Distributed Randomized PageRank Algorithm Based on Stochastic Approximation

Jinlong Lei and Han-Fu Chen, Fellow, IEEE

Abstract—A distributed randomized PageRank algorithm based on stochastic approximation (SA) is proposed to estimate the importance scores of web pages. Compared with the existing methods, the algorithm given here has wider applications in the sense that it can deal with a larger class of randomizations. The strong consistency of the estimates is proved, and the robustness of the PageRank value is analyzed as well. Numerical examples are given to verify the obtained theoretic results.

Index Terms—Distributed randomized PageRank algorithm, stochastic approximation, strong consistency.

I. INTRODUCTION

Since the quantity of information on the web has been rapidly growing up in recent years, the search engine becomes an indispensable tool for retrieving information. The PageRank algorithm [1] is the key technique of Google, which is one of the most successful search engines.

The mathematical description of the PageRank problem is as follows [2], [3]: Consider a web with $n$ pages, each of which is indexed by an integer $k$, $1 \leq k \leq n$. Without loss of generality, we assume $n \geq 2$. The web structure is described by a directed graph $G = (\mathcal{V}, \mathcal{E})$ with the node set $\mathcal{V} = \{1, 2, \ldots, n\}$ and the edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. If the web page $i$ has a hyperlink pointing to the web page $j$, then $(i, j) \in \mathcal{E}$. We call $i$ the inlink of $j$ and $j$ the outlink of $i$. Denote by $N_i$ the set of all inlinks of $i$ and by $n_j$ the number of outlinks of $j$. Define the link matrix of the directed graph $G$ as follows:

$$A = [a_{ij}] \in \mathbb{R}^{n \times n}, \quad a_{ij} = \begin{cases} \frac{1}{n_j} & \text{if } j \in N_i \\ 0 & \text{otherwise.} \end{cases}$$ (1)

It is clear that the sum of all $a_{ij}$ equals either 0 or 1. Denote by $x_i^* \in [0, 1]$ the importance score of the web page $i$ and assume $\sum_{i=1}^n x_i^* = 1$. If $x_i^* > x_j^*$, we say that the web page $i$ is more important than the web page $j$. The importance score of each web page is determined by the number of inlinks and their corresponding importance scores, i.e.,

$$x_i^* = \sum_{j \in N_i} \frac{x_j^*}{n_j}. \quad (2)$$

Then $x^* = (x_1^*, \ldots, x_n^*)^T$ satisfies the following equation:

$$x^* = Ax^*, \quad x^* \in S_p^n$$ (3)

where $S_p^n = \{x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n : x_i \geq 0, i = 1, \ldots, n, \sum_{i=1}^n x_i = 1\}$.

The importance score vector $x^*$ is unique when $G$ is strongly connected. This is not always true in practice, e.g., there may exist dangling nodes which have no outlinks. To ensure the uniqueness of $x^*$, the matrix $A$ in (3) is modified in [2] in such a way that all web pages are artificially set to be the outlinks of all dangling nodes. This means that all $A_0$ columns of $A$ are replaced by $1_n/n$ to make $A$ to be a column-stochastic matrix $H$, where and hereafter $A_0$ and $A_n$ denote the $n$-dimensional vectors with all entries being 0 and 1, respectively, and where a column-stochastic matrix $X = [x_{ij}] \in \mathbb{R}^{n \times n}$ is defined as $x_{ij} \geq 0$, $i, j = 1, \ldots, n$ and $\sum_{i=1}^n x_{ij} = 1$ for $j = 1, \ldots, n$. In order to guarantee that the algebraic multiplicity [4] of the eigenvalue 1 of the matrix $H$ is 1, a further modification is adopted in [2]. Let $\alpha \in (0, 1]$ be the damping factor, and define the modified link matrix $M \in \mathbb{R}^{n \times n}$ as

$$M = (1 - \alpha)H + \alpha \frac{1}{n} 1_n 1_n^T. \quad (4)$$

In [1], $\alpha$ is chosen as $\alpha = 0.15$. The issue on how to choose the damping factor $\alpha$ is discussed in [3], [14], [15]. Notice that the matrix $H$ has the following properties [2], [3]: $M$ is a positive column-stochastic matrix; the algebraic multiplicity of the eigenvalue 1 of the matrix $M$ is 1; the eigenvector corresponding to the eigenvalue 1 is positive or negative, where we say a matrix or a vector is positive (negative) if all its entries are positive (negative). It is noticed that the matrix $H$ considered here and the matrix $A$ discussed in [5]–[7] both are column-stochastic and they have similar properties. However, $H$ is the link matrix of the web after introducing artificial outlinks to all dangling nodes, while $A$ is directly assumed to be column-stochastic in [5]–[7]. Both the matrix $M$ considered here and $M$ discussed in [5]–[7] are the modified link matrices and they share the same properties.

By replacing $A$ and $x^*$ with $M$ and $x^*_\alpha$, respectively, in (3), the PageRank problem is defined as finding the vector $x^*_\alpha = (x^*_1\alpha, \ldots, x^*_n\alpha)^T$ that satisfies the following equation:

$$x^*_\alpha = Mx^*_\alpha, \quad x^*_\alpha \in S_p^n. \quad (5)$$

The PageRank problem can be considered as computing the eigenvector corresponding to the eigenvalue 1 of the matrix $M$. It can be solved by using the power method [2], [3], which suggests recursively computing

$$x(k + 1) = Mx(k) = (1 - \alpha)Hx(k) + \alpha \frac{1}{n} 1_n$$

where $x(1) \in S_p^n$. It is shown in [2], [3] that

$$x(k) \xrightarrow[k \to \infty]{} x^*_\alpha.$$

Since, the dimension of the web is extremely high, the computation burden of each iteration is very heavy. To reduce the computation burden, a random-walk-based algorithm is proposed in [9], and a randomized algorithm based on the mirror descent method of convex stochastic optimization is given in [8]. Moreover, the distributed algorithm, widely used in multi-agent systems (see [10]–[12] among others), is introduced to the PageRank algorithm in [5], where each web page is considered as an agent that can exchange information with its inlinks and outlinks to locally update its importance score.
In particular, several distributed randomized algorithms are proposed in [5], where convergence of the algorithm is established in the mean-square sense, while the strong consistency of the algorithm is further proved in [13]. The distributed PageRank algorithms under the unreliable communication channel are considered in [6], [7].

In this note, we propose a distributed algorithm based on SA to estimate the importance score and show the condition under which the distributed schemes can lead to the strongly consistent estimates. In contrast to the biased estimates given in [7], when the unreliable communication channel is modeled by a Markov chain, the almost sure convergence is established here. It is also shown that our algorithm covers those given in [5]–[7] as special cases. Furthermore, the robustness of the PageRank value with respect to changes of the web structure and the influence of the damping factor are investigated as well.

The rest of the note is organized as follows: The distributed randomized scheme is introduced in Section II. A distributed algorithm based on SA is proposed, and the strong consistency of the estimates is proved in Section III. The robustness of the PageRank value is analyzed in Section IV. Three numerical examples, two of which are with real link structure downloaded from the web, are given in Section V. Some concluding remarks are given in Section VI.

II. RANDOMIZATION

The basic idea of the distributed randomized scheme as is follows [5]: If the web page is triggered at time , then a web page sends its importance score to all its outlinks or requires the importance scores from all its inlinks, and all the page web pages renew their importance scores according to the newest information they have received.

Randomization consists in designing i) the random process (k) and ii) the distributed randomized matrices (k) and (k). The process (k) is designed to determine which web pages are triggered at time k, and (k) is designed on the basis of (k), while (k), obtained from (4) with replaced by (k) for some appropriately chosen , will be used in the distributed algorithm.

We require the designed randomized matrices (k) satisfy the following condition:

A1: \( \{M(k)\}_{k \geq 1} \) is a sequence of column-stochastic matrices with \( (1/k) \sum_{i=1}^{1} M(i, k) \rightarrow M \) and \( x_{n}^{*} = Mx_{n}^{*} \), where \( x_{n}^{*} \) is defined by (5).

Now we show that the distributed randomized schemes proposed in [5], [7] satisfy A1.

S1: The distributed randomized scheme \( \theta(k) = (\theta_{1}(k), \ldots, \theta_{n}(k)) \) designed in [5] is such that for all \( i \in V \), \( \{\theta_{i}(k)\}_{k \geq 1} \) is a sequence of independent identically distributed (i.i.d) random variables with the probability distribution

\[
\mathbb{P}\{\theta_{i}(k) = 1\} = \beta
\]
\[
\mathbb{P}\{\theta_{i}(k) = 0\} = 1 - \beta, \quad \beta \in (0, 1).
\]

If \( \theta_{i}(k) = 1 \), then the web page sends its importance score to all its outlinks and requires the importance scores from all its inlinks at time k. Otherwise, it transmits no information. \( [H(\theta(k))]_{ij} \) is defined as

\[
[H(\theta(k))]_{ij} = \begin{cases} [H]_{ij}, & \text{if } \theta_{i}(k) = 1 \text{ or } \theta_{j}(k) = 1 \\ 1 - \sum_{p: \theta_{p}(k) = 1} [H]_{pj}, & \text{if } \theta_{i}(k) = 0 \text{ and } i = j \\ 0, & \text{otherwise} \end{cases}
\]

where and hereafter \( [H]_{ij} \) always denotes the \((i,j)\)th entry of \( H \).

Let \( \alpha = (1 - \beta)^{2} / (1 - \alpha(1 - \beta)^{2}) \) and set \( M_{\theta}(k) = (1 - \alpha)H_{\theta}(k) + (\alpha/n)1_{n}1_{n}^{T} \). Then, \( \{M_{\theta}(k)\}_{k \geq 1} \) is a sequence of i.i.d random matrices. Denote by \( E[X] \) the expectation of the random variable X. It is shown in [5] that \( x_{n}^{*} = Mx_{n}^{*} \), where \( M = E[M(\theta(k))] \).

By the Khintchine-Kolmogorov convergence theorem [17], we have

\[
\sum_{k=1}^{\infty} \frac{1}{k} \left( M(\theta(k)) - \bar{M} \right) < \infty \quad \text{a.s.}
\]

which by the Kronecker lemma [17] leads to

\[
\frac{1}{k} \sum_{i=1}^{k} (M(\theta(i)) - \bar{M}) \xrightarrow{k \to \infty} 0.
\]

Therefore, A1 holds.

S2: The distributed randomized PageRank algorithm under the unreliable communication channel modeled by a Markov chain is discussed in [7]. A random variable \( \eta(k) \in \{0, 1\} \) with the probability distribution

\[
\mathbb{P}\{\eta(k) = 1\} = \beta
\]
\[
\mathbb{P}\{\eta(k) = 0\} = 1 - \beta, \quad \beta \in (0, 1)
\]
determines whether or not the web page sends its importance score to all its outlinks at time k. If \( \eta(k) = 1 \), then i sends its importance score to all its outlinks at time k. Otherwise, it does not. And for \( \forall i \in V \), \( \{\eta(k)\}_{k \geq 1} \) is a sequence of i.i.d random variables.

Since the communication channel is unreliable, some outlinks of the web page can receive the information sent by i. For simplicity, we assume that all the outlinks of i can or cannot receive the information sent by i at time k. This is described by a random variable \( \mu_{i}(k) \in \{0, 1\} \) as follows: If \( \mu_{i}(k) = 1 \), then all outlinks of i can receive the information sent by i. Otherwise, they cannot. We model \( \mu_{i}(k) \) as a Markov chain valued on \{0, 1\} with transition matrix \( R \in \mathbb{R}^{2 \times 2} \), and assume that \( R \) has a positive stationary distribution \( \gamma = (\gamma_{1}, \gamma_{2}) \).

Define by \( r_{mn} \) the \((m,n)\)th entry of \( R \). Then for \( \forall i \in V \), we have

\[
\mathbb{P}\{\mu_{i}(k+1) = j2| \mu_{i}(k) = j1\} = r_{j1+j2}, \quad j1, j2 \in \{0, 1\}.
\]

Denote by \( \theta(k) := (\eta_{1}(k), \mu_{1}(k), \ldots, \eta_{n}(k), \mu_{n}(k)) \in \prod_{i \in V} \{0, 1\} \times \{0, 1\} \) the combined process. It is clear that \( \theta(k) \) can take \( n \) different values. It is shown in [7] that \( \theta(k) \) is a Markov chain on the simplified value space \( I = \{1, 2, \ldots, n\} \) with transition matrix \( P \in \mathbb{R}^{n \times n} \), which has a positive stationary distribution \( \pi = (\pi_{1}, \ldots, \pi_{n}) \).

Define \( [H(\theta(k))]_{ij} \) as follows:

\[
[H(\theta(k))]_{ij} = \begin{cases} [H]_{ij}, & \text{if } \eta_{i}(k) = 1 \text{ and } \mu_{j}(k) = 1 \\ 1, & \text{if } I[\eta_{i}(k)-1] \cap I[\mu_{j}(k)-1] = \emptyset \\ 0, & \text{otherwise} \end{cases}
\]

where \( I[x] \) is the indicator function of A. Let \( \bar{\alpha} = \alpha \beta / (1 - \alpha(1 - \beta)^{2}) \) and set \( M_{\theta}(k) = (1 - \bar{\alpha})H_{\theta}(k) + (\bar{\alpha}/n)1_{n}1_{n}^{T} \). It is proved in [7] that \( \bar{M} = \sum_{i=1}^{n} \pi_{i}M_{i} \) and \( x_{n}^{*} = Mx_{n}^{*} \), where \( \bar{M} = \lim_{k \to \infty} E[M(\theta(k))] \).

Since \( P \) is irreducible, the ergodic property of Markov chains (Theorem 1.10.2 of [18]), we have

\[
\mathbb{P}\left( \frac{1}{k} \sum_{i=1}^{k} M_{\theta}(i) \xrightarrow{k \to \infty} \bar{M} \right) = 1
\]

i.e., \( (1/k) \sum_{i=1}^{k} M_{\theta}(i, k \to \infty) \bar{M} \), and A1 holds. The randomized matrices \( \{M(\theta(k))\}_{k \geq 1} \) designed in [7] have the same properties as the simplified randomized matrices \( \{M_{\theta}(k)\}_{k \geq 1} \) designed here, and hence A1 also holds for the randomized matrices \( \{M(\theta(k))\}_{k \geq 1} \) designed in [7].

III. DISTRIBUTED ALGORITHM BASED ON SA

Denote by \( x(k) \) the estimate of the importance score vector at step k. Based on the randomized matrix \( M_{\theta}(k) \), we propose the following
distributed PageRank algorithm based on SA

\[ x(k+1) = x(k) + \frac{1}{k} (M_{\theta(k)} x(k) - x(k)). \]  

(6)

Let us first introduce a result from SA and formulate it as a lemma. Let the root set of a function \( f() \) be denoted by \( J \triangleq \{x \in \mathbb{R}^n : f(x) = 0\} \), and let \( x_k \) and \( y_{k+1} \) denote the estimate of the root at time \( k \) and the error-corrupted observation at time \( k+1 \), respectively. The SA algorithm is recursively defined by

\[ x_{k+1} = x_k + a_k y_{k+1}, \quad y_{k+1} = f(x_k) + \epsilon_{k+1} \]  

(7)

where \( a_k \) is the step size. Assume that \( \{x_k\} \) given by algorithm (7) evolves in the subspace \( S \subset \mathbb{R}^n \). We list the conditions to be used:

**C1** \( a_k > 0, \ a_k \to 0, \) and \( \sum_{k=0}^{\infty} a_k = \infty. \)

**C2** There exists a continuously differentiable function \( v() : \mathbb{R}^n \to \mathbb{R} \) such that

\[ \sup_{x \leq \alpha(x) : x \leq \Delta} f^T v(x) < 0 \]  

(8)

for any \( \Delta > \delta > 0 \), where \( d(x, J \cap S) = \inf_{y \in J \cap S} \|x - y\| : y \in J \cap S \}, v(\cdot) \) denotes the gradient of \( v() \). Further, \( v(J \cap S) \triangleq \{v(x) : x \in J \cap S \} \) is nowhere dense.

**C3** The observation noise \( \{\epsilon_i\} \) satisfies the following condition:

\[ \lim \sup_{T \to 0} \sum_{k=0}^{\infty} a_k \epsilon_{i+1} = 0, \quad \forall T_k \in [0, T]. \]  

(9)

where \( m(k, T) = \max\{m : \sum_{i=0}^{m} a_i \leq T\}, T > 0. \)

**C4** \( f() \) is measurable and locally bounded.

**Lemma 3.1:** ([19] Theorem 2.2.3 and Remark 2.2.6): If \( x_k \) generated by (7) evolves in the subspace \( S \) for any initial value \( x_1 \in S \) and conditions **C1**, **C2**, and **C4** hold, then \( d(x_k, J \cap S) \to 0 \) for the sample path \( \omega \) for which **C3** holds.

Remark 3.2: The SA algorithm (7) is the well-known Robbins-Monro algorithm [16] for searching roots of the function \( f \). The PageRank problem can also be thought of as seeking roots of the function \( Mx - x \) in the subspace \( S^\alpha \). Thus we can use the SA type algorithm to solve the PageRank problem. In the distributed randomized algorithm (6), \( M_{\theta(k)} x(k) - x(k) \) can be thought as the observation of the function \( Mx - x \), which has the same root as the function \( Mx - x \) in the subspace \( S^\alpha \). The corresponding observation noise at \( x(k) \) is \( \epsilon_{k+1} = (M_{\theta(k)} - M)x(k) \).

**Theorem 3.3:** If the distributed randomized matrices \( \{M_{\theta(k)}\}_{k \geq 1} \) satisfy **A1**, then for any initial value \( x(1) \in S^\alpha \), \( x(k) \) generated by (6) converges to \( x^\alpha \) with probability 1, i.e.,

\[ x(k) \xrightarrow{a.s.}{k \to \infty} x^\alpha. \]  

(10)

Proof: Since \( x(1) \in S^\alpha \) and \( M_{\theta(k)} \) is column-stochastic for any \( k \), from

\[ x(k+1) = x(k) + \frac{1}{k} \big( M_{\theta(k)} x(k) - x(k) \big) \]

\[ = \left( 1 - \frac{1}{k} \right) x(k) + \frac{1}{k} M_{\theta(k)} x(k) \]

we conclude that \( x(k) \in S^\alpha \). Set \( a_k = 1/k \), and rewrite (6) as

\[ x(k+1) = x(k) + a_k (Mx(k) - x(k)) + a_k \big( M_{\theta(k)} - M \big) x(k). \]  

(11)

We notice that \( \epsilon_{k+1}, f(), \) and \( J \) for (7) correspond to \( \epsilon_{k+1} = (M_{\theta(k)} - M)x(k), f(x) = Mx - x, \) and \( J \triangleq \{x \in \mathbb{R}^n : Mx - x = 0\} \) for (11), respectively. By **A1**, we see \( J \triangleq \{c \alpha(x) : c \in \mathbb{R}\} \). By taking \( v(x) = x^T (x - Mx)/2 \), we have \( v_x(x) = x - Mx, \) and

\[ f(x)^T v_x(x) = (Mx - x)^T (x - Mx) = -(Mx - x)^T (Mx - x) \leq 0 \]

where the equality holds if and only if \( Mx = x \). Hence, we have

\[ f(x)^T \frac{\partial f}{\partial x} < 0, \quad \text{for all } x \in S^\alpha \text{ but } x \notin J. \]

Thus, conditions **C1**, **C2**, and **C4** hold. We now verify

\[ \lim \sup_{T \to 0} \sum_{k=0}^{\infty} a_k \big( M_{\theta(k)} - M \big) x(k) = 0, \quad \forall T_k \in [0, T]. \]

(12)

Setting \( \zeta(k) = \sum_{i=1}^{k+1} (M_{\theta(i)} - M), \zeta(0) = 0, \) by **A1** we have \( (\zeta(k)/k) \xrightarrow{k \to \infty} 0, \) and hence

\[ \sum_{i=0}^{m(k, T)} a_i \big( M_{\theta(i)} - M \big) \]

\[ = \sum_{i=0}^{m(k, T) - 1} \zeta(i - 1) + \sum_{i=m(k, T)}^{m(k, T) - 1} \zeta(i) \]

\[ \leq \zeta(m(k, T))/k + \sum_{i=0}^{m(k, T) - 1} \zeta(i) \]

\[ \leq \zeta(m(k, T))/k + \sum_{i=0}^{m(k, T) - 1} \frac{1}{i + 1} \]

\[ \xrightarrow{k \to \infty} 0. \]

(13)

Noticing \( \|x(k)\| \leq 1, \) by (13), we derive

\[ \lim \sup_{T \to 0} \sum_{k=0}^{\infty} a_k \big( M_{\theta(k)} - M \big) x(k) \]

\[ \leq \lim \sup_{T \to 0} \sum_{k=0}^{\infty} a_k \big( M_{\theta(k)} - M \big) \|x(k)\| \]

\[ \leq \lim \sup_{T \to 0} \sum_{k=0}^{\infty} a_k \|M_{\theta(k)} - M\| \leq 0 \]

\[ \forall T_k \in [0, T]. \]

(14)

Since \( M_{\theta(k)} \) is column-stochastic and \( x(k) \in S^\alpha \), there exists a constant \( c \) such that

\[ \|x(i) - x(k)\| \leq cT, \quad \forall i : k \leq i \leq m(k, T). \]

(15)

Since both \( M_{\theta(k)} \) and \( M \) are column-stochastic, there exists a constant \( c_0 \) such that \( \|(M_{\theta(k)} - M)\| \leq c_0, \forall i \geq 1. \) From (15), it follows that:

\[ \lim \sup_{T \to 0} \sum_{k=0}^{\infty} a_k \big( M_{\theta(k)} - M \big) \|x(i) - x(k)\| \]

\[ \leq \lim \sup_{T \to 0} \sum_{k=0}^{\infty} a_k \|M_{\theta(k)} - M\| \]

\[ \leq c_0 \lim \sup_{T \to 0} \sum_{k=0}^{\infty} a_k \|M_{\theta(k)} - M\| \]

\[ \leq c_0 \lim \sup_{T \to 0} T_k = 0. \quad \forall T_k \in [0, T]. \]

(16)

which incorporating with (14) verifies **C3**. By Lemma 3.1, we have

\[ d(x(k), J \cap S^\alpha) \xrightarrow{k \to \infty} 0. \]

By definition of \( J \) and \( S^\alpha \), it is clear that \( J \cap S^\alpha = x^\alpha. \) Therefore, we derive

\[ x(k) \xrightarrow{k \to \infty} x^\alpha. \]
Remark 3.4: Based on the randomized matrices \( \{ M_{\theta(k)} \} \), the following distributed update scheme
\[
y(k+1) = M_{\theta(k)} y(k)
\]  
(17)
is proposed in [5]–[7], where the time average of \( y(1), \ldots, y(k) \) serves as the estimate of the importance score vector, denoted by \( \bar{x}(k) \), i.e.,
\[
\bar{x}(k+1) = \frac{1}{k+1} \sum_{i=1}^{k+1} y(i) = \frac{1}{k+1} \left( \sum_{i=1}^{k} M_{\theta(i)} y(i) + y(1) \right).
\]  
(18)
In contrast to this, the distributed algorithm (6) can be rewritten as
\[
x(k+1) = \frac{1}{k} \sum_{i=1}^{k} M_{\theta(i)} x(i).
\]  
(19)
Notice that both \( \bar{x}(k) \) in (18), and \( x(k) \) in (19) are steady, while \( y(k) \) in (17) is oscillating as shown in [5]. This determines the different properties of the algorithms (18) and (19).

For a sequence of i.i.d randomized matrices \( \{ M_{\theta(k)} \} \), both the algorithms (17), (18), and (19) converge to the true PageRank value almost surely as shown by [13] and by Theorem 3.3, respectively. Under the unreliable communication channel modeled by a Markov chain, the algorithm (17), (18) gives a biased estimate as shown in [7], while (19) still gives the strongly consistent estimate. This implies that the algorithm (6) may have wider applications in comparison with algorithm (17), (18).

In summary, compared with other methods, the SA algorithm for analyzing the PageRank problem is characterized by two aspects. First, the estimates of the PageRank values generated by the SA algorithm are strongly consistent. Second, the SA algorithm can deal with a wider class of randomizations satisfying the general condition A1.

Remark 3.5: Although the SA algorithm is not a fast algorithm, the SA-based distributed randomized algorithms are adopted since the computation burden of each iteration is relatively low. While the computation burden of each iteration of the power method is high, the computation burden of each iteration at the cost of slowing down the convergence rate.

IV. THE ROBUSTNESS OF THE PAGE RANK VALUE

In this section, we study the robustness of the PageRank value with respect to changes of the web structure as well as the influence of damping factor.

Denote by \( \tilde{H} \) the perturbed matrix of \( H \) and by \( \tilde{x}^*_{\alpha} \) the importance score vector after perturbation. From (4) and (5) it follows that
\[
x^*_{\alpha} = (1 - \alpha) H x^*(\alpha) + \frac{\alpha}{n} 1_n
\]
and hence
\[
\tilde{x}^*_{\alpha} - x^*_{\alpha} = (1 - \alpha) \tilde{H} (\tilde{x}^*_{\alpha} - x^*_{\alpha}) - (1 - \alpha) (H - \tilde{H}) x^*_{\alpha}
\]  
(20)
and
\[
\tilde{x}^*_{\alpha} - x^*_{\alpha} = -(1 - \alpha) (I_n - (1 - \alpha) \tilde{H})^{-1} (H - \tilde{H}) x^*_{\alpha}.
\]  
(21)
Since \( \alpha (I_n - (1 - \alpha) \tilde{H})^{-1} \) is column-stochastic [3], we conclude that
\[
\| \tilde{x}^*_{\alpha} - x^*_{\alpha} \|_1 \leq \frac{1 - \alpha}{\alpha} \| (H - \tilde{H}) x^*_{\alpha} \|_1.
\]  
(22)
This gives a rough bound of the change of PageRank value caused by perturbation.

Fig. 1. Estimate errors \( \| e(k) \|_1 \) and \( \| e(k) \|_\infty \).

Denote by \( \tilde{V} \) the set of web pages which have some "bad" outlinks temporarily inaccessible. This may be caused by various reasons, e.g., some outlinks are removed or the sever of the web does not properly work. The proportion of "bad" links is denoted by \( \delta \in (0, 1) \), and the numbers of outlinks of the web page \( i \) before and after perturbation are denoted by \( n_i \) and \( \tilde{n}_i \), respectively. If \( i \in \tilde{V} \), then \( \tilde{n}_i \leq n_i \). Denote by \( H^{(i)} \) the ith column of \( H \). Thus, \( H^{(i)} \) differs from \( H^{(i)} \) only for \( i \in \tilde{V} \). For \( i \in \tilde{V} \), we obtain
\[
\| H^{(i)} - \tilde{H}^{(i)} \|_1 = \tilde{n}_i \left( \frac{1}{n_i} - \frac{1}{\tilde{n}_i} \right) + (n_i - \tilde{n}_i) \leq 2 \left( 1 - \frac{\tilde{n}_i}{n_i} \right) \leq 2 \delta.
\]

From (22) it follows that
\[
\| \tilde{x}^*_{\alpha} - x^*_{\alpha} \|_1 \leq \frac{1 - \alpha}{\alpha} \| H - \tilde{H} \|_1 \| \tilde{x}^*_{\alpha} \| \leq \frac{2 \delta (1 - \alpha)}{\alpha} \| \tilde{x}^*_{\alpha} \|.
\]  
(23)

From (23), it is seen that the upper bound of the change of the PageRank value is proportional to the proportion of "bad" links and to the sum of the PageRank values corresponding to the webpages with "bad" outlinks. This means that the changes of the authoritative webpages have significant influence on the changes of the PageRank value of the web. Moreover, the robustness of the PageRank value is also influenced by the damping factor. To be specific, a small \( \alpha \) makes \( x^*_{\alpha} \) sensitive to the changes of the web structure.

V. SIMULATIONS

In this section, we give three numerical examples to exemplify the theoretic analysis.

Example 1: We use MATLAB to run a web crawler starting from the url root www.amss.ac.cn and following the hyperlink of web pages to download 1000 web pages. Thus, a link matrix \( A \) of dimension 1000 \times 1000 is formed from the downloaded pages. Based on \( A \) we form \( M \) with \( \alpha = 0.15 \) in (4). In the distributed PageRank algorithm (6), the random variable \( \theta(k) \) and the distributed randomized link matrix \( M_{\theta(k)} \) are generated by applying the distributed randomized scheme S1 with \( \beta = 0.05 \). The estimate \( x(k) \) of the PageRank value is generated by (6) with the initial state \( x(1) \) satisfying \( x(1) \in S^o_0 \).

Let the estimation error be \( e(k) = x(k) - \tilde{x}^*_{\alpha} \). It is shown in Fig. 1 that the estimation errors \( \| e(k) \|_1 \) and \( \| e(k) \|_\infty \) asymptotically converge to zero as expected. In Fig. 2 the convergence of some components of \( x(k) \) is demonstrated, where the real lines represent the estimates of the PageRank values, and the dotted lines represent the real PageRank values.
Example 2: Let us consider a web with 6 pages, and let the link matrix $A$ be given by

$$\begin{pmatrix}
0 & 0.5 & 0 & 0 & 0.5 \\
0.5 & 0 & 0.5 & 0 & 0 \\
0.5 & 0 & 0 & 0.5 & 0 \\
0 & 0 & 0 & 0 & 0.5 \\
0 & 0 & 0 & 0 & 0.5
\end{pmatrix}.$$  

Choose $\alpha = 0.15$ in (4). In the distributed PageRank algorithm (6), the random variable $\theta(k)$ and the distributed randomized link matrix $M_{\theta(k)}$ are generated by applying the distributed randomized scheme S2 with $\beta = 3/16$ and with the transition matrix

$$R = \begin{pmatrix}
0.1 & 0.9 \\
0.3 & 0.7
\end{pmatrix}. $$

The estimate $x(k)$ of the PageRank value is generated by (6) with the initial state $x(1)$ satisfying $x(1) \in S_6$. Let the estimation error be $e(k) = x(k) - x_*$. In Fig. 3 it is shown that both $\|e(k)\|_1$ and $\|e(k)\|_\infty$ asymptotically converge to zero as expected, and in Fig. 4 the convergence of each component of $e(k)$ is demonstrated.

Example 3: We discuss the robustness of the PageRank value by using the same link matrix as in the example 1. Let $\alpha = 0.15$ and let $V$ contain 100 web pages. We change the proportion of “bad” links, and observe its influence on the change of the PageRank value caused by perturbation. The horizontal coordinate of Fig. 5 is $\delta$ and the vertical coordinate is the change of the PageRank value $e = \|x_* - \tilde{x}_*\|_1$. 
From the figure it is seen that $e$ increases as $\delta$ increases, which is consistent with the explanation given in Section IV.

Let $\alpha = 0.15$ and $\delta = 0.3$. We divide the web pages into 100 groups according to their importance scores, where the $k$th group contains 10 web pages with importance scores ranking from the $(10k-9)$th to the $10k$th. We observe how do the groups of webpages with “bad” outlinks influence the change of the PageRank value. The horizontal coordinate $k$ of Fig. 6 means that the web pages of the $k$th experiment and the vertical coordinate is the change of the PageRank value $e = \|x^n - \bar{x}^n\|_1$. It is seen from Fig. 6 that the authoritative webpages have significant influence on the changes of the PageRank value.

Let $\delta = 0.5$ and let $\hat{V}$ contain 100 web pages. We change the value of the damping factor, and observe how does $\alpha$ influence the change of the PageRank value. The horizontal coordinate of Fig. 7 is $\alpha$ and the vertical coordinate is the change of the PageRank value $e = \|x^n - \bar{x}^n\|_1$. In Fig. 7 it is shown that $e$ decreases as $\alpha$ increases. Therefore, the increase of $\alpha$ enhances the robustness of $x^n_*$ with respect to the changes of the web structure.

Fix $\hat{V}$ and let $e = \|x^n_* - \bar{x}^n_*\|_1$. We change the value of the damping factor and the proportion of “bad” links simultaneously, and observe how do these two factors influence the change of the PageRank value. Fig. 8 shows that the change of the Pagerank value increases with $\delta$ increasing and decreases with $\alpha$ increasing.

Fig. 7. Influence of $\alpha$.

Fig. 8. Influence of $\delta$ and $\alpha$.

VI. CONCLUSIONS

In this note, a distributed randomized PageRank algorithm based on SA is proposed, and its strong consistency is proved. The robustness of the PageRank value is analyzed as well. Numerical examples are given to verify the theoretic analysis.

It is of interest to discuss the convergence rate of the algorithm and to consider the dynamic rather than the static web structure since the web is self-organized and the web pages are updating fast. It is also of interest to consider the web aggregation since it may save the computation time.

ACKNOWLEDGMENT

The authors would like to thank Dr. W. Zhao for his helpful discussions.

REFERENCES