Internet Mathematics Vol. 5, No. 1–2: 3–22

Local Partitioning for Directed Graphs Using PageRank

Reid Andersen, Fan Chung, and Kevin Lang

Abstract. A local partitioning algorithm finds a set with small conductance near a specified seed vertex. In this paper, we present a generalization of a local partitioning algorithm for undirected graphs to strongly connected directed graphs. In particular, we prove that by computing a personalized PageRank vector in a directed graph, starting from a single seed vertex within a set S that has conductance at most α, and by performing a sweep over that vector, we can obtain a set of vertices S' with conductance $\Phi_M(S') = O(\sqrt{\alpha \log |S|})$. Here, the conductance function Φ_M is defined in terms of the stationary distribution of a random walk in the directed graph. In addition, we describe how this algorithm may be applied to the PageRank Markov chain of an arbitrary directed graph, which provides a way to partition directed graphs that are not strongly connected.

1. Introduction

In directed networks like the World Wide Web, it is critical to develop algorithms that utilize the additional information conveyed by the direction of the links. Algorithms for web crawling, web mining, and search ranking all depend heavily on the directedness of the graph. For the problem of graph partitioning, it is extremely challenging to develop algorithms that effectively utilize the directed links.

Spectral algorithms for graph partitioning have natural obstacles for generalizations to directed graphs. Nonsymmetric matrices do not have a spectral decomposition, meaning that there does not necessarily exist an orthonormal basis of eigenvectors. The stationary distribution for random walks on directed

1542-7951/08 $0.50 per page

graphs is no longer determined by the degree sequences. In the earlier work of Fill [Fill 91] and Mihail [Mihail 89], several generalizations for directed graphs were examined for regular graphs. Lovász and Simonovits established a bound for the mixing rate of an asymmetric ergodic Markov chain in terms of its conductance [Lovász and Simonovits 90]. When applied to the Markov chain of a random walk in a strongly connected directed graph, their results can be used to identify a set of states of the Markov chain with small conductance. Algorithms for finding sparse cuts, based on linear and semidefinite programming and metric embeddings, have also been generalized to directed graphs [Charikar et al. 06, Chuzhoy and Khanna 06]. A Cheeger inequality for directed graphs that relies on the eigenvalues of a normalized Laplacian for directed graphs can also be used to find cuts of small conductance [Chung 05].

This paper is concerned with a different type of partitioning algorithm, called a *local partitioning algorithm*. A local partitioning algorithm finds a set with small conductance near a specified seed vertex, and can produce such a cut by examining only a small portion of the input graph. In a recent paper, the authors introduced a local partitioning algorithm, for undirected graphs, that finds a cut with small conductance by performing a sweep over a personalized PageRank vector. Personalized PageRank traditionally has been applied and studied in directed web graphs, so it is natural to ask whether this local partitioning algorithm can be generalized to find sets with small conductance in a directed graph by sweeping over a personalized PageRank vector computed in a directed graph.

In this paper, we generalize the basic local partitioning results from [Andersen et al. 06] to strongly connected directed graphs. We prove that by computing a personalized PageRank vector in a directed graph, and sorting the vertices of the graph according to their probability in this vector divided by their probability in the stationary distribution, we can identify a set with small conductance, where the notion of conductance must be generalized appropriately. Directed graphs that arise in practice are typically not strongly connected, and this generalized local partitioning algorithm cannot be applied directly to such a graph. We address this problem by describing how our algorithm may be applied to the PageRank Markov chain of a directed graph, which is ergodic even when the underlying graph is not strongly connected. When applied to the PageRank Markov chain, the generalized local partitioning algorithm has a natural interpretation: we compute a personalized PageRank vector with a single starting vertex, and a global PageRank vector with a uniform starting vector, and sort the vertices of the graph according to the ratio of their entries in the personalized PageRank vector and global PageRank vector. We prove that by sorting the vertices of the graph according to this ratio, our algorithm finds a set with small conductance in the PageRank Markov chain. We also show that the required computation can be carried out efficiently.

The generalized local partitioning algorithm has advantages and disadvantages when compared to the undirected algorithm. One advantage is that our algorithm follows outlinks exclusively, and does not travel backward over inlinks. This ensures that all the vertices in the resulting cut are reachable from the starting vertex, and is particularly useful in settings where outlinks are more easily accessible than inlinks. One disadvantage is that the appropriate generalization of conductance to directed graphs requires reweighting the edges of the graph according to the amount of probability moving over them in the stationary distribution π of a random walk, which is more complicated in a directed graph than in the undirected case. The generalized local partitioning algorithm is guaranteed to find a cut for which the total weight of outlinks crossing the cut is small, but this weight depends on π, and the cut may have a large number of outlinks with small weight.

Here is an outline of the paper. In the next section, we define the generalizations of the key ingredients of the local partitioning algorithm from [Andersen et al. 06] to strongly connected directed graphs, including personalized PageRank, conductance, sweeps, and the Lovász–Simonovits potential function. In the main section, we prove a generalization of our basic local partitioning results to strongly connected directed graphs. We prove that a sweep over a personalized PageRank vector in the directed graph produces a set with small conductance. In Section 6, we describe how to apply our algorithm to the PageRank matrix of an arbitrary directed graph, which is always strongly connected. We will show that our local algorithm can find sets with small conductance by computing personalized PageRank vectors in the original directed graph, provided we compute two global PageRank vectors offline.

2. Preliminaries

Let G be a directed graph consisting of a vertex set V and a set of directed edges E, each of which is an ordered pair (u, v) of vertices from V. Let n be the number of vertices, and m the number of directed edges. We write $d_{\text{out}}(v)$ for the out-degree of a vertex v.

Let $v_1, \ldots, v_n$ be a fixed (arbitrary) ordering of the vertices. The adjacency matrix $A = A(G)$ is the $n \times n$ matrix in which $A_{i,j} = 1$ if there is a directed edge (v_i, v_j), and zero otherwise. The out-degree matrix $D = D(G)$ is the $n \times n$ diagonal matrix in which $D_{i,i} = d_{\text{out}}(v_i)$.

For a given directed graph, we will consider several different Markov chains. For our purposes, a Markov chain M is the matrix of a random walk on a weighted directed graph on the vertex set V. Equivalently, it is an $n \times n$ probability matrix for which the sum of each row is 1. A Markov chain is said to be *ergodic* if the corresponding random walk converges to a unique stationary distribution, that is, if there exists a vector π that is nonzero at each vertex that satisfies $\pi = \pi M$ and such that for every vertex v in V, we have $\lim_{t \to \infty} 1_v M^t = \pi$. The vector π is the *stationary distribution* of M. We remark that a Markov chain is ergodic if and only if it is a random walk on a graph that is strongly connected and aperiodic. Efficient numerical methods for computing the stationary distribution of an ergodic Markov chain M are described in [Stewart 97].

Let p be a probability distribution on the vertices of V, and let M be a Markov chain. For each set $S \subseteq V$, we define the sum of p over S to be

$$p(S) = \sum_{u \in S} p(u).$$

For each edge (u, v), we define

$$p(u, v) = p(u)M(u, v).$$

This is the amount of probability that moves from u to v when a step of the Markov chain is applied to the vector p. For each set A of directed edges, we define

$$p(A) = \sum_{(u,v) \in A} p(u, v),$$

which is the total amount of probability moving over the set of directed edges. This notation is overloaded, but it is unambiguous if the type of input is known.

2.1. Conductance and Sweeps

We now assume that the Markov chain M is ergodic with a unique stationary distribution π, and define the generalizations to ergodic Markov chains of conductance, of the sweep procedure for finding cuts with small conductance (which is often used in spectral partitioning [Chung 97, Spielman and Teng 96]), and of the potential function $p[x]$ (which was introduced by Lovász and Simonovits to bound the mixing rate of random walks). In the case of ergodic Markov chains, all of these are normalized by the stationary distribution π.

Given a set S of states, we define $\bar{\pi}(S) = \min(\pi(S), 1 - \pi(S))$ to be the measure of the smaller side of the partition induced by S, and define the outgoing edge border $\partial(S)$ as follows:

$$\partial(S) = \{(u, v) \in E \mid u \in S \text{ and } v \in \bar{S}\}.$$

Definition 2.1. Let M be an ergodic Markov chain, and let π be its unique stationary distribution. We define the M-conductance $\Phi_M(S)$ of a set of vertices S to be

$$\Phi_M(S) = \frac{\pi(\partial(S))}{\bar{\pi}(S)}.$$

Definition 2.2. Let M be an ergodic Markov chain with stationary distribution π, and let p be a probability distribution on the vertices. Let $v_1, \ldots, v_n$ be an ordering of the vertices such that

$$\frac{p(v_i)}{\pi(v_i)} \geq \frac{p(v_{i+1})}{\pi(v_{i+1})}.$$

For each integer j in $\{1, \ldots, n\}$, we define $S_j^p = \{v_1, \ldots, v_j\}$ to be the set containing the top j vertices in this ordering. We define $\Phi_M(p)$ to be the smallest M-conductance among the sets $S_1^p, \ldots, S_n^p$,

$$\Phi_M(p) = \min_{j \in [1,n]} \Phi_M(S_j^p).$$

The process of sorting the vertices according to this ordering and choosing the set of smallest M-conductance is called a *sweep*.

Definition 2.3. Let M be an ergodic Markov chain with stationary distribution π, and let p be a probability distribution on the vertices. We define $p\,[x]$ to be the unique function from $[0, 1]$ to $[0, 1]$ such that

$$p\left[\pi(S_j^p)\right] = p(S_j^p) \quad \text{for each } j \in [0, n]$$

and such that $p\,[x]$ is piecewise linear between these points.

Proposition 2.4. *We have the following facts about the function* $p\,[x]$:

1. *The function* $p\,[x]$ *is concave.*
2. *For any set* S *of vertices,*
$$p(S) \leq p\left[\pi(S)\right].$$
3. *For any set of directed edges* A*, we have*
$$p(A) \leq p\left[\pi(A)\right].$$

The facts in this proposition are proved in [Lovász and Simonovits 90] and are not difficult to verify.

2.2. Global PageRank and Personalized PageRank

Definition 2.5. Given a Markov chain M, the PageRank vector $\mathrm{pr}_M(\alpha, s)$, defined by Brin and Page [Page et al. 98], is the unique solution of the linear system

$$\mathrm{pr}_M(\alpha, s) = \alpha s + (1-\alpha)\mathrm{pr}_M(\alpha, s)M. \tag{2.1}$$

Here, α is a constant in $(0, 1]$ called the *jump probability* and s is a probability distribution called the *starting vector*.

We will use the following basic facts about PageRank.

Proposition 2.6. *For any Markov chain M, starting vector s, and jump probability $\alpha \in (0, 1]$, there is a unique vector $\mathrm{pr}_M(\alpha, s)$ satisfying*

$$\mathrm{pr}_M(\alpha, s) = \alpha s + (1-\alpha)\mathrm{pr}_M(\alpha, s)M.$$

Proposition 2.7. *For any Markov chain M and any fixed value of α in $(0, 1]$, there is a linear transformation R_α such that $\mathrm{pr}_M(\alpha, s) = sR_\alpha$. Furthermore, R_α is given by the matrix*

$$R_\alpha = \alpha I + \alpha \sum_{t=1}^{\infty}(1-\alpha)^t M^t. \tag{2.2}$$

We omit the proofs of these facts, which may be found in [Jeh and Widom 03].

We let $\psi = \frac{1}{n}1_V$ be the uniform distribution. If a PageRank vector has ψ for its starting vector, we call it a *global PageRank vector*. If a PageRank vector has for its starting vector the indicator vector 1_v, with all probability on a single vertex v, we call it a *personalized PageRank vector* and use the shorthand notation $\mathrm{pr}_M(\alpha, v) = \mathrm{pr}_M(\alpha, 1_v)$.

There are abundant algorithms for computing global PageRank and personalized PageRank, so we will treat the computation of PageRank as a primitive operation. We assume that we have the following two black-box algorithms:

- `GlobalPR`(M, α) computes the global PageRank vector $\mathrm{pr}_M(\alpha, \psi)$.
- `LocalPR`(M, α, v) computes the personalized PageRank vector $\mathrm{pr}_M(\alpha, v)$.

We make the distinction between these two black boxes because personalized PageRank can be computed more efficiently than global PageRank. One may use for `LocalPR` any of the algorithms described in [Jeh and Widom 03, Berkhin

06, Sarlós et al. 06, Gleich and Polito 07], each of which can compute an approximation of the personalized PageRank vector $\mathrm{pr}_M(\alpha, v)$ by examining only a small fraction of the input graph near v, provided that M is a sparse matrix. The global PageRank can be computed efficiently in numerous ways, for example the Arnoldi method described in [Golub and Greif 06], but this requires performing a computation over the entire graph. We will endeavor to use `LocalPR` instead of `GlobalPR` as much as possible.

3. Local Partitioning for Ergodic Markov Chains

We now state the main theorem of this paper, which shows that a sweep over a personalized PageRank vector in an ergodic Markov chain M can produce a set with small M-conductance. This is a natural generalization of the theorem proved for undirected graphs in [Andersen et al. 06].

Theorem 3.1. *Let M be an ergodic Markov chain with stationary distribution π. Let S be a set of vertices such that $\pi(S) \leq \frac{1}{2}$ and $\Phi_M(S) \leq \alpha/16$, for some constant α. If v is a vertex sampled from S according to the probability distribution $\pi(v)/\pi(S)$, then with probability at least $\frac{1}{2}$, we have $\Phi_M(\mathrm{pr}_M(\alpha, v)) = O(\sqrt{\alpha \log |S|})$.*

The proof of the theorem is given at the end of this section. Here is the outline of how we will proceed. Given a personalized PageRank vector $p = \mathrm{pr}_M(\alpha, s)$ in an ergodic Markov chain M, we place an upper bound on $p[x]$ that depends on α and $\Phi(p)$, and place a lower bound on $p[\pi(S)]$ that depends on the conductance of a certain set S near the starting vertex. These upper and lower bounds will be combined to show that $\Phi(p)$ is small. We establish the upper and lower bounds in the following lemmas.

Lemma 3.2. *Let M be an ergodic Markov chain with stationary distribution π, let $p = \mathrm{pr}_M(\alpha, v)$ be a personalized PageRank vector in M, and let $\phi = \Phi_M(p)$ be the smallest M-conductance found by the sweep over p. Then*

$$p[x] \leq x + \alpha t + \left(1 - \frac{\phi^2}{72}\right)^t \sqrt{x/\pi(v)} \quad \text{for all } x \in [0,1] \text{ and all } t \geq 0.$$

Lemma 3.3. *Let M be an ergodic Markov chain with stationary distribution π, and let S be a set of vertices and v a vertex sampled from S according to the* *probability distribution* $\pi(v)/\pi(S)$. *With probability at least* $3/4$,

$$\text{pr}_M(\alpha, v)(S) \geq 1 - 4\frac{\Phi_M(S)}{\alpha}.$$

These two lemmas will be proved in Section 7. We use them now to derive the main theorem.

Proof of Theorem 3.1. Let $p = \text{pr}_M(\alpha, v)$ and let $\phi = \Phi(p)$. If v is sampled from S with probability $\pi(v)/\pi(S)$, then Lemma 3.3 implies that the following bound holds with probability at least $3/4$:

$$\text{pr}_M(\alpha, v)(S) \geq 1 - 4\frac{\Phi_M(S)}{\alpha} \geq 1 - 4\frac{\alpha/16}{\alpha} \geq \frac{3}{4}. \tag{3.1}$$

We will now show that with probability at least $3/4$,

$$\frac{\pi(v)}{\pi(S)} \geq \frac{1}{4|S|}. \tag{3.2}$$

To see this, consider the set of vertices S' in S such that $\pi(v) \geq \frac{\pi(S)}{4|S|}$. Clearly $\pi(S \setminus S') < \pi(S)/4$, which shows that $\pi(S') > \frac{3}{4}\pi(S)$.

The probability that the two events described in (3.1) and (3.2) both occur is at least $1/2$. We will assume for the rest of the proof that both events hold.

Lemma 3.2 gives us the following upper bound on $\text{pr}_M(\alpha, v)(S)$:

$$\begin{aligned}
\text{pr}_M(\alpha, v)(S) &\leq \text{pr}_M(\alpha, v)[\pi(S)] \\
&\leq \frac{4}{3}\pi(S) + \alpha T + \left(1 - \frac{\phi^2}{72}\right)^T \sqrt{\pi(S)/\pi(v)} \\
&\leq \frac{4}{3}\frac{1}{2} + \alpha T + \left(1 - \frac{\phi^2}{72}\right)^T \sqrt{4|S|}.
\end{aligned}$$

If we let $T = (72/\phi^2)\ln 24\sqrt{4|S|}$, then

$$\text{pr}_M(\alpha, v)(S) \leq \frac{2}{3} + \alpha T + \frac{1}{24}.$$

This contradicts our lower bound from (3.1) if $\alpha < 1/25T$, so we have shown that $\alpha \geq 1/25T$, which implies the following bound:

$$\phi \leq \sqrt{72 \cdot 25 \cdot \alpha \ln 24\sqrt{4|S|}} = O(\sqrt{\alpha \log |S|}).$$

□

Algorithm 1. (Applying Theorem 3.1 to the lazy random-walk Markov chain of a strongly connected graph.)

We are given as input a strongly connected directed graph with lazy random-walk matrix $\mathcal{W}$. The following procedure may be used to apply Theorem 3.1 with several different starting vertices and values of α. The offline preprocessing must be done once, after which the local computation may be performed as many times as desired.

Offline preprocessing:

1. Compute the stationary distribution π of $\mathcal{W}$.

Local computation:

1. Pick a starting vertex v and a value of α.
2. Compute $p = \text{pr}_{\mathcal{W}}(\alpha, v)$, using `LocalPR`.
3. Sort the vertices in nonincreasing order of $p(x)/\pi(x)$.
4. Let S_j be the set of the top j vertices in this ranking.
5. Compute the $\mathcal{W}$-conductance of each set S_j^p, and output the set with the smallest $\mathcal{W}$-conductance.

4. Partitioning a Strongly Connected Graph

In the next two sections we describe two possible approaches to partitioning a directed graph. In this section, we describe the straightforward method that applies only when the directed graph is strongly connected.

If the graph is strongly connected, then we may apply Theorem 3.1 to the lazy random-walk Markov chain $\mathcal{W}$, which is defined to be

$$\mathcal{W} = \mathcal{W}(A) = \frac{1}{2}(I + AD^{-1}).$$

Here, D is the diagonal matrix whose nonzero elements are the out-degrees of the vertices. The laziness of the walk ensures that $\mathcal{W}$ is ergodic whenever A is strongly connected, which allows us to apply our main theorem to $\mathcal{W}$.

To apply Theorem 3.1 to the lazy random-walk Markov chain $\mathcal{W}$, we must compute and perform a sweep over a personalized PageRank vector. When performing the sweep, we must know the stationary distribution of $\mathcal{W}$ to sort the vertices into the proper order. The stationary distribution needs to be computed only once, and afterward we can find numerous cuts by computing a single personalized PageRank vector per cut. The necessary computation is summarized in Algorithm 1.

5. Partitioning the PageRank Markov Chain

The majority of directed graphs that arise in practice are not strongly connected, so we cannot directly apply the results of the previous section to such a graph. In this section, we describe how Theorem 3.1 can be applied to the PageRank Markov chain of an arbitrary graph, which is always ergodic. We show that the notion of conductance associated with this Markov chain has a natural interpretation in terms of PageRank. We describe how to find a large number of sets with low conductance in the PageRank Markov chain by performing a small number (two) of global PageRank computations as a preprocessing step, followed by any desired number of local computations.

5.1. The PageRank Markov Chain

We now define the PageRank Markov chain $M_\beta = M_\beta(A)$ in terms of the adjacency matrix A of an arbitrary directed graph. To do so, we first modify the adjacency matrix by adding a self-loop to each vertex, to ensure that no vertex has out-degree zero. This ensures that the random-walk matrix $W = D^{-1}A$ is a Markov chain, where D is the diagonal matrix containing the modified out-degrees after the self-loops have been added.

Let $\psi = \frac{1}{n}1_V$ be the uniform distribution, and let β be a constant in $[0, 1]$, which we will call the *global jump probability.* Recall that the global PageRank vector $\mathrm{pr}_W(\beta, \psi)$ is the unique solution of the linear system

$$\mathrm{pr}_W(\beta, \psi) = \beta\psi + (1 - \beta)\mathrm{pr}_W(\beta, \psi)W. \tag{5.1}$$

The PageRank Markov chain M_β is defined to be

$$M_\beta = \beta K_\psi + (1 - \beta)W,$$

where $K_\psi = \vec{1}^T\psi$ is the dense rank-1 matrix obtained by taking the outer product of ψ with the all-ones vector. The global PageRank vector $\mathrm{pr}_W(\beta, \psi)$ is the stationary distribution of the PageRank Markov chain M_β. In other words, we have $\mathrm{pr}_W(\beta, \psi) = \mathrm{pr}_W(\beta, \psi)M_\beta$. The PageRank Markov chain M_β is ergodic for any value of $\beta \in (0, 1]$.

The notion of conductance associated with the PageRank Markov chain M_β has a natural interpretation in terms of the global PageRank vector $\mathrm{pr}_W(\beta, \psi)$. To describe this, we will use the shorthand notation $\mathrm{pr}_\beta = \mathrm{pr}_W(\beta, \psi)$ for the global PageRank, and $\Phi_\beta(S) = \Phi_{M_\beta}(S)$ for the M_β-conductance. Then, for any a set of vertices S, we have

$$\Phi_\beta(S) = \frac{\mathrm{pr}_\beta(\partial(S))}{\mathrm{pr}_\beta(S)}.$$

This quantity can be interpreted as follows: If we choose a vertex x from S with probability proportional to its PageRank, and then select a new vertex x' by performing a single step in the PageRank Markov chain M_β, then $\Phi_\beta(S)$ is the probability that $x' \notin S$.

5.2. Computing Personalized PageRank in the PageRank Markov Chain

To apply our local partitioning theorem to M_β, we must compute a personalized PageRank vector in the Markov chain M_β. The personalized PageRank vector $\mathrm{pr}_{M_\beta}(\alpha, s)$ is the unique solution of the linear system

$$\mathrm{pr}_{M_\beta}(\alpha, s) = \alpha s + (1-\alpha)\mathrm{pr}_{M_\beta}(\alpha, s) M_\beta.$$

Although this is a personalized PageRank vector, the Markov chain M_β is dense because of its global random jump, so it is not possible to compute $\mathrm{pr}_{M_\beta}(\alpha, s)$ efficiently using $\mathtt{LocalPR}(M_\beta, \alpha, s)$. We will show that $\mathrm{pr}_{M_\beta}(\alpha, s)$ can be computed efficiently in another way, by taking a linear combination of a personalized PageRank vector and a global PageRank vector in the random-walk Markov chain W.

We now present two interpretations of the PageRank vector $\mathrm{pr}_{M_\beta}(\alpha, s)$. By definition, $\mathrm{pr}_{M_\beta}(\alpha, s)$ is a personalized PageRank vector in the Markov chain M_β. It can also be viewed as a PageRank vector in the random-walk Markov chain W. When viewed as a PageRank vector in W, its starting vector is a linear combination of the uniform distribution ψ and the starting vector s, and its jump probability is $\gamma = \alpha + \beta - \alpha\beta$:

$$\begin{aligned}
\mathrm{pr}_{M_\beta}(\alpha, s) &= \alpha s + (1-\alpha)\mathrm{pr}_\beta(\alpha, s) M_\beta \\
&= \alpha s + (1-\alpha)\beta\psi + (1-\alpha)(1-\beta)\mathrm{pr}_\beta(\alpha, s) W \\
&= \gamma\left(\frac{\alpha}{\gamma}s + \frac{(1-\alpha)\beta}{\gamma}\psi\right) + (1-\gamma)\mathrm{pr}_\beta(\alpha, s) W \\
&= \mathrm{pr}_W(\gamma, s').
\end{aligned}$$

Here $\gamma = \alpha + \beta - \alpha\beta$, and $s' = \frac{\alpha}{\gamma}s + \frac{(1-\alpha)\beta}{\gamma}\psi$. Using the fact that a PageRank vector is a linear function of its starting vector, we can write

$$\begin{aligned}
\mathrm{pr}_{M_\beta}(\alpha, s) &= \mathrm{pr}_W\left(\gamma, \frac{\alpha}{\gamma}s + \frac{(1-\alpha)\beta}{\psi}\right) \\
&= \frac{\alpha}{\gamma}\mathrm{pr}_W(\gamma, s) + \frac{(1-\alpha)\beta}{\gamma}\mathrm{pr}_W(\gamma, \psi).
\end{aligned}$$

In summary, we have taken a personalized PageRank vector $\mathrm{pr}_{M_\beta}(\alpha, s)$ from the PageRank Markov chain M_β, and written it as a linear combination of two

Algorithm 2. (Applying Corollary 5.1 to the PageRank Markov chain.)

We are given as input the adjacency matrix A of a directed graph (not necessarily strongly connected), the global jump probability β, and the local jump probability α. The following procedure may be used to apply Theorem 3.1 at several different starting vertices with these fixed values of α and β. The offline preprocessing must be done once, after which the local computation may be performed as many times as desired.

Offline preprocessing:

We must compute two global PageRank vectors:

1. Let $\gamma = \alpha + \beta - \alpha\beta$.
2. Let $W = W(A)$ be the random-walk matrix of A.
3. Compute the two global PageRank vectors $\mathrm{pr}_\beta = \mathrm{pr}_W(\beta, \psi)$ and $\mathrm{pr}_\gamma = \mathrm{pr}_W(\gamma, \psi)$ using the algorithm `GlobalPR`.

Local computation:

1. Pick a starting vertex v.
2. Compute $\mathrm{pr}_W(\gamma, v)$, using `LocalPR`.
3. Obtain $p = \mathrm{pr}_{M_\beta}(\alpha, v)$ by taking a linear combination of $\mathrm{pr}_W(\gamma, v)$ and $\mathrm{pr}_W(\gamma, \psi)$,
$$p = \mathrm{pr}_{M_\beta}(\alpha, v) = \frac{\alpha}{\gamma}\mathrm{pr}_W(\gamma, v) + \frac{(1-\alpha)\beta}{\gamma}\mathrm{pr}_W(\gamma, \psi).$$
4. Rank the vertices in nonincreasing order of $p(x)/\mathrm{pr}_\beta(x)$.
5. Let S_j be the set of the top j vertices in this ranking.
6. Compute the β-conductances $\Phi_\beta(S_j)$ for each set S_j, and output the set with the smallest β-conductance.

PageRank vectors from the random-walk Markov chain W. One of these is a personalized PageRank vector in W with starting vector s, and the other is a global PageRank vector in W with starting distribution ψ.

5.3. Local Partitioning in the PageRank Markov Chain

By applying our main theorem to the PageRank Markov chain, we obtain the following corollary, which shows that a sweep over the PageRank vector $\mathrm{pr}_{M_\beta}(\alpha, v)$ produces a set with small M_β-conductance.

Corollary 5.1. *Let S be a set of vertices such that $\mathrm{pr}_\beta(S) \le \frac{1}{2}$ and $\Phi_\beta(S) \le \alpha/16$, for some constants α and β. If a vertex v is sampled from S according to the* *probability distribution* $\mathrm{pr}_\beta(v)/\mathrm{pr}_\beta(S)$, *then with probability at least* $1/2$ *we have* $\Phi_\beta(\mathrm{pr}_{M_\beta}(\alpha, v)) = O(\sqrt{\alpha \log |S|})$.

Proof. The corollary is immediate, by applying Theorem 3.1 to the ergodic Markov chain M_β. □

To carry out the computation required by the corollary, we need to compute the stationary distribution of M_β, which is just the global PageRank vector $\mathrm{pr}_W(\beta, \psi)$. For each cut we want to find, we also need to compute a personalized PageRank vector $\mathrm{pr}_{M_\beta}(\alpha, v)$ in the Markov chain M_β. This can be done by computing $\mathrm{pr}_W(\gamma, v)$ and $\mathrm{pr}_W(\gamma, \psi)$, and then taking a linear combination of these two PageRank vectors, as described in the previous section. If we fix the values of α and β, we can compute the two global PageRank vectors $\mathrm{pr}_W(\beta, \psi)$ and $\mathrm{pr}_W(\gamma, \psi)$ ahead of time, and then compute a large number of personalized PageRank vectors $\mathrm{pr}_W(\gamma, v)$ using `LocalPR`. This procedure is summarized in Algorithm 2.

6. Concluding Remarks

6.1. When Is Partitioning the PageRank Markov Effective?

Corollary 5.1 can be applied to partition the PageRank Markov chain of an arbitrary directed graph, and to an arbitrary starting vertex. Because it may be applied to any graph (even the empty graph), the approximation guarantee that it provides may become vacuous for some graphs and starting vertices. In this section we will describe this concern in more detail, and give a positive result that describes when the approximation guarantee provided is strong rather than vacuous. We caution that this section contains high-level discussion rather than rigorous proofs.

As we increase β, we increase the probability of the global jump, which ensures that the β-conductance of every set in the graph is at least roughly β. If we partition the PageRank Markov chain of a graph with no edges, every subset of vertices will have conductance roughly β, so the approximation guarantee of Corollary 5.1 will be vacuous (which is what we should expect when partitioning a graph with no edges). On the other hand, if we partition the PageRank Markov chain of an undirected graph, using a very small value of β, the best partitions of the graph will have β-conductance larger than β, so the approximation guarantee of Corollary 5.1 will give a meaningful result.

Loosely speaking, we claim that partitioning the PageRank Markov chain M_β gives interesting results exactly when there are interesting partitions of the graph that have β-conductance larger than β. To provide evidence for this claim, we separate the β-conductance $\Phi_\beta(S)$ into two parts: the contribution $\Psi_\beta(S)$ from real graph edges in W and the contribution from the random jump. We define

$$\Psi_\beta(S) = \frac{\sum_{(u,v)\in S\times\bar{S}} \mathrm{pr}_\beta(u)W(u,v)}{\mathrm{pr}_\beta(S)}.$$

Then $\Phi_\beta(S)$ and $\Psi_\beta(S)$ are related by the following equation:

$$\Phi_\beta(S) = (1-\beta)\Psi_\beta(S) + \beta\frac{|\bar{S}|}{n}.$$

It is not hard to see that if a set S has β-conductance significantly larger than β, our algorithm finds a set S' for which $\Psi_\beta(S')$ is nearly as small as $\Psi_\beta(S)$. In particular, if S is a set of vertices for which $\Psi_\beta(S) = \Omega(\Phi_\beta(S))$, and S' is a set of vertices for which $\Phi_\beta(S') = O(\sqrt{\Phi_\beta(S)\log n})$, which is the conductance guaranteed by Corollary 5.1, then we have

$$\Psi_\beta(S') = O(\sqrt{\Psi_\beta(S)\log n}).$$

6.2. Cuts from Approximate PageRank Vectors

For the case of undirected graphs, it has been proved that a cut with small conductance can be found efficiently by sweeping over an *approximate* personalized PageRank vector. This was proved in [Andersen et al. 06], and requires a careful error analysis. We remark that a similar error analysis may be carried out for the directed case, although we have not described such an analysis in this paper.

7. Appendix: Proof of the Mixing Bounds for Personalized PageRank

In this section, we prove upper and lower bounds on the curve $p[x]$ of a PageRank vector $p = \mathrm{pr}_M(\alpha, s)$.

Lemma 7.1. *Let M be an ergodic Markov chain with stationary distribution π, let $p = \mathrm{pr}_M(\alpha, v)$ be a personalized PageRank vector in M, and let $S_j = S_j^p$. For each $j \in [1, n-1]$, we have*

$$\begin{aligned} p[\pi(S_j)] \leq\ & \frac{\alpha}{2-\alpha} s\left[\pi(S_j)\right] \\ & + \left(1 - \frac{\alpha}{2-\alpha}\right)\left(p\left[\pi(S_j) + \pi(\partial(S_j))\right] + p\left[\pi(S_j) - \pi(\partial(S_j))\right]\right). \end{aligned}$$

Proof. For any set S of vertices, we define the set of directed edges whose heads are in S,

$$\text{in}(S) = \{(u, v) \in E \mid v \in S\},$$

and the set of edges whose tails are in S,

$$\text{out}(S) = \{(u, v) \in E \mid u \in S\}.$$

The following describes the amount of probability from pM on a set S, in terms of the amount of probability moving across the edges in the sets $\text{in}(S)$ and $\text{out}(S)$:

$$\begin{aligned} pM(S) &= p(\text{in}(S)) \\ &= p\left(\text{in}(S) \cap \text{out}(S)\right) + p\left(\text{in}(S) \setminus \text{out}(S)\right) \\ &= p\left(\text{in}(S) \cap \text{out}(S)\right) + p\left(\text{in}(S) \cup \text{out}(S)\right) - p(S). \end{aligned}$$

We will now calculate the total measure of the edges in the sets $(\text{in}(S) \cup \text{out}(S))$ and $(\text{in}(S) \cap \text{out}(S))$. The following holds because π is the stationary distribution of M:

$$\pi(\text{in}(S)) = \pi(\text{out}(S)) = \pi(S).$$

It is not hard to observe the following two equalities:

$$\begin{aligned} \pi(\text{in}(S) \cup \text{out}(S)) + \pi(\text{in}(S) \cap \text{out}(S)) &= 2\pi(S), \\ \pi(\text{in}(S) \cup \text{out}(S)) - \pi(\text{in}(S) \cap \text{out}(S)) &= 2\pi(\partial(S)). \end{aligned}$$

Solving the system of equations above yields the following:

$$\begin{aligned} \pi\left(\text{in}(S) \cup \text{out}(S)\right) &= \pi(S) + \pi(\partial(S)), \\ \pi\left(\text{in}(S) \cap \text{out}(S)\right) &= \pi(S) - \pi(\partial(S)). \end{aligned}$$

Now let $p = \text{pr}_M(\alpha, s)$ be a personalized PageRank vector. For any set S of vertices, we have

$$\begin{aligned} p(S) &\leq \alpha s(S) + (1 - \alpha)pM(S) \\ &\leq \alpha s(S) + (1 - \alpha)\left(p\left(\text{in}(S) \cap \text{out}(S)\right) + p\left(\text{in}(S) \cup \text{out}(S)\right) - p(S)\right). \end{aligned}$$

By adding the term $(1 - \alpha)p(S)$ to both sides and then dividing by $2 - \alpha$, we obtain

$$p(S) \leq \frac{\alpha}{2 - \alpha} s(S) + \left(1 - \frac{\alpha}{2 - \alpha}\right)\left(\frac{1}{2} p\left(\text{in}(S) \cap \text{out}(S)\right) + \frac{1}{2} p\left(\text{in}(S) \cup \text{out}(S)\right)\right).$$

Now let $S_j = S_j^p$, and recall from Proposition 2.4 that $p\,[\pi(S_j)] = p(S_j)$ for any integer $j \in [0,n]$, and that for any set of directed edges A, we have the bound $p(A) \le p\,[\pi(A)]$. Therefore, we have

$$\begin{aligned}
p\,[\pi(S_j)] &= p(S_j)\\
&= \frac{\alpha}{2-\alpha} s(S_j)\\
&\quad + \left(1 - \frac{\alpha}{2-\alpha}\right)\left(\frac{1}{2}p\,(\mathrm{in}(S_j) \cap \mathrm{out}(S_j)) + \frac{1}{2}p\,(\mathrm{in}(S_j) \cup \mathrm{out}(S_j))\right)\\
&\le \frac{\alpha}{2-\alpha} s\,[\pi(S_j)]\\
&\quad + \left(1 - \frac{\alpha}{2-\alpha}\right)\left(\frac{1}{2}p\,[\pi\,(\mathrm{in}(S_j) \cap \mathrm{out}(S_j))] + \frac{1}{2}p\,[\pi\,(\mathrm{in}(S_j) \cup \mathrm{out}(S_j))]\right)\\
&\le \frac{\alpha}{2-\alpha} s\,[\pi(S_j)]\\
&\quad + \left(1 - \frac{\alpha}{2-\alpha}\right)\left(\frac{1}{2}p\,[\pi(S_j) - \pi(\partial(S_j))] + \frac{1}{2}p\,[\pi(S_j) + \pi(\partial(S_j))]\right).
\end{aligned}$$

□

Lemma 7.2. *For any ergodic Markov chain M, any starting vector s, any value $\alpha \in (0,1]$, and any $x \in [0,1]$, we have*

$$\mathrm{pr}_M\,(\alpha, s)\,[x] \le s\,[x]\,.$$

Proof. If we let $p = \mathrm{pr}_M(\alpha, s)$, then Lemma 7.1 implies that for each $j \in [1, n-1]$,

$$\begin{aligned}
p\left[\pi(S_j^p)\right] &\le \frac{\alpha}{2-\alpha} s\left[\pi(S_j^p)\right]\\
&\quad + \left(1 - \frac{\alpha}{2-\alpha}\right)\left(\frac{1}{2}p\left[\pi(S_j^p) - \pi(\partial(S_j^p))\right] + \frac{1}{2}p\left[\pi(S_j^p) + \pi(\partial(S_j^p))\right]\right)\\
&\le \frac{\alpha}{2-\alpha} s\left[\pi(S_j^p)\right] + \left(1 - \frac{\alpha}{2-\alpha}\right) p\left[\pi(S_j^p)\right].
\end{aligned}$$

The last line follows from the concavity of $p\,[k]$. This implies that $p\left[\pi(S_j^p)\right] \le s\left[\pi(S_j^p)\right]$ for each $j \in [1, n-1]$. The same equation then holds for all $x \in [0,1]$, because $s\,[x]$ is concave and $p\,[x]$ is linear between the points $\pi(S_j^p)$ and $\pi(S_{j+1}^p)$. □

We now prove the two lemmas we used in Section 3.

Proof of Lemma 3.2. We define the function

$$f_t(x) = \alpha t + \left(1 - \frac{\phi^2}{72}\right)^t \sqrt{x/\pi(v)}.$$

We will prove by induction that the following inequality holds for all $t \geq 0$:

$$p\,[x] \leq \frac{4}{3}x + f_t(x) \quad \text{for all } x \in [0,1]. \tag{7.1}$$

To prove the base case, notice that for any value of $x \in [0,1]$, we have $p[x] = \mathrm{pr}_M(\alpha, v)[x] \leq 1_v[x]$, by Lemma 7.2. This implies

$$p\,[x] \leq 1_v[x] \leq \min(1, x/\pi(v)) \leq x + \sqrt{x/\pi(v)},$$

which implies that (7.1) holds for $t = 0$.

We now assume that (7.1) holds for t, and prove that it holds for $t+1$. For each $j \in [0,n]$, let $x_j = \pi(S_j^p)$. It suffices to show that (7.1) holds for $t+1$ at the points $x_0, \ldots, x_n$, because $p\,[x]$ is piecewise linear between these points, and $f_{t+1}(x)$ is concave.

The inequality (7.1) for time $t+1$ holds trivially when $x = 0$, and also when $x \geq 3/4$, so it suffices to consider an arbitrary index j such that $j > 1$ and $\pi(S_j) \leq 3/4$. Because $\pi(S_j) \leq 3/4$, we have $\bar{\pi}(S_j) \geq (1/3)\pi$, and therefore

$$\pi(\partial(S_j)) = \Phi(S_j)\bar{\pi}(S_j) \geq \frac{1}{3}\Phi(S_j)\pi(S_j) \geq \frac{1}{3}\phi x_j.$$

We now apply Lemma 7.1:

$$\begin{aligned} p\,[\pi(S_j)] &\leq \frac{\alpha}{2-\alpha} s\,[\pi(S_j)] \\ &\quad + \left(1 - \frac{\alpha}{2-\alpha}\right)\left(p\,[\pi(S_j) - \pi(\partial(S_j))] + p\,[\pi(S_j) + \pi(\partial(S_j))]\right) \\ &\leq \frac{\alpha}{2-\alpha} + \left(\frac{1}{2}p\,[\pi(S_j) - \pi(\partial(S_j))] + \frac{1}{2}p\,[\pi(S_j) + \pi(\partial(S_j))]\right) \\ &\leq \alpha + \left(\frac{1}{2}p\,[x_j - (1/3)\phi x_j] + \frac{1}{2}p\,[x_j + (1/3)\phi x_j]\right). \end{aligned}$$

The last step above follows from the concavity of $p\,[x]$ and the fact that $\pi(\partial(S_j)) \geq \frac{1}{3}\phi x_j$. We now use the induction assumption that (7.1) holds,

$$
\begin{aligned}
p\left[x_j\right] &\le \alpha + \frac{1}{2}\left(\frac{4}{3}(x_j - \frac{1}{3}\phi x_j) + f_t(x_j - \frac{1}{3}\phi x_j)\right) \\
&\quad + \frac{1}{2}\left(\frac{4}{3}(x_j + \frac{1}{3}\phi x_j) + f_t(x_j + \frac{1}{3}\phi x_j)\right) \\
&= \frac{4}{3}x_j + \alpha + \frac{1}{2}\left(f_t(x_j - \frac{1}{3}\phi x_j) + f_t(x_j + \frac{1}{3}\phi x_j)\right) \\
&\le \frac{4}{3}x_j + \alpha + \frac{1}{2}\left(f_t(x_j - \frac{1}{3}\phi x_j) + f_t(x_j + \frac{1}{3}\phi x_j)\right) \\
&= \frac{4}{3}x_j + \alpha(t+1) \\
&\quad + \frac{1}{2}\left(\sqrt{x_j - \frac{1}{3}\phi x_j} + \sqrt{x_j + \frac{1}{3}\phi x_j}\right)\frac{1}{\sqrt{\pi(v)}}\left(1 - \frac{\phi^2}{72}\right)^t .
\end{aligned}
$$

We now use the fact that for any $x \ge 0$ and $z \in [0,1]$,

$$
\frac{1}{2}\left(\sqrt{x - zx} + \sqrt{x + zx}\right) \le \sqrt{x}\left(1 - \frac{z^2}{8}\right).
$$

Applying this bound with $x = x_j$ and $z = \frac{1}{3}\phi$, we obtain the following:

$$
\begin{aligned}
p\left[x_j\right] &\le \frac{4}{3}x + \alpha(t+1) + \sqrt{x_j/\pi(v)}\left(1 - \frac{\phi^2}{72}\right)\left(1 - \frac{\phi^2}{72}\right)^t \\
&= \frac{4}{3}x + f_{t+1}(x_j).
\end{aligned}
$$

This completes the proof. □

Proof of Lemma 3.3. Let π_S be the probability distribution described in the statement of the lemma, the one obtained by sampling a vertex v from the distribution π, conditioned on the event that $v \in S$.

The amount of probability that moves from S to $\bar{S}$ in the step from $\mathrm{pr}_M(\alpha, \pi_S)$ to $\mathrm{pr}_M(\alpha, \pi_S)M$ is equal to $[\mathrm{pr}_M(\alpha, \pi_S)](\partial(S))$, so we have

$$
\begin{aligned}
\left[\mathrm{pr}_M(\alpha, \pi_S)M\right]\left(\bar{S}\right) &\le \left[\mathrm{pr}_M(\alpha, \pi_S)\right]\left(\bar{S}\right) + \left[\mathrm{pr}_M(\alpha, \pi_S)\right](\partial(S)) \\
&\le \left[\mathrm{pr}_M(\alpha, \pi_S)\right]\left(\bar{S}\right) + \mathrm{pr}_M(\alpha, \pi_S)\left[\pi(\partial(S))\right].
\end{aligned}
$$

By Lemma 7.2,

$$
\mathrm{pr}_M(\alpha, \pi_S)\left[\pi(\partial(S))\right] \le \pi_S\left[\pi(\partial(S))\right] = \frac{\pi(\partial(S))}{\pi(S)} = \Phi_M(S).
$$

We combine this observation with the personalized PageRank equation:

$$\begin{aligned}[\mathrm{pr}_M(\alpha,\pi_S)]\left(\bar{S}\right) &= [\alpha\pi_S + (1-\alpha)\mathrm{pr}_M(\alpha,\pi_S)M]\left(\bar{S}\right)\\ &= (1-\alpha)[\mathrm{pr}_M(\alpha,\pi_S)M]\left(\bar{S}\right)\\ &\le (1-\alpha)[\mathrm{pr}_M(\alpha,\pi_S)]\left(\bar{S}\right) + \Phi_M(S).\end{aligned}$$

This implies the following:

$$[\mathrm{pr}_M(\alpha,\pi_S)]\left(\bar{S}\right) \le \frac{\Phi_M(S)}{\alpha}.$$

If we sample a vertex from the distribution π_S, then at least three-fourths of the time, $\mathrm{pr}_M(\alpha, 1_v)(\bar{S})$ is at most four times its expected value of $\mathrm{pr}_M(\alpha,\pi_S)(\bar{S})$, and the result follows. □

Acknowledgments. We thank Zoltán Gyöngyi for his expert advice during the formative stages of this research.

References

[Andersen et al. 06] Reid Andersen, Fan Chung, and Kevin Lang. "Local Graph Partitioning Using PageRank Vectors." In *FOCS '06: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pp. 475–486. Washington, DC: IEEE Computer Society, 2006.

[Berkhin 06] Pavel Berkhin. "Bookmark-Coloring Algorithm for Personalized PageRank Computing." *Internet Math.* 3:1 (2006), 41–62.

[Charikar et al. 06] Moses Charikar, Konstantin Makarychev, and Yury Makarychev. "Directed Metrics and Directed Graph Partitioning Problems." In *SODA '06: Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 51–60. New York: ACM Press, 2006.

[Chung 97] F. Chung. *Spectral Graph Theory*, CBMS Regional Conference Series in Mathematics 92. Providence, RI: American Mathematical Society, 1997.

[Chung 05] Fan Chung. "Laplacians and Cheeger Inequalities for Directed Graphs." *Annals of Combinatorics* 9 (2005), 1–19.

[Chuzhoy and Khanna 06] Julia Chuzhoy and Sanjeev Khanna. "Hardness of Cut Problems in Directed Graphs." In *STOC '06: Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing*, pp. 527–536. New York, ACM Press, 2006.

[Fill 91] James Allen Fill. "Eigenvalue Bounds on Convergence to Stationarity for Nonreversible Markov Chains, with an Application to the Exclusion Process." *Ann. Appl. Probab.* 1:1 (1991), 62–87.

[Gleich and Polito 07] David Gleich and Marzia Polito. "Approximating Personalized PageRank with Minimal Use of Web Graph Data." *Internet Mathematics* 3:3 (2006), 257–294.

[Golub and Greif 06] G. Golub and C. Greif. "Arnoldi-Type Algorithms for Computing Stationary Distribution Vectors, with Application to PageRank." *10543 BIT Numerical Mathematics* 46:4 (2006), 759–771.

[Jeh and Widom 03] Glen Jeh and Jennifer Widom. "Scaling Personalized Web Search." In *Proceedings of the 12th International Conference on World Wide Web*, pp. 271–279. New York: ACM Press, 2003.

[Lovász and Simonovits 90] László Lovász and Miklós Simonovits. "The Mixing Rate of Markov Chains, an Isoperimetric Inequality, and Computing the Volume." In *Proceedings of the 31st Annual Symposium on Foundations of Computer Science*, pp. 346–354. Los Alamitos, CA: IEEE Press, 1990.

[Mihail 89] M. Mihail. "Conductance and Convergence of Markov Chains: A Combinatorial Treatment of Expanders." In *Proceedings of the 30th Annual Symposium on Foundations of Computer Science*, pp. 526–531. Los Alamitos, CA: IEEE Press, 1989.

[Page et al. 98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. "The PageRank Citation Ranking: Bringing Order to the Web." Technical report, Stanford Digital Library Technologies Project, 1998.

[Sarlós et al. 06] Tamás Sarlós, András A. Benczúr, Károly Csalogány, Dániel Fogaras, and Balázs Rácz. "To Randomize or Not to Randomize: Space Optimal Summaries for Hyperlink Analysis." In *Proceedings of the 12th International Conference on World Wide Web*, pp. 297–306. New York: ACM Press, 2006.

[Spielman and Teng 96] Daniel A. Spielman and Shang-Hua Teng. "Spectral Partitioning Works: Planar Graphs and Finite Element Meshes." In *Proceedings of the 37th Annual Symposium on Foundations of Computer Science*, pp. 96–105. Los Alamitos, CA: IEEE Press, 1996.

[Stewart 97] W. J. Stewart. *Introduction to the Numerical Solution of Markov Chains.* Princeton: Princeton University Press, 1994.

Reid Andersen, Microsoft Research, One Microsoft Way, Redmond, WA 98052 (reidan@math.ucsd.edu)

Fan Chung, University of California, San Diego, Dept. of Mathematics, La Jolla, CA 92093-0112 (fan@ucsd.edu)

Kevin Lang, Yahoo! Research, 2821 Mission College Blvd., Santa Clara, CA 95054 (langk@yahoo-inc.com)

Received February 14, 2008; accepted May 6, 2008.