ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)
82, *Part* 1, *pp.* 99–126

Targeted sampling from massive block model graphs with personalized PageRank

Fan Chen, Yini Zhang and Karl Rohe

University of Wisconsin—Madison, USA

[Received January 2019. Revised October 2019]

Summary. The paper provides statistical theory and intuition for personalized PageRank (called 'PPR'): a popular technique that samples a small community from a massive network. We study a setting where the entire network is expensive to obtain thoroughly or to maintain, but we can start from a seed node of interest and 'crawl' the network to find other nodes through their connections. By crawling the graph in a designed way, the PPR vector can be approximated without querying the entire massive graph, making it an alternative to snowball sampling. Using the degree-corrected stochastic block model, we study whether the PPR vector can select nodes that belong to the same block as the seed node. We provide a simple and interpretable form for the PPR vector, highlighting its biases towards high degree nodes outside the target block. We examine a simple adjustment based on node degrees and establish consistency results for PPR clustering that allows for directed graphs. These results are enabled by recent technical advances showing the elementwise convergence of eigenvectors. We illustrate the method with the massive Twitter friendship graph, which we crawl by using the Twitter application programming interface. We find that the adjusted and unadjusted PPR techniques are complementary approaches, where the adjustment makes the results particularly localized around the seed node, and that the bias adjustment greatly benefits from degree regularization.

Keywords: Community detection; Degree-corrected stochastic block model; Local clustering; Network sampling; Personalized PageRank

1. Introduction

Much of the literature on graph sampling has treated the entire graph, or all of the people in it, as the target population. However, in many settings, the target population is a small community in the massive graph. For example, a key difficulty in studying social media is to gather data that are sufficiently relevant for the scientific objective. A motivating example for this paper is to sample the Twitter friendship graph for accounts that report and discuss current political events. (See our website `http://murmuration.wisc.edu`, which does this.) This corresponds to sampling and identifying multiple communities, each a potentially small part of the massive network. In such an application, the graph is useful for two primary reasons. First, via link tracing, we can find potential members of the target population. Second, the graph connections are informative for identifying community membership. Throughout, we presume that the sampling is initiated around a 'seed node' that belongs to the target community of interest.

A personalized PageRank (called 'PPR') can be thought of as an alternative to snowball sampling, which is a popular technique for gathering individuals close to the seed node. For

Address for correspondence: Fan Chen, Department of Statistics, University of Wisconsin—Madison, 1300 University Avenue, Madison, WI 53706, USA.
E-mail: fan.chen@wisc.edu

 1369–7412/20/82099

some $d \geqslant 0$, snowball sampling gathers all individuals who are d friends away from the seed. This process has two competing flaws for our application which are addressed by PPR. First, snowball sampling fails to account for the density of common friendships. For example, perhaps i and j are both one friend removed from the seed, but i has 10 friends in common with the seed, whereas j has only one friend in common. It seems natural to suppose that i is closer than j to the seed. Hence, the metric for snowball sampling can be misleading. Second, the snowball sample size grows very quickly with d. For example, under the 'six degrees of separation' phenomenon (Watts and Strogatz, 1998; Newman *et al.*, 2006), snowballing gathers the entire graph if $d \geqslant 6$.

PPR gives a sample that is more localized around the seed node. The PPR vector is defined as the stationary distribution of what we call a *personalized random walk* (Page *et al.*, 1999). At each step of the personalized random walk, the random walker returns to the seed node with probability α, called the teleportation constant, and, with probability $1 - \alpha$, the random walker goes to an adjacent node that is chosen uniformly at random. Consider the stationary distribution of this process as giving the inclusion probability for a sample of size 1. This is the PPR vector. PPR naturally leads to a clustering algorithm, where the cluster is made up of the nodes with a large inclusion probability. To approximate the PPR vector quickly, Berkhin (2006) proposed an algorithm that examines only nodes with large inclusion probabilities (i.e. nodes near the seed). As such, PPR is particularly useful for its computational efficiency—the running time and the amount of data that it requires are nearly linear in the size of the output cluster, which is typically much smaller than that of the entire graph. Because of the local nature of the algorithm, it can be used to study large graphs such as Twitter where the entire graph is not available, but where one can query to find the connections to any small set of nodes.

One way to conduct local clustering is by exploring and ranking the nearby nodes of a seed node (Andersen and Lang, 2006; Andersen and Peres, 2009; Alamgir and Von Luxburg, 2010; Gharan and Trevisan, 2012). Spielman and Teng (2004) pioneered local clustering by defining nearness as the landing probability of a random walk starting from the seed node. Their algorithm's guarantee was improved in follow-up work by Andersen *et al.* (2006) which proposed the use of an approximate PPR vector. Local algorithms can be applied recursively to solve more complicated problems such as graph partitions (k-way partitions) (Spielman and Teng, 1996; Karypis and Kumar, 1998) and have many fruitful applications (Jeh and Widom, 2003; Macropol *et al.*, 2009; Liao *et al.*, 2009; Gupta *et al.*, 2013; Gleich, 2015), particularly when it comes to sampling and studying massive graphs.

Along with the widespread use of PPR, there has been recent work to study its statistical estimation properties under a statistical model with latent community structure. Beyond the scope of local clustering, Kloumann *et al.* (2017) showed that the PPR vector is asymptotically equivalent to optimal linear discriminant analysis under the stochastic block model (SBM) (Holland *et al.*, 1983), assuming a symmetry condition on the block structure. We add to this statistical understanding of PPR by providing a simple and more general representation for PPR vectors that allows for different block sizes, more than two blocks, degree heterogeneity and directed edges. To understand the effects of heterogeneous node degrees, this paper uses the degree-corrected stochastic block model (DCSBM) (Karrer and Newman, 2011) and examines when the PPR clustering recovers nodes within the same block as the seed node (local cluster). Breaking the symmetry that was imposed by Kloumann *et al.* (2017) reveals additional insight. In particular, given a seed node in the first block, we show that PPR is likely to contain high degree nodes outside that block. We study an adjustment that was previously proposed in Andersen *et al.* (2006). We show how this adjustment can correct for the bias. We illustrate these ideas with examples from the Twitter friendship graph.

1.1. An illustrative example in social media

Local clustering using PPR is particularly well suited to studying current political events on Twitter because

(a) the accounts that discuss politics or current events are a small part of the entire Twitter graph,
(b) it is reasonable to believe that the accounts in our target population are well connected to one another in the Twitter friendship graph and,
(c) although the entire Twitter graph is not publicly available, the way that PPR (algorithms 1 and 3 in Sections 2.1 and 3.2 respectively) queries the graph matches the Twitter application programming interface protocol which is the primary mode of access for researchers.

Although we do not suppose that the Twitter friendship graph is sampled from a DCSBM, Twitter does have all of the heterogeneities that our results identify as important. The Twitter friendship graph is composed of users who can freely follow others but will not necessarily be followed back, or friended. Such asymmetry between following and friending forms a directed graph where follower count indicates status—some popular or high status nodes command millions of followers whereas the majority of nodes are followed by far fewer.

The theoretical results in this paper suggest that such degree heterogeneities will make the PPR vector biased for detecting block memberships (theorem 1 in Section 3.1). We propose a way to adjust for this bias (algorithm 2 in Section 2.1) and show that it is a consistent estimator (corollary 1 in Section 4). Not surprisingly, this section demonstrates that PPRs with and without the bias adjustment give fundamentally different results on the Twitter graph. However, depending on the application, the biases in the PPR vector might be advantageous. In this way, PPRs with and without the bias adjustment are complementary, not competing, approaches.

Table 1. Top 15 handles by PPR clustering†

Rank	*@CNN*	*@BreitbartNews*	*@dailykos*
1	CNN Breaking News	Alex Marlow	Hillary Clinton
2	CNN International	AndrewBreitbart	Stephen Colbert
3	Wolf Blitzer	Big Hollywood	Rachel Maddow MSNBC
4	Anderson Cooper	Big Government	Jake Tapper
5	Christiane Amanpour	James O'Keefe	Joy Reid
6	Pope Francis	Sean Hannity	Chris Hayes
7	Dr Sanjay Gupta	Raheem	Emma Gonzlez
8	CNNMoney	Joel B. Pollak	Markos Moulitsas
9	Jake Tapper	Ann Coulter	Maggie Haberman
10	Brian Stelter	Allum Bokhari	Sarah Silverman
11	CNN Newsroom	Ben Kew	Lin-Manuel Miranda
12	Dana Bash	Brandon Darby	Elizabeth Warren
13	CNN Politics	Noah Dulis	Jon Favreau
14	BBC Breaking News	Michelle Malkin	Michelle Obama
15	Brooke Baldwin	Nate Church	Bill Clinton

†Column names represent seed nodes, and the sampled nodes are ranked by PPR values, with teleportation constant $\alpha = 0.15$ uniformly. Through the PPR vector, the top 15 handles returned to each of the three seed nodes fit well with the characteristics of the seed nodes. They are popular or high status handles either directly related to the seed nodes or align with their political leanings. This shows the effectiveness of clustering via the PPR vector. It also shows the PPR vector's preference for highly connected nodes.

Table 2. Top 15 handles by adjusted PPR (with regularization) sampling†

Rank	*@CNN*	*@BreitbartNews*	*@dailykos*
1	PowerZ	Robert	Two Thanks
2	Elissa Weldon	Lee Peace	Catherine Daligga
3	Tess Eastment	Wynn Marlow	exmearden
4	Chris_Dawson	Logan Churchwell	Faith Gardner
5	carol kinstle	Peter Schweizer	Andrew Thornton
6	erinmclaughlin	Breitbart Sports	UnreasonableFridays
7	Taylor Ward	Jon Fleischman	DKos Top Comments
8	Jennifer Z. Deaton	Nate Church	2016 relitigator
9	Pam Benson	Daniel Nussbaum	Daily Kos
10	amy entelis	Noah Dulis	Walter Einenkel
11	Grace Bohnhoff	Jon David Kahn	Candelaria Vargas
12	kate lazarus	Breitbart California	Mara Schechter
13	Newstron	Ken Klukowski	Emi Feldman
14	Becky Brittain	pam key	The Soulful Negress
15	CNN Ballot Bowl	Auntie Hollywood	Kim Soffen

†Column names represent seed nodes, and the sampled nodes are ranked by adjusted PPR values, with teleportation constant $\alpha = 0.15$ uniformly. After adjustment, PPR returns a more localized cluster. Instead of the highly visible public faces of the three seed organizations, the individuals in this table serve a central role to the internal organization (e.g. editors and writers). Depending on the application, one might prefer the results in Table 1 or Table 2.

To illustrate, Table 1 displays the top 15 handles ranked by the PPR vector (without adjustment) for three different seed nodes: @CNN, @BreitbartNews and @dailykos, which are the Twitter accounts of three types of media outlets that exhibit distinct political leanings (legacy broadcast news, on-line right wing and on-line left wing). For @CNN, all top 15 handles ranked by the PPR vector are its subsidiary accounts and its celebrity reporters and anchors (like Wolf Blitzer and Anderson Cooper), except for one account, Pope Francis, who enjoys an extremely larger following. The top 15 handles for @BreitbartNews are a mixed bag of influential conservatives (like Sean Hannity and Ann Coulter) and Breitbart's editors or writers. However, the top 15 handles returned to @dailykos by the PPR vector are all famous liberal personalities who are not directly affiliated with Daily Kos, except one: its founder Markos Moulitsas. Those people range from democratic politicians to liberal media personalities and journalists, such as Hillary Clinton, Stephen Colbert and Rachel Maddow. All the handles align with the characteristics of their respective media outlets, attesting to the clustering effectiveness. However, it is worth noting that the top handles ranked by the PPR vector tend to be popular handles with millions of followers. This shows that the PPR vector's preference is for high in-degree nodes.

In contrast, for each of the three seeds, adjusted PPR finds accounts that are more central to the internal functioning of these organizations. Table 2 lists those accounts. The bias adjustment also greatly benefits from a degree regularization (Qin and Rohe, 2013). For @CNN, those handles include primarily its own staff, producers or journalists (like Elissa Weldon, Chris_Dawson and Grace Bohnhoff) and a freelance journalist (Tess Eastment). The pattern is similar for @BreitbartNews and @dailykos, their top 15 handles including their own journalists and editors as well as related writers, campaigners and activists. The general pattern is that the adjustment returns editors, journalists and staff working within each media outlet. As such, the adjustment is useful for identifying a more localized cluster.

1.2. Main contributions

The main contributions of the paper are a simple and interpretable form for the PPR vector and a statistical guarantee for clustering with the adjusted PPR vector.

(a) This paper reveals a simple two-stage form of the PPR vector under the population (expectation) DCSBM. Consider the vth element of the PPR vector as the probability of sampling node v in a sample of size 1 from the stationary distribution of the personalized random walk. This inclusion probability is akin to stratified sampling:

> The inclusion probability for node v is the product of two separate probabilities: first, the probability that the personalized random walk samples any node in v's block; second, the probability that the personalized random walk selects node v, conditionally on sampling that block.

Both of these probabilities have simple expressions. If there are K blocks in the graph, then the blockwise probability comes from the PPR vector of a graph with K vertices, with edge weights specified by the 'block connection matrix' in the DCSBM. The second probability is proportional to the degree of node v. In addition to the population results, theorem 2 in Section 4 demonstrates that, when the graph is random, the PPR vector concentrates around its population (expectation) under certain conditions.

(b) This paper identifies two sources of bias of using a PPR vector for local clustering under the DCSBM—the ancillary effects of heterogeneous node degrees and block degrees. With this finding, the paper examines a simple bias adjustment that remedies the two biases simultaneously and suggests conditions when the adjusted PPR can be used to return the correct local cluster. In other words,

> PPR clustering with the adjustment achieves the precise identification of the local cluster, provided that the graph is sufficiently dense.

These results establish statistical performance (consistency) of PPR clustering under the DCSBM, in the sparse regime where the minimum expected degree grows logarithmically with the number of nodes in the network. Our results provide an elementwise perturbation bound for PPR vectors, that allows the number of clusters to grow with the size of graphs, and generalize to a directed graph setting as Google PageRank does.

The rest of the paper proceeds as follows. Section 2 formally introduces the PPR method and some of the known results. Section 2 also introduces the DCSBM. Section 3 gives a population analysis of the PPR clustering under directed block model graphs. Section 4 provides concentration results for the PPR vector when the graph is random and provides a statistical guarantee on the PPR local clustering method. Section 5 presents several numerical results showing the effectiveness of the PPR clustering. Section 6 illustrates the PPR clustering through the massive Twitter friendship graph and demonstrates the benefits of a smoothing step in the PPR adjustment.

2. Preliminaries

Throughout this paper, $G = (V, E)$ denotes an unweighted and connected graph, where E is the edge set and V is the set of vertices indexed by $1, \ldots, N$. When G is an undirected and unweighted graph, encode E into a binary *adjacency* matrix $A \in \{0, 1\}^{N \times N}$ with $A_{uv} = A_{vu} = 1$ if and only if edge (u, v) appears in E. Define a diagonal matrix $D = \mathrm{diag}(d_1, \ldots, d_N)$ and the *graph transition* matrix P as follows:

$$d_u = \sum_{v \in V} A_{uv},$$
$$P = D^{-1} A.$$

When G is a directed graph, the adjacency matrix $A \in \{0,1\}^{N \times N}$ accordingly becomes asymmetric with $A_{uv} = 1$ if and only if edge $(u, v) \in E$, and the graph transition matrix is defined as

$$P = (D^{\text{out}})^{-1} A,$$

where $D^{\text{out}} = \text{diag}(d_1^{\text{out}}, \ldots, d_N^{\text{out}})$ and $d_u^{\text{out}} = \Sigma_{v \in V} A_{uv}$ is the number of edges leaving from node u. In addition, define $D^{\text{in}} = \text{diag}(d_1^{\text{in}}, \ldots, d_N^{\text{in}})$ where $d_v^{\text{in}} = \Sigma_{u \in V} A_{uv}$ is the number of edges pointing to node v.

2.1. Personalized PageRank and the local clustering algorithm

PPR is an extension of Google's PageRank (Brin and Page, 1998; Haveliwala, 2003). To illustrate, consider a personalized random walk (or originally called 'surfing') on the graph $G = (V, E)$ with a *seed node* $v_0 \in V$. At each step, the random walker either restarts from the seed node v_0 with probability α (called the *teleportation constant*) or continues the random walk from the current node to a neighbour uniformly at random. The *PPR vector* $p \in [0,1]^N$ is the stationary distribution of this process, and thus the solution to the equation

$$p^{\text{T}} = \alpha \pi^{\text{T}} + (1 - \alpha) p^{\text{T}} P, \tag{1}$$

where P is the graph transition matrix, and π is the elementary unit vector in the direction of seed node v_0. Here p is a column vector normalized by a positive scalar such that its elements sum to 1 and, without loss of generality, we set $v_0 = 1$ and thus $\pi = (1, 0, \ldots, 0)^{\text{T}}$.

In general, the *preference vector* π does not have to be an elementary unit vector, but any probability distribution on V. For example, when $\pi = (1/N, \ldots, 1/N)^{\text{T}}$, PPR is equivalent to ordinary PageRank. Moreover, the PPR vector is a linear function of the preference vector, i.e. let $p(\pi_1)$ and $p(\pi_2)$ be two PPR vectors corresponding to two preference vectors π_1 and π_2 respectively. Then, for a new preference vector that is a convex combination of π_i, the resulting PPR vector is constructive of $p(\pi_i)$:

$$p(w_1 \pi_1 + w_2 \pi_2) = w_1\, p(\pi_1) + w_2\, p(\pi_2),$$

where $w_i \geqslant 0$ and $w_1 + w_2 = 1$. Define Π to be an $N \times N$ matrix with repeating rows π^{T}, and let $Q = \alpha \Pi + (1 - \alpha) P$; then Q is the Markov transition matrix for the stochastic process and equation (1) becomes $p^{\text{T}} = p^{\text{T}} Q$. Below are some useful properties of the PageRank vector (also see Haveliwala (2003), Jeh and Widom (2003) and Appendix A).

Proposition 1. For any fixed $\alpha \in (0, 1]$, the PPR vector p is

(a) the left leading eigenvector of Q, associated with the simple eigenvalue 1, and
(b) the infinite sum of landing probability $\{(P^s)^{\text{T}} \pi\}_{s=0}^{\infty}$ with weights $\phi = \{\alpha(1-\alpha)^s\}_{s=0}^{\infty}$,

$$p^{\text{T}} = \alpha \sum_{s=0}^{\infty} (1 - \alpha)^s \pi^{\text{T}} P^s. \tag{2}$$

Berkhin (2006) gave an iterative algorithm based on proposition 1 to approximate the PPR vector (that scales to large graphs); each update requires only neighbourhood information of one visited vertex. A few lines of linear algebra show that the PPR vector is equivalent to the solution to the linear system

$$p^{\text{T}} = \alpha' \pi^{\text{T}} + (1 - \alpha') p^{\text{T}} W,$$

Table 3. Algorithm 1: approximate PPR vector (undirected) (Andersen *et al.*, 2006)

Require: undirected graph G, preference vector π, teleportation constant α and tolerance ϵ
Initialize $p \leftarrow 0$, $r \leftarrow \pi$, $\alpha' \leftarrow \alpha/(2-\alpha)$
while $\exists\, u \in V$ such that $r_u \geqslant \epsilon d_u$ *do*
uniformly sample a vertex u satisfying $r_u \geqslant \epsilon d_u$
$p_u \leftarrow p_u + \alpha' r_u$
for $v : (u, v) \in E$ *do*
$r_v \leftarrow r_v + (1-\alpha') r_u/(2d_u)$
end for
$r_u \leftarrow (1-\alpha') r_u/2$
end while
Return: ϵ-approximate PPR vector p

Table 4. Algorithm 2: PPR clustering (undirected)

Require: undirected graph G, seed node v_0 and the desired size of local cluster n
Step 1: calculate the approximate PPR vector p (algorithm 1)
Step 2: adjust the PPR vector p by node degrees, $p^*_v \leftarrow p_v/d_v$
Step 3: rank all vertices according to the adjusted PPR vector p^*
Return: local cluster—n top ranking nodes

where $W = (I + P)/2$ is the lazy graph transition matrix and $\alpha' = \alpha/(2-\alpha)$. Using this fact, algorithm 1 (Table 3) approximates the PPR vector in running time of order $\mathcal{O}\{1/(\epsilon\alpha)\}$, by reaching at most $2/\{\epsilon(1-\alpha)\}$ vertices. The following proposition gives a guarantee on the approximation error for this algorithm in terms of the *tolerance* parameter and the degrees of visited nodes.

Proposition 2 (entrywise approximation error (Andersen *et al.*, 2006)). Let p be a PPR vector, and let $p^\epsilon \in [0,1]^N$ be an approximate PPR vector computed by algorithm 1 with a tolerance $\epsilon > 0$. For any vertex u that is sampled in algorithm 1,

$$|p_u - p^\epsilon_u| \leqslant \epsilon d_u.$$

Proposition 2 ensures that, for any fixed graph, the approximate PPR vector is arbitrarily close to the exact PPR vector, as long as the tolerance $\epsilon > 0$ is sufficiently small. Appendix A contains a proof of this proposition for completeness. Given a seed node in the graph, algorithm 2 (Table 4) uses the approximate PPR vector from algorithm 1 and returns a set of nodes with the largest corresponding values in the *adjusted personalized PageRank* (called 'APPR') vector, which is defined as

$$p^*_v = \frac{p_v}{d_v}, \qquad \text{for } v = 1, 2, \ldots, N.$$

The APPR vector was previously proposed in Andersen *et al.* (2006). Algorithm 1 and algorithm 2 operate on undirected graphs. We shall generalize them to directed graphs in Section 3 thanks to a simplified and interpretable form for the PPR vector.

2.2. Stochastic block model

In an SBM, each node belongs to one of K blocks. The presence of each edge corresponds to an independent Bernoulli random variable, where the probability of an edge between any two

nodes depends only on the block memberships of two nodes (Holland *et al.*, 1983). The formal definition is as follows.

Definition 1. For a vertex set $V = \{1, 2, \ldots, N\}$, let $z : \{1, 2, \ldots, N\} \to \{1, 2, \ldots, K\}$ partition the N nodes into K blocks, so $z(v)$ is the block membership of vertex v. Let $\mathbf{B}$ be a $K \times K$ matrix with all entries' range in $[0, 1]$. Under the SBM, the probability of an edge between u and v is $\mathbf{B}_{z(u)z(v)}$, i.e.

$$A_{uv}|z(u), z(v) \overset{\text{ind.}}{\sim} \text{Bernoulli}(\mathbf{B}_{z(u)z(v)}), \qquad \text{for any } u, v \in \{1, 2, \ldots, N\}.$$

Under the ordinary SBM, nodes in the same block have the same expected degree. One extension is the DCSBM, which adds a series of parameters ($\theta_v > 0$ for every vertex v) to create more heterogeneous node degrees (Karrer and Newman, 2011). Let $\mathbf{B}$ be a $K \times K$ matrix with $\mathbf{B}_{ij} > 0$ for any i and j. Then the probability of an edge between u and v is $\theta_u \theta_v \mathbf{B}_{z(u)z(v)}$, i.e.

$$A_{uv}|z(u), z(v) \overset{\text{ind.}}{\sim} \text{Bernoulli}(\theta_u \theta_v \mathbf{B}_{z(u)z(v)}),$$

for $u, v \in \{1, 2, \ldots, N\}$. Since the θ_vs are arbitrary to a multiplicative constant which can be absorbed into $\mathbf{B}$, Karrer and Newman (2011) suggested imposing the constraint that the θ_vs sum to 1 within each block, i.e. $\Sigma_{v:z(v)=i} \theta_v = 1$ for all $i = 1, 2, \ldots, K$. With this constraint, $\mathbf{B}_{ij}$ represents the expected number of edges between block i and j if $i \neq j$, and twice that if $i = j$. Throughout this paper, we presume that $\mathbf{B}$ is positive definite. This prevents scenarios where edges are unlikely within blocks and more likely between blocks. (In such scenarios, local clustering needs to be reimagined cautiously; see the on-line supplementary materials section S2 for additional details about generalizations.) We also presume that all blocks are connected (we ignore any blocks that are isolated from the seed). The DCSBM can be generalized to directed graphs by giving each node two parameters, θ_v^{in} and θ_v^{out}, controlling its in-degree and out-degree respectively (Zhu *et al.*, 2013). Then, the presence of a directed edge from u to v, given the block memberships, corresponds to an independent Bernoulli random variable:

$$A_{uv}|z(u), z(v) \overset{\text{ind.}}{\sim} \text{Bernoulli}(\theta_u^{\text{out}} \theta_v^{\text{in}} \mathbf{B}_{z(u)z(v)}).$$

To make the model identifiable, we need to impose a structural constraint on the θ^{in}s and θ^{out}s, that both of them sum to 1 within each block:

$$\sum_{v:z(v)=i} \theta_v^{\text{in}} = \sum_{v:z(v)=i} \theta_v^{\text{out}} = 1, \qquad \text{for any } i = 1, 2, \ldots, K.$$

Because the off-diagonal elements of $\mathbf{B}$ can be interpreted as the expected number of edges between blocks, we define the block in-degree and block out-degree to be the total number of incoming edges and outgoing edges respectively, i.e. $\mathbf{d}_j^{\text{in}} = \Sigma_{i=1}^K \mathbf{B}_{ij}$, and $\mathbf{d}_i^{\text{out}} = \Sigma_{j=1}^K \mathbf{B}_{ij}$.

3. Population analysis of PageRank

In this section, we analyse the PPR vector of the expected adjacency matrix under the DCSBM. This provides a simple representation of the PPR vector that motivates

(a) the bias adjustment and
(b) the generalization of algorithm 1 and 2 to directed graphs.

We use three distinct typefaces to denote three classes of objects. A script typeface is given to the population version of any observable quantities in random graphs, such as graph adjacency matrix and node degrees (e.g. equation (3)). The normal typeface is given to unobserved model

parameters, such as block membership and degree parameters θ_i. Bold is given to all block level quantities and parameters like $\mathbf{B}$ and $\mathbf{d}_i^{\text{out}}$.

Define the population graph adjacency matrix,

$$\mathscr{A} = \mathbb{E}\{A|z(1), z(2), \ldots, z(N)\}, \tag{3}$$

to be the expectation of random adjacency matrix A. Let $Z \in \{0,1\}^{N\times K}$ be the block membership matrix with $Z_{vi} = 1$ if and only if vertex v belongs to block i, and define diagonal matrices Θ^{in} and Θ^{out} with entries θ^{in} and θ^{out} respectively. Then, under the directed DCSBM with K blocks and parameters $\{\mathbf{B}, Z, \Theta^{\text{in}}, \Theta^{\text{out}}\}$, $\mathscr{A} \in \mathbb{R}^{K\times K}$ can be compactly expressed as

$$\mathscr{A} = \Theta^{\text{out}} Z \mathbf{B} Z^{\mathrm{T}} \Theta^{\text{in}}.$$

Accordingly, we define the population node degrees and the population transition matrix, $\mathscr{d}_u^{\text{in}} = \Sigma_{v\in V}\mathscr{A}_{uv}$, $\mathscr{d}_v^{\text{out}} = \Sigma_{u\in V}\mathscr{A}_{uv}$ and $\mathscr{P} = (\mathscr{D}^{\text{out}})^{-1}\mathscr{A}$, where $\mathscr{D}^{\text{in}}$ and $\mathscr{D}^{\text{out}}$ are the diagonal matrices of the population node in-degrees $\mathscr{d}_u^{\text{in}}$ and out-degrees $\mathscr{d}_v^{\text{out}}$ respectively. Let $\mathscr{p}$ be the population PPR vector (i.e. the solution to the equation $\mathscr{p}^{\mathrm{T}} = \alpha\pi^{\mathrm{T}} + (1-\alpha)\mathscr{p}^{\mathrm{T}}\mathscr{P}$) and let $\mathscr{p}^* = (\mathscr{D}^{\text{in}})^{-1}\mathscr{p}$ be the population APPR vector.

In addition, define the *block transition matrix* $\mathbf{P} \in \mathbb{R}^{K\times K}$ as

$$\mathbf{P} = (\mathbf{D}^{\text{out}})^{-1}\mathbf{B}, \tag{4}$$

where $\mathbf{D}^{\text{in}} \in \mathbb{R}^{K\times K}$ and $\mathbf{D}^{\text{out}} \in \mathbb{R}^{K\times K}$ are diagonal matrices of the block in-degrees $\mathbf{d}_i^{\text{in}}$ and out-degrees $\mathbf{d}_i^{\text{out}}$.

3.1. A representation of personalized PageRank vectors

This section provides a simple and interpretable form for PPR vectors under the population DCSBM. For this, we define the *'blockwise' PPR vector* $\mathbf{p} \in \mathbb{R}^K$ to be the unique solution to the linear system

$$\mathbf{p}^{\mathrm{T}} = \alpha\boldsymbol{\pi}^{\mathrm{T}} + (1-\alpha)\mathbf{p}^{\mathrm{T}}\mathbf{P}, \tag{5}$$

where $\boldsymbol{\pi} = Z^{\mathrm{T}}\pi \in \mathbb{R}^K$ is the blockwise preference vector and $\mathbf{P}$ is the block transition matrix in equation (4). This treats the block connectivity matrix $\mathbf{B}$ as a weighted adjacency matrix of blocks and the block of seed nodes as a seed block. To build up the relationship between PPR and the blockwise PPR, the following theorem gives an explicit form for PPR vectors which also reveals the sources of bias for local clustering.

Theorem 1 (explicit form of PPR vectors). Under the population directed DCSBM with K blocks and parameters $\{\mathbf{B}, Z, \Theta^{\text{in}}, \Theta^{\text{out}}\}$,

(a) the population PPR vector $\mathscr{p} \in \mathbb{R}^N$ has elements

$$\mathscr{p}_u = \theta_u^{\text{in}}\mathbf{p}_{z(u)}$$

where $\mathbf{p}$ is the blockwise PPR vector in equation (5), and

(b) the population APPR vector $\mathscr{p}^* \in \mathbb{R}^N$ has elements

$$\mathscr{p}_u^* = \mathbf{p}_{z(u)}^* \tag{6}$$

where $\mathbf{p}^* = (\mathbf{D}^{\text{in}})^{-1}\mathbf{p}$.

Theorem 1 demonstrates that the PPR vector $\mathscr{p}$ decomposes into block-related information $\mathbf{p}$ and node-specific information Θ. Within each block, the PPR values are proportional to the node

degree parameters θ_v and sum to the blockwise PPR value of the block. The proof of theorem 1 (Appendix A) relies on a key observation that the powers of the population transition matrix, $\mathscr{P}^s$ for $s = 1, 2, \ldots,$ have a similarly simple form and the node-specific information components (i.e. $z(v)$ and θ_v) are invariant in s.

To justify the adjustment (step 2) in algorithm 2, we observe that the seed always has the highest population APPR score. This turns out to be a key feature that facilitates the APPR vector to recover a local cluster correctly, so we state it in the following lemma.

Lemma 1 (the largest entry of APPR vector). Under the population DCSBM, assume that the minimum expected degree is positive, i.e. $\min_{v \in V} d_v > 0$. Then, for any fixed $\alpha > 0$, the population APPR vector π^* has the strictly largest entry corresponding to the seed node,

$$\pi^*_{v_0} > \pi^*_v, \qquad \text{for any } v \neq v_0.$$

In contrast, this is not generally true for a PPR vector.

When $\alpha = 0$ (i.e. no teleportation), the PPR vector becomes the limiting distribution of a standard random walk and all entries of the APPR vector are equal (Appendix A). Lemma 1 (applied to blockwise PPR vectors) and theorem 1 together identify two sources of bias for PPR vectors and suggest a justification for the degree adjustment, which we discuss in order.

(a) Both node degree heterogeneity Θ and block size imbalance $\mathbf{D}$ confound the identification of local clusters by the PPR vector. In particular, suppose that vertex v belongs to a block $z(v) = i$ other than 1. The PPR vector assigns it a score $\theta_v \mathbf{p}_i$, where $\mathbf{p}_i$ is the blockwise PPR of block i, and θ_v is the parameter specifically controlling the degree of v. Then, node v may rank at the top, if θ_v is sufficiently large. Furthermore, lemma 1 implies that $\mathbf{p}_1$ is not necessarily the largest because of block degree heterogeneity. Specifically, if block i has an exceedingly high block degree, it is likely that π fails to downrank node v *vis-à-vis* those nodes of block 1.

(b) APPR removes the node and the block degree heterogeneity simultaneously and perfectly recovers the local cluster. To see this, note that $\mathbf{p}^*$ is the adjusted version of the blockwise PPR vector. From lemma 1, $\mathbf{p}^*_1$ is the largest entry of $\mathbf{p}^*$. From equation (6), the APPR vector assigns any vertex v a score $\mathbf{p}^*_{z(v)}$. Hence, nodes with the highest value of π^* belong to block 1, which is precisely the local cluster desired.

Note that the PPR vector can still be biased for local clustering even under the classical SBM. To see this, set the matrix Θ to the identity matrix in theorem 1. In this case, the heterogeneous block degrees still confound the PPR vector (Section 5.2); there is generally no guarantee for $\mathbf{p}_1$ to appear on the top (because of lemma 1), unless there are further symmetry conditions. Kloumann *et al.* (2017) used such a scenario. As a by-product of our analysis, we extend their results under the DCSBM with the symmetric conditions (see the on-line supplementary materials section S3 to the paper).

3.2. Local clustering on directed graphs

In light of the clean form of PPR vectors under the DCSBM, one can modify algorithms 1 and 2 to operate on a directed graph accordingly. For this, note that the transition matrix of a directed graph requires node out-degrees; hence algorithm 1 examines only the edges leaving visited nodes. Consequently it suffices to replace the d_us in algorithm 1 by d_u^{out}s (algorithm 3: Table 5). Proposition 2 applies to algorithm 3 as well, and one can approximate the PPR vector provided that the out-degrees of visited nodes can be observed and the tolerance parameter $\epsilon > 0$ is sufficiently small.

Table 5. Algorithm 3: approximate PPR vector (directed)

Require: directed graph G, preference vector π, teleportation constant α and tolerance ϵ
Initialize $p \leftarrow 0$, $r \leftarrow \pi$, $\alpha' \leftarrow \alpha/(2-\alpha)$
while $\exists\, u \in V$ such that $r_u \geqslant \epsilon d_u^{\text{out}}$ *do*
sample a vertex u uniformly at random, satisfying $r_u \geqslant \epsilon d_u^{\text{out}}$
$$p_u \leftarrow p_u + \alpha' r_u$$
for $v : (u, v) \in E$ *do*
$$r_v \leftarrow r_v + (1-\alpha') r_u/(2d_u^{\text{out}})$$
end for
$$r_u \leftarrow (1-\alpha') r_u/2$$
end while
Return: ϵ-approximate PPR vector p

Table 6. Algorithm 4: PPR clustering (directed)

Require: directed graph G, seed node v_0, the desired size of local cluster n and an optional regularization parameter τ

Step 1: calculate the approximate PPR vector p (algorithm 3)
Step 2: adjust the PPR vector p with two options

(a) node in-degrees, $p_v^* \leftarrow p_v/d_v^{\text{in}}$,
(b) regularized node in-degrees, $p_v^\tau \leftarrow p_v/(d_v^{\text{in}} + \tau)$

Step 3: rank all vertices according to the APPR vector p^* or p^τ

Return: local cluster—n top ranking nodes

To perform local clustering on a directed graph, algorithm 4 (Table 6) adjusts the approximate PPR vectors from algorithm 3 by node in-degrees, i.e.

$$p_v^* = \frac{p_v}{d_v^{\text{in}}}, \qquad \text{for } v = 1, 2, \ldots, N.$$

Another option is regularized adjustment, which produces the *regularized* PPR (RPPR) vector,

$$p_v^\tau = \frac{p_v}{d_v^{\text{in}} + \tau}, \qquad \text{for } v = 1, 2, \ldots, N,$$

where $\tau > 0$ is the regularization parameter. The regularized adjustment greatly stabilizes the PPR clustering in practice, by removing nodes with extremely low in-degrees (see Section 6 for more details). APPR for directed graphs is a local algorithm so long as d^{in} is available with a local query, e.g. the Twitter friendship graph.

4. Personalized PageRank in random graphs

This section establishes several concentration results for the local clustering algorithm using the APPR vector (algorithms 2 and 4) under the DCSBM. The results show that, if the graph is generated from the DCSBM, then PPR clustering returns the desired local cluster with high probability. Since, in algorithm 4, the calculation for PPR vectors relies on only node out-degrees and the adjustment step solely utilizes node in-degrees, it is not difficult to distinguish d^{in} and d^{out}. Thus, we state the results in undirected graphs for simplicity. One can draw the analogous conclusions for directed graphs by tracing the proof step by step.

We first present a useful tool that controls the entrywise errors of a PPR vector in random graphs. Recall that π is the stationary distribution of probability transition matrix $Q = \alpha\Pi +$

$(1-\alpha)\mathscr{P}$. For any vector $x \in \mathbb{R}^n$, define the vector infinity norm as $\|x\|_\infty = \max_i |x_i|$. The following theorem bounds the entrywise error of the stationary distribution of $\mathscr{Q}$.

Theorem 2 (concentration of the PPR vectors). Let $G = (V, E)$ be a graph of N vertices generated from the DCSBM with K blocks and parameters $\{\mathbf{B}, Z, \Theta\}$. Let p and $\mathscr{p}$ be the PPR vector corresponding to random transition matrix P and its population version $\mathscr{P}$ respectively, with the same teleportation constant α. Let $p^*, \mathscr{p}^* \in [0,1]^N$ be the adjusted PPR vector of p and $\mathscr{p}$. Let δ be the average expected node degrees, i.e. $\delta = \Sigma_{v \in V} \mathscr{d}_v / N$. Assume that $\rho = \max_{v \in V} \mathscr{d}_v / \min_{v \in V} \mathscr{d}_v$ is bounded by some finite constant and that

$$\delta > c_0 (1-\alpha)^2 \log(N), \tag{7}$$

for some sufficiently large constant $c_0 > 0$. Then, with probability at least $1 - \mathscr{O}(N^{-5})$,

$$\frac{\|p - \mathscr{p}\|_\infty}{\|\mathscr{p}\|_\infty} \leqslant c_1 (1-\alpha) \sqrt{\left\{ \frac{\log(N)}{\delta} \right\}}$$

and

$$\frac{\|p^* - \mathscr{p}^*\|_\infty}{\|\mathscr{p}^*\|_\infty} \leqslant c_2 (1-\alpha) \sqrt{\left\{ \frac{\log(N)}{\delta} \right\}},$$

for some sufficiently large constants $c_1, c_2 > 0$.

The proof of theorem 2 invokes the elementary eigenvector perturbation bound for asymmetric matrices, an analogue to the celebrated Davis–Kahan $\sin(\Theta)$ theorem (Davis and Kahan, 1970), and the novel leave-one-out technique due to Chen *et al.* (2019). A detailed proof is given in the on-line supplementary materials section S1 to the paper.

Theorem 2 demonstrates that, if the expected average degree δ exceeds $(1-\alpha)^2 \log(N)$ to some sufficiently large extent, then, with high probability, the random APPR vector concentrates around the population APPR vector in terms of all entries. In fact, the concentration statement holds for any valid preference vector π. Hence, the classic PageRank vector and some other variants also enjoy the entrywise error bounds, so long as they can be written as the solution to the linear system (1).

Next, we introduce a separation measure of the DCSBM. Recall that we can conduct a local clustering task by selecting nodes ranked by the APPR vector p^*. In the population version, it is equivalent to distinguishing between $\mathbf{p}_1^*$ and $\mathbf{p}_k^*$, for all $k = 2, 3, \ldots, K$, which also characterizes the distance from the desired local cluster (block 1) to its complement set (the other blocks). Only if they are sufficiently separated can the local cluster be identifiable in the sample. Because of lemma 1, we assume without loss of generality that the second block has the second highest value in the 'blockwise' APPR vector, i.e. $\mathbf{p}_1^* > \mathbf{p}_2^* \geqslant \mathbf{p}_k^*$ for $k = 3, 4, \ldots, K$. Then, we define the *separation measure* $\Delta_\alpha \in (0, 1]$:

$$\Delta_\alpha = \frac{\mathbf{p}_1^* - \mathbf{p}_2^*}{\mathbf{p}_1^*},$$

which turns out to be crucial in determining the sample complexity that is required to guarantee the exact recovery. We remark that Δ_α is an increasing function of the teleportation constant: hence the subscript α.

With theorem 2 and the separation measure, we then give the following corollary that bounds the accuracy of algorithm 2, in terms of graph edge density.

Corollary 1 (exact recovery by APPR vector). For any seed nodes, let $C \subset V$ be the local cluster of n nodes returned by algorithm 2 with teleportation constant α and tolerance ϵ, and $\mathscr{C} \subset V$ be the nodes in the seed node's block. Assume that $\rho < c_0$, $\epsilon \leqslant c_1(1-\alpha)\mathbf{p}_1^* \surd\{\log(N)/\delta\}$, and that

$$\delta > 16c_2\left(\frac{1-\alpha}{\Delta_\alpha}\right)^2 \log(N), \tag{8}$$

for some sufficiently large constants $c_0, c_1, c_2 > 0$. If the desired size of the local cluster $n = |\mathscr{C}|$, then, with probability at least $1 - \mathscr{O}(N^{-5})$, we have $C = \mathscr{C}$.

A proof of corollary 1 is presented in Appendix A. We make a few remarks.

(a) Corollary 1 demonstrates that algorithm 2 works under a sparse scenario, where the number of edges is exceedingly small in proportion to the number of possible edges in the network. To reach the entrywise control of the APPR vector and the sufficient separation of local cluster from others, theorem 2 calls for the expected node degree δ to grow with only a fraction (for any fixed teleportation constant α) of the logarithm of the size of the network, $\log(N)$. In other words, algorithm 2 requires a sample complexity (the number of edges) of order

$$\left(\frac{1-\alpha}{\Delta_\alpha}\right)^2 N \log(N).$$

(b) The results show that α leverages between the sampling complexity and statistical performance of PPR clustering. To see this, rearrange condition (8),

$$\left(\frac{1-\alpha}{\Delta_\alpha}\right)^2 < \frac{c'\delta}{\log(N)},$$

for some sufficiently small constant $c' > 0$. As α increases, the left-hand side decreases to 0, thus making the condition more likely to hold. In contrast, as α increases, the tolerance ϵ must decrease at rate $\mathscr{O}(1-\alpha)$ to guarantee an entrywise control of p^ϵ that is analogous to the form in theorem 2 (Appendix A). More intuitively, if ϵ does not decrease, then, as $\alpha \to 1$, algorithm 1 may terminate early without reaching all vertices in the desired local cluster. In sum, algorithms 1 and 3 need at least $\mathscr{O}[1/\{\alpha(1-\alpha)\}]$ queries (see the on-line supplementary materials section S2 for an example). This implies that we can approach the conditions in corollary 1 by setting the teleportation constant sufficiently large, whereas the computational burden can increase as $\alpha \to 1$.

5. Simulation studies

This section compares the PPR vector and the APPR vector. The results show the effectiveness and robustness of the APPR vector in detecting a local cluster. Experiment 1 utilizes the DCSBM with a power law degree distribution and investigates the effects of heterogeneous node degrees. Experiment 2 uses the SBM with unequal block sizes to study the influences of heterogeneous block degrees. Experiment 3 generates networks from the SBM with equal block sizes and varying edge density to examine the efficacy of PPR methods in sparse graphs.

In all simulations, we employ the block connectivity matrix $\mathbf{B}$ with homogeneous diagonal elements $\mathbf{B}_{ii} = b_1$ and homogeneous off-diagonal elements $\mathbf{B}_{ij} = b_2$ for any $i \neq j$. Define the signal-to-noise ratio to be the expected number of in-block edges divided by the expected number

of out-block edges, i.e. $b_1/\{b_2(K-1)\}$, where K is the number of blocks. In particular, we set the signal-to-noise ratio to 1.5 and choose a teleportation constant of $\alpha = 0.15$ throughout the section. Additional simulation results (illustrating theorem 2) are available from the on-line supplementary materials section S2.

5.1. Experiment 1

Experiment 1 illustrates how node degree heterogeneity affects the discriminant power in identifying a local cluster by using a PPR vector or an APPR vector. The results also illustrate the advantages of having multiple seed nodes. The Θ-parameters from the DCSBM are drawn from the power law distribution with lower bound $x_{\min} = 1$ and shape parameter $\beta = 2.5$. Random networks were sampled from the DCSBM with $K = 3$, $N = 1500$ and equal block sampling proportions:

$$z(v) \overset{\text{IID}}{\sim} \text{multinomial}(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}),$$

for vertex $v = 1, 2, \ldots, N$, whose expected average degree δ is set to 105. The PPR vector is calculated with one or 10 seeds randomly chosen from block 1.

Fig. 1 plots PPR values (Figs 1(a) and 1(b)) and APPR values (Figs 1(c) and 1(d)) of a random graph generated from the DCSBM, excluding seed node(s). Figs 1(a) and 1(c) contrast the PPR and APPR when there is only one seed node and Figs 1(b) and 1(d) compare two vectors when 10 seed nodes are used. The vertices from the local block in the SBM are coloured in blue and the others are in yellow. The nodes are ordered first by block and then by node degree parameters θ (left is larger). A horizontal line is drawn for each block indicating the median of the APPR values within that block.

With one seed node (Figs 1(a) and 1(c)), the scatter plots have two clouds within each block. The upper cloud contains the immediate neighbours of the seed node. This separation disappears when multiple seed nodes are used (Figs 1(b) and 1(d)). To see the effect of node heterogeneity, the skewed distribution of PPR values in each block demonstrates its bias towards high degree nodes inside and outside the seed nodes block in the SBM. In contrast, APPR values are evenly distributed within blocks, verifying that the APPR vector removes the effects of node degree heterogeneity.

5.2. Experiment 2

Experiment 2 compares PPR and APPR under the SBM with block degree heterogeneity. A number of random networks were sampled from the SBM with $K = 3$, $N = 900$ and geometric block sampling proportions:

$$z(v) \overset{\text{IID}}{\sim} \text{multinomial}(1, b, b^2), \tag{9}$$

where $b \in \{1.0, 1.2, 1.4, 1.6, 1.8, 2.0\}$. When b is larger, the population of nodes in each block becomes more unbalanced and thus induces greater block degree heterogeneity. The block connectivity matrix $\mathbf{B}$ is configured as described at the beginning of this section. The expected average degree δ is set to 70. For each sampled network, the size of the first block is assumed known to algorithm 2. The PPR vector is calculated exactly in place of the approximation PPR vector (step 1), with one seed randomly chosen from the first block.

Fig. 2(a) displays the PPR vector on an example network with $b = 1.4$, demonstrating its preference towards the high degree block (the third block) over local clusters. Fig. 2(b) depicts the APPR vector on the same network. Given the size of the first block, we measure the accuracy by the proportion of vertices belonging to the first block in the cluster returned. Fig. 2(c) shows

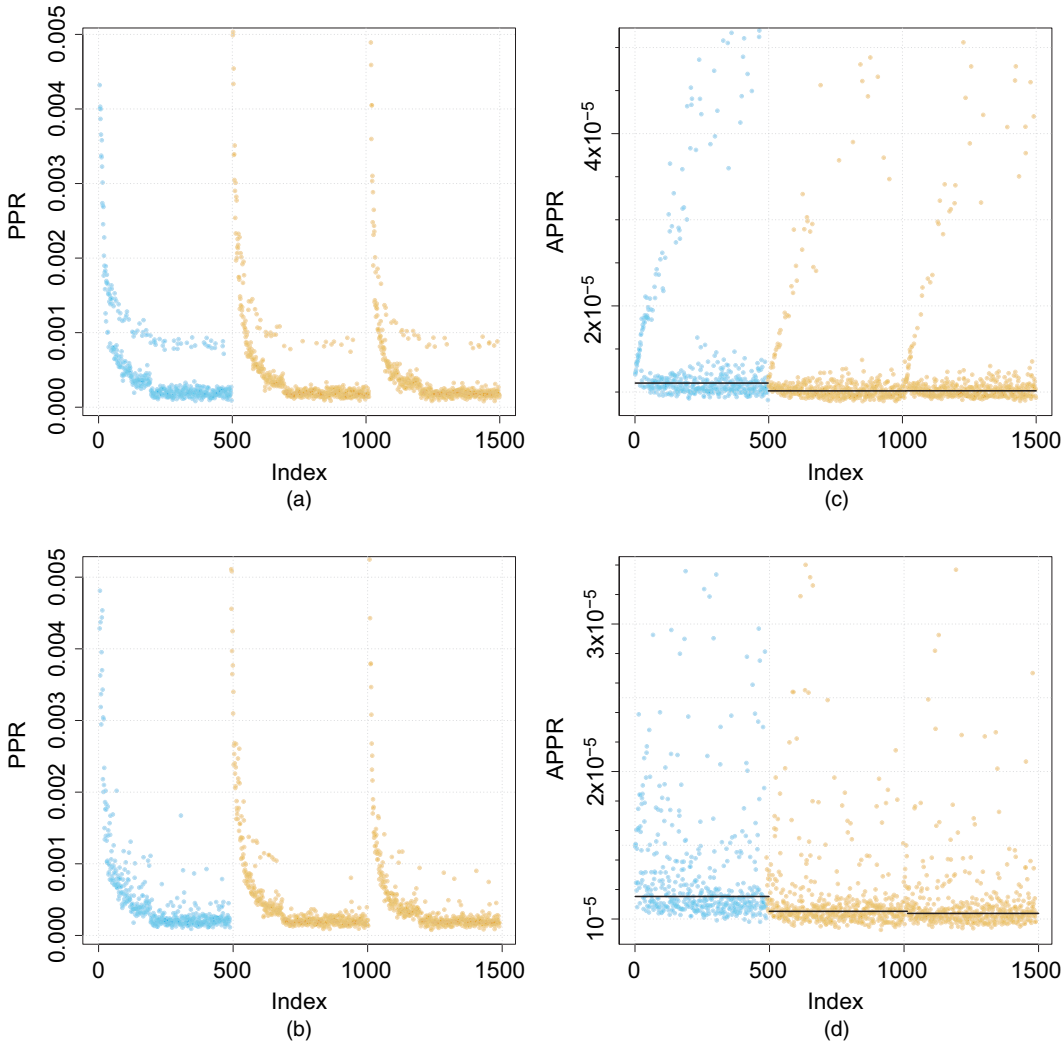

Fig. 1. Comparison of (a), (b) PPR and (c), (d) APPR under the DCSBM with (a), (c) one seed node and (b), (d) 10 seed nodes: ●, local cluster; ●, other clusters; ———, median of APPR values within each cluster

the accuracy of PPR and APPR for six values of b (i.e. the geometric ratio in distribution (9)) where each point is the average of 100 sampled networks. The comparison demonstrates that the APPR vector corrects the bias of PPR caused by block heterogeneity. Moreover, block degree heterogeneity degrades the performance of both PPR and APPR. Note that APPR outperforms PPR even when $b = 1$; this is probably because, even when nodes have equal expected degrees in the SBM, the actual node degrees will be heterogeneous because of the randomness in the sampled graph. In a finite graph, this variability is enough to give APPR an advantage over PPR. Asymptotically, this advantage should fade away (Kloumann *et al.*, 2017).

5.3. Experiment 3

Experiment 3 investigates the performance of PPR and APPR under the SBM where there is no

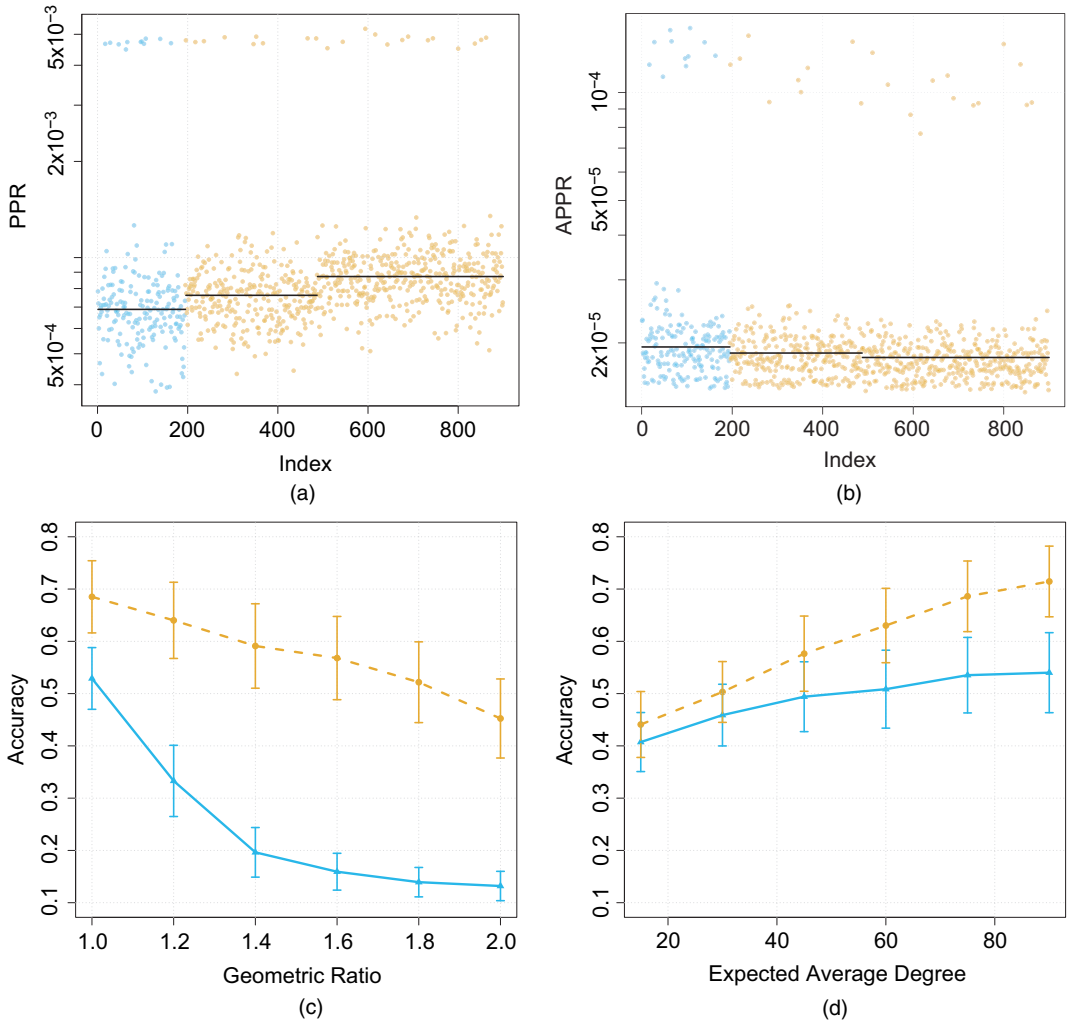

Fig. 2. (a) Simulated network generated from the classic SBM of three blocks with block degree heterogeneity (———, median of PPR values within each cluster), (b) comparison of performance for PPR (▲) and APPR (●) under the SBM with different levels of block degree heterogeneity and (c) comparison of performance for PPR (▲) and APPR (●) under the four-parameters SBM with different sparsity: error bars are drawn by using the standard deviation

heterogeneity in the expected node degrees or block degrees. A number of random networks were sampled from the four-parameter SBM ($K = 3, N = 900, b_1 = 0.6, b_2 = 0.2$) (Rohe *et al.*, 2011). Under the four-parameter SBM, each of K blocks has equal size in expectation, N/K, and the probability of a connection between two nodes is b_2 if they are in two separate blocks, or b_1 if in the same block. In addition, the expected average degree varies: $\delta \in \{15, 30, 45, 60, 75, 90\}$. For every setting, the results are averaged over 100 samples of the network. The PPR vector is calculated with one seed randomly chosen from block 1. Fig. 2(d) contrasts the accuracy of PPR and APPR against six values of expected average degree, showing that, when the sampled graph has minimal degree heterogeneity, the adjusted PPR vector has only slightly higher accuracy than the PPR vector.

6. A sample of Twitter

In this section, we provide a more detailed case-study to illustrate the properties of different PPR vectors. We obtain a local cluster of nodes around the seed node @NBCPolitics (NBC Politics) in the Twitter friendship graph. In the Twitter graph, the nodes are called handles or accounts (e.g. @NBCPolitics) and, if Twitter handle i follows Twitter handle j, then we define this as a directed edge (i, j) pointing from i to j. Affiliated with NBC News, NBC Politics specializes in political news coverage and had over 470000 followers on Twitter (in-degree) and follows 145 handles (out-degree) in December 2018. A brief look through @NBCPolitics's following list reveals that it follows a wide range of accounts, from television programmes, reporters and editors affiliated with the National Broadcasting Company (NBC), to media accounts and journalists of other news outlets as well as politicians.

Data on following and handle profile information were collected through the standard Twitter search application programming interface. We queried the Twitter friendship graph starting from the seed node @NBCPolitics, using algorithm 3 with teleportation constant $\alpha = 0.15$ and termination parameter $\epsilon = 10^{-7}$, ending up with 5840 surrounding handles. Through this exercise, we intend to illustrate the properties and applications of local clustering by using PPR, APPR and RPPR vectors, where we set the regularization parameter τ to 100.

We first present the results of PPR. As Table 7 shows, the top 30 handles (except @NBCPolitics) with the highest PPR values are a combination of

(a) NBC's news-related programmes such as NBC News, TODAY and Meet the Press,
(b) NBC's political reporters, anchors and editors, from well-known figures like Chuck Todd and Andrea Mitchell to less-known figures like Pete Williams (Justice Correspondent) and Mark Murray (Senior Political Editor);
(c) other mainstream news outlets such as the *Wall Street Journal*, *POLITICO* and *TIME*, and
(d) prominent public figures and politicians like Melania Trump, Bill Clinton and John McCain.

In light of NBC's status as a mainstream news outlet and the political focus of @NBCPolitics, such results make sound sense. It must also be noted that all the top 30 handles are direct friends of @NBCPolitics's and have at least tens of thousands of followers. The median follower count is 1.4 million, suggesting high in-degrees. In fact, the pattern that is observed in the top 30 extends to the top 200 handles with the highest PPR values, which include NBC's own programmes, journalists, editors and staff, fellow mainstream media outlets and their staff, and prominent public figures, politicians and government institutions (see the on-line supplementary materials section S4). The median in-degree of the top 200 handles is around 184000, though there are four handles with fewer than 1000 followers. One important thing to note is that, among the top 200 handles, the first 139 are all directly followed by @NBCPolitics, with handles having high in-degrees generally ranked higher than those having low in-degrees (although @NBCPolitics follows 145 handles, six of them might have privacy protection that has prevented us from accessing their information). The remaining handles that are on the list, although not directly followed by @NBCPolitics, include five handles that are associated with NBC, from its news anchor Lester Holt to its News International President. However, the majority of those indirectly followed by @NBCPolitics are mainly high profile political and public figures (like President Trump, Vice-President Pence, Hillary Clinton and Stephen Colbert), government organizations (like the White House Office of Cabinet Affairs and National Security Council) and mainstream news outlets (like the *New York Times*, Cable News Network and the Associated Press) and well-known journalists (like John Dickerson and Anderson Cooper). We can thus conclude that the

Table 7. Top 30 handles of PPR with seed node @NBCPolitics and the teleportation constant $\alpha = 0.15$ in December 2018†

Rank	*Name*	*Followers*	*Description*
1	Melania Trump	11242283	This account is run by the Office of First Lady Melania…
2	The White House	17625630	Welcome to @WhiteHouse!: follow for the latest from…
3	Chuck Todd	2032038	Moderator of @meetthepress and @nbcnews political…
4	NBC News	6280551	The leading source of global news and info for more than…
5	NBC Nightly News	962290	Breaking news, in-depth reporting, context on news from…
6	Andrea Mitchell	1737764	NBC News Chief Foreign Affairs Correspondent/…
7	Savannah Guthrie	881669	Mom to Vale & Charley, TODAY Co-Anchor,…
8	Joe Scarborough	2521215	With Malice Toward None
9	MSNBC	2261911	The place for in-depth analysis, political commentary…
10	Rachel Maddow MSNBC	9498076	I see political people…
11	Breaking News	9223158	
12	NBC News First Read	53847	The first place for news and analysis from the @NBC…
13	TODAY	4276453	America's favorite morning show | Snapchat: todayshow
14	Meet the Press	566713	Meet the Press is the longest-running television show…
15	The *Wall Street Journal*	16188842	Breaking news and features from the WSJ
16	Pete Williams	70062	NBC News Justice Correspondent: covers US Supreme…
17	Mark Murray	97571	Mark Murray is the senior political editor for NBC…
18	*POLITICO*	3695835	Nobody knows politics like POLITICO: got a news tip…
19	Katy Tur	587474	MSNBC anchor @2pm, NBC News correspondent,…
20	Bill Clinton	10697521	Founder, Clinton Foundation and 42nd President of the…
21	Kasie Hunt	381704	@NBCNews Capitol Hill Correspondent: host,…
22	*TIME*	15584815	Breaking news and current events from around the globe:…
23	Kelly O'Donnell	195765	White House Correspondent @NBCNews Veteran of…
24	John McCain	3181773	Memorial account for U.S. Senator John McCain,…
25	Peter Alexander	283522	@NBCNews White House Correspondent/Weekend…
26	Hallie Jackson	359099	Chief White House Correspondent/@NBCNews/…
27	Kristen Welker	182244	@NBCNews White House Correspondent: links and…
28	Carrie Dann	37119	@NBCNews / @NBCPolitics: RTs not endorsements
29	Willie Geist	807536	Host @NBC #SundayTODAY, Co-Host @Morning_Joe…
30	Morning Joe	563650	Live tweet during the show!: links to must-read op-eds…

†Through the PPR vector, the top 30 handles returned to @NBCPolitics include NBC's news-related programmes and celebrity reporters and comparable mainstream media outlets, as well as prominent political and public figures and institutions. Such results line up with its status as a mainstream political news source, demonstrating clustering effectiveness. Those Twitter handles tend to have millions of followers, showing the PPR vector's bias towards high in-degree.

PPR vector is biased towards popular accounts followed directly by the seed node or indirectly by its friends, reflecting the popular Twitter handles that are followed by them. This property of the PPR vector can be harnessed by researchers who are interested in identifying the upstream of a handle, i.e. those Twitter elites who are followed by and might influence the seed node and by extension its followers.

In contrast, the APPR vector upweights handles that are much less popular (i.e. those with low in-degrees). As shown in Table 8, the 30 handles with the highest APPR values include NBC's reporters, writers, editors, producers and programmes, all of whom have a few hundred to a few thousand followers. The 30 handles also include those unaffiliated with NBC, such as the director of a non-profit (Enroll America), the director of digital programming at National Geographic and @CNNPolitics's editor. All of them are professionally related to the seed node. This testifies to the applicability of APPR for locating an idiosyncratic local cluster around a seed node. However, more than half (17) of the 30 handles are obscure and not directly followed by @NBCPolitics. The reason why they appear on the list is probably that they have just one and

Table 8. Top 30 handles of APPR with seed node @NBCPolitics and the teleportation constant $\alpha = 0.15$ in December 2018†

Rank	*Name*	*Followers*	*Description*
1	Stephanie Palla	198	Enroll America National Regional Director…
2	Jennifer Sizemore	386	
3	Alissa Swango	441	Director of Digital Programming at @natgeo: all things…
4	Making a Difference	670	@NBCNightlyNews' popular feature profiles ordinary…
5	Ron Whittemore	1	
6	Svante Stockselius	3	
7	Greg Martin	1161	Political Booking Producer at @nbcnews @todayshow
8	Area Man	1	I am Area Man. I pwn your news feed
9	CELESTIA ROBINSON	2	
10	NBC Field Notes	1390	NBC News correspondents and http://t.co/1eSopOQt8s…
11	rob adams	2	
12	JL	2	
13	David Kelsey	1	
14	Hank Morris	1	
15	Jesse Marks	1	
16	Brayden Rainey	1	
17	child of the tiger	3	Yet another activist twitter, fighting all those fun…
18	Julie Swango	4	
19	Author Dianne Kube	7	Dianne Kube is an Author with a passion, for family,…
20	Consider the Source	7	
21	Adam Edelman	2341	Political reporter @nbcnews: Wisconsin native,…
22	Phil McCausland	2519	@NBCNews Digital reporter focused on the rural-urban…
23	Corky Siemaszko	2538	Senior Writer at NBC News Digital (former NY Daily…
24	Sam Petulla	2588	Editor @cnnpolitics: Usually looking for datasets…
25	Ken Strickland	2693	NBC News Washington Bureau Chief
26	Mike Mullen	7	
27	Elyse PG	2697	White House producer @nbcnews |@USCAnnenberg alum…
28	A. Johnson	2	Change your thoughts & you change your world: -Normal…
29	Steve Fenton	4	
30	Dobe Pitty Mami	13	

†Through the APPR vector, the top 30 handles returned to @NBCPolitics include some relevant handles (NBC's news team and their counterparts in other mainstream news organizations) and many obscure handles (handles with few followers and no profile descriptions). This results from the APPR vector's bias towards extreme low degree and introduces noise to the clustering results.

at most a dozen followers (recall that APPR divides by in-degree). In fact, 160 of the top 200 handles are not direct friends of @NBCPolitics; the median in-degree of the top 200 handles is merely 8 (on-line supplementary materials section S4). Those handles might have ended up on the list through a combination of luck and, more importantly, their extremely low in-degrees. In this regard, noise can be introduced by the APPR vector because it prioritizes handles with extremely low in-degrees that are possibly several degrees separated from the seed node.

To reduce noise, we applied a regularization step to the APPR vector to remove those 'distant' and small nodes while preserving the close and relevant nodes. In Table 9, the majority of the top 30 handles with the highest regularized APPR (i.e. RPPR) values have three- or four-digit numbers of followers. Similarly to the APPR results, they include NBC's news crew. But the difference is that the overwhelming majority (18) of the top 30 handles work at NBC. Some handles who work for other news organizations (e.g. Sam Petulla at @cnnpolitics and Emmanuelle Saliba at @Euronews) might have previously worked at NBC or have a close connection with its news team. Even the four handles that are not directly followed by @NBCPolitics are interesting—they are non-profit organizations (NYC Clothing Bank and Voices United) and a news-related

Table 9. Top 30 handles of RPPR with seed node @NBCPolitics and the teleportation constant $\alpha = 0.15$ in December 2018†

Rank	*Name*	*Followers*	*Description*
1	Stephanie Palla	198	Enroll America National Regional Director http://t.co/X6jJIE…
2	Jennifer Sizemore	386	
3	Alissa Swango	441	Director of Digital Programming at @natgeo: all things food…
4	Making a Difference	670	@NBCNightlyNews' popular feature profiles ordinary people do…
5	Greg Martin	1161	Political Booking Producer at @nbcnews @todayshow
6	NBC Field Notes	1390	NBC News correspondents and http://t.co/1eSopOQt8s reporters…
7	Adam Edelman	2341	Political reporter @nbcnews: Wisconsin native, Bestchester…
8	Phil McCausland	2519	@NBCNews Digital reporter focused on the rural-urban divide…
9	Corky Siemaszko	2538	Senior Writer at NBC News Digital (former NY Daily News…
10	Sam Petulla	2588	Editor @cnnpolitics: usually looking for datasets; you can…
11	Ken Strickland	2693	NBC News Washington Bureau Chief
12	Elyse PG	2697	White House producer @nbcnews |@USCAnnenberg alum | LA kid…
13	Hasani Gittens	3002	Level 29 Mage: senior News Ed. @NBCNews; sheriff of Nattahna…
14	Scott Foster	3464	Senior Producer, Washington @NBCNEWS @TODAYshow
15	Zach Haberman	3693	Lead Breaking News Editor, @NBCNews: previously had other jobs…
16	Emmanuelle Saliba	4004	Head of Social Media Strategy @Euronews | Launched #THECUBE…
17	Alex Johnson	4371	News, data and analysis for @NBCNews; data geek;…
18	Savannah Sellers	4637	News junkie: host of NBC's "Stay Tuned" on Snapchat…
19	NYC Clothing Bank	154	We distribute new, never-worn clothing and merchandise…
20	Shaquille Brewster	5362	@NBCNews Producer/Politics | @HowardU Alum| Journalist…
21	Joey Scarborough	6277	NBC News Social Media Editor: New York Daily News Alum; RTs…
22	Jane C. Timm	6478	@nbcnews political reporter and fact checker: more fun than…
23	Anthony Terrell	6827	Emmy Award winning journalist: political observer; covered…
24	NBC News Videos	7838	The latest video from http://t.co/xPyvMOTEF6
25	Libby Leist	7946	Executive Producer @todayshow
26	Voices United	310	Voices United is a non profit educational organization…
27	Social Headlines	344	Daily roundup of top social media and networking stories
28	James Miklaszewski	337	Writer, Photographer, Editor, Director, Producer, Newshound…
29	Courtney Kube	9494	NBC News National Security & Military Reporter…
30	Bob Corker	10042	Serving Tennesseans in the U.S. Senate

†Through the RPPR vector, the top 30 handles returned to @NBCPolitics include much fewer low in-degree and obscure handles and many more moderately connected nodes that are relevant to @NBCPolitics, including its reporters and editors and media professionals from other organizations.

individual or organization (James Miklaszewski and Social Headlines). This pattern can also be observed in the top 200 handles, 72 of whom are directly followed by @NBCPolitics. The overwhelming majority of those who are directly followed by it are affiliated with NBC, comprising its day-to-day news team, who enjoy much less publicity than the celebrity reporters. The remaining 128 of them, who are not directly followed by @NBCPolitics, actually also include 20 of NBC's journalists and staff, such as Ray Farmer (NBC News photographer) and Jim Miklaszewski (Chief Pentagon Correspondent for NBC News). Others are non-profit organizations like Vets Helping Heroes and professionals from other news organizations or companies such as the *Wall Street Journal*, National Football League Network and Microsoft, who might have worked for NBC or have a close connection with it. Although there still appear to be obscure handles with few followers, they decrease significantly in number—the median in-degree of the top 200 handles is 340 (on-line supplementary materials section S4): a precipitous drop from that of the top PPR handles yet not too small compared with that of the top APPR handles. We thus conclude that the regularized APPR vector returns a local cluster with little noise, reflecting a seed node's close circles, either directly or indirectly related.

To evaluate the influence of the desired cluster size n on the results based on different PPR vectors, we compare the local clusters of PPR, APPR and RPPR by varying sample size. Define the *in-and-out ratio* of local cluster $C \subset V$ as the proportion of edges inside C among all edges that are connected to C:

$$\frac{2 \sum_{u,v \in C} A_{uv}}{\sum_{u \in C} d_u^{\text{in}} + d_u^{\text{out}}}.$$

A higher in-and-out ratio indicates a more internally connected sample. Fig. 3(b) shows the effectiveness of APPR and RPPR in producing a compact local cluster. When the sample size is bigger than 100, the connectedness of the local cluster produced by RPPR stabilizes; the greater the sample size, the more densely connected a cluster that APPR would produce. However, PPR is easily susceptible to the inclusion of popular nodes. In this case, a sharp drop of the in-and-out ratio for PPR when the sample size reaches around 140 is caused by inclusions of highly popular accounts @POTUS (President Trump) and @realDonaldTrump (Donald J. Trump).

The PPR clustering is fairly robust to the choice of teleportation constant, despite the size of local cluster. To illustrate this, we also performed the same pipeline of analysis with the seed @NBCPolitics while varying the value of α (e.g. 0.05, 0.25 and 0.33) in parallel. We observed that those local clusters returned by algorithm 4 all share a great portion of members in common. For example, there are 280 (93.3%) overlapping members between two targeted samples of size $n = 300$, using $\alpha = 0.15$ and $\alpha = 0.25$ respectively. These suggest a low sensitivity to the teleportation constant (see the on-line supplementary materials section S2).

Fig. 3(a) depicts the behaviours of PPR, APPR and RPPR. Each handle queried in this sampling is displayed as a dot, with the y-axis representing the PPR value and x-axis the number of followers (i.e. in-degree). Top handles with the highest PPR values are above the blue broken line, which tend to concentrate on the right-hand end of the x-axis and thus are biased towards high in-degrees. Top handles with the highest APPR values are dots to the left of the yellow chain curve, which gather on the left-hand end of the x-axis and thus in favour of low in-degrees. Regularized APPR, by purple dots, excludes the very low degree nodes and very high degree nodes. As the empirical results show, these three vectors can be thought of as lenses through which we view the local structure of a given Twitter handle with varying foci, rendering high, moderate and low in-degree blocks and serving different needs and purposes.

7. Discussion

This paper studies the PPR vector under the DCSBM and PPR clustering in massive block model graphs. We establish some consistency results for this method and examine its performance through analysis of Twitter friendship graph. As shown in the results, the PPR vectors with and without adjustment have distinct properties and can be used to sample a massive graph effectively for various purposes. However, there are limitations that are worthy of future investigations.

In Section 3, we provide a representation of the PPR vector under the DCSBM and its extension into directed graphs. The result does not impose extra structural restrictions on the model parameters, except that $\mathbf{B}$ corresponds to a strongly connected blockwise graph. We consider a positive definite connectivity matrix particularly so that it is intuitive to conceive the notion of a local cluster. In practice (and many of our experiments; see the on-line supplementary materials section S2), however, a PPR-type algorithm appears to continue working for a broader range of $\mathbf{B}$ (e.g. singular or indefinite), provided that the teleportation constant is sufficiently

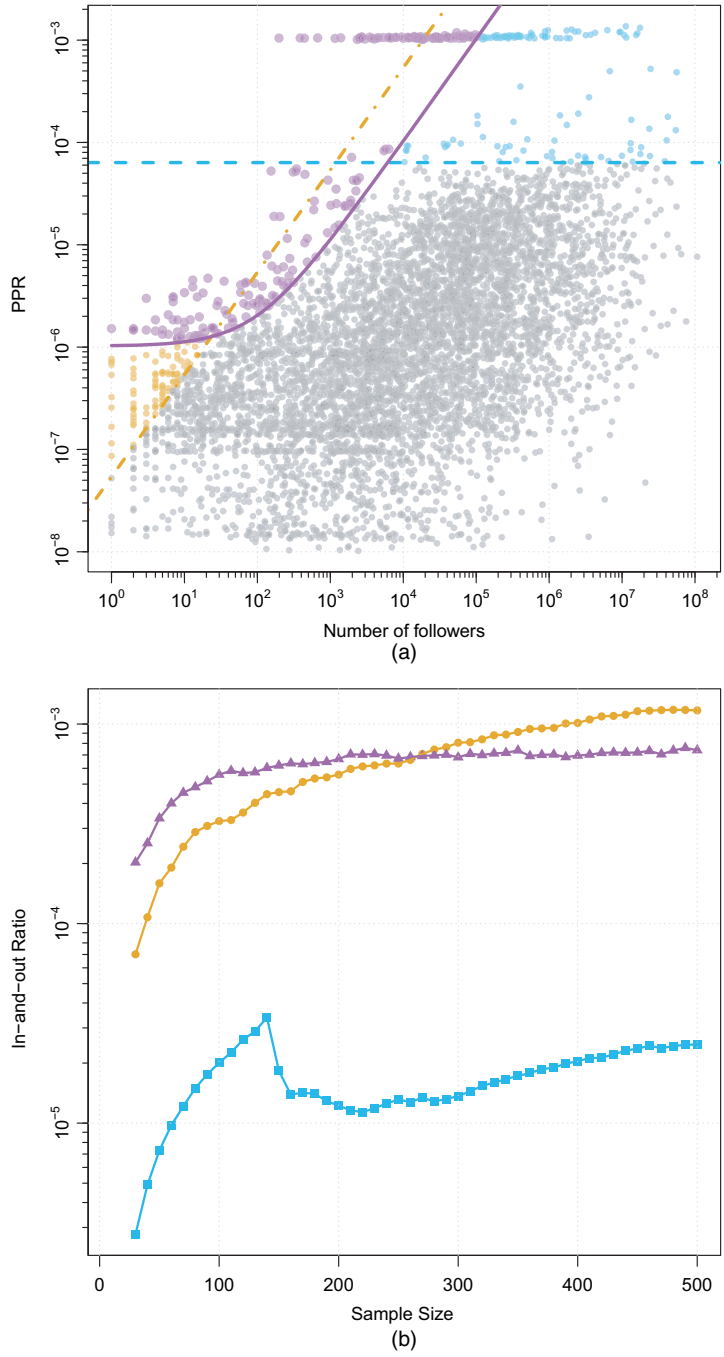

Fig. 3. (a) Illustration of 5840 Twitter handles examined by algorithm 3 and three samples of size 200 by PPR, APPR and RPPR (each dot represents a user in Twitter) (– – –, top 200 handles by PPR vector (vertices above the line are PPR's sample); – · –, sample returned by algorithm 4 given $n = 200$ (vertices above this boundary correspond to APPR's sample); ●, sample of RPPR; ———, boundary of this sample) and (b) in-and-out ratio of local clusters identified by PPR (■), APPR (●) and RPPR (▲), as the sample sizes vary (a higher in-and-out ratio indicates a more internally connected cluster)

large (e.g. $\alpha > 0.1$). It is unclear yet what is the minimum constraint needed on $\mathbf{B}$ for the PPR clustering to function. In addition, the DCSBM does have its limits. For example, the model fails to capture either mixed block membership or popularity features which are potentially informative in real world networks. The behaviour of a PPR vector under other extensions of SBMs, such as the mixed membership SBM and popularity-adjusted block model, remains unknown (Airoldi *et al.*, 2008; Sengupta and Chen, 2018). Future studies on the PPR vector under these models could shed further light on the PPR clustering and offer more practical guidelines on their application.

In Section 4, we proved the consistency of PPR clustering, requiring the average expected node degree to grow of the order of $\log(N)$, which hits the boundary between the theoretical guarantees and the realistic observation. In contrast, scale-free networks such as the preferential attachment model (Barabási and Albert, 1999) have finite expected node degrees. Future investigations into variants of PPR that could possibly overcome this limitation yet ensure a fine local cluster discovery would be particularly interesting and useful.

In Section 6, we introduce the regularized version of the APPR (the RPPR) vector, with a series of empirical evidence showing its efficacy in targeted sampling. Although the results appear promising, theoretical guarantees for this technique remain unexplored. For some mathematical analyses, one may resort to the techniques that were used in Le *et al.* (2016). It has previously been shown that the regularized graph Laplacian (or transition matrix) enjoys finite sample convergence properties, which facilitate the consistency of many regularized spectral methods. It thus is a reasonable conjecture that RPPR vectors are also suitable for local clustering.

An R implementation of PPR clustering is available from `https://github.com/RoheLab/aPPR`.

Acknowledgements

This research is supported by National Science Foundation grant DMS-1612456 and DMS-1916378 and Army Research Office grant W911NF-15-1-0423. We thank Yuling Yan and E. Auden Krauska for their helpful comments. We thank Alex Hayes for kindly advising on the software development.

Appendix A: Technical proofs

A.1. Proof of proposition 1

We apply the Perron–Frobenius theorem for the first part (Perron, 1907; Frobenius, 1912) and complete the proof by construction.

(a) First, we show that Q is a Markov transition matrix by modifying $G = (V, E)$. For this, we shrink the weights of every existing edge by a factor $1 - \alpha$ and add an edge-weighted α between seed node v_0 and all nodes in the graph. Then Q represents the new graph $G'(V, E')$, which is strongly connected by construction. Hence Q is irreducible.

The PPR vector p is all positive. To see this, note that the equation $p^{\mathrm{T}} = p^{\mathrm{T}} Q$ implies that p is a stationary distribution for the standard random walk on G'. Since G' is strongly connected, it follows that the stationary distribution must be all positive.

From the Perron–Frobenius theorem, the only all positive eigenvector of a non-negative irreducible matrix is associated with the leading eigenvalue, which is 1 in our case. Since the leading eigenvalue of a non-negative irreducible matrix is simple, we conclude that p is unique.

(b) We finish the proof by constructing an explicit form of the PPR vector. Let $R_\alpha = \alpha \Sigma_{s=0}^{\infty} (1-\alpha)^s P^s$. The infinite sum converges for $\alpha \in (0, 1]$. Then, $p = R_\alpha^{\mathrm{T}} \pi$ satisfies the definition of the PPR vector,

$$\alpha \pi^{\mathrm{T}} + (1-\alpha)\pi^{\mathrm{T}} R_\alpha P = \alpha \pi^{\mathrm{T}} + (1-\alpha)\pi^{\mathrm{T}} \left\{ \alpha \sum_{s=0}^{\infty} (1-\alpha)^s P^s \right\} P$$

$$= \alpha\pi^{\mathrm{T}} + \alpha\sum_{s=1}^{\infty}(1-\alpha)^s\pi^{\mathrm{T}}P^s$$
$$= \pi^{\mathrm{T}}R_\alpha.$$

Since the solution is unique, we have $p = R_\alpha^{\mathrm{T}}\pi$.

A.2. Proof of proposition 2

Algorithm 1 maintains two vectors, p^ϵ and r, by transporting probability mass from r to p^ϵ at each updating step. Note that the termination criterion implies that $r_u < \epsilon d_u$ for any u sampled; thus it suffices to prove that

$$|p_u - p_u^\epsilon| \leqslant r_u.$$

For a fixed α, let $p(x)$ be the PPR vector with preference vector $x \in \mathbb{R}^N$ satisfying $x_i \geqslant 0$ and $\|x\|_1 \leqslant 1$. Then $p(\pi)$ is the exact PPR vector as in equation (2). Since $p(x)^{\mathrm{T}}P = p(x^{\mathrm{T}}P)$, we have (Jeh and Widom, 2003)

$$p(x) = \alpha x + (1-\alpha)p(P^{\mathrm{T}}x). \tag{10}$$

We argue that $p^\epsilon + p(r)$ is invariant in updating steps. To see this, suppose that $(p^\epsilon)'$ and r' are the results of performing one update on p^ϵ and r after sampling node u. We have

$$(p^\epsilon)' = p^\epsilon + \alpha r_u e_u,$$
$$r' = r - r_u e_u + (1-\alpha)r_u P^{\mathrm{T}}e_u,$$

where e_u is the unit vector on the direction of u. Then,

$$\begin{aligned} p(r) &= p(r - r_u e_u) + p(r_u e_u) \\ &\overset{\text{(i)}}{=} p(r - r_u e_u) + \alpha r_u e_u + (1-\alpha)p(r_u P^{\mathrm{T}}e_u) \\ &\overset{\text{(ii)}}{=} p\{r - r_u e_u + (1-\alpha)r_u P^{\mathrm{T}}e_u\} + \alpha r_u e_u \\ &= p(r') + (p^\epsilon)' - p^\epsilon, \end{aligned}$$

where equality (i) is applying equation (10) at $x = r_u e_u$ and equality (ii) comes from the linearity of a PPR vector in the preference vector.

The desired result follows from recognizing that $p^\epsilon + p(r)$ is initially $\mathbf{0} + p(\pi)$ and that, when the algorithm terminates, $[p(r)]_u \leqslant r_u$ for any sampled u.

Remark 1. If $\epsilon d_1 > 1$, algorithm 1 terminates after the first round and simply outputs $p = \mathbf{0}$. Under this circumstance, proposition 2 still holds, because $|p_u - p_u^\epsilon| \leqslant |p_u| + |p_u^\epsilon| \leqslant 1$.

A.3. Lemmas for the degree-corrected stochastic block model

Lemma 2 (properties of the DCSBM). Under the population directed DCSBM with K blocks and parameters $\{\mathbf{B}, Z, \Theta^{\text{in}}, \Theta^{\text{out}}\}$,

(a) $\mathbf{D}^{\text{in}} = Z^{\mathrm{T}}\mathscr{D}^{\text{in}}Z$, and $\mathbf{D}^{\text{out}} = Z^{\mathrm{T}}\mathscr{D}^{\text{out}}Z$, and
(b) $\mathscr{A}_v^{\text{in}} = \theta_v^{\text{in}}\mathbf{d}_{z(v)}^{\text{in}}$, and $\mathscr{A}_v^{\text{out}} = \theta_v^{\text{out}}\mathbf{d}_{z(v)}^{\text{out}}$.

Proof. Result (a) is an alternative way of writing the definition. For result (b), we prove the first equation. Recall that, for any i, $\Sigma_{u:z(u)=i}\theta_u^{\text{out}} = 1$; then, by definition,

$$\mathscr{A}_v^{\text{in}} = \sum_u \theta_u^{\text{out}}\theta_v^{\text{in}}B_{z(u)z(v)} = \theta_v^{\text{in}}\sum_{j=1}^{K}\left(\mathbf{B}_{jz(v)}\sum_{u:z(u)=j}\theta_u^{\text{out}}\right) = \theta_v^{\text{in}}\mathbf{d}_{z(v)}^{\text{in}}.$$

Remark 2. Since $Z^{\mathrm{T}}\Theta^{\text{in}}Z = I_K$, result (a) implies that $(\mathscr{D}^{\text{in}})^{-1}\Theta^{\text{in}}Z = Z(\mathbf{D}^{\text{in}})^{-1}$.

Lemma 3 (explicit form of $\mathscr{P}$ and its powers). Under the population directed DCSBM with K blocks and parameters $\{\mathbf{B}, Z, \Theta^{\text{in}}, \Theta^{\text{out}}\}$, the population graph transition is the product

$$\mathscr{P} = Z\mathbf{P}Z^{\mathrm{T}}\Theta^{\mathrm{in}},$$

and its matrix powers are

$$\mathscr{P}^k = Z\mathbf{P}^k Z^{\mathrm{T}}\Theta^{\mathrm{in}}.$$

Proof. By definition and lemma 2, part (b), for any $u, v \in V$,

$$\mathscr{P}_{uv} = (\theta_u^{\mathrm{out}}\mathbf{d}_{z(u)}^{\mathrm{out}})^{-1}\theta_u^{\mathrm{out}}\theta_v^{\mathrm{in}}\mathbf{B}_{z(u)z(v)} = \theta_v^{\mathrm{in}}\mathbf{B}_{z(u)z(v)}/\mathbf{d}_{z(u)}^{\mathrm{out}} = \theta_v^{\mathrm{in}}\mathbf{P}_{z(u)z(v)}.$$

For the powers of $\mathscr{P}$, noting that $Z^{\mathrm{T}}\Theta^{\mathrm{in}}Z = I_K$,

$$\mathscr{P}^2 = Z\mathbf{P}Z^{\mathrm{T}}\Theta^{\mathrm{in}}Z\mathbf{P}Z^{\mathrm{T}}\Theta^{\mathrm{in}} = Z\mathbf{P}^2Z^{\mathrm{T}}\Theta^{\mathrm{in}}.$$

The result desired follows from the principle of induction on the kth power.

A.4. Proof of theorem 1

By proposition 1 and lemma 3, we have

$$\begin{aligned}
\pi &= \alpha\sum_{s=0}^{\infty}(1-\alpha)^s(\mathscr{P}^s)^{\mathrm{T}}\pi \\
&= \alpha\sum_{s=0}^{\infty}(1-\alpha)^s\Theta^{\mathrm{in}}Z(\mathbf{P}^s)^{\mathrm{T}}Z^{\mathrm{T}}\pi \\
&= \Theta^{\mathrm{in}}Z\left\{\alpha\sum_{s=0}^{\infty}(1-\alpha)^s(\mathbf{P}^s)^{\mathrm{T}}\boldsymbol{\pi}\right\} \\
&= \Theta^{\mathrm{in}}Z\mathbf{p}.
\end{aligned}$$

In addition, it follows from lemma 2, part (a), that

$$\pi^* = (\mathscr{D}^{\mathrm{in}})^{-1}\pi = (\mathscr{D}^{\mathrm{in}})^{-1}\Theta^{\mathrm{in}}Z\mathbf{p} = Z(\mathbf{D}^{\mathrm{in}})^{-1}\mathbf{p} = Z\mathbf{p}^*.$$

This completes the proof.

A.5. Proof of lemma 1

For any $\alpha > 0$, the PPR vector with seed node $v_0 = 1$ is the solution to the equation $\pi^{\mathrm{T}} = \pi^{\mathrm{T}}\mathbb{Q}$, where $\mathbb{Q} = \alpha\Pi + (1-\alpha)\mathscr{P}$. Define a sequence of probability distributions $\pi^s \in \mathbb{R}^N$ such that $\pi^s = (\mathbb{Q}^s)^{\mathrm{T}}\pi^0$, where π^0 is an arbitrary initial probability distribution. Then, $\lim_{s\to\infty}\pi^s = \pi$. For simplicity, we assume that π^0 is close to π, i.e., for any $\varepsilon > 0$ and $s \geqslant 0$,

$$\|\pi^s - \pi\|_\infty < \varepsilon/2. \tag{11}$$

This can be achieved by finding an integer $S(\varepsilon)$ that is sufficiently large and setting $\pi^0 = \pi^S$.

We first claim that

$$\max_{u\neq 1}\frac{\pi_u^{s+1}}{d_u} \leqslant (1-\alpha)\max_{u\in V}\frac{\pi_u^s}{d_u}. \tag{12}$$

In fact, for any $u \neq 1$,

$$\begin{aligned}
\pi_u^{s+1} &= \alpha\mathbb{1}_{\{u=1\}} + (1-\alpha)\sum_{v\in V}\frac{A_{vu}}{d_v}\pi_v^s \\
&\leqslant (1-\alpha)\left(\sum_{v\in V}A_{vu}\right)\max_{v\in V}\frac{\pi_v^s}{d_v} \\
&= (1-\alpha)d_u\max_{v\in V}\frac{\pi_v^s}{d_v}.
\end{aligned}$$

We then show that $\pi_1^s/d_1 > \pi_v^s/d_v$ for any $v \neq 1$ by contradiction. Suppose otherwise that $\pi_1^s/d_1 \leqslant \max_{u\neq 1}\pi_u^s/d_u$; then equation (11) implies that, for any s',

$$\frac{p_1^{s'}}{d_1} \leqslant \frac{p_1^s + \varepsilon}{d_1} \leqslant \max_{u \neq 1} \frac{p_u^s}{d_u} + \frac{\varepsilon}{d_1} \leqslant \max_{u \neq 1} \frac{p_u^{s'} + \varepsilon}{d_u} + \frac{\varepsilon}{d_1} \leqslant \max_{u \neq 1} \frac{p_u^{s'}}{d_u} + \frac{2\varepsilon}{d_{\min}},$$

where $d_{\min} = \min_{v \in V} d_v$. Hence, $\max_{u \in V} p_u^{s'}/d_u \leqslant \max_{u \neq 1} p_u^{s'}/d_u + 2\varepsilon/d_{\min}$. In addition, applying equation (12) recursively we have

$$\begin{aligned}
\max_{u \in V} \frac{p_u^s}{d_u} &= \max_{u \neq 1} \frac{p_u^s}{d_u} \\
&\leqslant (1-\alpha) \max_{u \in V} \frac{p_u^{s-1}}{d_u} \\
&\leqslant (1-\alpha) \left(\max_{u \neq 1} \frac{p_u^{s-1}}{d_u} + \frac{2\varepsilon}{d_{\min}} \right) \\
&\leqslant (1-\alpha)^s \max_{u \in V} \frac{p_u^0}{d_u} + \frac{2\varepsilon}{d_{\min}} \sum_{t=1}^{s-1} (1-\alpha)^t.
\end{aligned}$$

The inequality means that, if $d_{\min} > 0$ is fixed, p_u^s can be arbitrarily small when $s \to \infty$, which contradicts the fact that p is a probability distribution. This completes the proof.

Remark 3. When the teleportation constant is 0, the PPR vector becomes the stationary probability distribution of a standard random walk:

$$\left(\frac{d_1}{\sum_i d_i}, \frac{d_2}{\sum_i d_i}, \ldots, \frac{d_N}{\sum_i d_i} \right).$$

After adjusting by node degrees, every entry becomes identical ($1/\Sigma_i d_i$). Lemma 1 is intuitive, recognizing that the teleportation introduces a particular favour of the seed node.

Remark 4. When the edges are weighted (non-negative), the stationary distribution of a random walk is still proportional to node degrees, if one defines the degree as the sum of edge weights incident to the node (Lovász, 1993). Note also that the stationary distribution of a random walk in a directed graph is characterized by the in-degree of nodes (Ghoshal and Barabási, 2011; Lu *et al.*, 2013). The conclusion and a modified proof apply to directed or weighted graphs.

A.6. Proof of corollary 1

The algorithm ranks all vertices according to $p^{\epsilon *}$, and the population local cluster can be explicitly written as

$$\mathcal{C} = \{ v \in V : p_v^* = \mathbf{p}_1^* \}.$$

It suffices to show that

$$p_v^{\epsilon *} > p_u^{\epsilon *}, \qquad \text{for } \forall\, v \in \mathcal{C}, u \in V \backslash \mathcal{C},$$

where $p_v^{\epsilon *} = p_v^{\epsilon}/d_v$. For this, we apply the triangle inequality and obtain

$$\begin{aligned}
\frac{p_v^{\epsilon *} - p_u^{\epsilon *}}{\|p^*\|_\infty} &\geqslant \frac{p_v^* - p_u^*}{\|p^*\|_\infty} - \frac{|p_v^* - p_v^{\epsilon *}|}{\|p^*\|_\infty} - \frac{|p_u^* - p_u^{\epsilon *}|}{\|p^*\|_\infty} - \frac{|p_u^{\epsilon *} - p_u^*|}{\|p^*\|_\infty} - \frac{|p_v^{\epsilon *} - p_v^*|}{\|p^*\|_\infty} \\
&\geqslant \Delta - \frac{2\|p^* - p^*\|_\infty}{\|p^*\|_\infty} - \frac{2\|p^{\epsilon *} - p^*\|_\infty}{\|p^*\|_\infty}.
\end{aligned}$$

Since $\Delta_\alpha \leqslant 1$, assumption (8) contains condition (7) in theorem 2, which together with proposition 2 implies that

$$\frac{\|p^* - p^*\|_\infty}{\|p^*\|_\infty} < \frac{1}{4}\Delta,$$

$$\frac{\|p^{\epsilon *} - p^*\|_\infty}{\|p^*\|_\infty} < \frac{1}{4}\Delta,$$

if $\Delta^2\delta/\log(N)$ is sufficiently large. These collectively imply that $p_v^* > p_u^*$ as desired.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E. and Xing, E. P. (2008) Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, **9**, 1981–2014.

Alamgir, M. and Von Luxburg, U. (2010) Multi-agent random walks for local clustering on graphs. In *Proc. 10th Int. Conf. Data Mining* (eds G. I. Webb, B. Liu, C. Zhang, D. Gunopulos and X. Wu), pp. 18–27. Piscataway: Institute of Electrical and Electronics Engineers.

Andersen, R., Chung, F. and Lang, K. (2006) Local graph partitioning using pagerank vectors. In *Proc. 47th A. Symp. Foundations of Computer Science*, pp. 475–486. Washington DC: Institute of Electrical and Electronics Engineers Computer Society.

Andersen, R. and Lang, K. J. (2006) Communities from seed sets. In *Proc. 15th Int. Conf. World Wide Web, Edinburgh*, pp. 223–232. New York: Association for Computing Machinery.

Andersen, R. and Peres, Y. (2009) Finding sparse cuts locally using evolving sets. In *Proc. Symp. Theory of Computing, Bethesda*, pp. 235–244. New York: Association for Computing Machinery.

Barabási, A.-L. and Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Berkhin, P. (2006) Bookmark-coloring algorithm for personalized pagerank computing. *Internet Math.*, **3**, 41–62.

Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual web search engine. In *Proc. Int. Conf. World Wide Web*, pp. 107–117.

Chen, Y., Fan, J., Ma, C. and Wang, K. (2019) Spectral method and regularized MLE are both optimal for top-K ranking. *Ann. Statist.*, **47**, 2204–2235.

Davis, C. and Kahan, W. M. (1970) The rotation of eigenvectors by a perturbation: iii. *SIAM J. Numer. Anal.*, **7**, 1–46.

Frobenius, F. G. (1912) *Über Matrizen aus Nicht Negativen Elementen*. Königliche Akademie der Wissenschaften.

Gharan, S. O. and Trevisan, L. (2012) Approximating the expansion profile and almost optimal local graph clustering. In *Proc. 53rd A. Symp. Foundations of Computer Science*, pp. 187–196. Washington DC: Institute of Electrical and Electronics Engineers Computer Society.

Ghoshal, G. and Barabási, A.-L. (2011) Ranking stability and super-stable nodes in complex networks. *Nat. Communs*, **2**, article 394.

Gleich, D. F. (2015) Pagerank beyond the web. *SIAM Rev.*, **57**, 321–363.

Gupta, P., Goel, A., Lin, J., Sharma, A., Wang, D. and Zadeh, R. (2013) WTF: the who to follow service at twitter. In *Proc. 22nd Int. Conf. World Wide Web*, pp. 505–514. New York: Association for Computing Machinery.

Haveliwala, T. H. (2003) Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search. *IEEE Trans. Knowl. Data Engng*, **15**, 784–796.

Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983) Stochastic blockmodels: first steps. *Socl Netwrks*, **5**, 109–137.

Jeh, G. and Widom, J. (2003) Scaling personalized web search. In *Proc. 12th Int. Conf. World Wide Web, Budapest*, pp. 271–279. New York: Association for Computing Machinery.

Karrer, B. and Newman, M. E. (2011) Stochastic blockmodels and community structure in networks. *Phys. Rev.* E, **83**, article 016107.

Karypis, G. and Kumar, V. (1998) Multilevel k-way partitioning scheme for irregular graphs. *J. Para. Distrbtd Comput.*, **48**, 96–129.

Kloumann, I. M., Ugander, J. and Kleinberg, J. (2017) Block models and personalized PageRank. *Proc. Natn. Acad. Sci. USA*, **114**, 33–38.

Le, C. M., Levina, E. and Vershynin, R. (2016) Optimization via low-rank approximation for community detection in networks. *Ann. Statist.*, **44**, 373–400.

Liao, C.-S., Lu, K., Baym, M., Singh, R. and Berger, B. (2009) Isorankn: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, **25**, i253–i258.

Lovász, L. (1993) Random walks on graphs: a survey. In *Combinatorics, Paul Erdos is Eighty*, vol. 2 (eds D. Miklós, V. T. Sós and T. Szőnyi), pp. 1–46. Budapest: János Bolyai Mathematical Society.

Lu, X., Malmros, J., Liljeros, F. and Britton, T. (2013) Respondent-driven sampling on directed networks. *Electron. J. Statist.*, **7**, 292–322.

Macropol, K., Can, T. and Singh, A. K. (2009) RRW: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinform.*, **10**, article 283.

Newman, M., Barabási, A.-L. and Watts, D. J. (2006) *The Structure and Dynamics of Networks*. Princeton: Princeton University Press.

Page, L., Brin, S., Motwani, R. and Winograd, T. (1999) The pagerank citation ranking: bringing order to the web. *Technical Report.* Stanford InfoLab, Stanford.

Perron, O. (1907) Zur Theorie der Matrices. *Math. Ann.*, **64**, 248–263.
Qin, T. and Rohe, K. (2013) Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Proc. 26th Int. Conf. Advances in Neural Information Processing Systems, Lake Tahoe*, vol. 2, pp. 3120–3128. Red Hook: Curran Associates.
Rohe, K., Chatterjee, S. and Yu, B. (2011) Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.*, **39**, 1878–1915.
Sengupta, S. and Chen, Y. (2018) A block model for node popularity in networks with community structure. *J. R. Statist. Soc.* B, **80**, 365–386.
Spielman, D. A. and Teng, S.-H. (1996) Spectral partitioning works: planar graphs and finite element meshes. In *Proc. 37th A. Symp. Foundations of Computer Science*, pp. 96–105. New York: Institute of Electrical and Electronics Engineers.
Spielman, D. A. and Teng, S. H. (2004) Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proc. 36th A. Symp. Theory of Computing*, pp. 81–90. New York: Association for Computing Machinery.
Watts, D. J. and Strogatz, S. H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.
Zhu, Y., Yan, X. and Moore, C. (2013) Oriented and degree-generated block models: generating and inferring communities with inhomogeneous degree distributions. *J. Complx Netwrks*, **2**, 1–18.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary materials to "Targeted sampling from massive Blockmodel graphs with personalized PageRank"'.