Internet Mathematics Vol. 5, No. 1–2: 23–44

Local Computation of PageRank Contributions

Reid Andersen, Christian Borgs, Jennifer Chayes, John Hopcroft, Vahab Mirrokni, and Shang-Hua Teng

Abstract. Motivated by the problem of detecting link-spam, we consider the following graph-theoretic primitive: Given a webgraph G, a vertex v in G, and a parameter $\delta \in (0, 1)$, compute the set of all vertices that contribute to v at least a δ-fraction of v's PageRank. We call this set the δ-contributing set of v. To this end, we define the contribution vector of v to be the vector whose entries measure the contributions of every vertex to the PageRank of v. A local algorithm is one that produces a solution by adaptively examining only a small portion of the input graph near a specified vertex. We give an efficient local algorithm that computes an ϵ-approximation of the contribution vector for a given vertex by adaptively examining $O(1/\epsilon)$ vertices. Using this algorithm, we give a local approximation algorithm for the primitive defined above. Specifically, we give an algorithm that returns a set containing the δ-contributing set of v and at most $O(1/\delta)$ vertices from the $\delta/2$-contributing set of v, and that does so by examining at most $O(1/\delta)$ vertices. We also give a local algorithm for solving the following problem: If there exist k vertices that contribute a ρ-fraction to the PageRank of v, find a set of k vertices that contribute at least a $(\rho - \epsilon)$-fraction to the PageRank of v. In this case, we prove that our algorithm examines at most $O(k/\epsilon)$ vertices.

1. Introduction

In numerous applications of PageRank one needs to know, in addition to the rank of a given web page, which pages or sets of pages contribute most to its rank. These PageRank contributions have been used for link-spam detection [Benczúr et al. 05, Gyöngyi et al. 06a] and in the classification of web pages [Gyöngyi et al. 06b]. A set of pages that contributes significantly to the PageRank of a page is often called a *contribution set* or *supporting set* of the page [Benczúr et al. 05, Gyöngyi et al. 06a].

1542-7951/08 $0.50 per page

The contribution that a vertex u makes to the PageRank of a vertex v is defined rigorously in terms of personalized PageRank. For a webgraph $G = (V, E)$ and a *teleportation constant* α (sometimes called the restart probability), let P_α be the matrix whose uth row is the personalized PageRank vector of u. The PageRank contribution of u to v, written $\mathrm{pr}_\alpha(u \to v)$, is defined to be the entry (u, v) of this matrix. The PageRank of a vertex v is the sum of the vth column of the matrix P_α, and thus the PageRank of a vertex can be viewed as the sum of the contributions from all other vertices. The *contribution vector* of v is defined to be the vth column of the matrix P_α, whose entries are the contributions of every vertex to the PageRank of v.

Given that the web graph is massive and getting larger at a substantial rate, it is essential to compute contribution vectors and identify supporting sets by examining as small a fraction of the graph as possible. In particular, it is helpful to design a *local* algorithm for computing the supporting sets of a particular vertex. Local algorithms search for a solution near a specified vertex by adaptively examining only a small subset of the input graph. They have been studied previously in distributed computing [Naor and Stockmeyer 95] and in graph partitioning and clustering [Spielman and Teng 04, Andersen et al. 06]. Personalized PageRank vectors can be approximated locally. Using one of several possible algorithms [Jeh and Widom 03, Berkhin 06, Sarlós et al. 06], it is possible to compute an approximation of the personalized PageRank vector of a vertex u by examining only $O(1/\epsilon)$ vertices, where ϵ is the desired amount of error at each vertex.

1.1. Problem Formulation

Inspired by local algorithms for computing personalized PageRank, and motivated by the importance of supporting sets in link-spam detection, we consider the problem of directly computing the contribution vector of a given vertex to quickly identify its supporting sets. In particular, we consider the following graph-theoretic primitive: Given a webgraph G, a vertex v in G, and a parameter $\delta \in (0, 1)$, compute the set of all vertices each contributing at least a δ-fraction to the PageRank of v. We call this set the *δ-contributing set of v*.

Such a primitive is useful for spam detection, since given a webpage whose PageRank has recently increased suspiciously, we can quickly identify the set of pages that contribute significantly to the PageRank of that suspicious page. The above primitive may also be useful for analyzing social networks. In social networks in which the links capture the influence of vertices on each other, we can identify the nodes with the most influence to a given node.

1.2. Our Results

We give an efficient local algorithm for computing an ϵ-approximation of the contribution vector for a given vertex v, a vector whose difference from the

contribution vector is at most ϵ at each vertex. We prove that the number of vertices examined by the algorithm is $O(1/\epsilon)$. The algorithm performs a sequence of probability-pushing operations on vertices of the graph, which we call pushback operations. When the pushback operation is applied to a vertex u, we perform a small amount of computation for each in-neighbor of u. Particularly, we add a fraction of a number stored at u to a number stored at each in-neighbor of u. The number of such operations that our algorithm performs is $O(1/\epsilon)$, and its running time can be bounded by the sum of the in-degrees of the vertices from which these operations were performed. To derive this algorithm, we adapt Jeh and Widom's technique for computing personalized PageRank vectors [Jeh and Widom 03] to directly compute contribution vectors. To analyze the algorithm's running time and error bounds, we use techniques developed for the local clustering algorithm in [Andersen et al. 06].

Using our algorithm for approximating contribution vectors, we give an approximation algorithm to the primitive defined above. Explicitly, we give a local algorithm that returns a set containing the δ-contributing set of v and at most $O(1/\delta)$ vertices from the $\delta/2$-contributing set of v. Our algorithm applies at most $O(1/\delta)$ pushback operations. We also give a local algorithm for solving the following problem: If there are k vertices that contribute a ρ-fraction to the PageRank of v, find a set of k vertices that contribute at least a $(\rho - \epsilon)$-fraction to the PageRank of v. In this case, we prove that our algorithm needs at most $O(k/\epsilon)$ pushback operations.

After presenting our local algorithm and its applications to computing contributing sets, we introduce the notion of PageRank traffic, which provides a different way to measure the influence of a node on the PageRank of another node. We prove a result that shows that PageRank traffic, while seemingly quite different from our definition of PageRank contributions, is closely related to a certain weighted version of PageRank contributions.

Finally, we remark that in principle, one could directly compute the contribution vector for a vertex v by approximating the personalized PageRank vector of v in the time-reversal of the random-walk Markov chain. We describe the computation required for this approach, and argue that for most graphs it is not as efficient as the method we propose.

1.3. Related Work

Supporting sets and PageRank contributions have been studied before as a tool for spam detection, notably in the SpamRank algorithm of Benczúr et al. [Benczúr et al. 05], and in the Spam Mass algorithm of Gyöngyi et al. [Gyöngyi et al. 06a]. However, none of these papers developed a local algorithm for computing the contribution vector or supporting set. In the SpamRank algorithm, the contribution vectors are computed in the following way. One computes an approximation of each personalized PageRank vector in the graph to create an

approximate PageRank matrix, and then takes the transpose of this matrix to obtain the approximate contribution vectors. This method is efficient for the task of computing the contribution vectors for every vertex in the graph, and it leverages fast algorithms for computing many personalized PageRank vectors simultaneously [Fogaras and Racz 04, Sarlós et al. 06], but it does not provide an efficient way to compute the contribution vectors of a few selected suspicious vertices. Furthermore, the relative error in the resulting approximate contribution vectors may be larger than the relative error in the computed personalized PageRank vectors, since this is not preserved by the transpose operation.

PageRank contributions have also been used to estimate the PageRank of a target vertex. The algorithm in [Chen et al. 04] heuristically identifies the top contributors to a vertex v by adaptively choosing vertices with high likelihood of being large contributors, and then locally computes personalized PageRank from those vertices. This is different from our approach of directly computing the contribution vector, and more difficult to analyze rigorously.

Local algorithms have been studied in distributed computing [Naor and Stockmeyer 95] and in graph partitioning and clustering [Spielman and Teng 04, Andersen et al. 06]. Personalized PageRank vectors can be computed locally using a number of methods [Berkhin 06, Andersen et al. 06, Sarlós et al. 06], many of which are based on the algorithm of Jeh and Widom [Jeh and Widom 03]. None of these algorithms can be used directly to compute a contribution vector or supporting set.

There are numerous methods for detecting link spam besides the SpamRank-type algorithms we have mentioned here. Examples include applying machine learning to link-based features [Becchetti et al. 06], the analysis of page content [Mishne et al. 05, Ntoulas et al. 06], TrustRank [Gyöngyi et al. 04] and Anti-TrustRank [Krishnan and Raj 06], and statistical analysis of various page features [Fetterly et al. 04].

1.4. Organization

This paper will be organized as follows. In Section 2 we review the basic concepts used in this paper, including PageRank, personalized PageRank, and PageRank contribution vectors. In Section 3 we derive an alternative formula for the PageRank contribution vector. Using this formula, we present an efficient local algorithm for computing PageRank contribution and analyze its performance. In Section 4 we consider several notions of supporting sets, which are sets of vertices that contribute significantly to the PageRank of a target vertex, and show how to efficiently compute approximate supporting sets. In Section 5 we introduce PageRank traffic, and relate this new concept to a weighted variation of PageRank contributions. In Section 6 we make a few concluding remarks. We also show that in principle, the time-reverse Markov chain can be used to compute the contribution vector, but argue that our method is more efficient.

2. Preliminaries

The web can be modeled by a directed graph $G = (V, E)$, where V are webpages and a directed edge $(u \to v) \in E$ represents a hyperlink in u that references v. Although the web graph is usually viewed as an unweighted graph, our discussion can be extended to weighted models. To deal with the problem of dangling nodes with no out-edges, we assume that an artificial node with a single self-loop has been added to the graph, and an edge has been added from each dangling node to this artificial node. Let A denote the adjacency matrix of G. For each $u \in V$, let $d_{\text{out}}(u)$ denote the out-degree of u and let $d_{\text{in}}(u)$ denote the in-degree of u. Let D_{out} be the diagonal matrix of out-degrees.

We will now define PageRank vectors and contribution vectors. For convenience, we will view all vectors as row vectors, unless explicitly stated otherwise.

For a teleportation constant α, the PageRank vector $\mathbf{pr}_\alpha$ defined by Brin and Page [Brin and Page 98] satisfies the following equation:

$$\mathbf{pr}_\alpha = \alpha \cdot \mathbf{1} + (1 - \alpha) \cdot \mathbf{pr}_\alpha \cdot M, \tag{2.1}$$

where M is the random walk transition matrix given by $M = D_{\text{out}}^{-1} A$ and $\mathbf{1}$ is the row vector of all ones (always of proper size). The PageRank of a page u is then $\text{pr}_\alpha(u)$. When there is no danger of confusion, we may drop the subscript α. Note that the above definition corresponds to the normalization $\sum_u \text{pr}_\alpha(u) = |V|$.

Similarly, the personalized PageRank vector $\mathbf{ppr}(\alpha, u)$ of a page $u \in V$, defined by Haveliwala [Haveliwala 03], satisfies the following equation:

$$\mathbf{ppr}(\alpha, u) = \alpha \cdot \mathbf{e}_u + (1 - \alpha) \cdot \mathbf{ppr}(\alpha, u) \cdot M, \tag{2.2}$$

where $\mathbf{e}_u$ is the row unit vector whose uth entry is equal to 1.

Let P_α denote the (personalized) PageRank matrix, whose uth row is the personalized PageRank vector $\mathbf{ppr}(\alpha, u)$. The (global) PageRank vector $\mathbf{pr}_\alpha$ is then $\mathbf{1} \cdot P_\alpha$, the sum of all the personalized PageRank vectors. The *PageRank contribution* of u to v is defined to be the (u, v)th entry of P_α, and will be written $\text{ppr}_\alpha(u \to v)$. The contribution vector $\mathbf{cpr}(\alpha, v)$ for the vertex v is defined to be the row vector whose transpose is the vth column of P_α. If $\mathbf{c} = \mathbf{cpr}(\alpha, v)$ is the contribution vector for v, then we denote by $\mathbf{c}(S)$ the total contribution of the vertices in S to the PageRank of v. In particular, we have $\mathbf{c}(V) = \text{pr}_\alpha(v)$ and $\mathbf{c}(u) = \text{ppr}_\alpha(u \to v)$.

3. Local Approximation of PageRank Contributions

In this section, we describe an algorithm for computing an approximation of the contribution vector $\mathbf{c} = \mathbf{cpr}(\alpha, v)$ of a vertex v.

Definition 3.1. (Approximate contribution.) A vector $\tilde{\mathbf{c}}$ is an *ϵ-approximation* of the contribution vector $\mathbf{c} = \mathbf{cpr}(\alpha, v)$ if $\tilde{\mathbf{c}} \geq 0$ and, for all vertices u,

$$\mathbf{c}(u) - \epsilon \cdot \mathrm{pr}_\alpha(v) \leq \tilde{\mathbf{c}}(u) \leq \mathbf{c}(u).$$

A vector $\tilde{\mathbf{c}}$ is an *ϵ-absolute-approximation* of the contribution vector $\mathbf{c} = \mathbf{cpr}(\alpha, v)$ if $\tilde{\mathbf{c}} \geq 0$ and, for all vertices u,

$$\mathbf{c}(u) - \epsilon \leq \tilde{\mathbf{c}}(u) \leq \mathbf{c}(u).$$

Clearly, an ϵ-approximation of $\mathbf{cpr}(\alpha, v)$ is an $(\epsilon \cdot \mathrm{pr}_\alpha(v))$-absolute-approximation of $\mathbf{cpr}(\alpha, v)$. In the algorithm below, we will focus on the computation of an ϵ-absolute-approximation of the contribution vector.

The *support* of a nonnegative vector $\tilde{\mathbf{c}}$, denoted by $\mathrm{Supp}(\tilde{\mathbf{c}})$, is the set of all vertices whose entries in $\tilde{\mathbf{c}}$ are strictly positive. The vector $\mathbf{c}$ has a canonical ϵ-absolute-approximation. Let $\bar{\mathbf{c}}$ denote the vector

$$\bar{\mathbf{c}}(u) = \begin{cases} \mathbf{c}(u) & \text{if } \mathbf{c}(u) > \epsilon, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, $\bar{\mathbf{c}}$ is the ϵ-absolute-approximation of $\mathbf{c}$ with the smallest support. Moreover, $\|\bar{\mathbf{c}}\|_1 \leq \|\mathbf{c}\|_1$, and thus $|\mathrm{Supp}(\bar{\mathbf{c}})| \leq \|\mathbf{c}\|_1/\epsilon$. Our local algorithm attempts to find an approximation $\tilde{\mathbf{c}}$ of $\mathbf{c}$ that has a similar support structure to that of $\bar{\mathbf{c}}$.

3.1. High-Level Idea of the Local Algorithm

It is well known that for each α, the personalized PageRank vector that satisfies (2.2) also satisfies

$$\mathbf{ppr}(\alpha, u) = \alpha \sum_{t=0}^{\infty} (1-\alpha)^t \cdot \left(\mathbf{e}_u M^t\right). \tag{3.1}$$

The contribution of u to v can then be written in the following way:

$$\mathrm{ppr}_\alpha(u \to v) = \langle\, \mathbf{ppr}(\alpha, u)\,,\, \mathbf{e}_v \,\rangle \tag{3.2}$$

$$= \left\langle\, \alpha \sum_{t=0}^{\infty} (1-\alpha)^t (\mathbf{e}_u M^t)\,,\, \mathbf{e}_v \,\right\rangle \tag{3.3}$$

$$= \left\langle\, \mathbf{e}_u\,,\, \alpha \sum_{t=0}^{\infty} (1-\alpha)^t (\mathbf{e}_v M^T)^t \,\right\rangle. \tag{3.4}$$

The standard way to compute the contribution of u to v is based on (3.3). We refer to this approach as the time-forward calculation of $\mathrm{ppr}_\alpha(u \to v)$. Recall

that $\mathbf{e}_u M^t$ is the t-step random walk distribution starting from u. In the time-forward calculation, we emulate the random walk from u step by step and add up the walk distributions scaled by the power sequence of $(1-\alpha)^t$. Without knowing in advance which vertices u make large contributions to v, one may have to perform the time-forward calculation of $\mathbf{ppr}(\alpha, u)$ for many vertices u to obtain a good approximation of $\mathbf{cpr}(\alpha, v)$.

To overcome this difficulty, we can directly calculate $\mathbf{cpr}(\alpha, v)$ in the manner suggested by (3.4). This equation implies that

$$\mathbf{cpr}(\alpha, v) = \alpha \sum_{t=0}^{\infty} (1-\alpha)^t \cdot \left(\mathbf{e}_v (M^T)^t\right). \tag{3.5}$$

Thus, the contribution vector can be computed by starting with $\mathbf{e}_v$, iteratively computing $\mathbf{e}_v(M^T)^t$, and adding up the resulting vectors scaled by the power sequence of $(1-\alpha)^t$. Note that the matrix M^T is no longer a random walk matrix, since the sum of each row will in general not be equal to 1. Unlike the time-forward calculation, the direct calculation of $\mathbf{cpr}(\alpha, v)$ is no longer an emulation of the random walk starting from v. This fact complicates the error analysis of the next subsection.

We remark that (3.5) can also be derived quickly and easily from the fact that the contribution vector is a column of P_α, and from the matrix equation $P_\alpha = \alpha \sum_{t=0}^{\infty} (1-\alpha)^t M^t$. We derived it using (3.4) to highlight the difference between directly computing the contribution vector and computing several personalized PageRank vectors.

The discussion above provides one way to directly compute $\mathbf{cpr}(\alpha, v)$, but our local algorithm will perform a different calculation. Instead of iteratively computing the vectors $\mathbf{e}_v(M^T)^t$, we adapt the technique of Jeh and Widom [Jeh and Widom 03] for computing personalized PageRank to the task of computing contribution vectors. Using this method, we can compute the contribution vector in a decentralized way, and avoid spending computational effort manipulating small numerical values. This enables us to bound the running time required to obtain a fixed level of error.

The formula in (3.5) also enables us to compute the vector of contributions to a specified subset S of vertices, which we define to be $\mathbf{cpr}(\alpha, S) = \sum_{v \in S} \mathbf{cpr}(\alpha, v)$. Let $\mathbf{e}_S = \sum_{v \in S} \mathbf{e}_v$. Then,

$$\mathbf{cpr}(\alpha, S) = \alpha \sum_{t=0}^{\infty} (1-\alpha)^t \cdot \left(\mathbf{e}_S (M^T)^t\right). \tag{3.6}$$

To further abuse notation, for any nonnegative vector $\mathbf{s}$, we define

$$\mathbf{cpr}(\alpha, \mathbf{s}) = \alpha \sum_{t=0}^{\infty} (1-\alpha)^t \cdot \left(\mathbf{s} (M^T)^t\right). \tag{3.7}$$

3.2. The Local Algorithm and Its Analysis

The theorem below describes our algorithm `ApproxContributions` for computing an ϵ-absolute-approximation of the contribution vector of a target vertex v. We give an upper bound on the number of vertices examined by the algorithm that depends on $\mathrm{pr}_\alpha(v)$, ϵ, and α, but is otherwise independent of the number of vertices in the graph. The algorithm performs a sequence of operations, which we call pushback operations. Each pushback operation is performed on a single vertex of the graph, and requires time proportional to the in-degree of that vertex. We place an upper bound on the *number* of pushback operations performed by the algorithm, rather than the total running time of the algorithm. The total running time of the algorithm depends on the in-degrees of the sequence of vertices on which the pushback operations were performed. The number of pushback operations is an upper bound on the number of vertices in the support of the resulting approximate contribution vector.

Theorem 3.2. *The algorithm* `ApproxContributions`$(v, \alpha, \epsilon, \mathbf{p}_{\max})$ *has the following properties. The input is a vertex v, two constants α and ϵ in the interval $(0, 1]$, and a real number $\mathbf{p}_{\max}$. The algorithm computes a vector $\tilde{\mathbf{c}}$ such that $0 \leq \tilde{\mathbf{c}} \leq \mathbf{c}$, and either*

1. *$\tilde{\mathbf{c}}$ is an ϵ-absolute approximation of $\mathbf{cpr}(\alpha, v)$, or*
2. *$\|\tilde{\mathbf{c}}\|_1 \geq \mathbf{p}_{\max}$.*

The number of pushback operations P performed by the algorithm satisfies the following bound:

$$P \leq \frac{\min\left(\mathrm{pr}_\alpha(v), \mathbf{p}_{\max}\right)}{\alpha\epsilon} + 1.$$

The proof of Theorem 3.2 is based on a series of facts that we describe below. The starting point is the following observation, which is easy to verify from (3.7). For any vector $\mathbf{s}$,

$$\mathbf{cpr}(\alpha, \mathbf{s})M^T = \mathbf{cpr}(\alpha, \mathbf{s}M^T). \tag{3.8}$$

We can further derive the following equation:

$$\mathbf{cpr}(\alpha, \mathbf{s}) = \alpha\mathbf{s} + (1-\alpha)\cdot\mathbf{cpr}(\alpha, \mathbf{s})M^T = \alpha\mathbf{s} + (1-\alpha)\cdot\mathbf{cpr}(\alpha, \mathbf{s}M^T). \tag{3.9}$$

This is the transposed version of the equation that was used by Jeh and Widom to compute approximate personalized PageRank vectors [Jeh and Widom 03]. Very naturally, we will use it to compute approximate contribution vectors.

The algorithm `ApproxContributions`$(v, \alpha, \epsilon, \mathbf{p}_{\max})$ maintains a pair of vectors $\mathbf{p}$ and $\mathbf{r}$ with nonnegative entries, starting with the trivial approximation $\mathbf{p} = \vec{0}$

Algorithm 1. (pushback (u))

Let $\mathbf{p}' = \mathbf{p}$ and $\mathbf{r}' = \mathbf{r}$, except for these changes:

1. $\mathbf{p}'(u) = \mathbf{p}(u) + \alpha\mathbf{r}(u)$.
2. $\mathbf{r}'(u) = 0$.
3. For each vertex w such that $w \to u$:
   $\mathbf{r}'(w) = \mathbf{r}(w) + (1-\alpha)\mathbf{r}(u)/d_{\text{out}}(w)$.

and $r = \mathbf{e}_v$, and applies a series of pushback operations that increase $\|\mathbf{p}\|_1$ while maintaining the invariant $\mathbf{p} + \mathbf{cpr}(\alpha, \mathbf{r}) = \mathbf{cpr}(\alpha, v)$. Each pushback operation picks a single vertex u, moves an α-fraction of the mass at $\mathbf{r}(u)$ to $\mathbf{p}(u)$, and then modifies the vector $\mathbf{r}$ by replacing $\mathbf{r}(u)\mathbf{e}_u$ with $(1-\alpha)\mathbf{r}(u)\mathbf{e}_u M^T$. Note that $\|\mathbf{r}\|_1$ may increase or decrease during this operation. We will define the pushback operation more formally as Algorithm 1, and then verify that each pushback operation does indeed maintain the invariant.

Lemma 3.3. (Invariant.) *Let $\mathbf{p}'$ and $\mathbf{r}'$ be the result of performing* pushback(u) *on $\mathbf{p}$ and $\mathbf{r}$. If $\mathbf{p}$ and $\mathbf{r}$ satisfy the invariant $\mathbf{p} + \mathbf{cpr}(\alpha, \mathbf{r}) = \mathbf{cpr}(\alpha, v)$, then $\mathbf{p}'$ and $\mathbf{r}'$ satisfy the invariant $\mathbf{p}' + \mathbf{cpr}(\alpha, \mathbf{r}') = \mathbf{cpr}(\alpha, v)$.*

Proof. After the pushback operation, we have, in vector notation,

$$\mathbf{p}' = \mathbf{p} + \alpha\mathbf{r}(u)\mathbf{e}_u.$$
$$\mathbf{r}' = \mathbf{r} - \mathbf{r}(u)\mathbf{e}_u + (1-\alpha)\mathbf{r}(u)\mathbf{e}_u M^T.$$

We will apply equation (3.9) to $\mathbf{r}(u)\mathbf{e}_u$ to show that $\mathbf{p} + \mathbf{cpr}(\alpha, \mathbf{r}) = \mathbf{p}' + \mathbf{cpr}(\alpha, \mathbf{r}')$:

$$\begin{aligned}
\mathbf{cpr}(\alpha, \mathbf{r}) &= \mathbf{cpr}(\alpha, \mathbf{r} - \mathbf{r}(u)\mathbf{e}_u) + \mathbf{cpr}(\alpha, \mathbf{r}(u)\mathbf{e}_u) \\
&= \mathbf{cpr}(\alpha, \mathbf{r} - \mathbf{r}(u)\mathbf{e}_u) + \alpha\mathbf{r}(u)\mathbf{e}_u + \mathbf{cpr}(\alpha, (1-\alpha)\mathbf{r}(u)\mathbf{e}_u M^T) \\
&= \mathbf{cpr}(\alpha, \mathbf{r} - \mathbf{r}(u)\mathbf{e}_u + (1-\alpha)\mathbf{r}(u)\mathbf{e}_u M^T) + \alpha\mathbf{r}(u)\mathbf{e}_u \\
&= \mathbf{cpr}(\alpha, \mathbf{r}') + \mathbf{p}' - \mathbf{p},
\end{aligned}$$

which completes the proof. □

During each pushback operation, the quantity $\|\mathbf{p}\|_1$ increases by $\alpha\mathbf{r}(u)$. The quantity $\|\mathbf{p}\|_1$ can never exceed $\|\mathbf{cpr}(\alpha, v)\|_1$, which is equal to $\text{pr}_\alpha(v)$. By performing pushback operations only on vertices where $\mathbf{r}(u) \geq \epsilon$, we can ensure that $\|\mathbf{p}\|_1$ increases by a significant amount at each step, which allows us to bound the

Algorithm 2. (`ApproxContributions`$(v, \alpha, \epsilon, \mathbf{p}_{\max})$)

1. Let $\mathbf{p} = \vec{0}$, and $\mathbf{r} = \mathbf{e}_v$.
2. While $\mathbf{r}(u) > \epsilon$ for some vertex u:
   (a) Pick any vertex u where $\mathbf{r}(u) \geq \epsilon$.
   (b) Apply `pushback` (u).
   (c) If $\|\mathbf{p}\|_1 \geq \mathbf{p}_{\max}$, halt and output $\tilde{\mathbf{c}} = \mathbf{p}$.
3. Output $\tilde{\mathbf{c}} = \mathbf{p}$.

number of pushes required to compute an ϵ-absolute-approximation of the contribution vector. This is the idea behind the algorithm `ApproxContributions` (Algorithm 2).

This algorithm can be implemented by maintaining a queue containing those vertices u satisfying $\mathbf{r}(u) \geq \epsilon$. Initially, v is the only vertex in the queue. At each step, we take the first vertex u in the queue, remove it from the queue, and perform a pushback operation from that vertex. If the pushback operation raises the value of $\mathbf{r}(x)$ above ϵ for some in-neighbor x of u, then x is added to the back of the queue. This continues until the queue is empty, at which point all vertices satisfy $\mathbf{r}(u) < \epsilon$, or until $\|\mathbf{p}\|_1 \geq \mathbf{p}_{\max}$. We now show that this queue-based algorithm has the properties promised in Theorem 3.2.

Proof of Theorem 3.2. Let T be the total number of push operations performed by the algorithm, and let $\mathbf{p}_t$ and $\mathbf{r}_t$ be the states of the vectors $\mathbf{p}$ and $\mathbf{r}$ after t pushes. The initial setting of $\mathbf{p}_0 = \vec{0}$ and $\mathbf{r}_0 = \mathbf{e}_v$ satisfies the invariant $\mathbf{p}_t + \mathbf{cpr}(\alpha, \mathbf{r}_t) = \mathbf{cpr}(\alpha, v)$, which is maintained throughout the algorithm. Since $\mathbf{r}_t$ is nonnegative at each step, the error term $\mathbf{cpr}(\alpha, \mathbf{r}_t)$ is also nonnegative, so we have $\mathbf{cpr}(\alpha, v) - \mathbf{p}_t \geq 0$. In particular, this implies $\|\mathbf{p}_t\|_1 \leq \|\mathbf{cpr}(\alpha, v)\|_1 = \mathrm{pr}_\alpha(v)$.

Let $\tilde{\mathbf{c}} = \mathbf{p}_T$ be the vector output by the algorithm. When the algorithm terminates, we must have either $\|\tilde{\mathbf{c}}\|_1 \geq \mathbf{p}_{\max}$ or $\|\mathbf{r}_T\|_\infty \leq \epsilon$. In the latter case, the following calculation shows that $\tilde{\mathbf{c}}$ is an ϵ-absolute-approximation of $\mathbf{cpr}(\alpha, v)$:

$$\|\mathbf{cpr}(\alpha, v) - \tilde{\mathbf{c}}\|_\infty = \|\mathbf{cpr}(\alpha, \mathbf{r}_T)\|_\infty \leq \|\mathbf{r}_T\|_\infty \leq \epsilon.$$

The inequality $\|\mathbf{cpr}(\alpha, \mathbf{r}_T)\|_\infty \leq \|\mathbf{r}_T\|_\infty$ holds because $\mathbf{r}_T$ is nonnegative and each row of M sums to 1.

The vector $\mathbf{p}_{T-1}$ must have satisfied $\|\mathbf{p}_{T-1}\|_1 < \mathbf{p}_{\max}$, since the algorithm decided to push one more time. We have already observed that $\|\mathbf{p}_{T-1}\|_1 \leq$

$\mathrm{pr}_\alpha(v)$. Each push operation increased $\|\mathbf{p}\|_1$ by at least $\alpha\epsilon$, so we have

$$\alpha\epsilon(T-1) \leq \|\mathbf{p}_{T-1}\|_1 \leq \min\left(\|\mathbf{cpr}(\alpha, v)\|_1, \mathbf{p}_{\max}\right).$$

This gives the desired bound on T. □

It is possible to perform a pushback operation on the vertex u, and to perform the necessary queue updates, in time proportional to $d_{\mathrm{in}}(u)$. Therefore, the running time of the algorithm is proportional to the sum over all pushback operations of the in-degree of the pushed vertex. We remark that a useful heuristic for choosing the next vertex to push is to maintain a priority queue rather than a queue, and to push from the vertex u with the highest value of $\mathbf{r}(u)$. That approach was used previously by Berkhin [Berkhin 06] for the related task of computing personalized PageRank vectors.

We can compute an ϵ-approximation of $\mathbf{cpr}(\alpha, v)$, provided that $\mathrm{pr}_\alpha(v)$ is known, by calling the algorithm `ApproxContributions`$(v, \alpha, \epsilon \cdot \mathrm{pr}_\alpha(v), \mathrm{pr}_\alpha(v))$.

Corollary 3.4. (ϵ-approximation of contribution vectors.) *Given $\mathrm{pr}_\alpha(v)$, an ϵ-approximation of $\mathbf{cpr}(\alpha, v)$ can be computed with $\frac{1}{\alpha\epsilon} + 1$ pushback operations.*

We also observe that using (3.6), our algorithm can be easily adapted to compute an ϵ-absolute-approximation and ϵ-approximation of $\mathbf{cpr}(\alpha, S)$ for a group S of vertices, with a similar bound on the number of pushback operations.

3.3. The Support of the Approximate Contribution Vector

The number of vertices in the support of the ϵ-approximate contribution vector $\tilde{\mathbf{c}}$ is bounded above by the number of pushback operations used to compute it, which is at most $\frac{1}{\alpha\epsilon}+1$. In this section, we introduce a small modification to the algorithm in the previous section, and show that for this modified algorithm we can prove a stronger upper bound on the size of the support of the approximate contribution vector. In particular, consider the following modified pushback operation: instead of moving all the mass from $\mathbf{r}(u)$ during the pushback operation, move all but $\epsilon/2$ units of mass, and leave $\epsilon/2$ units on $\mathbf{r}(u)$. This will give us a lower bound on the value of $\mathbf{r}(u)$ on each vertex u from which a push operation has ever been performed, which will in turn allow us to bound the size of the support.

For the remainder of this section, assume that `ApproxContributions` uses the modified pushback operation described above. The number of pushback operations performed by the modified algorithm is at most $\frac{2}{\alpha\epsilon}+1$, which is twice the bound we proved for the unmodified algorithm, and which can be proved by the same argument. The modified algorithm ensures that $\mathbf{r}(x) \geq \epsilon/2$ at each vertex in $\mathrm{Supp}(\tilde{\mathbf{c}})$. Below, we will use this fact to give a family of bounds on the size of $\mathrm{Supp}(\tilde{\mathbf{c}})$. These bounds do not necessarily hold for the unmodified

algorithm, so one must implement the modified version for the stronger support bounds to hold.

We will abuse our notation a bit by defining

$$\mathrm{pr}_\alpha(\mathbf{x} \to \mathbf{y}) = \langle\, \mathbf{x}M_\alpha\,,\, \mathbf{y}\,\rangle,$$

where $M_\alpha = P_\alpha$ is the PageRank matrix. In particular, $\mathrm{pr}_\alpha(\mathbf{x} \to \mathbf{e}_S)$ is the amount of probability from the PageRank vector with starting distribution $\mathbf{x}$ on the set S.

Proposition 3.5. *Let $\tilde{\mathbf{c}}$ be the ϵ-approximate contribution vector for v computed by the modified algorithm described above, and let $S = \mathrm{Supp}(\tilde{\mathbf{c}})$. For any nonnegative vector $\mathbf{z}$, we have the following upper bound on S:*

$$\mathrm{pr}_\alpha(\mathbf{z} \to \mathbf{e}_S) \le \frac{2}{\epsilon}\mathrm{pr}_\alpha(\mathbf{z} \to \mathbf{e}_v).$$

Proof. Note that $\mathbf{ppr}(\alpha, v) = \mathbf{e}_v M_\alpha$ and $\mathbf{cpr}(\alpha, v) = \mathbf{e}_v M_\alpha^T$. We know that $\mathbf{cpr}(\alpha, \mathbf{r}) \le \mathbf{cpr}(\alpha, \mathbf{e}_v)$, which can also be written $\mathbf{r}M_\alpha^T \le \mathbf{e}_v M_\alpha^T$. Let $S = \mathrm{Supp}(\tilde{\mathbf{c}})$ and recall that $\mathbf{r}(x) \ge \epsilon/2$ for any vertex $x \in S$. Then,

$$\langle\, \mathbf{z}M_\alpha\,,\, \mathbf{e}_v\,\rangle = \left\langle\, \mathbf{z}\,,\, \mathbf{e}_v M_\alpha^T\,\right\rangle \ge \left\langle\, \mathbf{z}\,,\, \mathbf{r}M_\alpha^T\,\right\rangle = \langle\, \mathbf{z}M_\alpha\,,\, \mathbf{r}\,\rangle \ge (\epsilon/2)\langle\, \mathbf{z}M_\alpha\,,\, \mathbf{e}_S\,\rangle.$$

In the second step we needed $\mathbf{z}$ to be nonnegative, and in the last step we needed $\mathbf{z}M_\alpha$ to be nonnegative, which is true whenever $\mathbf{z}$ is nonnegative. □

In words, this proposition states that for any starting vector $\mathbf{z}$, the amount of probability from the PageRank vector $\mathbf{ppr}(\alpha, \mathbf{z})$ on the set $S = \mathrm{Supp}(\tilde{\mathbf{c}})$ is at most $2/\epsilon$ times the amount on the vertex v. If we let $\mathbf{z} = \mathbf{e}_V$, then we obtain a bound on the amount of global PageRank on the set S,

$$\mathrm{pr}_\alpha(S) \le \frac{2}{\epsilon}\mathrm{pr}_\alpha(v).$$

To see that this bound is at least as strong as what we knew before, recall that the PageRank of any given vertex is at least α. If we make the pessimistic assumption that $\mathrm{pr}_\alpha(u) = \alpha$ for each $u \in \mathrm{Supp}(\tilde{\mathbf{c}})$, then the bound we have just proved reduces to our earlier bound on the number of pushback operations,

$$|\mathrm{Supp}(\tilde{\mathbf{c}})| \le 2\mathrm{pr}_\alpha(v)/\alpha\epsilon.$$

4. Computing Supporting Sets

In this section, we use our local algorithm for approximating contribution vectors to compute approximate supporting sets, sets of vertices that contribute

significantly to the PageRank of a target vertex. There are several natural notions of supporting sets, which we define below. For a vertex v, let π_v be the permutation that orders the entries $\mathbf{cpr}(\alpha, v)$ from the largest to the smallest. Ties may be broken arbitrarily.

- **top k contributors**: the first k pages of π_v.
- **δ-significant contributors**: $\{u \mid \mathrm{ppr}_\alpha(u \to v) > \delta\}$.
- **ρ-supporting set**: a set S of pages such that

$$\mathrm{ppr}_\alpha(S \to v) \geq \rho \cdot \mathrm{pr}_\alpha(v).$$

In addition, let $k_\rho(v)$ be the smallest integer such that

$$\mathrm{ppr}_\alpha(\pi_v(1 : k_\rho(v)) \to v) \geq \rho \cdot \mathrm{pr}_\alpha(v).$$

Clearly the set of the first $k_\rho(v)$ pages of π_v is the minimum-size ρ-supporting set for v. Also, we define $\rho_k(v) = \mathrm{ppr}_\alpha(\pi_v(1 : k) \to v)/\mathrm{pr}_\alpha(v)$ to be the fraction of v's PageRank contributed by its top k contributors.

4.1. Approximating Supporting Sets

Without precisely computing $\mathbf{cpr}(\alpha, v)$ it might be impossible to identify supporting sets exactly, so we consider approximate supporting sets. For a precision parameter ϵ, we define the following:

- **ϵ-precise top k contributors**: a set of k pages that contains all pages whose contribution to v is at least $\mathrm{ppr}_\alpha(\pi_v(k) \to v) + \epsilon \cdot \mathrm{pr}_\alpha(v)$, but no page with contribution to v less than $\mathrm{ppr}_\alpha(\pi_v(k) \to v) - \epsilon \cdot \mathrm{pr}_\alpha(v)$.
- **ϵ-precise δ-significant contributors**: a set that contains the set of δ-significant contributors and is contained in the set of $(\delta - \epsilon)$-significant contributors.

Later in this section, we will also consider the computation of approximate ρ-supporting sets. The results in the remainder of this section assume that $\mathrm{pr}_\alpha(v)$ is known.

Theorem 4.1. *An ϵ-precise set of top k contributors of a vertex v can be found by performing $\frac{1}{\alpha\epsilon} + 1$* `pushback` *operations.*

Proof. Call $\tilde{\mathbf{c}} = \texttt{ApproxContributions}(v, \alpha, \epsilon \cdot \mathrm{pr}_\alpha(v), \mathrm{pr}_\alpha(v))$. Let $C = \mathrm{Supp}(\tilde{\mathbf{c}})$. If $|C| > k$, then return the vertices with the top k entries in $\tilde{\mathbf{c}}$; otherwise, return C augmented with $k - C$ arbitrarily chosen vertices not in C. Consider a page u with $\mathbf{cpr}(u, v) \geq \mathbf{cpr}(\pi_v(k), v) + \epsilon \cdot \mathrm{pr}_\alpha(v)$. Clearly $u \in C$ because

$\tilde{\mathbf{c}}(u) \geq \mathbf{cpr}(\pi_v(k), v)$, implying that $\tilde{\mathbf{c}}(u)$ is among the top k entries in $\tilde{\mathbf{c}}$. On the other hand, $\tilde{\mathbf{c}}(\pi_v(j))$ is at least $\mathbf{cpr}(\pi_v(k), v) - \epsilon \cdot \mathrm{pr}_\alpha(v)$ for all $j \in [1:k]$. Thus, each of the vertices with the top k entries in $\tilde{\mathbf{c}}$ must contribute at least $\mathbf{cpr}(\pi_v(k), v) - \epsilon \cdot \mathrm{pr}_\alpha(v)$ to v. □

Theorem 4.2. *An ϵ-precise δ-significant contributing set of a vertex v can be found by performing $\frac{1}{\alpha\epsilon} + 1$* `pushback` *operations.*

Proof. Call $\tilde{\mathbf{c}} =$ `ApproxContributions`$(v, \alpha, \epsilon \cdot \mathrm{pr}_\alpha(v), \mathrm{pr}_\alpha(v))$ and return the vertices whose entries in $\tilde{\mathbf{c}}$ are at least $(\delta - \epsilon) \cdot \mathrm{pr}_\alpha(v)$. Clearly, the set contains the δ-contributing set of v and is contained in the $(\delta - \epsilon)$-supporting set of v. Moreover, the number of pages not in the δ-supporting set that are included is at most $1/(\delta - \epsilon)$. □

In the remainder of this section, we consider the computation of approximate ρ-supporting sets. We give two different algorithms, one for finding a supporting set on a fixed number of vertices with the largest contribution possible, and one for finding a supporting set with a fixed contribution on as few vertices as possible.

Theorem 4.3. *Given a vertex v and an integer k, a set of k vertices that is a $(\rho_k - \epsilon)$-supporting set for v can be found by performing $\frac{k}{\alpha\epsilon} + 1$* `pushback` *operations.*

Proof. Compute $\tilde{\mathbf{c}} =$ `ApproxContributions`$(v, \alpha, \epsilon \mathrm{pr}_\alpha(v)/k, \mathrm{pr}_\alpha(v))$. Let S_k be the set of k top contributors to v, which are the k vertices with the highest values in $\mathbf{c}$, and let $\tilde{S}_k$ be the set of k vertices with the highest values in $\tilde{\mathbf{c}}$. The set $\tilde{S}_k$ meets the requirements of the theorem, since we have

$$\tilde{\mathbf{c}}(\tilde{S}_k) \geq \mathbf{c}(S_k) - k(\epsilon \mathrm{pr}_\alpha(v)/k) \geq \rho_k \cdot \mathrm{pr}_\alpha(v) - \epsilon \cdot \mathrm{pr}_\alpha(v) = \mathrm{pr}_\alpha(v)(\rho_k - \epsilon).$$

The proof is complete. □

Theorem 4.4. *Assume that we are given ρ but not k_ρ. A set of at most k_ρ vertices that is a $(\rho - \epsilon)$-supporting set for v can be found by performing $O(k_\rho \log k_\rho / \alpha\epsilon)$* `pushback` *operations.*

Proof. The challenge here is that we do not know k_ρ, so we need to use a binary search procedure to find a proxy for k_ρ. We will proceed in two phases. In the first phase, we guess a value of k, starting with $k = 1$, and compute $\tilde{\mathbf{c}} =$ `ApproxContributions`$(v, \alpha, \epsilon \cdot \mathrm{pr}_\alpha(v)/k, \mathrm{pr}_\alpha(v))$. As in Theorem 4.3, let $\tilde{S}_k$ be the set of k vertices with the highest values in $\tilde{\mathbf{c}}$, which we know satisfies $\tilde{\mathbf{c}}(\tilde{S}_k) \geq (\rho_k - \epsilon)$. If we observe that $\tilde{\mathbf{c}}(\tilde{S}_k) < (\rho - \epsilon)$, then we double k and repeat the procedure. If we observe that $\tilde{\mathbf{c}}(\tilde{S}_k) \geq (\rho - \epsilon)$, then we halt and proceed to the

second phase, and set k_1 to be the value of k for which this happens. We must have $k_1 \leq 2k_\rho$, since we are guaranteed to halt if $k \geq k_\rho$.

Let $k_0 = k_1/2$ be the value of k from the step before the first phase halted. In the second phase, we perform binary search within the interval $[k_0, k_1]$ to find the smallest integer $k_{\min}$ for which $\tilde{\mathbf{c}}(\tilde{S}_{k_{\min}}) \geq (\rho - \epsilon)$, which must satisfy $k_{\min} \leq k_\rho$. We output $\tilde{S}_{k_{\min}}$.

Each time we call the subroutine

$$\tilde{\mathbf{c}} = \texttt{ApproxContributions}(v, \alpha, \epsilon \mathrm{pr}_\alpha(v)/k, \mathrm{pr}_\alpha(v)),$$

it requires $\frac{k}{\alpha\epsilon}+1$ push operations. In the first phase we call this subroutine with a sequence of k values that double from 1 up to at most $2k_\rho$, so the number of push operations performed is $O(k_\rho/\alpha\epsilon+\log k_\rho)$. In the second phase, the binary search makes at most $\log k_\rho$ calls to the subroutine, with k set to at most $2k_\rho$ in each step, so the number of push operations performed is $O(k_\rho \log k_\rho/\alpha\epsilon+\log k_\rho)$. The total number of push operations performed in both phases is $O(k_\rho \log k_\rho/\alpha\epsilon)$. □

4.2. Local Estimation of PageRank

Up to this point, we have assumed when computing the supporting set of a vertex that its PageRank is known. We now consider how to apply our approximate contribution algorithm when nothing is known about the PageRank of the target vertex. In particular, we consider the problem of computing a lower bound on the PageRank of a vertex using local computation.

A natural lower bound on the PageRank $\mathrm{pr}_\alpha(v)$ is provided by the contribution to v of its top k contributors, $p_k = \mathbf{cpr}(\pi_v(1:k), v)$. The theorem below shows that given k, we can efficiently certify that $\mathrm{pr}_\alpha(v)$ is approximately as large as p_k without prior knowledge of $\mathrm{pr}_\alpha(v)$ or p_k. This should be contrasted with the algorithms from the previous section, for which we needed to know the value $\mathrm{pr}_\alpha(v)$ in order to set ϵ appropriately to obtain the stated running times. This result is useful for determining the amount of contribution a node receives from its top contributors. The fraction of a node's PageRank that it receives from its top k contributors can vary greatly between nodes, and depends on the structure of the graph.

Theorem 4.5. *Given k and δ, we can compute a real number p such that*

$$p_k(1+\delta)^{-2} \leq p \leq \mathrm{pr}_\alpha(v),$$

where $p_k = \mathbf{cpr}(\pi_v(1:k), v)$, by performing $10k\log(k/\alpha\delta)/\alpha$ `pushback` operations.

Proof. Fix k and δ, choose a value of p, and compute

$$\tilde{\mathbf{c}} = \texttt{ApproxContributions}(v, \alpha, \epsilon, p)$$

with $\epsilon = \delta p/k$. The number of `pushback` operations performed is at most

$$1 + \frac{p}{\alpha\epsilon} = 1 + \frac{p}{\alpha(\delta p/k)} = 1 + \frac{10k}{\alpha}.$$

When the algorithm halts, either we have $\|\tilde{\mathbf{c}}\|_1 \geq p$, in which case we have certified that $\mathrm{pr}_\alpha(v) \geq p$, or else we have $\|\tilde{\mathbf{c}} - \mathbf{cpr}(\alpha, v)\|_\infty \leq \delta p/k$, in which case we have certified that $p_k \leq (1+\delta)p$, by the following calculation:

$$p_k = \mathbf{cpr}(\pi_v(1:k), v) \leq \tilde{\mathbf{c}}(\pi_v(1:k), v) + (\delta p/k)k \leq p + \delta p.$$

We now perform binary search over p in the range $[\alpha, k]$. Let p_{low} be the largest value of p for which we have certified that $\mathrm{pr}_\alpha(v) \geq p$, and let p_{high} be the smallest value of p for which we have certified that $p_k \leq (1+\delta)p$. We perform binary search until $p_{\mathrm{high}} \leq p_{\mathrm{low}}(1+\delta)$, which requires at most $\log(k/\alpha\delta)$ steps. Then, p_{low} has the property described in the theorem,

$$\mathrm{pr}_\alpha(v) \geq p_{\mathrm{low}} \geq p_{\mathrm{high}}(1+\delta)^{-1} \geq p_k(1+\delta)^{-2}.$$

The total number of `pushback` operations performed during the calls to `ApproxContributions` during the binary search is at most $10k\log(k/\alpha\delta)/\alpha$. □

5. Weighted Contributions and PageRank Traffic

In this section we introduce two alternative ways to measure the effect a given node has on the PageRank of another node: weighted contributions and PageRank traffic. Weighted contributions capture the idea that contributions from nodes with high PageRank should count more than contributions from nodes with low PageRank. PageRank traffic is a natural way to define the amount of PageRank that is contributed *through* a given node. We derive an equation that shows that these two concepts are closely related, and show that both can be computed by computing PageRank contributions. This helps to unify seemingly different notions of how PageRank is contributed and transmitted within a graph.

We first define the weighted contribution from u to v, which is simply the contribution from u to v multiplied by the PageRank of u.

Definition 5.1. (Weighted contributions.) The *weighted contribution* from u to v is defined to be the product $\mathrm{pr}_\alpha(u) \cdot \mathrm{ppr}_\alpha(u \to v)$.

While weighted contributions are a natural way to take into account the importance of nodes, the significance of this quantity in terms of the graph is not immediately clear. We will give a concrete interpretation of weighted contributions by introducing the concept of PageRank traffic.

PageRank traffic is defined in terms of the paths followed by a collection of random surfers. To motivate this definition, we first review the standard definition of PageRank contributions in terms of these random surfers. Consider a random surfer beginning at a specified node x_0. At each step, the surfer moves to a random neighbor of the current node with probability $1 - \alpha$, and with probability α the surfer stops at the current node. The path followed by the random surfer is the sequence of nodes $p = x_0 \dots x_l$, where $l = l(p)$ is the length of the path and x_l is the node at which the surfer stops. We define the path weight $w_\alpha(p)$ to be the probability that a random surfer starting from x_0 follows the path p. This probability can be written as follows:

$$w_\alpha(p) = \alpha(1-\alpha)^{l(p)} \prod_{i=0}^{l(p)-1} \left(\frac{1}{d_{\text{out}}(x_i)} \right).$$

We let $P_{(u,v)}$ be the set of paths that begin at u and end at v. The contribution $\text{ppr}_\alpha(u \to v)$ can be viewed as the probability that a random surfer that starts from u stops at v, which is the sum of the path weights in $P_{(u,v)}$,

$$\text{ppr}_\alpha(u \to v) = \sum_{p \in P_{(u,v)}} w_\alpha(p). \tag{5.1}$$

It can be seen from (3.1) that this definition of contributions is equivalent to the definition we presented earlier.

We can now define PageRank traffic. The PageRank traffic from u to v through x, which will be written $\text{ppr}_\alpha(u \to x \to v)$, is defined to be the expected number of times a random surfer that starts from u travels through x and stops at v, taking into account the multiplicity of times the surfer travels through x.

Definition 5.2. (PageRank traffic.) Let $I(x, p)$ be the number of times node x appears in path p. The *PageRank traffic* from u to v through x is written $\text{ppr}_\alpha(u \to x \to v)$, and is defined to be

$$\text{ppr}_\alpha(u \to x \to v) = \sum_{p \in P_{(u,v)}} w_\alpha(p) I(x, p). \tag{5.2}$$

We believe that PageRank traffic is a natural way to define the amount of PageRank that is contributed from u to v *through* a given node x. As an important special case, we define *global PageRank traffic*.

Definition 5.3. (Global PageRank traffic.) The *global PageRank traffic* routed through x to v is defined to be

$$\text{ppr}_\alpha(\mathbf{1} \to x \to v) = \sum_{u \in V} \text{ppr}_\alpha(u \to x \to v). \tag{5.3}$$

If we consider a collection of random surfers, one starting from each node, the global PageRank traffic measures the expected number of these random surfers that travel through x and end up at v. It is a natural way to measure the *total* amount of PageRank that is contributed to v through a given node x. The global PageRank traffic through x provides an alternative to the PageRank contribution from x for measuring how the PageRank of v is influenced by the presence of node x.

We now show that PageRank traffic can be written as the product of two PageRank contributions.

Proposition 5.4.

$$\mathrm{ppr}_\alpha(u \to x \to v) = \frac{1}{\alpha}\mathrm{ppr}_\alpha(u \to x) \cdot \mathrm{ppr}_\alpha(x \to v). \tag{5.4}$$

Proof. Given two paths p and q, where the last node in p is the first node in q, let $p \circ q$ denote the composition of the two paths:

$$\begin{aligned}
\mathrm{ppr}_\alpha(u \to x \to v) &= \sum_{p \in P_{(u,v)}} w_\alpha(p) I(x,p) \\
&= \sum_{(p,q) \in P_{(u,x)} \times P_{(x,v)}} w_\alpha(p \circ q) \\
&= \sum_{(p,q) \in P_{(u,x)} \times P_{(x,v)}} \frac{1}{\alpha} w_\alpha(p) w_\alpha(q) \\
&= \frac{1}{\alpha}\Big(\sum_{p \in P_{(u,x)}} w_\alpha(p)\Big)\Big(\sum_{q \in P_{(x,v)}} w_\alpha(q)\Big) \\
&= \frac{1}{\alpha}\mathrm{ppr}_\alpha(u \to x) \cdot \mathrm{ppr}_\alpha(x \to v),
\end{aligned}$$

which completes the proof. □

As an immediate corollary we arrive at the main result of this section, which relates the global PageRank traffic through a node to the weighted PageRank contributions from that node.

Corollary 5.5.

$$\mathrm{ppr}_\alpha(\mathbf{1} \to x \to v) = \frac{1}{\alpha}\mathrm{pr}_\alpha(x) \cdot \mathrm{ppr}_\alpha(x \to v).$$

6. Concluding Remarks

6.1. Improving the Dependency on In-Degrees

In our performance analysis, we give a bound of $\mathrm{pr}_\alpha(v)/(\alpha\epsilon) + 1$ on the total number of `pushback` operations performed by our algorithm. In a pushback at a vertex u, we update the entry for u in the vector $\mathbf{p}$ as well as the entries in $\mathbf{r}$ for all vertices that point to u. As a result, the overall time complexity of our algorithm is proportional to the sum of the in-degrees of the sequence of vertices that we push back from. A possible direction for future research is to devise an algorithm whose running time can be bounded in terms of the total in-degree of the supporting set that the algorithm attempts to approximate. This type of bound would offer stronger control over the running time than the result obtained in this paper, where the number of pushback operations is bounded in terms of the number of vertices in the supporting set, but the running time depends on the in-degrees of the vertices from which the sequence of push operations is performed.

6.2. Computing Contribution Vectors via the Time-Reverse Chain

As noted earlier, the matrix M^T in the formula of (3.5) may not be Markov. It is natural ask whether the time-reverse Markov chain of the random-walk matrix M may be used to compute the contribution vector for a vertex v, and, if so, whether this method is efficient.

For the following discussion, we assume that M has a unique stationary distribution, which will not be true for general directed graphs. Recall the following definition of the time-reverse Markov chain.

Definition 6.1. (Time-reverse chain.) Given a Markov chain M with transition probability m_{ij}, and stationary distribution π, the *time-reverse chain* is the Markov chain R with transition probability $r_{ij} = \pi(j)m_{ji}/\pi(i)$.

In other words, let Π be the matrix whose (i, j) entry is $\pi(j)/\pi(i)$. Then $R = \Pi \cdot * M^T$, where the operation $\cdot *$ is the componentwise multiplication of two matrices. The time-reverse chain has the following properties:

- R has the same stationary distribution as M,
- for all i, k, and t, consider the t-step random walk starting from i in M and k in R; then

$$\left\langle \mathbf{e}_i M^t \, , \, \mathbf{e}_k \right\rangle = \left(\frac{\pi(k)}{\pi(i)} \right) \left\langle \mathbf{e}_k R^t \, , \, \mathbf{e}_i \right\rangle \tag{6.1}$$

Recall that $\langle \mathbf{e}_i M^t , \mathbf{e}_k \rangle$ is equal to the probability that k is the vertex reached by a t-step random walk from i. Let $\mathrm{ppr}_\alpha^M(u \to v)$ denote the personalized PageRank contribution from u to v in a Markov chain M.

Theorem 6.2. *Suppose a Markov chain M has a stationary distribution π and R is its time-reverse chain. Then*

$$\mathrm{ppr}_\alpha^M(u \to v) = \left(\frac{\pi(v)}{\pi(u)}\right) \mathrm{ppr}_\alpha^R(v \to u). \tag{6.2}$$

Proof. The result follows from (3.3) and (6.1). □

Thus, if the stationary distribution exists, we can in principle compute the contribution vector of M by computing the personalized PageRank vector for v in the time-reverse chain. We argue that the method we presented in Section 3 is preferable to the time-reverse Markov chain method for the following reasons. Our method does not require that M have a stationary distribution. Computing a personalized PageRank vector in the time-reverse Markov chain requires that we first compute the stationary distribution π of M, which may be computationally expensive. Perhaps most important is the difference in the error analysis. If the stationary distribution exists, one can compute an ϵ-approximate contribution vector by computing a personalized PageRank vector in R for which the error at each vertex i is at most $\epsilon\pi(i)$. If $\pi(i)$ is extremely small at some vertices, and it may be exponentially small in the number of vertices in the graph, this will require a large amount of computation.

We prefer the method presented in Section 3 to the time-reverse method for most graphs that are likely to be encountered in practice. However, there are special cases in which the time-reverse method will be efficient. In particular, if the Markov chain has a stationary distribution that is nearly proportional to the in-degrees of the vertices, as it would be in an undirected graph, then computing a personalized PageRank vector in the time-reverse chain is an efficient way to compute a contribution vector.

References

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

[Becchetti et al. 06] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. "Link-Based Characterization and Detection of Web Spam." Paper presented at Second International Workshop on Adversarial Information Retrieval on the Web

(AIRWeb), Seattle, WA, August 10, 2006. Available at http://airweb.cse.lehigh.edu/2006/becchetti.pdf.

[Benczúr et al. 05] A. A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher. "Spamrank: Fully Automatic Link Spam Detection." Paper presented at First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), Chiba, Japan, May 10–14, 2005. Available at http://airweb.cse.lehigh.edu/2005/benczur.pdf.

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

[Brin and Page 98] S. Brin and L. Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine." *Computer Networks and ISDN Systems* 30:1–7 (1998), 107–117.

[Chen et al. 04] Y. Chen, Q. Gan, and T. Suel. "Local Methods for Estimating PageRank Values." In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, pp. 381–389. New York: ACM Press, 2004.

[Fetterly et al. 04] D. Fetterly, M. Manasse, and M. Najork. "Spam, Damn Spam, and Statistics: Using Statistical Analysis to Locate Spam Web Pages." In *Proceedings of the 7th International Workshop on the Web and Databases*, pp. 1–6. New York: ACM Press, 2004.

[Fogaras and Racz 04] D. Fogaras and B. Racz. "Towards Scaling Fully Personalized PageRank." In *Algorithms and Models for the Web-Graph: Third International Workshop, WAW 2004, Rome, Italy, October 16, 2004, Proceeedings*, Lecture Notes in Computer Science 3243, pp. 105–117. Berlin: Springer, 2004.

[Gyöngyi et al. 04] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. "Combating Web Spam with Trustrank." In *Proceedings of the Thirtieth International Conference on Very Large Data Bases*, pp. 576–587. San Francisco: Morgan Kaufmann, 2004.

[Gyöngyi et al. 06a] Z. Gyöngyi, P. Berkhin, H. Garcia-Molina, and J. Pedersen. "Link Spam Detection Based on Mass Estimation." In *Proceedings of the 32nd International Conference on Very Large Databases*, pp. 439–450. New York, ACM Press, 2006.

[Gyöngyi et al. 06b] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. "Web Content Categorization Using Link Information." Technical report, Stanford University, 2006.

[Haveliwala 03] T. H. Haveliwala. "Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search." *IEEE Trans. Knowl. Data Eng.* 15:4 (2003), 784–796.

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

[Krishnan and Raj 06] V. Krishnan and R. Raj. "Web Spam Detection with Anti-trust Rank." Paper presented at Second International Worshop on Adversarial Information Retreival on the Web (AIRweb), Seattle, WA, August 10, 2006. Available at http://airweb.cse.lehigh.edu/2006/krishnan.pdf.

[Mishne et al. 05] G. Mishne, D. Carmel, and R. Lempel. "Blocking Blog Spam with Language Model Disagreement." Paper presented at First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), Chiba, Japan, May 10–14, 2005. Available at http://airweb.cse.lehigh.edu/2005/mishne.pdf.

[Naor and Stockmeyer 95] M. Naor and L. Stockmeyer. "What Can Be Computed Locally?" *SIAM J. Comput.* 24:6 (2005), 1259–1277.

[Ntoulas et al. 06] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. "Detecting Spam Web Pages through Content Analysis." In *Proceedings of the 15th International Conference on World Wide Web*, pp. 83–92. New York: ACM Press, 2006.

[Sarlós et al. 06] T. Sarlós, A. A. Benczúr, K. Csalogány, D. Fogaras, and B. Rácz. "To Randomize or Not to Randomize: Space Optimal Summaries for Hyperlink Analysis." In *Proceedings of the 15th International Conference on World Wide Web*, pp. 297–306. New York: ACM Press, 2006.

[Spielman and Teng 04] D. A. Spielman and S.-H. Teng. "Nearly-Linear Time Algorithms for Graph Partitioning, Graph Sparsification, and Solving Linear Systems." In *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing*, pp. 81–90. New York: ACM Press, 2004.

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

Christian Borgs, Microsoft Research New England, One Memorial Drive, Cambridge, MA 02142 (borgs@microsoft.com)

Jennifer Chayes, Microsoft Research New England, One Memorial Drive, Cambridge, MA 02142 (jchayes@microsoft.com)

John Hopcroft, Computer Science Department, Cornell University, 5144 Upson Hall, Ithaca, NY 14853 (jeh@cs.cornell.edu)

Vahab Mirrokni, Google Inc., Research Group, NYC, 76 9th Ave, Room 4E310C, New York, NY, 10011 (mirrokni@google.com)

Shang-Hua Teng, Department of Computer Science, Boston University, 111 Cummington Street, Boston, MA (steng@cs.bu.edu)

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