# Mean Field Analysis of Personalized PageRank with Implications for Local Graph Clustering

Konstantin Avrachenkov, Arun Kadavankandy, Nelly Litvak

# Mean Field Analysis of Personalized PageRank with Implications for Local Graph Clustering

Konstantin Avrachenkov* Arun Kadavankandy† Nelly Litvak‡

**Abstract**

We analyse a mean-field model of Personalized PageRank on the Erdős-Rényi random graph containing a denser planted Erdős-Rényi subgraph. We investigate the regimes where the values of Personalized PageRank concentrate around the mean-field value. We also study the optimization of the damping factor, the only parameter in Personalized PageRank. Our theoretical results help to understand the applicability of Personalized PageRank and its limitations for local graph clustering.

**Keywords:** Personalized PageRank; Mean Field; Concentration; Local Graph Clustering

## 1 Introduction

Personalized PageRank (PPR) can be described as a random walk on a weighted graph. With probability $\alpha$ the random walk chooses one of the neighbour nodes, with probabilities proportional to the edge weights, and with the complementary probability $1-\alpha$ the random walk restarts from a set of seed nodes [20]. The analytical study of Personalized PageRank is challenging because of its non-local nature. Interestingly enough, it is easier to analyse Personalized PageRank on directed graphs. In [6], the expected values of the standard PageRank [29] with uniform restart and Personalized PageRank have been analyzed for directed preferential attachment graphs. In [32] a stochastic recursive equation has been derived for the Personalized PageRank on directed configuration model. This equation has been thoroughly analyzed in [11, 22] and in the works mentioned therein.

On the other hand, the analysis of Personalized PageRank on undirected random graph models is more difficult because a simple random walk on an undirected graph can pass through an edge in both directions, thus creating many short cycles and loops. To the best of our knowledge, [5] is the only work studying Personalized PageRank on undirected Erdős-Rényi (ER) random graphs and stochastic block models. For the analysis of [5] to hold, the personalization vector or the restart distribution has to be sufficiently delocalized. In [13] a mean-field model for the standard PageRank has been proposed without a formal justification. In the recent work [26] a mean-field model has been proposed for a modification of Personalized PageRank where the contributions from all paths are same. The authors of [26] have carried out their analysis in dense stochastic block models when the edge probabilities are fixed, i.e., they do not scale with the size of the graph.

In the present work we analyze Personalized PageRank with a localized restart distribution. As a graph model, we consider an ER random graph with a smaller denser ER graph planted within. We establish conditions for concentration and non-concentration of PPR under different

*Inria Sophia Antipolis, France
†CentraleSupélec, France
‡University of Twente, The Netherlands

scaling laws of the edge probabilities. In particular, we show that when the graph is not too sparse there is a concentration to the mean field model of PPR when the size of subgraph scales linearly and the number of seeds scales sufficiently fast with the graph size. In other words, we establish sufficient conditions for the convergence of PPR to its mean-field form in medium dense graphs. In addition, we show that these conditions are also necessary in a class of sparse graphs with tree-like local structure; i.e., if the number of seed nodes is too small, PPR does not concentrate in a class of sparse graphs. We also show that when there is concentration, the values of PPR can be well approximated by a simple mean-field model and this model can be used for instance for the optimal choice of the damping factor.

An ER graph with a planted denser ER subgraph is the simplest model of a random graph with heterogeneity. It is also a good benchmark model for testing various local graph clustering algorithms. Local graph clustering algorithms are gaining importance since often in practice one would like to recover one particular cluster of a graph using as a guide a few representative seed nodes. One of the first efficient local clustering algorithms is the Nibble algorithm [30, 31] with quasi-linear complexity. The Nibble algorithm is truncation based approximation of a few steps of a lazy random walk. In [2, 3] a modification of the Nibble algorithm using Approximate Personalized PageRank (APPR) has been proposed and evaluated. APPR has lighter computational complexity than the Nibble algorithm. In [15] it has been shown that APPR can be obtained as a solution of an optimization problem with $l_1$-regularization, which ensures sparsity in APPR elements. Both Nibble and APPR try to keep the probability mass localized on few important elements. Recently, in [4] a further improvement to the Nibble algorithm has been proposed based on the technique of the evolving sets.

Our results imply that one needs a significant number of seed nodes to obtain a high quality local cluster. If there are only a few seed nodes, both PPR and APPR suffer from non-concentration. Specifically, the main reason for the non-concentration of PPR and APPR is the significant leakage of probability mass via the seed nodes' neighbours which are outside of the target community.

The methods in [30, 31, 2, 3, 4] aim to find a local cut with target conductance. However, in [10] and [34] significant limitations of the random walks and PPR based local clustering methods are presented in terms of graph conductance and related quantities. As a by-product of our analysis, in Subsection 6.3 we show that the natural cluster in our random graph model also does not really correspond to the problem of conductance minimization. This observation can be viewed as complementary to the results in [10] and [34].

We would like to note that in [18] semidefinite relaxation is used to recover a hidden subgraph without seed nodes and in [24] the belief propagation based algorithm is used to recover a subgraph with seed nodes. These two methods appear to be superior to Personalized PageRank based methods on the considered random graph models but require graph parameters as an input. It is interesting to observe that in the semi-supervised clustering [1, 33] the detectability transition disappears when any linear fraction of seed nodes is introduced.

This paper is organized as follows. In the following section we formally define the random graph model and describe the mean field approximation of Personalized PageRank. In Section 3 we show that there is a concentration in the mean field model of PPR when the size of subgraph and the number of seeds scale linearly with the graph size. However, as demonstrated in Section 4, if the number of seed nodes is too small, there is no concentration. Then, in Section 5 using the mean field model we provide a recommendation for setting the restart probability. Section 6 concludes the technical part of the paper with numerical illustrations and discussions about possible limitations of the PPR and APPR based local graph clustering methods. Finally, in Section 7 we recall our main results and outline promising avenues for further research.

## 2 Graph model and notations

In this section we introduce the model and notations. The notations are summarised in Table 3 at the end of the paper.

We consider an ER random graph $G(n,p)$ with a planted ER subgraph $G(m,q)$. We are interested in the case when the planted ER subgraph is denser than the background ER graph, i.e., when $q > p$. Without loss of generality, we assume that the indices of the subgraph nodes coincide with the first $m$ indices of the background graph $G(n,p)$. Denote this set of vertices corresponding to the planted subgraph by $\mathcal{C}$. At the moment we do not specify any scaling for $m$, $p$ and $q$ and shall discuss various scalings whenever it is needed. We denote by $A = \{a_{ij}\}$ the adjacency matrix of the resulting graph, i.e.,

$$a_{ij} = \begin{cases} 1, & \text{if } i \text{ is a neighbour of } j, \\ 0, & \text{otherwise.} \end{cases}$$

Denote by $\mathbf{1}_n$ the column vector of ones of dimension $n$ and by $\mathbf{J}_{m,n}$ the matrix of ones of dimensions $m$-by-$n$. Also denote by $\mathbf{0}_n$ the column vector of zeros of dimension $n$. Let $d = A\mathbf{1}_n$ be the vector of nodes' degrees and $D = diag\{d\}$ is the diagonal matrix of nodes' degrees. Then, $P = D^{-1}A$ is the transition probability matrix of the standard random walk when the walker chooses the next node to visit uniformly among the neighbours of the current node. Let $k$ nodes of the planted subgraph be disclosed to us. Of course, $k \le m$. Again without the loss of generality, we can assume that these $k$ seed nodes correspond to the first $k$ nodes of the background graph. Denote the set of seed nodes by $\mathcal{S}$. Then the personalization vector or the restart distribution $\nu$ is given by

$$\nu = \left[\frac{1}{k}\mathbf{1}_k^T \ \mathbf{0}_{n-k}^T\right].$$

Personalized PageRank $\pi$ can be expressed as follows:

$$\pi = (1-\alpha)\nu[I - \alpha P]^{-1}. \tag{1}$$

Now let us define the mean field model of Personalized PageRank. It is based on the expected adjacency matrix:

$$\bar{A} = \begin{bmatrix} q\mathbf{J}_{m,m} & p\mathbf{J}_{m,n-m} \\ p\mathbf{J}_{n-m,m} & p\mathbf{J}_{n-m,n-m} \end{bmatrix},$$

and the associated mean field transition probability matrix $\bar{P} = \bar{D}^{-1}\bar{A}$. The mean field Personalized PageRank is given by

$$\bar{\pi} = (1-\alpha)\nu[I - \alpha\bar{P}]^{-1}. \tag{2}$$

Note that due to symmetry, the mean field Personalized PageRank has the following structure

$$\bar{\pi} = \left[\bar{\pi}_0\mathbf{1}_k^T \ \bar{\pi}_1\mathbf{1}_{m-k}^T \ \bar{\pi}_2\mathbf{1}_{n-m}^T\right],$$

where $\bar{\pi}_i, i = 0,1,2$, are determined by the system of linear equations:

$$\bar{\pi}_0 - \bar{\pi}_0\frac{\alpha kq}{mq+(n-m)p} - \bar{\pi}_1\frac{\alpha(m-k)q}{mq+(n-m)p} - \bar{\pi}_2\frac{\alpha(n-m)}{n} = \frac{1-\alpha}{k}, \tag{3}$$

$$-\bar{\pi}_0\frac{\alpha kq}{mq+(n-m)p} + \bar{\pi}_1 - \bar{\pi}_1\frac{\alpha(m-k)q}{mq+(n-m)p} - \bar{\pi}_2\frac{\alpha(n-m)}{n} = 0, \tag{4}$$

$$-\bar{\pi}_0\frac{\alpha kp}{mq+(n-m)p} - \bar{\pi}_1\frac{\alpha(m-k)p}{mq+(n-m)p} + \bar{\pi}_2 - \bar{\pi}_2\frac{\alpha(n-m)}{n} = 0. \tag{5}$$

These equations can easily be solved in explicit form. For instance, subtracting equation (4) from equation (3) we obtain

$$\bar{\pi}_0 = \frac{1-\alpha}{k} + \bar{\pi}_1. \tag{6}$$

Multiplying equation (4) by $p$ and equation (5) by $q$, respectively, and then subtracting one from another, we get

$$\bar{\pi}_1 = \bar{\pi}_2 \left( \frac{q}{p} - \frac{\alpha(n-m)}{n} \frac{q}{p} + \frac{\alpha(n-m)}{n} \right). \tag{7}$$

Then, substituting subsequently (6) and (7) into (5) yields

$$\bar{\pi}_2 = \frac{(1-\alpha)\alpha p}{(mq + (n-m)p)\left(1 - \frac{\alpha(n-m)}{n}\right) - \alpha m \left(q - \frac{\alpha(n-m)}{n}(q-p)\right)}. \tag{8}$$

Using (6) and (7), one easily retrieves $\bar{\pi}_1$ and $\bar{\pi}_2$. Namely, we have

$$\bar{\pi}_1 = \frac{(1-\alpha)\alpha \left(q - \frac{\alpha(n-m)}{n}(q-p)\right)}{(mq + (n-m)p)\left(1 - \frac{\alpha(n-m)}{n}\right) - \alpha m \left(q - \frac{\alpha(n-m)}{n}(q-p)\right)}, \tag{9}$$

and

$$\bar{\pi}_0 = \frac{1-\alpha}{k} + \frac{(1-\alpha)\alpha \left(q - \frac{\alpha(n-m)}{n}(q-p)\right)}{(mq + (n-m)p)\left(1 - \frac{\alpha(n-m)}{n}\right) - \alpha m \left(q - \frac{\alpha(n-m)}{n}(q-p)\right)}. \tag{10}$$

We would like to note that there is a simple bound on the expectated value of PPR. Clearly, we have $\sum_{i \in \mathcal{C} \setminus \mathcal{S}} \pi_i \leq 1$. By taking expectation of both sides and using symmetry, we obtain

$$\mathbb{E}(\pi_i) \leq \frac{1}{m-k}, \tag{11}$$

for $i \in \mathcal{C} \setminus \mathcal{S}$. Similarly, we have

$$\mathbb{E}(\pi_i) \leq \frac{1}{n-m}, \tag{12}$$

for $i \in \{1, 2, \ldots, n\} \setminus \mathcal{C}$.

# 3 Conditions for concentration of the PPR

Let us study the conditions when Personalized PageRank concentrates around its mean-field model. In order to investigate different regimes, we shall emphasize the dependence of the key parameters on the size of the graph $n$, that is $k := k(n)$, $m := m(n)$, $p := p(n)$ and $q := q(n)$. The following result states the $L^2$ convergence of the relative error of the mean field Personalized PageRank.

**Theorem 1** *Assume that $nq(n) = \omega(\log(n))$ and $p(n)/q(n) = \Theta(1)$. Then, the relative $L^2$ distance between $\pi$ and $\bar{\pi}$ converges in probability to zero. More precisely, there exists $C > 0$ such that*

$$\frac{||\pi - \bar{\pi}||_2}{||\bar{\pi}||_2} \leq \frac{\alpha C}{(1-\alpha)\sqrt{\frac{np(n)}{\log(n)}} - \alpha C}, \quad a.a.s. \tag{13}$$

**Proof:** It follows from the sensitivity analysis of the system of linear equations [16], $(A + \Delta A)(x + \Delta x) = b$, that the following inequality takes place

$$\frac{||\Delta x||_2}{||x||_2} \leq \frac{||A^{-1}||_2||\Delta A||_2}{1 - ||A^{-1}||_2||\Delta A||_2}. \tag{14}$$

In our case, this general inequality becomes

$$\frac{||\pi - \bar{\pi}||_2}{||\bar{\pi}||_2} \leq \frac{|||[I - \alpha\bar{P}]^{-1}||_2|| - \alpha[P - \bar{P}]||_2}{1 - |||[I - \alpha\bar{P}]^{-1}||_2|| - \alpha[P - \bar{P}]||_2}. \tag{15}$$

Since one is the maximal modulus eigenvalue of $\bar{P}$, we have

$$|||[I - \alpha\bar{P}]^{-1}||_2 = \frac{1}{1 - \alpha}. \tag{16}$$

From Lemma 1, which we provide below, it follows that there is $C > 0$ such that

$$||\alpha[P - \bar{P}]||_2 \leq \alpha C\sqrt{\frac{\log(n)}{np(n)}}, \quad \text{a.a.s.} \tag{17}$$

The combination of (15), (16) and (17) yields the result. □

We would like to note that the inequality (13) indicates very slow convergence. Indeed, if we consider the standard moderately sparse regime with

$$p(n) = \frac{\log^c(n)}{n}, \quad c > 1, \tag{18}$$

the rate of convergence will be of the order $1/\log^{\frac{c-1}{2}}(n)$.

We will now provide Lemma 1, which is crucial for the proof of Theorem 1.

**Lemma 1** *Assume that $nq(n) = \omega(\log(n))$, and $p(n)/q(n) = \Theta(1)$. Then for some $C > 0$,*

$$||P - \bar{P}||_2 \leq C\sqrt{\frac{\log(n)}{np}}, \quad \textit{a.a.s.}$$

**Proof:** We denote by $\bar{A}$ the expected adjacency matrix and by $\bar{D}$ the diagonal matrix of expected degrees. From Lemma 10 in [5] we have

$$\|A - \bar{A}\|_2 \leq K\sqrt{\log(n)nq(n)}, \quad \text{a.a.s.}, \tag{19}$$

where we used the fact that $q(n) > p(n)$. Then, since $p(n)/q(n) = \Theta(1)$, we also have $np(n) = \omega(\log(n))$. Therefore, from Lemma 8 of [5] for some $C_1 > 0$ we have

$$\|\bar{D}^{-1}D - I\|_2 \leq C_1\sqrt{\frac{\log(n)}{np(n)}}, \quad \text{a.a.s.}, \tag{20}$$

since $\bar{A}$ is a rank two matrix with all entries in the upper-left sub-matrix of size $|\mathcal{C}| \times |\mathcal{C}|$ equal to $q(n)$ and all other entries being $p(n)$. Now, from (19) we obtain

$$\|A\|_2 \leq \|A - \bar{A}\|_2 + \|\bar{A}\|_2 \leq K\sqrt{\log(n)nq(n)} + nq(n), \quad \text{a.a.s.} \tag{21}$$

Using the above bounds, we get

$$\begin{aligned}\|D^{-1}A - \bar{D}^{-1}\bar{A}\|_2 &= \|(D^{-1}\bar{D} - I)\bar{D}^{-1}A + \bar{D}^{-1}(A - \bar{A})\|_2 \\ &= \|D^{-1}\bar{D} - I\|_2\|\bar{D}^{-1}\|_2\|A\|_2 + \|\bar{D}^{-1}\|_2\|A - \bar{A}\|_2 \\ &\leq \frac{1}{np(n)}\left(C_1\sqrt{\frac{\log(n)}{np(n)}}(K\sqrt{\log(n)nq(n)} + nq(n)) + K\sqrt{\log(n)nq(n)}\right) \\ &\leq C\sqrt{\frac{\log(n)}{np(n)}}\end{aligned}$$

for some $C > 0$. □

Now let $U$ be a uniformly randomly sampled integer between 1 and $n$ and

$$\eta(i) = \begin{cases} 0, & 1 \leq i \leq k, \\ 1, & k+1 \leq i \leq m, \\ 2, & m+1 \leq i \leq n. \end{cases}$$

Then, using Theorem 1 we can establish that the difference between $\pi_U$ and $\bar{\pi}_{\eta(U)}$ is vanishing with high probability when $k(n)$ is large enough. The result is formally stated in the next theorem.

**Theorem 2** *Let conditions of Theorem 1 hold. Furthermore, let $k = k(n)$ be such that $k(n)p(n) = \omega(\log(n))$ and let $U$ be the index of a randomly sampled node $1, 2, \ldots, n$. Then, for any $\varepsilon > 0$ the following result holds*

$$\lim_{n\to\infty} \mathbb{P}\left(|\pi_U - \bar{\pi}_{\eta(U)}| \geq \varepsilon n^{-1}\right) = 0. \tag{22}$$

**Proof:** Denote by $B_n$ the event that the inequality in (13) holds:

$$B_n = \left\{ \frac{||\pi - \bar{\pi}||_2}{||\bar{\pi}||_2} \leq \frac{\alpha C}{(1-\alpha)\sqrt{\frac{np(n)}{\log(n)}} - \alpha C} \right\}$$

for an appropriate value of $C$. The idea of the proof is to use this inequality to bound our probability of interest on $B_n$ and then use the fact that $\lim_{n\to\infty} P(B_n) = 1$.

To this end, we need to bound the probability in (22) using $L^2$-norms. We do this by first conditioning on the realization of $G$. We will denote by $\mathbb{P}_n$ the probability measure conditioned on $G$. Then the randomness is only in the choice of $U$. By Markov's inequality, we have

$$\begin{aligned}\mathbb{P}_n\left(|\pi_U - \bar{\pi}_{\eta(U)}| \geq \varepsilon n^{-1}\right) &\leq \frac{n\mathbb{E}_n(|\pi_U - \bar{\pi}_{\eta(U)}|)}{\varepsilon} \\ &= \frac{n}{\varepsilon}\left(\frac{1}{n}\sum_i |\pi_i - \bar{\pi}_{\eta(i)}|\right) \\ &\leq \frac{n}{\varepsilon}\left(\frac{1}{n}\sum_i |\pi_i - \bar{\pi}_{\eta(i)}|^2\right)^{1/2} \\ &= \frac{\sqrt{n}}{\varepsilon}||\pi - \bar{\pi}||_2.\end{aligned}$$

Next, by the full probability formula, we have

$$\begin{aligned}\mathbb{P}\left(|\pi_U - \bar{\pi}_{\eta(U)}| \geq \varepsilon n^{-1}\right) &= \mathbb{E}\left[\mathbb{P}_n\left(|\pi_U - \bar{\pi}_{\eta(U)}| \geq \varepsilon n^{-1}\right)\right] \\ &= \mathbb{E}\left[P_n\left(|\pi_U - \bar{\pi}_{\eta(U)}| \geq \varepsilon n^{-1}\right)\mathbf{1}\{B_n\}\right] + \mathbb{E}\left[P_n\left(|\pi_U - \bar{\pi}_{\eta(U)}| \geq \varepsilon n^{-1}\right)\mathbf{1}\{\bar{B}_n\}\right] \\ &\leq \mathbb{E}\left[\frac{\sqrt{n}}{\varepsilon}\,||\pi - \bar{\pi}||_2\,\mathbf{1}\{B_n\}\right] + P(\bar{B}_n) \\ &\leq \frac{\alpha C\sqrt{n}\,||\bar{\pi}||_2}{(1-\alpha)\sqrt{\frac{np(n)}{\log(n)}} - \alpha C}\,P(B_n) + P(\bar{B}_n), \end{aligned} \tag{23}$$

where we used (13) for the last step. Since $P(B_n)$ converges to one as $n \to \infty$, the statement of the theorem follows when $\frac{\alpha C\sqrt{n}||\bar{\pi}||_2}{(1-\alpha)\sqrt{\frac{np(n)}{\log(n)}}-\alpha C}$ converges to zero. It remains to verify that this is indeed the case when $\frac{k(n)p(n)}{\log(n)} \to \infty$. Now note that (23) together with the fact that $\bar{\pi}_0 = \Theta((k(n))^{-1})$, $\bar{\pi}_1 = \Theta((m-k(n))^{-1})$, $\bar{\pi}_2 = \Theta((n-m)^{-1})$, and $k = O(m)$, implies that

$$||\bar{\pi}||_2 = \left(k(n)\bar{\pi}_0^2 + (m-k(n))\bar{\pi}_1^2 + (n-m)\bar{\pi}_2^2\right)^{1/2} = \Theta\left(k(n)^{-1/2}\right),$$

which gives the result. □

The practical implication of Theorem 2 is that the condition $\frac{k(n)p(n)}{\log(n)} \to \infty$ is sufficient for $\pi$ to be well approximated by $\bar{\pi}$. In other words, $\pi$ is concentrated around $\bar{\pi}$. Notice that the result of Theorem 2 holds for a large range of regimes. Indeed, the requirement $k(n)p(n) = \omega(\log(n))$ means that the number of seed node neighbours of a node $i \notin \mathcal{C}$ is of the order larger than $\log(n)$. This condition is satisfied in a dense regime as well, when $p(n)$ and $q(n)$ are constants. For example, the above analysis is also applicable in the setting of [26], but without the artificial modification of PPR. In the next section we will focus on the regimes where the local tree approximation of the graph is valid. This does not include a dense regime or any regime where $np(n)$, $nq(n)$ are powers of $n$. In this class of regimes, we will obtain conditions, under which concentration does not occur.

# 4 Non-concentration conditions for PPR

In this section we will, as before, assume that $p(n)/q(n) = \Theta(1)$ and consider the regime when $nq(n)$ is smaller than a power of $n$, more precisely, $nq(n) = o(n^{\varepsilon})$ for all $\varepsilon > 0$. Note this includes our regime of interest (18). We will show that in this range of parameters the condition $k(n)q(n) \to \infty$ is necessary for concentration of $\pi$ around $\bar{\pi}$ to occur. Specifically, when this condition is violated, then $\pi$ does not concentrate at all, that is, the coefficient of variation of $\pi_i$, $i = 1, 2, \ldots, n$, is non-vanishing.

Our argument relies on the local tree approximation of our random graph model constructed as follows. For any node $i$ in $G$, we will say that $i$ is of type $\mathcal{C}$ if $i \in \mathcal{C}$ and of type $\bar{\mathcal{C}}$ otherwise. Consider a rooted Galton-Watson tree $\mathcal{T}_i^t$ of depth $t$ with root $i$. Assume that each node has Poisson($mq(n)$) number of offspring of type $\mathcal{C}$ and Poisson($(n-m)p(n)$) number of offspring of type $\bar{\mathcal{C}}$, each type independent of each other. The following lemma from [17] states that $\mathcal{T}_i^t$ can be coupled with high probability with the $t$-hop neighborhood of a random node, $G_i^t$.

**Lemma 2** *[17, Lemma 10] Assume that $p(n)/q(n) = \Theta(1)$ and $nq(n) = o(n^{\varepsilon})$ for all $\varepsilon > 0$. Then for any node $i = 1, 2, \ldots, n$ and $t := t(n) \to \infty$ such that $(nq(n))^t = n^{o(1)}$, there exists a coupling such that $(G_i^t, \sigma^t) = (\mathcal{T}_i^t, \tau^t)$ with probability $1 - n^{-1+o(1)}$, where $G_i^t$ is the subgraph induced by the set of nodes at distance $t$ from $i$ and $\sigma^t$ is the vector of the types of the nodes of the graph. Also, $\mathcal{T}_i^t$ is a Galton-Watson tree with Poisson offspring distribution and $\tau^t$ is the vector of types on $\mathcal{T}_i^t$.*

We want to show that if $k(n)q(n) = O(1)$, then, with positive probability, the difference between $\pi_i$ and $E(\pi_i)$ is of the same order of magnitude as $\mathbb{E}(\pi_i)$ itself. The prove consists of two steps. First, in Lemma 3 we will show that $\pi_i$ is well approximated by a $t$-neighborhood. Then, in Theorem 3, we will use this result together with Lemma 2 to demonstrate the non-concentration.

Denote by $\pi^t$ the contribution of paths shorter than $t$:

$$\pi^t = (1-\alpha)\nu \sum_{l=0}^{t-1} \alpha^l P^l.$$

Now we can easily prove Lemma 3 below.

**Lemma 3** *Take $t := t(n) \to \infty$. Then for any $i \notin \mathcal{S}$,*

$$\mathbb{E}(|\pi_i - \pi_i^t|) = o(n^{-1}). \tag{24}$$

**Proof:** First, we split the PPR in (1) as follows:

$$\pi = \pi^t + (1-\alpha)\nu\alpha^t P^t \sum_{l=0}^{\infty} \alpha^l P^l.$$

Now, proceeding exactly as in [11], for the second term we get

$$\|\pi_i - \pi_i^t\|_1 = \alpha^t.$$

Assume that $i \in \mathcal{C}\backslash\mathcal{S}$, and note that for the nodes outside $\mathcal{C}$ the argument is exactly the same. Since all $m$ nodes in $\mathcal{C}$ are symmetric, for any $t = t(n) \to \infty$ we immediately obtain

$$\mathbb{E}(|\pi_i - \pi_i^t|) \leq \frac{1}{m}\,\alpha^t = o(n^{-1}).$$

This gives (24). □

Now using Lemma's 2 and 3 we can prove the non-concentration result, stated in the next theorem. We will prove the result in the regime when $k(n)$ is at least a power of $n$ (however, such power can be arbitrarily small).

**Theorem 3** *(Non-concentration of PPR) Let $G$ be rooted at node $i \notin \mathcal{S}$. If $k(n)q(n) = O(1)$ and there exists $\xi > 0$ such that $k(n) \geq n^{\xi}$, then*

$$\frac{Var(\pi_i)}{\mathbb{E}^2(\pi_i)} = \Omega(1). \tag{25}$$

**Proof:** We will again prove the result for node $i \in \mathcal{C}\backslash\mathcal{S}$. As in Lemma 3, the argument for $i \notin \mathcal{C}$ is exactly the same. First of all, note that

$$\mathbb{E}(\pi_i) \leq \frac{1}{m} = \Theta(n^{-1})$$

because $\mathbb{E}(\pi_i) \leq \frac{1}{m} = \Theta(n^{-1})$. Next, taking into account only neighbors of $i$ from $\mathcal{S}$ and using Jensen's inequality, we can write

$$\begin{aligned}\mathbb{E}(\pi_i) &\geq \frac{\alpha}{k(n)}\,\mathbb{E}\left(\sum_{j\in\mathcal{S}}\frac{1\{a_{i,j}=1\}}{d_j}\right)\\ &= \frac{\alpha}{k(n)}\,k(n)q(n)\,\mathbb{E}\left(\frac{1}{d_j}\middle|\, a_{i,j}=1\right)\\ &\geq \frac{\alpha q(n)}{\mathbb{E}(d_j|a_{i,j}=1)}\\ &= \frac{\alpha q(n)}{((m-1)q(n)+np(n)+1)} = \Theta(n^{-1}),\end{aligned}$$

where we recall that $a_{i,j}$ is the element of the adjacency matrix $A$. In the rest of the proof we will evaluate $\mathrm{Var}(\pi_i)$. For that, we will use the decomposition of PageRank from [7]. Consider a simple random walk $(X_l)_{l\geq 0}$ on $G$ such that at each step the walk continues with probability $\alpha$ and terminates with probability $1-\alpha$. Let $T$ be the termination time, which has a geometric distribution with parameter $(1-\alpha)$. Denote by $\mathbb{P}_{(j)}$ the conditional probability given the event $\{X_0 = j\}$. Then, for any realization of the graph, from [7] we have:

$$\begin{aligned}\pi_i &= \frac{(1-\alpha)}{|\mathcal{S}|}\left(\sum_{s\in\mathcal{S}}\mathbb{P}_{(s)}(X_l = i \text{ for some } l\leq T)\right)\mathbb{E}_{(i)}\left[\sum_{l=0}^{\infty}1\{X_l=i\}\right]. \qquad (26)\\ &= (I)\times(II)\times(III).\end{aligned}$$

Here (I), (II) and (III) are random variables that depend on a realization of the random graph. Note that (III) is the average number of visits to $i$, starting from $i$ and before termination of the random walk. It is easy to see that this number is not smaller than 1 (at least one visit at the initial step) and not greater than $(1-\alpha^2)^{-1}$ (there is a geometric number of returns, while $\alpha^2$ is the maximal possible return probability). Thus, it is sufficient to consider the variance of (II).

We will do this by using the local tree approximation. Choose $t := t(n)$ as in Lemma 2. Consider only $t$-neighborhood of $i$, denoted by $G_i^t$, and let $\tilde{\pi}_i^t$ be such that the contribution of (II) in (26) is restricted only by paths in $G_i^t$:

$$\begin{aligned}\tilde{\pi}_i^t &= \frac{(1-\alpha)}{|\mathcal{S}|}\left(\sum_{s\in\mathcal{S}}\mathbb{P}_{(s)}(X_l = i \text{ for some } l\leq T,\, X_0,\ldots,X_{l-1}\in G_i^t)\right)\mathbb{E}_{(i)}\left[\sum_{l=0}^{\infty}1\{X_l=i\}\right] \qquad (27)\\ &= (I)\times(II)'\times(III),\end{aligned}$$

Note that

$$\tilde{\pi}_i^t \leq \pi_i, \qquad (28)$$

because $\pi_i$ includes all paths of length $t$ plus some paths of lengths longer than $t$, which make loops in $G_i^t$ on the way to $i$, plus the paths that include a loop from $i$ back to $i$. Thus, if we write $\pi_i = \tilde{\pi}_i^t + \delta_i^t$ with $\delta_i^t \geq 0$ then $\mathbb{E}(\delta^t) = o(n^{-1})$ due to (28) and Lemma 3. It follows that

$$\begin{aligned}\mathrm{Var}(\pi_i) &= \mathrm{Var}(\tilde{\pi}_i^t) + \mathrm{Var}(\delta^t) + 2\mathbb{E}(\tilde{\pi}_i^t\delta^t) - 2\mathbb{E}(\pi_i)\mathbb{E}(\delta^t) \qquad (29)\\ &\geq \mathrm{Var}(\tilde{\pi}_i^t) - 2\mathbb{E}(\pi_i)o(n^{-1}) = \mathrm{Var}(\tilde{\pi}_i^t) + o\left([\mathbb{E}(\pi_i)]^2\right). \qquad (30)\end{aligned}$$

Therefore, it is sufficient to bound $\mathrm{Var}(\tilde{\pi}_i^t)$ from below by a term of the order at least $[\mathbb{E}(\pi_i)]^2$. To this end, it follows from the same argument as after equation (26), we need to analyze $\mathrm{Var}((II)')$.

Conditioning on the last step before reaching $i$, we get:

$$(II)' = \sum_{j:j \text{ offspring of } i} \frac{\alpha}{d_j} \sum_{s\in\mathcal{S}} \mathbb{P}_{(s)}(X_l = j \text{ for some } l \le T-1,\, j \text{ reached before } i,\, X_0,\dots,X_{l-1} \in G_i^t). \tag{31}$$

Next, denote by $C_n$ the event that the $t$-neighborhood of $i$ coincides with the Galton-Watson tree $(\mathcal{T}_i^t, \tau^t)$. Conditioned on $C_n$, the terms in the external summation in $(II)'$ are independent. In particular, $\mathrm{Var}((II)'|C_n)$ is a sum of three independent contributions: from the neighbors of $i$ in $\mathcal{S}$, $\mathcal{C}\backslash\mathcal{S}$ and $\bar{\mathcal{C}}$. We will lower bound $\mathrm{Var}((II)'|C_n)$ by considering only the contribution of the neighbors of $i$ that are seed nodes. We number such neighbors as $i_1, i_2, \dots, i_{N_0}$. Then we obtain

$$\begin{aligned}
&\mathrm{Var}((II)'|C_n)\\
&\ge \mathrm{Var}\left(\sum_{j\in\{i_1,\dots,i_{N_0}\}} \frac{\alpha}{d_j}\sum_{s\in\mathcal{S}} \mathbb{P}_{(s)}(X_l = j \text{ for some } l \le T-1,\, j \text{ reached before } i,\, X_0,\dots,X_{l-1} \in G_i^t)|C_n\right)\\
&:= \mathrm{Var}\left(\sum_{j\in\{i_1,\dots,i_{N_0}\}} Z_j|C_n\right)\\
&= \mathbb{E}(N_0|C_n)\mathrm{Var}(Z_j|C_n) + \mathrm{Var}(N_0|C_n)(\mathbb{E}Z_j|C_n)^2.
\end{aligned}$$

Motivated by the above expression, we will evaluate the moments of $N_0$ given $C_n$. Recall that in the original graph, $N_0$ has Binomial$(k(n), q(n))$ distribution. Now, for $r > 0$ and some $\epsilon < \xi/r$ we split $\mathbb{E}(N_0^r)$ as follows:

$$\mathbb{E}(N_0^r) = \mathbb{E}(N_0^r|C_n)\mathbb{P}(C_n) + E(N_0^r 1\{N_0 < n^\epsilon\}1\{\bar{C}_n\}) + E(N_0^r 1\{N_0 > n^\epsilon\}1\{\bar{C}_n\}). \tag{32}$$

By Lemma 2, the second term in (32) is bounded from above by $n^{r\epsilon}\mathbb{P}(\bar{C}_n) = O(n^{-1+r\epsilon+o(1)}) = o(k(n)q(n))$. The third term in (32) is bounded by $k^r(n)\mathbb{P}(N_0 > n^\varepsilon)$. Using the bound from Theorem 2.21 in [21], we obtain that

$$\begin{aligned}
k^r(n)\mathbb{P}(N_0 > n^\varepsilon) &\le k^r(n) e^{-\frac{(n^\varepsilon - k(n)q(n))^2}{2(2k(n)q(n)/3 + n^\varepsilon/3)}}\\
&= k^r(n) O(e^{-n^\varepsilon/2}) = o(k(n)q(n)).
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}(N_0|C_n)\mathbb{P}(C_n) &= k(n)q(n)(1+o(1)),\\
\mathrm{Var}(N_0|C_n) &= k(n)q(n)(1-q(n))(1+o(1)).
\end{aligned}$$

From this and $q(n) = o(1)$, we conclude that for some $0 < \gamma < 1$ we have

$$\mathrm{Var}((II)'|C_n) \ge \gamma k(n)q(n)\mathbb{E}(Z_j^2|C_n). \tag{33}$$

Note that for every $j \in \mathcal{S}$ we have the trivial lower bound

$$Z_j \ge \frac{\alpha}{d_j}\,\mathbb{P}_{(j)}(X_l = j \text{ for some } l \le T-1,\, j \text{ reached before } i,\, X_0,\dots,X_{l-1} \in G_i^t\,) = \frac{\alpha}{d_j}.$$

Further, recall that given $C_n$, $d_j \stackrel{d}{=} 1 + \text{Poisson}(mq(n) + (n-m)q(n))$. It follows that

$$\begin{aligned}\mathbb{E}(Z_j^2|C_n) &\geq \alpha^2 \mathbb{E}\left(\frac{1}{d_j^2}\right) \geq \alpha^2 \left(\frac{1}{[\mathbb{E}(d_j)]^2}\right) \\ &= \frac{\alpha^2}{(1+mq(n)+(n-m)p(n))^2} \\ &\geq \frac{\alpha^2}{4n^2q(n)^2}, \end{aligned} \tag{34}$$

where in the second inequality we used Jensen's inequality. From (33), (34) it follows that

$$\text{Var}((II)'|C_n) \geq \frac{\gamma\alpha^2 k(n)q(n)}{4n^2q(n)^2}.$$

Hence, since $(III) \geq 1$, from (27), we get

$$\text{Var}(\tilde{\pi}_i^t|C_n) \geq \frac{\gamma\alpha^2(1-\alpha)^2}{4k(n)q(n)n^2}.$$

Finally, using that $\mathbb{E}(\pi_i) = \Theta(n^{-1})$ and $\lim_{n\to\infty} \mathbb{P}(C_n) = 1$, we obtain

$$\frac{\text{Var}(\tilde{\pi}_i^t)}{[\mathbb{E}(\pi_i)]^2} \geq \frac{\text{Var}(\tilde{\pi}_i^t|C_n)P(C_n)}{[\mathbb{E}(\pi_i)]^2} = \Omega\left(\frac{1}{k(n)q(n)}\right), \tag{35}$$

which, together with (29), gives the result. □

**Remark 1** *It should be possible to relax the condition $k(n) = \omega(n^\xi)$ to $k(n) = \omega(1)$. For that, we need either a stronger coupling than in Lemma 2 or another way to evaluate $E(N_0|C_n)$ instead of (32).*

Let us now discuss some implications of Theorem 3. Suppose, for example, that $q(n) = (1+a)p(n)$ for some $a > 0$. The non-vanishing coefficient of variation means that $\pi_i$ has finite spreading around its mean. Then, in practice, if $a$ is small, we will not be able to distinguish many nodes in $\mathcal{C}$ from the nodes outside of $\mathcal{C}$, even in a very large network. We will provide an illustration for this scenario in Figures 2, 3 in Section 6.

The necessary condition $k(n)q(n) \to \infty$ has the following very intuitive interpretation. Note that $k(n)q(n)$ is the average number of neighbors from $\mathcal{S}$ of a node in $\mathcal{C}$. Recall that each seed node (in $\mathcal{S}$) receives a certain large probability mass. When node $i$ has a finite average number of neighbors from $\mathcal{S}$, then their total contribution to $\pi_i$ is a finite random variable, so there is no concentration. Moreover, when $k(n)q(n) \to \infty$ then contributions of $\mathcal{S}$ to $\pi_i$ is a sum of asymptotically infinite number of terms, so the concentration should occur.

In the proof we used the coupling of the graph with a tree. Note that recent work [14] allows one to pass the distribution of PageRank to the limit when the graph converges (possibly to a tree) in the 'local weak convergence' sense. However, such convergence is defined only for sparse graphs, i.e., with asymptotically finite degrees, and does not apply to our 'medium dense' case (18). Also, Theorems 1 and 2 are applicable to the dense regime when $p$ and $q$ do not depend on the size of the graph.

## 5 Optimization with respect to the damping factor

In clustering applications, the performance of personalized PageRank is influenced by the choice of the parameter $\alpha$. In [2], the authors choose $\alpha$ as a function of the conductance of the desired smallest cut. Typically, the conditions of [2] lead to the values of $\alpha$ very close to one. In community detection applications, the probability of an error is a function of the difference between PageRank scores within and outside the community. In Theorem 2, we have identified a regime, where PPR $\pi$ is concentrated around its mean-field proxy $\bar{\pi}$. In such regime we can use the expressions for $\bar{\pi}_1$ and $\bar{\pi}_2$ from Section 2 to find an *optimal* parameter $\alpha$ that maximizes the difference between PageRank inside and outside community $\mathcal{C}$. Thus, the optimal $\alpha$ can be found as the solution to the following optimization problem:

$$\alpha_{opt} = \arg\max_{\alpha}(\bar{\pi}_1(\alpha) - \bar{\pi}_2(\alpha)).$$

Let us denote $\rho = \frac{q}{p}$ and $\beta = \frac{n-m}{n}$. Then by (7) we have

$$\begin{aligned}\bar{\pi}_1 - \bar{\pi}_2 &= (\rho-1)(1-\alpha\beta)\bar{\pi}_2 \\ &= \frac{\alpha(1-\alpha)(\rho-1)(1-\alpha\beta)}{m\left(\alpha^2\beta(\rho-1) - \alpha(\rho\beta+\rho+\frac{\beta^2}{1-\beta}) + \rho + \frac{\beta}{1-\beta}\right)}.\end{aligned}$$

The optimum $\alpha$ is such that $\frac{d}{d\alpha}(\bar{\pi}_1 - \bar{\pi}_2)\big|_{\alpha=\alpha_{\rm opt}} = 0$. Thus, we find the optimum $\alpha$ as the solution of the following equation:

$$\begin{aligned}&\alpha^4\beta^2(\rho-1) - 2\beta\left(\rho\beta+\rho+\frac{\beta^2}{1-\beta}\right)\alpha^3 \\ &\quad + \alpha^2\left(3\beta(\rho+\frac{\beta}{1-\beta}) + (1+\beta)(\rho\beta+\rho+\frac{\beta^2}{1-\beta}) - \beta(\rho-1)\right) \\ &\quad - 2\alpha(1+\beta)\left(\rho+\frac{\beta}{1-\beta}\right) + \rho + \frac{\beta}{1-\beta} = 0.\end{aligned}$$

With straightforward algebra, we can simplify the above equation as follows:

$$(\alpha-1)^2\left(\alpha^2\beta^2(\rho-1) - 2\alpha\beta\left(\rho+\frac{\beta}{1-\beta}\right) + \rho + \frac{\beta}{1-\beta}\right) = 0.$$

Since $\alpha < 1$, the optimum is the solution of the following quadratic equation

$$\alpha^2\beta^2(\rho-1) - 2\alpha\beta\left(\rho+\frac{\beta}{1-\beta}\right) + \rho + \frac{\beta}{1-\beta} = 0.$$

The solutions to the above equation are

$$\alpha_i^* = \frac{\rho - \beta(\rho-1) + (2i-1)\sqrt{\rho-\beta(\rho-1)}}{\beta(1-\beta)(\rho-1)},$$

for $i = 0, 1$. The solution corresponding to $i = 1$ can be shown to be greater than 1 for any $\rho > 1, \beta < 1$ and hence the only feasible solution is given by

$$\alpha_{opt} = \min\left(1, \frac{\rho - \beta(\rho-1) - \sqrt{\rho-\beta(\rho-1)}}{\beta(1-\beta)(\rho-1)}\right). \tag{36}$$

From the above equation, we can glean the following insight. Notice that if $x := \rho - \beta(\rho - 1) = (1-\beta)\rho + \beta$, then after some elementary algebraic manipulations we have

$$\alpha_{opt} = \frac{\sqrt{x}}{(1+\sqrt{x})(\rho - x)} = \frac{\sqrt{x}}{\beta(1+\sqrt{x})}.$$

Thus, for a fixed $\rho$ or $\beta$, $\alpha_{opt}$ is an increasing function of $x$, while $x$ itself is an increasing (or decreasing) function of $\rho$ (or $\beta$). In other words, the more distinguishable the community is (larger $\rho$ or smaller $\beta$), the larger is the optimum $\alpha$. This conforms to the intuition that the RW starting from the seed nodes should explore the graph more before termination when we have a denser or larger community.

# 6 Numerical examples and implications for local graph clustering

The theoretical results of the two preceding sections have important implications for PPR based local graph clustering. Our main theoretical result is that the parameter $k$ should scale linearly and the parameter $m$ sufficiently fast with the size of the graph $n$ in order to ensure the concentration of PPR. In practice, this means that the number of seed nodes should be significant to guarantee high quality clustering results and the target community should not be too small.

As an aside, one could pose the following natural question: Can the sparsity-enforcing nature of the Approximate PPR (APPR) algorithm help to avoid the leakage of probability mass to the nodes outside of the target community? Unfortunately, as we will demonstrate below in Subsection 6.2, APPR suffers from the same non-concentration phenomenon as the original PPR.

For the purpose of illustration let us consider a specific numerical example. We take $n = 10000$, $m = 2000$, and the edge probabilities as follows:

$$p(n) = \frac{5\,\log^2(n)}{n}, \quad q(n) = \frac{10\,\log^2(n)}{n}, \tag{37}$$

We first consider the case $\alpha = 0.8$. If we set $k = 200$, we observe a reasonably good concentration (see Figure 1) even though the values of $p(n)$ and $q(n)$ set by (37) imply very slow convergence with the rate $1/\sqrt{\log(n)}$, according to (13). We can also calculate the percentage of nodes wrongly assigned to the community $\mathcal{C}$ according to the rank of PPR, which we denote by $\mathcal{E}$ defined below

$$\mathcal{E} = \frac{|\overline{\mathcal{C}} \bigcap \widehat{\mathcal{C}}|}{|\mathcal{C}|},$$

where $\mathcal{C}$ is the target community, $\overline{\mathcal{C}}$ is its compliment and $\widehat{\mathcal{C}}$ is an algorithm output. For the above chosen parameters we obtain $\mathcal{E} = 3.6\%$.

In the next experiment we decrease the number of seed nodes to $k = 20$. Then the error increases by an order of magnitude and becomes $\mathcal{E} = 44.2\%$, and we can observe the effect of non-concentration in Figure 2. Curiously enough, if we decrease the number of seeds further to $k = 2$, the error actually improves a bit to $\mathcal{E} = 34.5\%$ but still remains very high. We can explain the slight decrease in the error by the fact that most misclassified nodes are the neighbours of seed nodes. Notice that this is in accordance with the proof of Theorem 3 where the neighbors that are seed nodes played a crucial role. Indeed, the spikes in Figure 3 correspond to the neighbours of the seed nodes. Thus, if we decrease the number of seed nodes, we also subdue the main source of errors. Of course, there is a fine trade off and one cannot eliminate completely the strong effect from non-concentration.

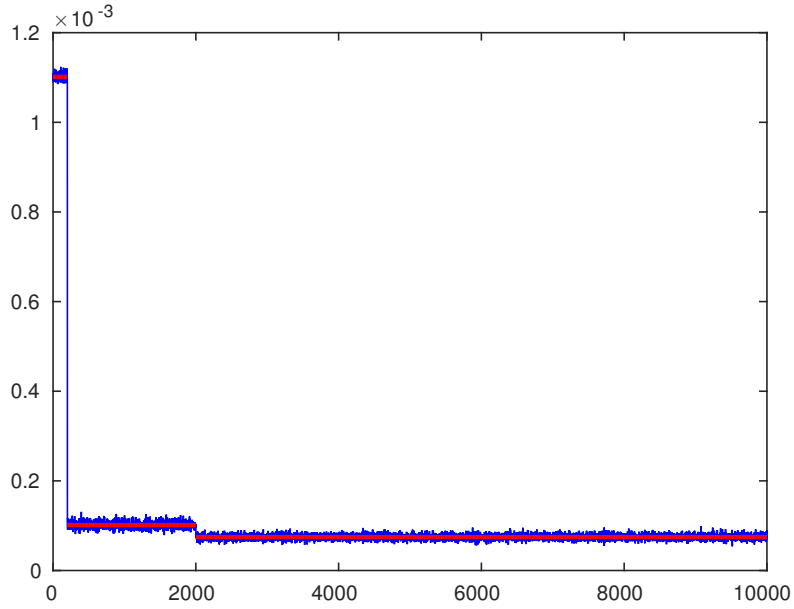


Figure 1: PPR (blue) and its mean-field model (red) for k=200. On the $x$-axis are the indices of the nodes. Nodes with indices $1, 2, \ldots, 2000$ belong to $\mathcal{C}$.

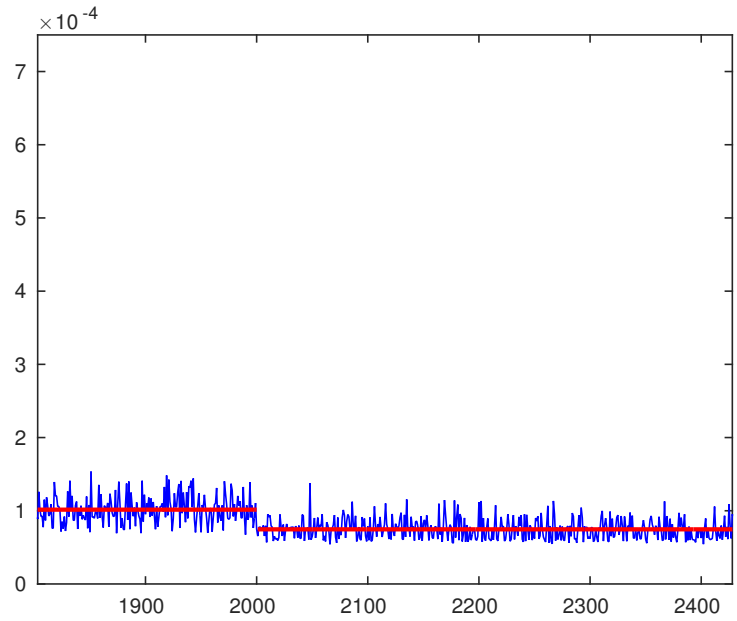


Figure 2: PPR (blue) and its mean-field model (red) for k=20. On the $x$-axis are the indices of the nodes. Nodes with indices $1, 2, \ldots, 2000$ belong to $\mathcal{C}$.

### 6.1 Optimum value of $\alpha$

In this section we investigate numerically the dependence of the optimum $\alpha$ derived from the mean-field version of PPR on the graph parameters. In Figure 4 we plot the difference $\bar{\pi}_1(\alpha) - \bar{\pi}_2(\alpha)$ as a function of $\alpha$ for $n = 10000$, $m = 3000$, and $p$ and $q$ as in (37). We see that the

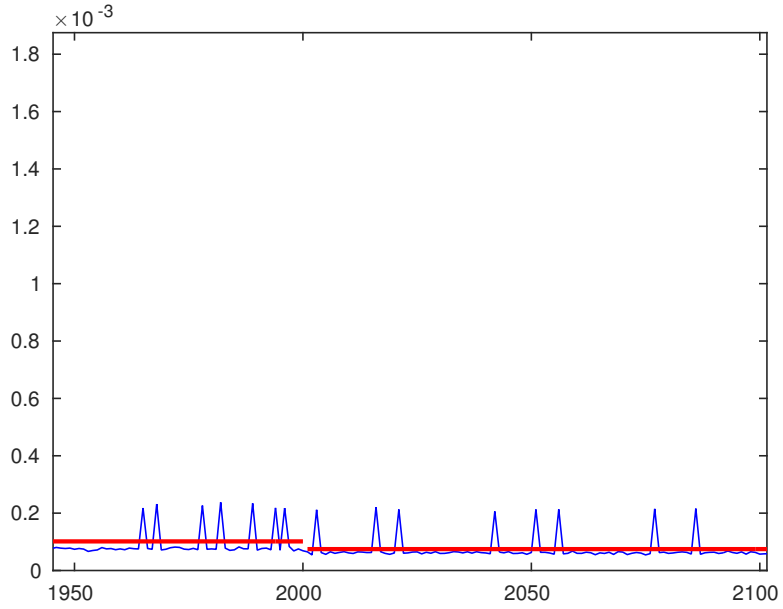


Figure 3: PPR (blue) and its mean-field model (red) for k=2. On the $x$-axis are the indices of the nodes. Nodes with indices $1, 2, \ldots, 2000$ belong to $\mathcal{C}$.

curves for $k = 2$ and $k = 200$ coincide. This is the case because $\bar{\pi}_1$ and $\bar{\pi}_2$ in fact do not depend on $k$. It is interesting to observe that for reasonably large communities the optimum value of $\alpha$ is quite close to the default value 0.85 set by Google. Now, if we decrease the community size from 3000 to 300, the optimum value of $\alpha$ decreases towards 0.5 (see Figure 5). The decrease is expected, since to identify a smaller community, PPR needs shorter walks. It might not be a coincidence that the optimum value of $\alpha$ decreased towards 0.5, which was a value recommended in [8, 9] by some other considerations.

### 6.2 Non-concentration of Approximate Personalized PageRank

The Approximate Personalized PageRank (APPR) algorithm in [2] is used to find a set of nodes $S$ with a given target conductance $\phi$. In order to have a set with conductance at most $\phi$, we need to choose the parameters of the algorithm $\alpha, \epsilon$ [2] such that $1 - \alpha = \frac{\phi^2}{225 \log(100\sqrt{|E|})}$ and $\epsilon = \frac{2^{-b}}{48B}$, where $B = \lceil \log_2(|E|) \rceil$ and $b \in [1, B]$ (Note: In our numerical experiments we take $b = 13$; larger values of $b$ decrease $\epsilon$ and thus increase the time to convergence of APPR, without significant gain in performance.)

For the graph parameter values considered in this section the 'mean form' conductance is $\overline{\phi}(\mathcal{C}) = 0.66$, see equation (39) below. The conditions of [2] give $\alpha = 0.999$. Using these values we run a clustering algorithm based on the exact PPR where nodes are ranked according to their pagerank scores and the output community is the set of first $m$ nodes. We also run the APPR algorithm [2] with $\epsilon = 10^{-7}$. The algorithm is quoted here for the sake of completeness in Algorithm 1. Note that the final step of the clustering algorithm based on the approximate PageRank is to perform a "sweep operation" on the nodes ordered in decreasing order of a ranking function on the nodes. The ranking function proposed in [2] is the values of APPR divided by the node degrees. In our simulations we also investigate the performance of the algorithm where the ranking function is the approximate PageRank without this degree scaling.

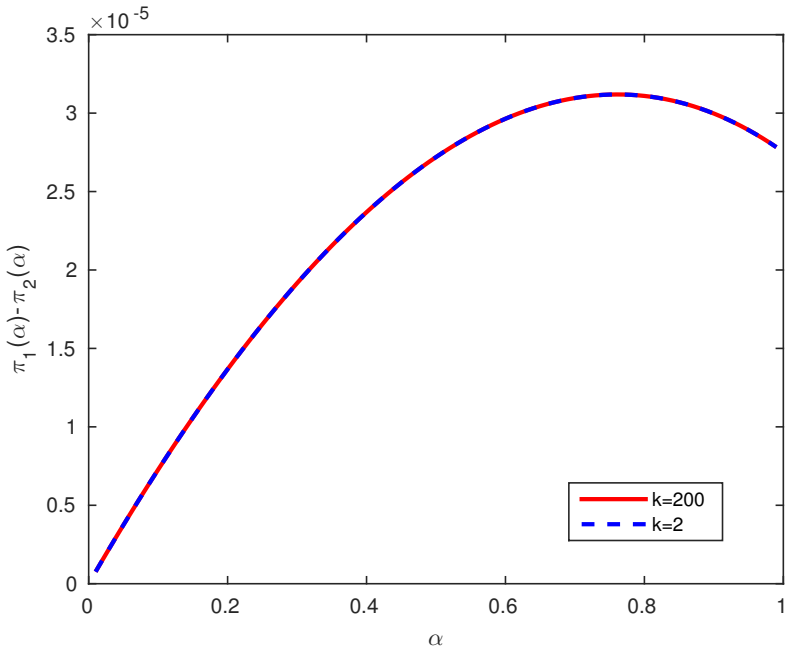


Figure 4: Optimum value of the restart probability for $m = 3000$.

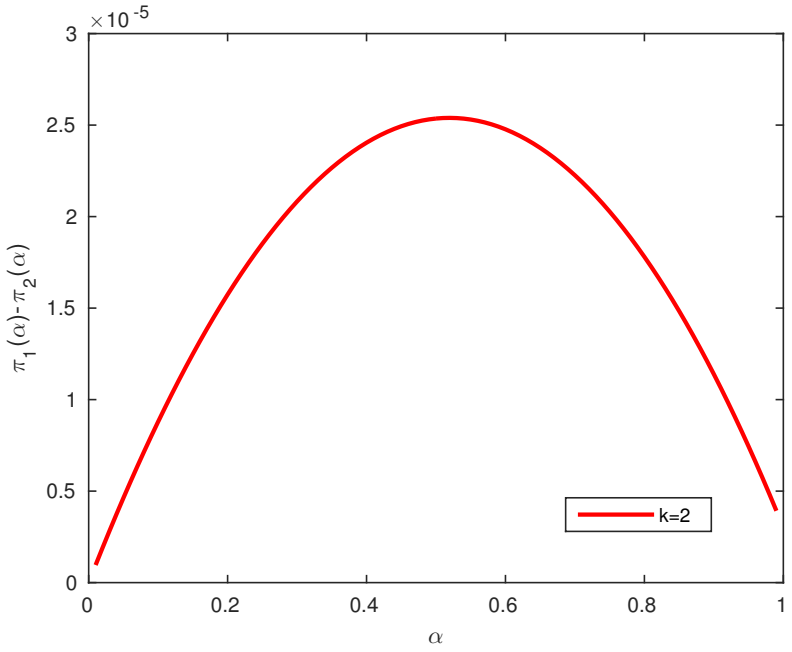


Figure 5: Optimum value of the restart probability for $m = 300$.

The two cases are denoted as 'with degree scaling' and 'without degree scaling' respectively. Finally, if the output of the sweep has more nodes than $m$, we take the first $m$ nodes (i.e., nodes with largest values of the ranking function).

We summarize the values of the error $\mathcal{E}$ in Tables 1 and 2. In Table 1, we choose $\alpha = 0.85$ (initially used by Google webranking [27]). In Table 2, we use the values computed based on the formulae in [2], but with $\alpha$ chosen to be 0.99 (the algorithm does not converge for $\alpha = 0.999$).

| $\alpha = 0.85$, $\epsilon = 10^{-8}$ | without degree scaling | with degree scaling |
|---|---|---|
| PPR | 0.35 | - |
| APPR | 0.49 | 0.724545 |

Table 1: The error $\mathcal{E}$ of Approximate and Exact PPRs for $\alpha = 0.85$

| $\alpha = 0.99$, $\epsilon = 10^{-7}$ | without degree scaling | with degree scaling |
|---|---|---|
| PPR | 0.044 | - |
| APPR | 0.069 | 0.72 |

Table 2: The error $\mathcal{E}$ of Approximate and Exact PPRs for $\alpha = 0.99$

We make the following observations from our simulations. It is clear that APPR is also impacted by the same non-concentration phenomenon as the exact PPR. This is demonstrated in Figure 6 where we plot a realisation of APPR for $k = 2$ and $\alpha = 0.85$. The spikes again correspond to the neighbours of the seed nodes.

When $\alpha = 0.99$ the PPR solution is very close to the stationary distribution of a standard RW which is proportional to degrees. Hence we get almost perfect reconstruction in this case, since the expected degrees of the nodes can be used to cluster the graph nodes efficiently.

This observation stems from the fact that we chose $m$ that grows linearly with $n$ and hence the degrees of the nodes inside the community are sufficiently different from the degrees of the nodes outside it. A more interesting scenario is a situation where $m = o(n)$. In this case, asymptotically the degrees of nodes outside the community and inside the community converge to the same value, making it impossible to detect the community only using the node degrees.

Let us take $m = 200, n = 10000, p = 5\frac{\log^2(n)}{n}$ and $q = 10\frac{\log^2(n)}{n}$. By simply ranking by degrees and choosing the first $m$ nodes, we get error $\mathcal{E} = 0.935$. Notice that if we do random guessing we get an error value $\mathcal{E} = 1 - \frac{m}{n} = 0.98$. Hence ranking based on degrees is almost as bad as random guessing! But using PPR we can get an error of 0.77 with $\alpha = 0.7$ just with 20 seed nodes.

**Algorithm 1** Clustering Algorithm using APPR

1: Compute approximate pagerank vector $\mathrm{pr}(v, \alpha, \epsilon)$ as in [2].
2: Do sweep operation:
3: Sort vertices in decreasing order of $\frac{\mathrm{pr}(v,\alpha,\epsilon)_i}{d_i}$ for $1 \leq i \leq N_p$, where $N_p$ is the maximum size of the subgraph.
4: For the recursive node-set $S_i = \{1, 2, \ldots i\}$ at step $i$, let $\phi_i$ be the conductance. Then $S_{\text{out}} = \arg\min_{i \leq N_p} \phi_i$.
5: Return $S_{\text{out}}$ if $\phi(S_{\text{out}}) < \phi$.

### 6.3 Minimum conductance set

Notice that the conductance of community $\mathcal{C}$, denoted by $\mathrm{cond}(\mathcal{C})$ is given by

$$\begin{aligned} \mathrm{cond}(\mathcal{C}) &= \frac{|\delta E|}{\min(\mathrm{vol}(\mathcal{C}), \mathrm{vol}(\overline{\mathcal{C}}))} \\ &= \frac{\sum_{i=1}^{m} \sum_{j=m+1}^{n} A_{ij}}{\min\left(\sum_{i=1}^{m} d_i, \sum_{i=m+1}^{n} d_i\right)}. \end{aligned} \tag{38}$$

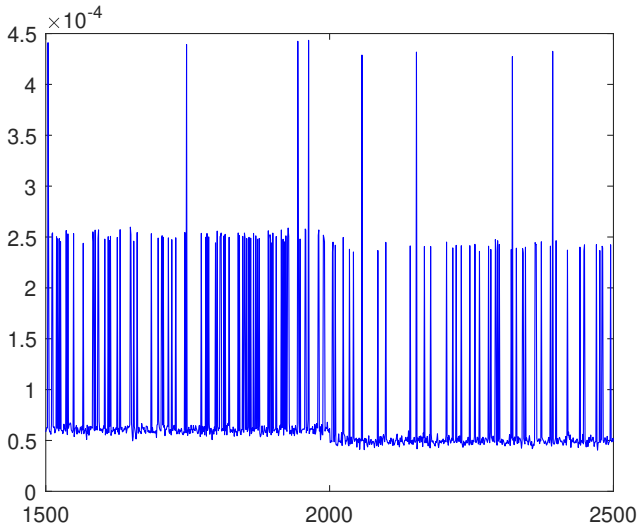


Figure 6: APPR for $k = 2$ and $\alpha = 0.85$.

By virtue of Bernstein's inequality applied to the numerator and the degree concentration lemma applied to the denominator, we can see that the conductance of the community $\mathcal{C}$ converges to and can be well-approximated for a finite $n$ by $\overline{\phi}$ given below.

$$\overline{\phi}(\mathcal{C}) = \frac{\kappa(1-\kappa)p}{\min((\kappa)^2 q, (1-\kappa)^2 p) + \kappa(1-\kappa)p}, \tag{39}$$

where $\kappa := \frac{m}{n}$.

In [10] and [34] it has been observed that the graph conductance has significant limitations as a criterion for graph clustering. We show that in our random graph model the minimization of the graph conductance does not lead to the determination of the natural cluster.

Now, suppose we are looking for a set $\mathcal{A}$ with minimal conductance, so we want to assign fraction $\gamma$ of nodes to $\mathcal{A}$. Assume first that an edge between any two nodes is present with equal probability $p$. Then we have a very simple expression for the 'mean' conductance of $\mathcal{A}$:

$$\overline{\phi}(\mathcal{A}) = \frac{\gamma(1-\gamma)p}{\min(\gamma^2 p, (1-\gamma)^2 p) + \gamma(1-\gamma)p}. \tag{40}$$

It is easy to see that $\overline{\phi}(\mathcal{A})$ is minimized when $\gamma = 1/2$, so that in the denominator we get $\gamma^2 = (1-\gamma)^2$. Now, assume that $q = (1+c)p$, and fraction $\kappa$ of nodes are in a hidden community $\mathcal{C}$. For simplicity of understanding, consider the case when $\gamma > \kappa$, and $\mathcal{C} \subseteq S$. Then in (40) only one term in the denominator will change, namely, it will increase by $\kappa^2 c$:

$$\overline{\phi}(S) = \frac{\gamma(1-\gamma)p}{\min((\gamma^2 p + \kappa^2\, c), (1-\gamma)^2 p) + \gamma(1-\gamma)p}.$$

Clearly, the equality $\gamma^2 p + \kappa^2\, c = (1-\gamma)^2$ will now hold for $\gamma < 1/2$, so the set of minimal conductance will reduce. However, note that the value of $\gamma$, which minimizes conductance, is a continuous function of $c$ and $\kappa$. For example, when $c$ and/or $\kappa$ are small, the conductance will be minimized on a set $S$ that contains almost $1/2$ of the nodes. This explains why the sets with minimal conductance, which we find in our experiments according to Algorithm 1, are typically much larger than $\mathcal{C}$.

Specifically, the conductance of the community returned by exact PPR using only the first $m$ largest elements is 0.6748. The output conductance of the set returned by the sweep algorithm for APPR is 0.0024. However this set is much larger than the target community (its size is 4686). But when this set is truncated to 2000, its conductance becomes 0.6809.

We would like to mention that if one considers the conductance values of communities of size $m$ (like in [23, 28]), then such 'restricted' conductance is minimized on the natural community $\mathcal{C}$, given the community $\mathcal{C}$ is neither too large nor too dense.

Thus, in the context of PPR based local graph clustering, the size of the community (if available) can provide a better guidance than the conductance.

# 7 Conclusions and future research

We analysed a mean-field model of Personalized PageRank on the Erdős-Rényi random graph containing a denser planted Erdős-Rényi subgraph. We also studied the optimization of the damping factor, the only parameter in Personalized PageRank. Our main conclusion is that PPR concentrates in the regime when the community size scales linearly and the number of seed nodes scales sufficiently fast with the size of the graph. We also identify the regime where concentration does not occur. We have also demonstrated that the truncation of APPR does not mitigate the non-concentration of PPR. The main reason for non-concentration of PPR and APPR is the significant leakage of probability mass via the neighbours of the seed nodes. This raises concerns about obtaining high quality local clustering when the number of seed nodes is small. Of course, we have studied a very particular model of a network with community structure. At the same time this model appears to be a very natural benchmark for local graph clustering algorithms. Our concerns complement the limitations of PPR based clustering discussed in [10] and [34]. As in [10, 34, 23], we also note that the plain conductance might not be the best criterion for local graph clustering.

From [17, 18] we know that recovering a hidden community is easy and can be done by light complexity algorithms if the size of the community scales linearly with the size of the graph and this is possible even without using the seed nodes. As our analysis indicates, there are concerns about the applicability of PPR based method in the regime with sublinear scaling of the number of seed nodes. In contrast, belief propagation based algorithms can achieve good detection performance even with a few number of seeds; however, they require good quality seeds and, unlike PPR, they require the knowledge of the graph parameters [24]. Possibly, a combination of these ideas is needed to overcome the limitations of both PPR and belief propagation algorithms.

One more interesting research direction is the extension of the present results to the setting of multiplex networks [12, 19] when several networks represent one actual underlying phenomenon. We expect that using several instances of the same network will significantly improve concentration, and hence the performance, of the PPR based clustering methods.

# Acknowledgements

This work was partly funded by the French Government (National Research Agency, ANR) through the "Investments for the Future" Program reference #ANR-11-LABX-0031-01, Inria - IIT Bombay joint team (grant IFC/DST-Inria-2016-01/448) and by EU COST Project COST-NET (CA15109).

Table 3: Notation

| Symbol | Meaning |
|---|---|
| $V$ | set of nodes |
| $\mathcal{C}$ | planted subgraph (community) |
| $\mathcal{S}$ | set of seed nodes |
| $n$ | size of the graph |
| $m$ | size of the planted subgraph |
| $k$ | number of seed nodes |
| $p$ | probability of edge in the graph |
| $q$ | probability of edge in the subgraph |
| $\mathbf{1}_n$ | vector of ones of dimension $n$ |
| $\mathbf{J}_{m,n}$ | matrix of ones of dimension $m$-by-$n$ |
| $\mathbf{0}_n$ | vector of zeros of dimension $n$ |
| $A$ | adjacency matrix |
| $d = A\mathbf{1}_n$ | vector of nodes' degrees |
| $D = diag\{d\}$ | diagonal matrix of nodes' degrees |
| $P = D^{-1}A$ | transition probability matrix |
| $\pi$ | Personalized PageRank |
| $\alpha$ | damping factor |
| $\bar{A}$ | expected adjacency matrix |
| $\bar{d} = \bar{A}\mathbf{1}_n$ | vector of expected nodes' degrees |
| $\bar{D} = diag\{\bar{d}\}$ | diagonal matrix of expected nodes' degrees |
| $\bar{P} = \bar{D}^{-1}\bar{A}$ | mean-field transition probability matrix |
| $\nu$ | personalization vector or restart distribution |
| $\bar{\pi}$ | mean-field Personalized PageRank |
| $\bar{\pi}_0$ | mean-field Personalized PageRank of a seed node |
| $\bar{\pi}_1$ | mean-field Personalized PageRank of a subgraph node |
| $\bar{\pi}_2$ | mean-field Personalized PageRank of a node outside the subgraph |
| $\mathcal{E}$ | percentage of nodes in $\mathcal{C}$ that are misclassified by the algorithm |
| $f(n) = \omega(g(n))$ | $f$ dominates $g$ asymptotically |
| $f(n) = \Omega(g(n))$ | $f$ is bounded below by $g$ asymptotically |
| $f(n) = \Theta(g(n))$ | $f$ is bounded both above and below by $g$ asymptotically |

## References


[1] Allahverdyan, A.E., Ver Steeg, G. and Galstyan, A.: Community detection with and without prior information. EPL (Europhysics Letters) **90**(1), 18002 (2010).

[2] Andersen, R., Chung, F., Lang, K.: Local graph partitioning using PageRank vectors. In Proceedings of IEEE FOCS'06, 475–486 (2006)

[3] Andersen, R., Lang, K.J.: Communities from seed sets. In Proceedings of ACM WWW'06, 223–232 (2006)

[4] Andersen, R., Gharan, S.O., Peres, Y. and Trevisan, L.: Almost optimum local graph clustering using evolving sets. Journal of the ACM (JACM) **63**(2), p.15 (2016)

[5] Avrachenkov, K., Kadavankandy, A., Prokhorenkova, L.O., Raigorodskii, A.: PageRank in undirected random graphs. Internet Mathematics **13**(1), 10.24166/im.09.2017, (2017)

[6] Avrachenkov, K. and Lebedev, D.: PageRank of scale-free growing networks. Internet Mathematics **3**(2), 207–231 (2006)

[7] Avrachenkov, K., and Litvak, N.: The effect of new links on Google PageRank. Stochastic Models **22.2**, 319–331, (2006).

[8] Avrachenkov, K., Litvak, N., Pham, K.S.: Distribution of PageRank mass among principle components of the web. In Proceedings of WAW, 16–28, (2007)

[9] Avrachenkov, K., Litvak, N., Pham, K.S.: A singular perturbation approach for choosing the PageRank damping factor. Internet Mathematics, **5**(1-2), 47–69 (2008)

[10] Chan, S.O., Kwok, T.C. and Lau, L.C.: Random walks and evolving sets: Faster convergences and limitations. In Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 1849–1865 (2017)

[11] Chen, N., Litvak, N. and Olvera-Cravioto, M.: Generalized PageRank on directed configuration networks. Random Structures & Algorithms **51**(2), 237–274 (2017)

[12] Dickison, M.E., Magnani, M., Rossi, L.: Multilayer social networks. Cambridge University Press (2016)

[13] Fortunato, S., Boguñá, M., Flammini, A., Menczer, F.: On local estimations of PageRank: A mean field approach. Internet Mathematics **4**(2-3), 245–266 (2007)

[14] Garavaglia, A., van der Hofstad, R. and Litvak, N.: Local weak convergence for PageRank. arXiv preprint arXiv:1803.06146 (2018)

[15] Gleich, D., and Mahoney, M.: Anti-differentiating approximation algorithms: A case study with min-cuts, spectral, and flow. in Proceedings of International Conference on Machine Learning (ICML), 1018–1025 (2014)

[16] Golub, G.H., Van Loan, C.F.: Matrix computations. 4th ed. The Johns Hopkins University Press, Baltimore (2013)

[17] Hajek, B., Wu, Y., and Xu, J.: Recovering a hidden community beyond the spectral limit in $O(|E|\log^*|V|)$ time." arXiv preprint arXiv:1510.02786 (2015).

[18] Hajek, B., Wu, Y., Xu, J.: Semidefinite programs for exact recovery of a hidden community. in Proceedings of COLT'16, 1–44 (2016)

[19] Halu, A., Mondragón, R.J., Panzarasa, P. and Bianconi, G.: Multiplex pagerank. PloS one, **8**(10), e78293 (2013)

[20] Haveliwala, T.H.: Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. IEEE Trans. on Knowledge and Data Engineering, **15**(4), 784–796 (2003)

[21] Van Der Hofstad, R. Random graphs and complex networks, Cambridge University Press, (2016)

[22] Jelenković, P.R. and Olvera-Cravioto, M.: Information ranking and power laws on trees. Advances in Applied Probability **42**(4), 1057–1093 (2010)

[23] Jeub, L. G., Balachandran, P., Porter, M. A., Mucha, P. J., and Mahoney, M. W.: Think locally, act locally: Detection of small, medium-sized, and large communities in large networks. Physical Review E, **91**(1), 012821 (2015)

[24] Kadavankandy, A., Avrachenkov, K., Cottatellucci, L., Sundaresan, R.: The power of side-information in subgraph detection. IEEE Trans. Signal Processing, **66**(7), 1905–1919 (2018)

[25] Kadavankandy, A., Cottatellucci, L., Avrachenkov, K.: Characterization of $L^1$-norm statistic for anomaly detection in Erdős Rényi graphs. In Proceedings of IEEE CDC'16, 4600–4605 (2016)

[26] Kloumann, I.M., Ugander, J. and Kleinberg, J.: Block models and personalized PageRank. Proceedings of the National Academy of Sciences, p.201611275 (2016).

[27] Langville, A.N. and Meyer, C.D.: Deeper inside pagerank. Internet Mathematics **1**.3, 335-380 (2004)

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

[29] Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. Stanford InfoLab Research Report, (1999)

[30] Spielman, D.A., Teng, S.-H.: Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In Proceedings of STOC 2004, 81-90 (2004)

[31] Spielman, D.A., Teng, S.-H.: A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. SIAM J. Comput. **42**(1), 1-26 (2013)

[32] Volkovich, Y., Litvak, N.: Asymptotic analysis for personalized web search. Advances in Applied Probability **42**(2), 577–604 (2010)

[33] Zhang, P., Moore, C. and Zdeborova, L. Phase transitions in semisupervised clustering of sparse networks. Physical Review E, **90**(5), 052802 (2014).

[34] Zhu, Z.A., Lattanzi, S. and Mirrokni, V.S.: A Local Algorithm for Finding Well-Connected Clusters. In Proceedings of ICML, **3**, 396–404 (2013).