, recall, and F1score. Figure 2 shows average results over all nodes for each given ground truth cluster. We can see that APPR and ℓ_1-regularized PageRank produce output with nearly identical precision, recall and F1score. Any minor differences of the two methods are attributed to the minor differences between the termination criteria of ℓ_1-regularized PageRank and APPR. In addition we observe that performance of both methods decreases as the conductance of the target cluster increases. This is an expected outcome that is predicted by our theoretical result. Since APPR and ℓ_1-reg. PR are nearly identical, we only use ℓ_1-regularized PageRank in the subsequent experiments.



(a) F1score vs conductance (b) Recall vs conductance (c) Precision vs conductance

Figure 2: Results of APPR to ℓ_1-reg. PR illustrating F1score, recall and precision vs conductance of ground truth clusters plots for the Orkut dataset (282 ground truth clusters). This plot demonstrates that the methods ℓ_1-reg. PR and APPR get nearly identical results. We also observe that as conductance becomes larger then overall performance decreases.

Results for biology networks (Sfld and PPI-mips). The results for the biology datasets are shown in Table 1 (see [Ha et al., 2019, Section 5.2.4] for additional results). 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.

We note the consistent state-of-the-art performance of ℓ_1-regularized PageRank for all clusters in Table 1. 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 1). In most experiments, SimpleLocal did not improve the performance of the input of ℓ_1-regularized PageRank, 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-regularized PageRank in terms of the F1score. This is because SimpleLocal found clusters that have smaller conductance but 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 AMP 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 and SimpleLocal for the BFS-SL method, but the performance was poor for all settings of parameters that we tried. We charge the poor performance of BFS-SL in the BFS-type algorithm, which provides the input to SimpleLocal. In particular, the BFS-type Algorithm 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 Orecchia and Zhu [2014], Veldt et al. [2016], SimpleLocal requires as input the output of a local spectral method such as ℓ_1-regularized PageRank in order to perform well, which is also verified by the results in Table 1.

Table 1: 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	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**
	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

Results for social networks (FB-Johns55 and Colgate88). The results for the social network graphs are shown in Table 2 (see [Ha et al., 2019, Section 5.2.4] for additional results). There are a lot of interesting observations for this set of experiments. First, ℓ_1-regularized PageRank 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-regularized PageRank. 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-regularized PageRank, but better conductance is often not related to the ground truth cluster, especially in cases that the ground truth cluster itself has large conductance. We charge 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 is not a good approximation to the ground truth cluster, which is a required property of SL such that it performs well.

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 or on par to ℓ_1-regularized PageRank, with the exception of clusters year 2009 and major index 217 in FB-Johns55 and 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 often when the ground truth target cluster does not have small conductance value and this is also confirmed through our simulation study (see Figure 1(d) in Section 4.1). When ℓ_1-reg. PR-SL performs better it is because the target cluster has small conductance but not small enough such that ℓ_1-regularized PageRank performs well by itself. In particular, ℓ_1-regularized PageRank 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.

Table 2: 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

5 Conclusion

We have examined the ℓ_1-regularized PageRank optimization problem for local graph clustering, where the objective is to find a single target cluster given a seed node in the cluster. 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 demonstrate the state-of-the-art performance of ℓ_1-regularized PageRank on real data graphs.

Bibliography

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.

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.

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.

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.

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.

Wooseok Ha, Kimon Fountoulakis, and Michael W Mahoney. Statistical guarantees for local graph clustering. *arXiv preprint arXiv:1906.04863*, 2019.

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.

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.

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.

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.

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.