PageRank Nibble on the sparse directed stochastic block model*

Sayan Banerjee, Prabhanka Deka, and Mariana
Olvera-Cravioto[0000−0003−3335−759]

University of North Carolina at Chapel Hill, Chapel Hill NC 27514, USA
{sayan,deka,molvera}@email.unc.edu

Abstract. We present new results on community recovery based on the PageRank Nibble algorithm on a sparse directed stochastic block model (dSBM). Our results are based on a characterization of the local weak limit of the dSBM and the limiting PageRank distribution. This characterization allows us to estimate the probability of misclassification for any given connection kernel and any given number of seeds (vertices whose community label is known). The fact that PageRank is a local algorithm that can be efficiently computed in both a distributed and asynchronous fashion, makes it an appealing method for identifying members of a given community in very large networks where the identity of some vertices is known.

Keywords: PageRank Nibble · directed stochastic block model · local weak convergence · community detection.

1 Introduction

Many real-world networks exhibit community structure, where members of the same community are more likely to connect to each other than to members of different communities. Stochastic block models are frequently used to model random graphs with a community structure, and there are many problems where the goal is to identify the members of a given community, often based only on the graph structure, i.e,, on the vertices and the existing edges among them. A two community symmetric SBM is described by two parameters α and β, which determine the edge probabilities, with α corresponding to the probability that two members from the same community connect to each other, and β to the probability that two members from different communities connect to each other. In [6], the authors work on the semi-sparse regime $\alpha = a \log n/n$ and $\beta = b \log n/n$, where n is the number of vertices in the graph, and show that the exact recovery of communities is efficiently possible if $|\sqrt{a} - \sqrt{b}| > 2$ and impossible otherwise. When recovery is possible, the authors use spectral methods to get an initial guess of the partition and fine tune it to retrieve the communities. Similar

---

work has been done in the sparse regime, where $\alpha = a/n$ and $\beta = b/n$. In [7], the authors show that recovery is impossible when $(a-b)^2 < 2(a+b)$. In [8], it was proved that recovery is efficiently possible when $(a-b)^2 > 2(a+b)$ through the use of the spectral properties of a modified adjacency matrix B that counts the number of self avoiding paths of a given length l between two vertices in the graph. Further, the authors of [9] show that it is possible to recover a fraction $(1-\gamma)$ of the vertices of community 1 if a and b are sufficiently large and satisfy $(a-b)^2 > K_1 \log(\gamma^{-1})(a+b)$ for some constant K_1 . The clustering methods in [6,8,9] all rely on finding eigenvectors of the adjacency matrix (or a modified adjacency matrix), which is computationally expensive for large networks.

Although the literature on community detection is vast, and there are in fact many methods that work remarkably well, many of those methods become computationally costly for very large networks. In some important cases like the web graph and social media networks, the networks of interest are so large and constantly changing that it becomes difficult to implement some of these methods. Moreover, in many cases, one has more information about the network than just its structure, e.g., vertex attributes that tell us the community to which certain vertices belong to. The question is then whether one can leverage knowledge of such vertices to help identify other members of their community using a computationally efficient method that does not require information about the entire network. One such problem was studied in [12], where the authors consider community detection in a dense (average degree of vertices scale linearly with the size of the network) SBM in which information about the presence or absence of each edge was hidden at random. Here, we will analyze a setting where the labels of some prominent members of the community of interest are known.

The PageRank Nibble algorithm was introduced in [11] as a modification of the Nibble algorithm described in [10] that uses personalized PageRank. This algorithm provides a cheap method for identifying the members of one community when a number of individuals in that community have been identified. PageRank based clustering methods were also proposed in [4] for the two-commmunity SBM, as a special case of a more general method of combining random walk probabilities using a "discriminant" function.

The intuition behind PageRank Nibble is that random walks that start with the individuals that are known to belong to the community we seek will tend to visit more often members of that same community. PageRank Nibble works by choosing the personalization parameter of the known individuals, which we refer to as the "seeds", to be larger than for all other vertices in the network, and then choosing a damping factor c sufficiently far from either 0 or 1. This choice of the personalization values makes the PageRanks of close neighbors of the seeds to be larger, compared to those of individuals outside the community. Once the ranks produced by PageRank Nibble have been computed, a simple threshold rule can be used to identify the likely members of the community of interest. PageRank based methods can generally be executed quickly due to the availability of fast, distributed algorithms [13].

PageRank Nibble on the undirected SBM was studied in [1] under regimes where personalized PageRank (PPR) concentrates around its mean field approximation. The idea proposed there was to use the mean field approximation to identify vertices belonging to the same community as the seeds. In particular, the authors of [1] show that concentration occurs provided the average degrees grow as $a(n) \log n$ for some $a(n) \to \infty$ as $n \to \infty$, and is impossible for the sparse regime where average degrees remain constant as the network size grows. Our present work focuses on the directed stochastic block model (dSBM) in the sparse regime, and our results are based on the existence of a local weak limit and, therefore, of a limiting PageRank distribution. Once we have this characterization, we can compute the probability that an individual will be correctly or incorrectly classified, and choose the threshold that minimizes the misclassification probability.

2   Main Results

Let $\mathcal{G}_n = G(V_n, E_n)$ be a dSBM on the vertex set $V_n = \{1, \ldots, n\}$ with two communities. To start, each vertex $v \in V_n$ is assigned a latent label $C_v \in \{1, 2\}$ identifying its community. We assume that these labels are unknown to us. Denote by $\mathcal{C}_1$ and $\mathcal{C}_2$ the subsets of vertices in communities 1 and 2 respectively. Then, each possible directed edge is sampled independently according to:

$$p_{vw}^{(n)} := \mathbb{P}((v, w) \in E_n | C_v, C_w) = \begin{cases} \frac{a}{n} \wedge 1 & \text{if } C_v = C_w \\ \frac{b}{n} \wedge 1 & \text{if } C_v \neq C_w. \end{cases}$$

The edge probabilities can be written as $p_{vw}^{(n)} = (n^{-1} \kappa_{C_v, C_w}) \wedge 1$, where

$$\kappa = \begin{bmatrix} a & b \\ b & a \end{bmatrix}$$

is called the connection probability kernel for the dSBM.

For $i = 1, 2$, we define

$$\pi_i^{(n)} = \frac{1}{n} \sum_{v=1}^{n} 1(C_v = i)$$

to be the proportion of vertices belonging to each community. We focus specifically on the case where $\pi_1^{(n)} = \pi_2^{(n)} = 1/2$, but the techniques used here can be applied to more general dSBMs.

To describe the setting for our results, start by fixing a constant $0 < s < 1$, and assume there exists a subset $\mathcal{S} \subseteq \mathcal{C}_1$, with $|\mathcal{S}| = n\pi_1^{(n)} s$, for which the community labels are known. In other words, we assume that we know the identities of a fixed, positive proportion of the vertices in community 1. We refer to the vertices in $\mathcal{S}$ as the *seeds*. In a real-world social network one can think of the seeds as famous individuals whose community label or affiliation is known or

easy to infer. Given the seed set $\mathcal{S}$, the goal is to identify the vertices $v \in \mathcal{C}_1 \setminus \mathcal{S}$, i.e., to recover the remaining members of community 1.

In order to describe the PageRank Nibble algorithm, we start first with the definition of personalized PageRank. On a directed graph $G = (V, E)$, the PageRank of vertex $v \in V$ is given by:

$$r_v = c \sum_{w \in V : (w,v) \in E} \frac{1}{D_w^+} r_w + (1 - c)q_v, \tag{1}$$

where D_w^+ is the out-degree of vertex $w \in V$, q_v is the personalization value of vertex v, and $c \in (0, 1)$ is a damping factor.

PageRank is one of most popular measures of network centrality, due to both its computational efficiency (it can be computed in a distributed and asynchronous way), and its ability to identify *relevant* vertices. When $\mathbf{q} = (q_v : v \in V)$ is a probability vector, the PageRank vector $\mathbf{r} = (r_v : v \in V)$ is known to correspond to the stationary distribution of a random walk that, at each time step, chooses, with probability c, to follow an outbound edge uniformly chosen at random, or with probability $1 - c$, chooses its next destination according to $\mathbf{q}$ (if the current vertex has no outbound edges, the random walk always chooses its next destination according to $\mathbf{q}$). PageRank is known to rank highly vertices that either have a large in-degree, or that have close inbound neighbors whose PageRanks are very large [14], hence capturing both popularity and credibility. Since on large networks the PageRank scores will tend to be very small, it is often convenient to work with the scale-free (graph-normalized) PageRank vector $\mathbf{R} = |V|\mathbf{r}$ instead.

For the two community dSBM $G_n = (V_n, E_n)$ described above, let $Q_v = nq_v$ and define

$$\mu_n(B) = \frac{1}{n} \sum_{v=1}^{n} 1((C_v, Q_v) \in B)$$

for any measurable set B. We assume that there exists a limiting measure μ with $\pi_i := \mu(\{i\} \times \mathbb{R}_+) > 0$ for $i = 1, 2$ such that

$$\mu_n \Rightarrow \mu \tag{2}$$

in probability. Here and in the sequel, $\Rightarrow$ denotes weak convergence. Further, for any measurable A, let

$$\sigma_i^{(n)}(A) = \frac{1}{n\pi_i^{(n)}} \sum_{v \in V_n} 1(C_v = i, Q_v \in A), \qquad i = 1, 2, \tag{3}$$

denote the empirical distribution of Q_v conditionally on $C_v = i$ for $i = 1, 2$. Due to assumption (2), we get the existence of limiting distributions σ_i, given by

$$\sigma_i(A) = \frac{\mu(\{i\} \times A)}{\pi_i}, \qquad i = 1, 2,$$

such that $\sigma_i^{(n)} \Rightarrow \sigma_i$ in probability as $n \to \infty$.

As mentioned in the introduction, our analysis is based on the existence of a local weak limit for the dSBM, and the fact that if we let I be uniformly chosen in V_n, independently of $G(V_n, E_n)$, and let R_I denote the scale-free PageRank of vertex I, then R_I converges weakly to a random variable $\mathcal{R}$ as $n \to \infty$. In order to characterize the distribution of $\mathcal{R}$, first define $R^{(1)}$ and $R^{(2)}$ to be random variables satisfying

$$\mathbb{P}\left(R^{(i)} \in \cdot\right) = \mathbb{P}\left(R_I \in \cdot \mid C_I = i\right), \qquad i = 1, 2.$$

Our first result establishes the weak convergence of $R^{(i)}$ for $i = 1, 2$ and characterizes the limiting distributions as the solutions to a system of distributional fixed-point equations.

Theorem 1. *Let $G_n = (V_n, E_n)$ be a sequence of dSBM as described above such that (2) holds. Then, there exist random variables $\mathcal{R}^{(1)}$ and $\mathcal{R}^{(2)}$ such that for any $x \in \mathbb{R}$ that is a point of continuity of the limit,*

$$R^{(i)} \Rightarrow \mathcal{R}^{(i)} \qquad and \qquad \frac{2}{n} \sum_{v \in V_n} 1(R_v \leq x, C_v = i) \xrightarrow{P} \mathbb{P}\left(\mathcal{R}^{(i)} \leq x\right),$$

as $n \to \infty$, $i = 1, 2$. Moreover, the random variables $\mathcal{R}^{(1)}$ and $\mathcal{R}^{(2)}$ satisfy:

$$\mathcal{R}^{(1)} \overset{d}{=} c \sum_{j=1}^{\mathcal{N}^{(11)}} \frac{\mathcal{R}_j^{(1)}}{\mathcal{D}_j^{(1)}} + c \sum_{j=1}^{\mathcal{N}^{(12)}} \frac{\mathcal{R}_j^{(2)}}{\mathcal{D}_j^{(2)}} + (1-c)\mathcal{Q}^{(1)} \tag{4}$$

$$\mathcal{R}^{(2)} \overset{d}{=} c \sum_{j=1}^{\mathcal{N}^{(21)}} \frac{\mathcal{R}_j^{(1)}}{\mathcal{D}_j^{(1)}} + c \sum_{j=1}^{\mathcal{N}^{(22)}} \frac{\mathcal{R}_j^{(2)}}{\mathcal{D}_j^{(2)}} + (1-c)\mathcal{Q}^{(2)} \tag{5}$$

where $\mathcal{Q}^{(1)}$ and $\mathcal{Q}^{(2)}$ are random variables distributed according to σ_1 and σ_2 respectively, $\mathcal{N}^{(kl)}$ are Poisson random variables with means $\pi_l \kappa_{lk}$, $(\mathcal{D}_j^{(i)} - 1 : j \geq 1)$, $i = 1, 2$, are i.i.d. sequences of Poisson random variables with mean $\pi_1 \kappa_{i1} + \pi_2 \kappa_{i2}$, and $(\mathcal{R}_j^{(i)} : j \geq 1)$, $i = 1, 2$, are i.i.d. copies of $\mathcal{R}^{(i)}$, with all random variables independent of each other.

Remark 1. Note that the $(\mathcal{D}_j^{(i)})$ are size-biased Poisson random variables that represent the out-degrees of the inbound neighbors of the explored vertex I.

The above result holds in more generality for a degree-corrected dSBM with k-communities, but for the purposes of this paper, we restrict ourselves to the $k = 2$ case. We will only outline a sketch of the proof, and focus our attention instead on the following theorem about the classification of the vertices.

Equations (4) and (5) are the key behind our classification method. Observe that in the PageRank equations (1), the parameters within our control are the damping factor c and the personalization vector $\mathbf{Q} = (Q_v : v \in V_n)$. If we choose $\mathbf{Q}$ that results in $\mathcal{R}^{(1)} \geq_{\text{s.t.}} \mathcal{R}^{(2)}$, we can identify vertices in community 1 as the

ones having higher PageRank scores. With that in mind, we set $Q_v = 1(v \in \mathcal{S})$, choose an appropriate cutoff point x_0 (which might depend on c, s and κ), and classify as a member of community 1 any vertex $v \in V_n$ such that its scale-free PageRank, R_v, satisfies $R_v > x_0$. The algorithm requires that we choose c sufficiently bounded away from both zero and one, since from the random walk interpretation of PageRank, it is clear that we want to give the random surfer time to explore the local neighborhood, while at the same time ensuring that it returns sufficiently often to the seed set. In practice, a popular choice for the damping factor is $c = 0.85$. In the context of the dSBM, we have that when $a >> b$, the random surfer ends up spending more time exploring the vertices in community 1, and the probability that it escapes to community 2 before jumping back to the seeds is much smaller. As a result, the stationary distribution ends up putting more mass on the community 1 vertices, and the proportion of misclassified vertices diminishes when $a+b$ is large and b/a is close to zero. We formalize this in the theorem below. Note that Theorem 1 gives that the miscalssification probabilities satisfy:

$$\mathbb{P}\left(R_v \leq x_0 | v \in \mathcal{C}_1 \right) \approx \mathbb{P}\left(\mathcal{R}^{(1)} \leq x_0 \right) \quad \text{and}$$

$$\mathbb{P}\left(R_v > x_0 | v \in \mathcal{C}_2 \right) \approx \mathbb{P}\left(\mathcal{R}^{(2)} > x_0 \right).$$

Our local classification algorithm with input parameters c and x_0 is then described as follows:

1. Set $Q_v = 1$ if $v \in \mathcal{S}$, and zero otherwise.
2. Fix the damping factor $c \in (0,1)$ and compute the personalized scale-free PageRank vector $\mathbf{R}$.
3. For a threshold x_0, the estimated members of $\mathcal{C}_1$ are the vertices in the set $\hat{\mathcal{C}}_1(x_0, c) = \{v \in V_n : R_v > x_0\}$.

The theorem below can be used to quantify the damping factor c and the classification threshold x_0, and the corollary that follows shows that the proportion of misclassified vertices becomes small with high probability as $n \to \infty$.

Theorem 2. *Let $G_n = (V_n, E_n)$ be a 2-community dSBM with*

$$\kappa = \begin{bmatrix} a & b \\ b & a \end{bmatrix}$$

and $\pi_1 = \pi_2 = 1/2$. Assume a, b satisfy $8b/(a+b) < 1/2$ and $e^{-(a+b)/2} < b/4a$. Let $Q_v = 1(v \in \mathcal{S})$ for $v \in V_n$, and take any $c \in (1/2, 1 - 8b/(a+b)]$. Then, for $x_0 = 5s/8$, we have

$$\mathbb{P}\left(\mathcal{R}^{(1)} < \frac{5s}{8} \right) \leq \frac{256c^2}{(a+b)(1-c^2)} + \frac{64(1-c)(1-s)}{(1+c)s}, \tag{6}$$

$$\mathbb{P}\left(\mathcal{R}^{(2)} > \frac{5s}{8} \right) \leq \frac{256c^2}{(a+b)(1-c^2)} \left(1 + \frac{(1-c)(1-s)}{2(1+c)s} \right). \tag{7}$$

Naturally, the misclassification errors get smaller as s increases, i.e., as more members of community 1 are known. Also, we get better bounds for the misclassification errors when $a + b$ is large (strong connectivity within a community) and $b/(a+b)$ is small (equivalently, $a/(a+b)$ close to one), i.e., when the network is strongly assortative.

Note that the assumptions in Theorem 2 do not involve s (proportion of seeds). As the proof indicates, our classification errors involve Chebychev bounds which crucially depend on: (i) the mean PageRank scores of the two communities being sufficiently different, and (ii) the ratio of the variance of the PageRank scores of vertices in each community to the square of the mean community PageRank being small. By Lemma 1 below, the ratio of the mean community PageRank scores is independent of s and hence their separation required by (i) is ensured by conditions involving a, b but not s. Moreover, as seen in Lemma 2, the scaled fluctuations in (ii) depend more significantly on the 'sparsity' of the underlying network, quantified by $a + b$ (expected total degree of a vertex), than s. Thus, the dependence on s arises mainly through the choice of the threshold x_0 in our classification algorithm (see Corollary 1).

As a direct corollary to Theorem 2, we have

Corollary 1. *Let $x_0 = 5s/8$, $c \in (1/2, 1 - 8b/(a + b)]$,*

$$\delta_1 = \frac{256c^2}{(a + b)(1 - c^2)} + \frac{64(1 - c)(1 - s)}{(1 + c)s}$$

and

$$\delta_2 = \frac{256c^2}{(a + b)(1 - c^2)} \left(1 + \frac{(1 - c)(1 - s)}{2(1 + c)s} \right).$$

Then, under the hypothesis of Theorem 2, for $\delta = \delta_1 + \delta_2$ and any $\epsilon > 0$, we have

$$\lim_{n \to \infty} \mathbb{P}\left(|\mathcal{C}_1 \triangle \hat{\mathcal{C}}_1(x_0, c)| > \frac{(\delta + \epsilon)n}{2} \right) = 0.$$

Proof. For notational convenience, we drop the dependence of $\hat{\mathcal{C}}_1$ on x_0 and c. Observe that $|\mathcal{C}_1 \triangle \hat{\mathcal{C}}_1| = |\mathcal{C}_1 \backslash \hat{\mathcal{C}}_1| + |\hat{\mathcal{C}}_1 \cap \mathcal{C}_2|$, and we have $\mathcal{C}_1 \backslash \hat{\mathcal{C}}_1 = \{v \in \mathcal{C}_1 : R_v < 5s/8\}$ and $\hat{\mathcal{C}}_1 \cap \mathcal{C}_2 = \{v \in \mathcal{C}_2 : R_v > 5s/8\}$. So we get that for $x_0 = 5s/8$,

$$\mathbb{P}\left(|\mathcal{C}_1 \triangle \hat{\mathcal{C}}_1| > \frac{(\delta + \epsilon)n}{2} \right) = \mathbb{P}\left(\frac{2}{n} \sum_{v \in \mathcal{C}_1} 1(R_v < x_0) + \frac{2}{n} \sum_{v \in \mathcal{C}_2} 1(R_v > x_0) > \delta + \epsilon \right).$$

Then the result follows since

$$\frac{2}{n} \sum_{v \in \mathcal{C}_1} 1(R_v < x_0) + \frac{2}{n} \sum_{v \in \mathcal{C}_2} 1(R_v > x_0)$$

$$\xrightarrow{P} \mathbb{P}(\mathcal{R}^{(1)} < x_0) + \mathbb{P}(\mathcal{R}^{(2)} > x_0) = \delta$$

as $n \to \infty$.

Remark 2. Our proof of Theorem 2 uses Chebyshev's inequalities based on mean and variance bounds for the limiting (scale-free) personalized PageRank distribution obtained from the distributional fixed-point equations in Theorem 1. The choice of x_0 above is rather ad hoc and mainly for simplicity of the associated misclassification error bounds. One can check that the choice of x_0 which minimizes the sum of the Chebyshev error bounds is given by $x_0^* = (r_1 v_2^{1/3} + r_2 v_1^{1/3})/(v_1^{1/3} + v_2^{1/3})$, where r_1, r_2 are the expected limiting PageRank values obtained in Lemma 1 and v_1, v_2 are the corresponding variances obtained in Lemma 2. Further, $x_0 = 5s/8$ is independent of the kernel parameters a and b, which are often unknown in practice. Moreover, although the range of c depends on a, b, the results above hold for any c in the given range. Hence, in practice, when a, b are not known, then any $c > 1/2$ which is not too close to one should work provided the network is not too sparse ($b/(a + b)$ is sufficiently small).

3   Proofs

As mentioned earlier, Theorem 1 holds in considerably more generality than the one stated here, so we will only provide a sketch of the proof that suffices for the simpler dSBM considered here. The proof of Theorem 2 is given later in the section.

Proof. Theorem 1 (Sketch). The proof consists of three main steps.

1. **Establish the local weak convergence of the dSBM:** For the 2-community dSBM considered here, one can modify the coupling in [3] (which works for an undirected SBM) to the exploration of the in-component of a uniformly chosen vertex. The coupled graph is a 2-type Galton-Watson process, with the two types corresponding to the two communities in the dSBM, and all edges directed from offspring to parent. The number of offspring of type j that a node of type i has is a Poisson random variable with mean $m_{ij}^- = \pi_j \kappa_{ji}$ for $j = 1, 2$. For each node $\mathbf{i}$ in the coupled tree, denote by $C_{\mathbf{i}}$ its type, and assign it a mark $\boldsymbol{X_i} = (\mathcal{D}_{\mathbf{i}}, Q_{\mathbf{i}})$, where $(\mathcal{D}_{\mathbf{i}} - 1)|C_{\mathbf{i}} = j$ is a Poisson random variable with mean $m_j^+ = \pi_1 \kappa_{j1} + \pi_2 \kappa_{j2}$, and $Q_{\mathbf{i}}|C_{\mathbf{i}} = j$ has distribution σ_j as defined in (3). The construction of the coupling follows a two step exploration process similar to the one done for inhomogeneous random digraphs in [5]. First the outbound edges of a vertex are explored, followed by the exploration of its inbound neighbors, assigning marks to a vertex once we finish exploring both its inbound and outbound one-step neighbors. This establishes the local weak convergence in probability of the dSBM to the 2-type Galton-Watson process.

2. **Establish the local weak convergence of PageRank:** Once we have the local weak convergence of the dSBM, let $\mathcal{R}^*$ denote the personalized PageRank of the root node of the 2-type Galton-Watson process in the coupling. The local weak convergence in probability of the PageRanks on the dSBM to $\mathcal{R}^*$, i.e.,

$$\frac{1}{n} \sum_{v \in V_n} 1(R_v \leq x) \xrightarrow{P} \mathbb{P}(\mathcal{R}^* \leq x)$$

as $n \to \infty$, follows from Theorem 2.1 in [2]. Note that the random variables $\mathcal{R}^{(1)}$ and $\mathcal{R}^{(2)}$ correspond to the conditional laws of $\mathcal{R}^*$ given that the root has type 1 or type 2, respectively. And since the two communities are assumed to have the same size, the probability that the root has type 1 is $1/2$, hence,

$$\frac{1}{n} \sum_{v \in V_n} 1(R_v \leq x, \, C_v = i) \xrightarrow{P} \mathbb{P}(\mathcal{R}^{(i)} \leq x)\frac{1}{2},$$

as $n \to \infty$. The weak convergence result follows from the bounded convergence theorem.

3. **Derive the distributional fixed point equations:** If the nodes in the first generation of the 2-type Galton-Watson process are labeled $1 \leq j \leq \mathcal{N}$, where $\mathcal{N}$ denotes the number of offspring of the root node, then

$$\mathcal{R}^* = c \sum_{j=1}^{\mathcal{N}} \frac{\mathcal{R}_j}{\mathcal{D}_j} + (1 - c)\mathcal{Q},$$

where $\mathcal{Q}$ denotes the personalization value of the root, $(\mathcal{D}_j : j \geq 1)$ correspond to the out-degrees of the offspring, and the $(\mathcal{R}_j : j \geq 1)$ correspond to their PageRanks. Conditioning on the type of the root, as well as on the types of its offspring, gives the two distributional fixed-point equations (4) and (5). In particular, conditionally on the root having type i, $\mathcal{N}^{(ik)}$ corresponds to the number of offspring of type k, $\mathcal{Q}^{(i)}$ has distribution σ_i, and $\mathcal{D}_1^{(k)}$ and $\mathcal{R}_1^{(k)}$ are independent random variables having the distribution of $\mathcal{D}_1$ and $\mathcal{R}_1$ conditionally on node 1 having type k.

We prove Theorem 2 through the second moment method. First we prove the following lemmas establishing bounds on the mean and variance of $\mathcal{R}^{(1)}$ and $\mathcal{R}^{(2)}$.

Lemma 1. *Let $r_i = \mathbb{E}\left[\mathcal{R}^{(i)}\right]$, $\lambda = 1 - \mathrm{e}^{-(a+b)/2}$ and*

$$\Delta = \left(1 - \frac{c\lambda a}{a + b}\right)^2 - \left(\frac{c\lambda b}{a + b}\right)^2.$$

Then, we have

$$r_1 = \frac{\left(1 - \frac{c\lambda a}{a+b}\right) s(1 - c)}{\Delta} \tag{8}$$

$$r_2 = \frac{\left(\frac{c\lambda b}{a+b}\right) s(1 - c)}{\Delta}. \tag{9}$$

Further, if $1 - \lambda = \mathrm{e}^{-(a+b)/2} \leq b/4a$ and $c > 1/2$, we have the bounds

$$r_1 \geq s\left(1 - \frac{2b}{(1 - c)(a + b)}\right), \tag{10}$$

$$r_2 \leq \frac{s}{2}. \tag{11}$$

Proof. Recall the distributional equations satisfied by $\mathcal{R}^{(1)}$ and $\mathcal{R}^{(2)}$ from Theorem 1. Taking expectation on both sides gives us

$$\mathbb{E}[\mathcal{R}^{(1)}] = c\mathbb{E}\left[\sum_{j=1}^{\mathcal{N}^{(11)}} \frac{\mathcal{R}_j^{(1)}}{\mathcal{D}_j^{(1)}} + \sum_{j=1}^{\mathcal{N}^{(12)}} \frac{\mathcal{R}_j^{(2)}}{\mathcal{D}_j^{(2)}}\right] + (1-c)\mathbb{E}[\mathcal{Q}^{(1)}],$$

$$\mathbb{E}[\mathcal{R}^{(2)}] = c\mathbb{E}\left[\sum_{j=1}^{\mathcal{N}^{(21)}} \frac{\mathcal{R}_j^{(1)}}{\mathcal{D}_j^{(1)}} + \sum_{j=1}^{\mathcal{N}^{(22)}} \frac{\mathcal{R}_j^{(2)}}{\mathcal{D}_j^{(2)}}\right] + (1-c)\mathbb{E}[\mathcal{Q}^{(2)}].$$

First, note that with our choice of $\mathcal{Q}$, $\mathbb{E}[\mathcal{Q}^{(1)}] = s$ and $\mathbb{E}[\mathcal{Q}^{(2)}] = 0$. Further $(\mathcal{R}_j^{(i)}, \mathcal{D}_j^{(i)})_{j\geq 1}$ (resp. $(\mathcal{R}_j^{(i)}, \mathcal{D}_j^{(i)})_{j\geq 1}$) are independent of $\mathcal{N}^{(1i)}$ (resp. $\mathcal{N}^{(2i)}$), and of each other, for $i = 1, 2$. So the above expressions can be simplified to

$$r_1 = c\left(\mathbb{E}[\mathcal{N}^{(11)}]\mathbb{E}\left[\frac{1}{\mathcal{D}^{(1)}}\right]r_1 + \mathbb{E}[\mathcal{N}^{(12)}]\mathbb{E}\left[\frac{1}{\mathcal{D}^{(2)}}\right]r_2\right) + (1-c)s,$$

$$r_2 = c\left(\mathbb{E}[\mathcal{N}^{(21)}]\mathbb{E}\left[\frac{1}{\mathcal{D}^{(1)}}\right]r_1 + \mathbb{E}[\mathcal{N}^{(22)}]\mathbb{E}\left[\frac{1}{\mathcal{D}^{(2)}}\right]r_2\right),$$

where $\mathcal{N}^{(ij)}$ and $(\mathcal{D}^{(i)} - 1)$ are Poisson random variables with means as described in Theorem 1. Therefore we can further reduce the equations to

$$r_1 = c\left(\frac{a}{2} \cdot \frac{(1 - \mathrm{e}^{-(a+b)/2})}{(a+b)/2} \cdot r_1 + \frac{b}{2} \cdot \frac{(1 - \mathrm{e}^{-(a+b)/2})}{(a+b)/2} \cdot r_2\right) + (1-c)s,$$

$$r_2 = c\left(\frac{b}{2} \cdot \frac{(1 - \mathrm{e}^{-(a+b)/2})}{(a+b)/2} \cdot r_1 + \frac{a}{2} \cdot \frac{(1 - \mathrm{e}^{-(a+b)/2})}{(a+b)/2} \cdot r_2\right),$$

or in matrix form, and after substituting $\lambda = (1 - \mathrm{e}^{-(a+b)/2})$,

$$\begin{bmatrix} 1 - \frac{ca\lambda}{a+b} & -\frac{cb\lambda}{a+b} \\ -\frac{cb\lambda}{a+b} & 1 - \frac{ca\lambda}{a+b} \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \begin{bmatrix} (1-c)s \\ 0 \end{bmatrix}. \tag{12}$$

Solving (12), we get

$$\begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} (1 - c\lambda a/(a+b))\, s(1-c) \\ c\lambda bs(1-c)/(a+b) \end{bmatrix},$$

where

$$\Delta = \left(1 - \frac{c\lambda a}{a+b}\right)^2 - \left(\frac{c\lambda b}{a+b}\right)^2$$

as required. Note that since $c\lambda < 1$, we have $\Delta > 0$, and so the above quantities are well defined. From here, the bound for r_2 is a straightforward calculation.

$$r_2 = \frac{\frac{c\lambda b}{a+b}(1-c)s}{\Delta} \leq \frac{\frac{b}{a+b}(1-c)s}{(1-c\lambda)\left(1 - c\lambda\frac{a-b}{a+b}\right)} \leq \frac{sb}{(a+b)} \frac{1}{\left(1 - \frac{a-b}{a+b}\right)} = \frac{s}{2}.$$

To get the bound for r_1, we proceed as follows

$$r_1 = \frac{\left(1 - \frac{c\lambda a}{a+b}\right)(1-c)s}{\Delta}$$

$$\geq \frac{s(1-c)}{1 - \frac{c\lambda a}{a+b}} = \frac{s(1-c)}{\frac{a+b}{a+b} - \frac{c\lambda a}{a+b}} = \frac{s(1-c)}{\frac{b}{a+b} + \frac{a}{a+b}(1 - \lambda + \lambda(1-c))}$$

$$\geq \frac{s(1-c)}{\frac{b}{a+b} + \frac{a}{a+b}\left(e^{-(a+b)/2} + (1-c)\right)} = \frac{s(1-c)}{(1-c) + \frac{cb}{a+b} + \frac{a}{a+b}e^{-(a+b)/2}}$$

$$\geq \frac{s(1-c)}{1 - c + \frac{2bc}{a+b}} \geq s\left(1 - \frac{2b}{(1-c)(a+b)}\right),$$

where for the last inequality we used fact that $e^{-(a+b)/2}a/(a+b) \leq b/4(a+b) \leq cb/(a+b)$ due to our assumptions on λ and c, and $1 - x^2 \leq 1$ for all $x \in \mathbb{R}$. This completes the proof.

The next lemma provides a result for the variances.

Lemma 2. *Define* $v_i = \text{Var}(\mathcal{R}^{(i)})$ *for* $i = 1, 2$, *then, if we let* $\mathbf{v} = (v_1, v_2)'$, *and* $\mathbf{r^2} = (r_1^2, r_2^2)'$, *then*

$$\mathbf{v} = \frac{1}{2(1-g_1)(1-g_2)}\left(K\mathbf{r^2} + (1-c)^2 s(1-s)\mathbf{k}\right),$$

where $g_1 = c^2 \mathbb{E}[1/(\mathcal{D}^{(1)})^2](a-b)/2$, $g_2 = c^2 \mathbb{E}[1/(\mathcal{D}^{(1)})^2](a+b)/2$,

$$K = \begin{bmatrix} g_1 + g_2 - 2g_1 g_2, & g_2 - g_1 \\ g_2 - g_1, & g_1 + g_2 - 2g_1 g_2 \end{bmatrix}, \quad and \quad \mathbf{k} = \begin{bmatrix} 2 - g_1 - g_2 \\ g_2 - g_1 \end{bmatrix}.$$

Furthermore,

$$v_1 \leq \frac{4c^2 s^2}{(a+b)(1-c^2)} + \frac{1-c}{1+c}s(1-s),$$

$$v_2 \leq \frac{4c^2 s^2}{(a+b)(1-c^2)}\left(1 + \frac{(1-c)(1-s)}{2s(1+c)}\right).$$

Proof. To calculate the variance of $\mathcal{R}^{(1)}$ and $\mathcal{R}^{(2)}$, we will rely on the law of total variances, i.e., for any two random variables X and Y,

$$\text{Var}(X) = \text{Var}[\mathbb{E}(X|Y)] + \mathbb{E}[\text{Var}(X|Y)].$$

Applying this to equation (4), along with the fact that $r_i < 1$ for $i = 1, 2$, we get the following bound for $\text{Var}(\mathcal{R}^{(1)})$:

$$\text{Var}(\mathcal{R}^{(1)}) = c^2 \text{Var}\left(r_1 \mathcal{N}^{(11)} \mathbb{E}\left[\frac{1}{\mathcal{D}^{(1)}}\right] + r_2 \mathcal{N}^{(12)} \mathbb{E}\left[\frac{1}{\mathcal{D}^{(2)}}\right]\right)$$

$$+ c^2 \mathbb{E}\left[\mathcal{N}^{(11)} \text{Var}\left(\frac{\mathcal{R}^{(1)}}{\mathcal{D}^{(1)}}\right) + \mathcal{N}^{(12)} \text{Var}\left(\frac{\mathcal{R}^{(2)}}{\mathcal{D}^{(2)}}\right)\right] + (1-c)^2 \text{Var}(\mathcal{Q}^{(1)}).$$

Now use the observation that $\text{Var}(\mathcal{N}^{(11)}) = \mathbb{E}[\mathcal{N}^{(11)}] = a/2$, $\text{Var}(\mathcal{N}^{(12)}) = \mathbb{E}[\mathcal{N}^{(12)}] = b/2$, and $\text{Var}(\mathcal{Q}^{(1)}) = s(1-s)$, to obtain that for $v_i = \text{Var}(\mathcal{R}^{(i)})$,

$$
\begin{aligned}
v_1 &= c^2 \left(r_1^2 \cdot \frac{a}{2} \cdot \left(\mathbb{E}\left[\frac{1}{\mathcal{D}^{(1)}} \right] \right)^2 + r_2^2 \cdot \frac{b}{2} \cdot \left(\mathbb{E}\left[\frac{1}{\mathcal{D}^{(2)}} \right] \right)^2 \right) \\
&\quad + c^2 \left(\frac{a}{2} \text{Var}\left(\frac{\mathcal{R}^{(1)}}{\mathcal{D}^{(1)}} \right) + \frac{b}{2} \text{Var}\left(\frac{\mathcal{R}^{(2)}}{\mathcal{D}^{(2)}} \right) \right) + (1-c)^2 s(1-s) \\
&= c^2 \left(r_1^2 \cdot \frac{a}{2} \cdot \left(\mathbb{E}\left[\frac{1}{\mathcal{D}^{(1)}} \right] \right)^2 + r_2^2 \cdot \frac{b}{2} \cdot \left(\mathbb{E}\left[\frac{1}{\mathcal{D}^{(2)}} \right] \right)^2 \right) + (1-c)^2 s(1-s) \\
&\quad + \frac{c^2 a}{2} \left(\text{Var}\left(\frac{1}{\mathcal{D}^{(1)}} r_1 \right) + \mathbb{E}\left[\frac{1}{(\mathcal{D}^{(1)})^2} \text{Var}(\mathcal{R}^{(1)}) \right] \right) \\
&\quad + \frac{c^2 b}{2} \left(\text{Var}\left(\frac{1}{\mathcal{D}^{(2)}} r_2 \right) + \mathbb{E}\left[\frac{1}{(\mathcal{D}^{(2)})^2} \text{Var}(\mathcal{R}^{(2)}) \right] \right) \\
&= c^2 \left(r_1^2 \cdot \frac{a}{2} \cdot \left(\mathbb{E}\left[\frac{1}{\mathcal{D}^{(1)}} \right] \right)^2 + r_2^2 \cdot \frac{b}{2} \cdot \left(\mathbb{E}\left[\frac{1}{\mathcal{D}^{(2)}} \right] \right)^2 \right) + (1-c)^2 s(1-s) \\
&\quad + \frac{c^2 a}{2} \left(r_1^2 \text{Var}\left(\frac{1}{\mathcal{D}^{(1)}} \right) + \mathbb{E}\left[\frac{1}{(\mathcal{D}^{(1)})^2} \right] v_1 \right) \\
&\quad + \frac{c^2 b}{2} \left(r_2^2 \text{Var}\left(\frac{1}{\mathcal{D}^{(2)}} \right) + \mathbb{E}\left[\frac{1}{(\mathcal{D}^{(2)})^2} \right] v_2 \right) \\
&= (1-c)^2 s(1-s) + \frac{c^2 a}{2} \mathbb{E}\left[\frac{1}{(\mathcal{D}^{(1)})^2} \right] (r_1^2 + v_1) + \frac{c^2 b}{2} \mathbb{E}\left[\frac{1}{(\mathcal{D}^{(2)})^2} \right] (r_2^2 + v_2).
\end{aligned}
$$

Similarly, using $\mathcal{Q}^{(2)} \equiv 0$ and

$$
\begin{aligned}
v_2 &= c^2 \text{Var}\left(r_1 \mathcal{N}^{(21)} \mathbb{E}\left[\frac{1}{\mathcal{D}^{(1)}} \right] + r_2 \mathcal{N}^{(22)} \mathbb{E}\left[\frac{1}{\mathcal{D}^{(2)}} \right] \right) \\
&\quad + c^2 \mathbb{E}\left[\mathcal{N}^{(21)} \text{Var}\left(\frac{\mathcal{R}^{(1)}}{\mathcal{D}^{(1)}} \right) + \mathcal{N}^{(22)} \text{Var}\left(\frac{\mathcal{R}^{(2)}}{\mathcal{D}^{(2)}} \right) \right] + (1-c)^2 \text{Var}(\mathcal{Q}^{(2)}) \\
&= \frac{c^2 b}{2} \mathbb{E}\left[\frac{1}{(\mathcal{D}^{(1)})^2} \right] (r_1^2 + v_1) + \frac{c^2 a}{2} \mathbb{E}\left[\frac{1}{(\mathcal{D}^{(2)})^2} \right] (r_2^2 + v_2).
\end{aligned}
$$

Writing the above in matrix notation we obtain for $\mathbf{v} = (v_1, v_2)'$ and $\mathbf{r^2} = (r_1^2, r_2^2)'$,

$$
\mathbf{v} = c^2 M (\mathbf{v} + \mathbf{r^2}) + \mathbf{h},
$$

where (note that $\mathcal{D}^{(1)} \overset{d}{=} \mathcal{D}^{(2)}$),

$$
M = \frac{\mathbb{E}\left[1/(\mathcal{D}^{(1)})^2 \right]}{2} \begin{bmatrix} a & b \\ b & a \end{bmatrix} \qquad \text{and} \qquad \mathbf{h} = \begin{bmatrix} (1-c)^2 s(1-s) \\ 0 \end{bmatrix}.
$$

Moreover, use the observation that

$$M = BAB, \quad A = \frac{\mathbb{E}[1/(\mathcal{D}^{(1)})^2]}{2} \begin{bmatrix} (a-b) & 0 \\ 0 & (a+b) \end{bmatrix}, \quad B = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix},$$

so the maximum eigenvalue of M is $\mathbb{E}[1/(\mathcal{D}^{(1)})^2]\mathbb{E}[\mathcal{D}^{(1)} - 1]$. Since for a Poisson random variable N with mean μ we have that

$$\mathbb{E}[1/(N+1)^2]E[N] = \sum_{n=1}^{\infty} \frac{e^{-\mu}\mu^n}{n \cdot n!} = \mathbb{E}\left[\frac{1}{N \vee 1}\right] - e^{-\mu} \leq 1, \tag{13}$$

then the matrix $I - c^2 M$ is invertible, and we obtain

$$\mathbf{v} = (I - c^2 M)^{-1}(c^2 M \mathbf{r^2} + \mathbf{b}) = \sum_{k=0}^{\infty} c^{2k} M^k (c^2 M \mathbf{r^2} + \mathbf{b})$$

$$= B \begin{bmatrix} \frac{c^2 A_{11}}{1 - c^2 A_{11}} & 0 \\ 0 & \frac{c^2 A_{22}}{1 - c^2 A_{22}} \end{bmatrix} B \mathbf{r^2} + B \begin{bmatrix} \frac{1}{1 - c^2 A_{11}} & 0 \\ 0 & \frac{1}{1 - c^2 A_{22}} \end{bmatrix} B \mathbf{b}.$$

Setting $g_i = c^2 A_{ii}$ for $i = 1, 2$, and computing the product of matrices gives:

$$\mathbf{v} = \frac{1}{2(1 - g_1)(1 - g_2)} \left(K\mathbf{r^2} + (1-c)^2 s(1-s)\mathbf{k} \right),$$

for K and $\mathbf{k}$ defined in the statement of the lemma.

Further, if we let $\Delta_2 = 2(1 - g_1)(1 - g_2)$ and expand the above equation, we obtain

$$\mathbf{v} = \frac{1}{\Delta_2} \left(\begin{bmatrix} (g_1 + g_2 - 2g_1 g_2)r_1^2 + (-g_1 + g_2)r_2^2 \\ (-g_1 + g_2)r_1^2 + (g_1 + g_2 - 2g_1 g_2)r_2^2 \end{bmatrix} + (1-c)^2 s(1-s) \begin{bmatrix} (2 - (g_1 + g_2)) \\ (-g_1 + g_2) \end{bmatrix} \right).$$

From equations (8) and (9) we also get that $r_i \leq s$ for $i = 1, 2$, so we can reduce this to

$$\mathbf{v} \leq \frac{1}{\Delta_2} \begin{bmatrix} 2g_2(1 - g_1)s^2 + (2 - (g_1 + g_2))(1-c)^2 s(1-s) \\ 2g_2(1 - g_1)s^2 + (-g_1 + g_2)(1-c)^2 s(1-s) \end{bmatrix}.$$

Plugging in $\Delta_2 = 2(1 - g_1)(1 - g_2)$, and noting that $g_2 \geq g_1$, we get

$$v_1 \leq \frac{g_2 s^2}{1 - g_2} + \frac{1}{2}\left(\frac{1}{1 - g_2} + \frac{1}{1 - g_1}\right)(1-c)^2 s(1-s)$$

$$\leq \frac{g_2 s^2}{1 - g_2} + \frac{1}{1 - g_2}(1-c)^2 s(1-s),$$

and

$$v_2 \leq \frac{g_2 s^2}{1 - g_2} + \frac{1}{2}\left(\frac{1}{1 - g_2} - \frac{1}{1 - g_1}\right)(1-c)^2 s(1-s)$$

$$\leq \frac{g_2 s^2}{1 - g_2} + \frac{g_2}{2(1 - g_1)(1 - g_2)}(1 - c)^2 s(1 - s).$$

Finally, using $\mathbb{E}[1/(\mathcal{D}^{(1)})^2] \leq 8/(a+b)^2$ and (13), we have $g_2 \leq \min\{c^2, 4c^2/(a+b)\}$, and so

$$v_1 \leq \frac{4c^2 s^2}{(a + b)(1 - c^2)} + \frac{1 - c}{1 + c} s(1 - s),$$

$$v_2 \leq \frac{4c^2 s^2}{(a + b)(1 - c^2)}\left(1 + \frac{(1 - c)(1 - s)}{2s(1 + c)}\right).$$

We are now ready to prove Theorem 2.

Proof (Proof of Theorem 2). For any $z > 0$, Chebyshev's inequality gives

$$\mathbb{P}(\mathcal{R}^{(1)} \leq r_1 - z) = \mathbb{P}(\mathcal{R}^{(1)} - r_1 \leq -z)$$
$$\leq \frac{v_1}{z^2}$$
$$\leq \frac{1}{z^2}\left(\frac{4c^2 s^2}{(a + b)(1 - c^2)} + \frac{1 - c}{1 + c} s(1 - s)\right). \qquad (14)$$

A similar application of Chebyshev's inequality for any $w > 0$ with $\mathcal{R}^{(2)}$ gives

$$\mathbb{P}\left(\mathcal{R}^{(2)} > \frac{s}{2} + w\right) \leq \mathbb{P}(\mathcal{R}^{(2)} > r_2 + w) \leq \frac{v_2}{w^2}$$
$$\leq \frac{1}{w^2}\frac{4c^2 s^2}{(a + b)(1 - c^2)}\left(1 + \frac{(1 - c)(1 - s)}{2s(1 + c)}\right), \qquad (15)$$

where the first inequality follows from equation (11). Choosing $c \in (1/2, 1 - 8b/(a + b)]$ results in $r_1 \geq 3s/4$, so choosing $z = w = s/8$ and plugging into the bounds from equations (14) and (15) gives

$$\mathbb{P}\left(\mathcal{R}^{(1)} < \frac{5s}{8}\right) \leq \frac{256c^2}{(a + b)(1 - c^2)} + \frac{64(1 - c)}{1 + c} \cdot \frac{1 - s}{s},$$

$$\mathbb{P}\left(\mathcal{R}^{(2)} > \frac{5s}{8}\right) \leq \frac{256c^2}{(a + b)(1 - c^2)}\left(1 + \frac{(1 - c)(1 - s)}{2s(1 + c)}\right).$$

4   Results from simulations

We illustrate the algorithm with some simulation experiments. First, we calculated the personalized PageRank scores for a 2-community dSBM with $n = 20000$ vertices, $a = 150$, $b = 10$, $s = .2$ and $c = .85$. The plot shows a clear separation of the PPR scores of the seeds, the rest of community 1 and the vertices in community 2.

We also investigated the role of the damping factor c and the best way to choose it. One natural way of doing so is to find the value of c that maximizes
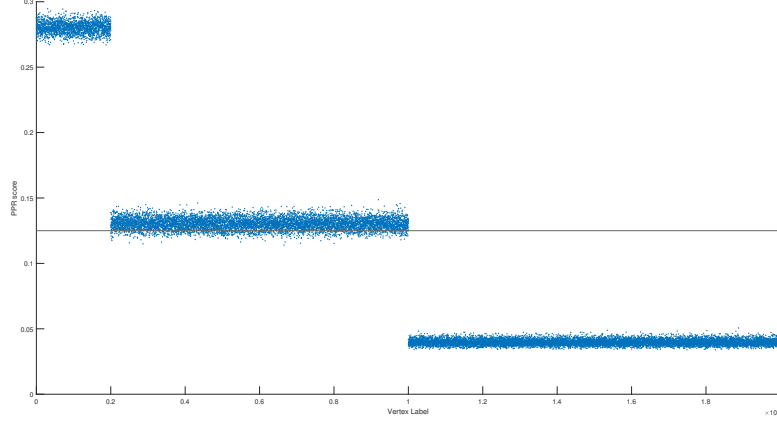
Fig. 1. A plot of the personalized PageRank scores for a 2-community dSBM with $a = 150$, $b = 10$, $n = 20000$, $s = 0.2$, and $c = .85$. The first 2000 vertices are the seeds, vertices 2001-10000 are the rest of community 1, and vertices 10001-20000 are community 2. The horizontal black line is our cutoff level $5s/8$. Proportion of misclassified community 1 vertices is 0.0935.

the difference between the mean PPR scores for the two communities. Note that $r_1 - r_2$ is strictly monotone in c, but if we let $\hat{r}_1$ to be the mean of the non-seeded members of community 1, we see in Fig. 2 that $\hat{r}_1 - r_2$ is strictly convex with a maximum attained at $c = .86$. We have a description for the optimal c^* as follows.

Lemma 3. *Let $\hat{r}_1$ and r_2 be as described above. Then*

$$c^* := \text{argmax}_c\{\hat{r}_1(c) - r_2(c)\} = \frac{1 - \sqrt{1 - E}}{E},$$

where

$$E = \frac{a - b}{a + b}\left(1 - e^{-(a+b)/2}\right).$$

Proof. To calculate $\hat{r}_1$, we consider the dSBM to have 3 communities, where we separate the seeds and the rest of the vertices in community 1. Then, Theorem 1 gives us a system of 3 distributional fixed-point equations. Using those, and calculations similar to the ones we did for Lemma 1, we get

$$\hat{r}_1 = (1 - c)s\left(\frac{1 - \frac{c\lambda a}{a+b}}{(1 - c\lambda)\left(1 - c\lambda\left(\frac{a-b}{a+b}\right)\right)} - 1\right) \tag{16}$$

$$r_2 = (1 - c)s\left(\frac{\frac{c\lambda b}{a+b}}{(1 - c\lambda)\left(1 - c\lambda\frac{a-b}{a+b}\right)}\right).$$

Now substitute $E = \lambda(a - b)/(a + b)$ to obtain that

$$\hat{r}_1 - r_2 = \frac{(1 - c)s}{1 - cE},$$

and use calculus to compute the optimal

$$c^* = \frac{1 - \sqrt{1 - E}}{E}.$$

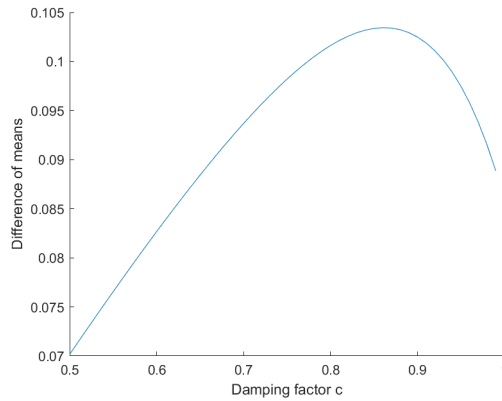Note that the value E is the second eigenvalue of the matrix on the left hand side of equation (12).

Fig. 2. Plot of $\hat{r}_1 - r_2$ as c varies from .5 to 1 for a smaller dSBM with $n = 2000$, $a = 100$, $b = 2$ and $s = .15$.

5   Remarks and conclusions

In the sparse regime, we have proposed a cutoff level to identify vertices of community 1 based on their personalized PageRank scores and provided theoretical bounds on the probability of misclassifying a vertex. Our bounds are not tight, and simulations indicate that we might be able to use a lower threshold to further reduce the error (see also Remark 2). Another possible threshold option in the case of the symmetric SBM ($\pi_1 = \pi_2$) is the median of PageRank scores. We also believe that the proposed method should work for asymmetric dSBMs with $\pi_1 \neq \pi_2$, but the expressions for the mean and variance of PageRank become too complicated to compute clean bounds. Possible future work could include trying to show that the π_1-th quantile of the limiting PageRank distribution is a good threshold in the case $\pi_1 \neq \pi_2$, or trying to find a threshold independent of π so that we can recover communities even when we do not have information about their sizes. Another interesting direction would be to investigate whether

the inference can be strengthened if the seed set contains members from both communities and/or the connectivity structure of the subgraph spanned by the seeds is fully or partially known.

References

1. Avrachenkov, K., Kadavankandy, A., & Litvak, N. (2018). Mean field analysis of personalized pagerank with implications for local graph clustering. *Journal of statistical physics, 173(3)*, 895-916.
2. Garavaglia, A., van der Hofstad, R., & Litvak, N. (2020). Local weak convergence for PageRank. *The Annals of Applied Probability, 30(1)*, 40-79.
3. Gulikers, L., Lelarge, M., & Massoulié, L. (2018). An impossibility result for reconstruction in the degree-corrected stochastic block model. *The Annals of Applied Probability, 28(5)*, 3002-3027.
4. Kloumann, I. M., Ugander, J., & Kleinberg, J. (2017). Block models and personalized PageRank. *Proceedings of the National Academy of Sciences, 114(1)*, 33-38.
5. Lee, J., & Olvera-Cravioto, M. (2020). PageRank on inhomogeneous random digraphs. *Stochastic Processes and their Applications, 130(4)*, 2312-2348.
6. Mossel, E., Neeman, J., & Sly, A. (2014). Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*, 3(5).
7. Mossel, E., Neeman, J., & Sly, A. (2015). Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3), 431-461.
8. Massoulié, L. (2014, May). Community detection thresholds and the weak Ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing* (pp. 694-703).
9. Chin, P., Rao, A., & Vu, V. (2015, June). Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference on Learning Theory* (pp. 391-423). PMLR.
10. Spielman, D. A., & Teng, S. H. (2013). A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM Journal on computing*, 42(1), 1-26.
11. Andersen, R., Chung, F., & Lang, K. (2007, December). Local partitioning for directed graphs using pagerank. In *International Workshop on Algorithms and Models for the Web-Graph* (pp. 166-178). Springer, Berlin, Heidelberg.
12. Dhara, S., Gaudio, J., Mossel, E., & Sandon, C. (2022). Spectral recovery of binary censored block models*. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)* (pp. 3389-3416). Society for Industrial and Applied Mathematics.
13. Das Sarma, A., Molla, A. R., Pandurangan, G., & Upfal, E. (2013, January). Fast distributed pagerank computation. In *International Conference on Distributed Computing and Networking* (pp. 11-26). Springer, Berlin, Heidelberg.
14. Olvera–Cravioto, M. (2021). PageRank's behavior under degree correlations. *The Annals of Applied Probability*, 31(3), 1403-1442.