ISSN 2220-5438

Reprint from

Moscow Journal of Combinatorics and Number Theory

Volume 4 • Issue 4 2014

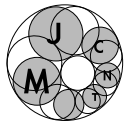
URSS

МФТИ

Яндекс

Moscow Journal of Combinatorics and Number Theory Volume 4 • Issue 4 2014 URSS

Moscow Journal
of Combinatorics
and Number Theory
2014, vol. 4, iss. 4,
pp. 5–20, [pp. 387–402]



Personalized PageRank with Node-dependent Restart

Konstantin Avrachenkov (Sophia Antipolis), Remco van der Hofstad (Eindhoven), Marina Sokol (Sophia Antipolis)

Abstract: Personalized PageRank is an algorithm to classify the importance of web pages on a user-dependent basis. We introduce two generalizations of Personalized PageRank with node-dependent restart. The first generalization, Occupation-Time Personalized PageRank, is based on the proportion of visits to nodes before the restart, whereas the second generalization, Location-of-Restart Personalized PageRank, is based on the proportion of time a node is visited just before the restart. In the original case of constant restart probability, the two measures coincide. We discuss interesting particular cases of restart probabilities and restart distributions. We show that both generalizations of Personalized PageRank have an elegant expression connecting the so-called direct and reverse Personalized PageRanks that yield a symmetry property of these Personalized PageRanks.

Keywords: Personalized PageRank; Markov Processes with Restart; Renewal-reward Theorem

AMS Subject classification: 60J10; 65C40; 68P20

Received: 14.10.2014

1. Introduction and definitions

PageRank has become a standard algorithm to classify the importance of nodes in a network. Let us start by introducing some notation. Let $G = (V, E)$ be a finite graph, where V is the node set and $E \subseteq V \times V$ the collection of (directed) edges. Then, PageRank can be interpreted as the stationary distribution of a random walk

on G that restarts from a uniform location in V at each time with fixed probability $1-\alpha \in (0,1)$. Thus, in the Standard PageRank centrality measure [10], the random walk restarts after a geometrically distributed number of steps, and the restart takes place from a uniform location in the graph, and otherwise jumps to any one of the neighbours in the graph with equal probability. Personalized PageRank [17] is a modification of the Standard PageRank where the restart distribution is not uniform. Both the Standard and Personalized PageRank have many applications in data mining and machine learning (see e. g., [3, 4, 10, 13, 16, 17, 19, 20]).

In the (standard) Personalized PageRank, the random walker restarts with a given fixed probability $1-\alpha$ at every step, independently of the node the walker presently is at. We suggest a generalization where a random walker restarts with probability $1-\alpha_i$ when it is at node $i \in V$. When the random walker restarts, it chooses a node to restart at with probability distribution v^T. In many cases, we let the random walker restart at a fixed location, say $j \in V$. Then the Personalized PageRank is a vector whose ith coordinate measures the importance of node i to node j.

The above random walks $(X_t)_{t\geq 0}$ can be described by a finite-state Markov chain with the transition matrix

$$\widetilde{P} = AD^{-1}W + (I-A)\underline{1}v^T, \tag{1.1}$$

where W is the (possibly non-symmetric) adjacency matrix, D is the diagonal matrix with diagonal entries $d_i = D_{ii} = \sum_{j=1}^{n} W_{ij}$, and $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$ is the diagonal matrix of damping factors. The case of undirected graphs corresponds to the case when W is a symmetric matrix. In general, D_{ii} is the out-degree of node $i \in V$. Throughout the paper, we assume that the graph is strongly connected and $\alpha_i < 1$ for at least one node. This assumption implies that the matrix $[I - AD^{-1}W]$ is invertible.

We propose two generalizations of the Personalized PageRank with node-dependent restart:

Definition 1 (Occupation-Time Personalized PageRank, OT PPR).
The Occupation-Time Personalized PageRank *with restart vector v is the vector*

whose ith coordinate is given by

$$\pi_i(v) = \lim_{t\to\infty} \mathbb{P}(X_t = i). \tag{1.2}$$

By the fact that $(\pi_i(v))_{i\in V}$ is the stationary distribution of the Markov chain, we can interpret $\pi_i(v)$ as a long-run frequency of visits to node i, i. e.,

$$\pi_i(v) = \lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \mathbf{1}_{\{X_s=i\}}. \tag{1.3}$$

Our second generalization is based on the location where the random walker restarts.

Definition 2 (Location-of-Restart Personalized PageRank, LOR PPR). *The* Location-of-Restart Personalized PageRank *with restart vector v is the vector whose ith coordinate is given by*

$$\rho_i(v) = \lim_{t\to\infty} \mathbb{P}(X_t = i \text{ just before restart}) = \lim_{t\to\infty} \mathbb{P}(X_t = i \mid \text{restart at time } t+1). \tag{1.4}$$

We can interpret $\rho_i(v)$ as a long-run frequency of visits to node i which are immediately followed by a restart, i. e.,

$$\rho_i(v) = \lim_{t\to\infty} \frac{1}{N_{t+1}} \sum_{s=1}^{t} \mathbf{1}_{\{X_t=i, X_{t+1} \text{ restarts}\}}, \tag{1.5}$$

where N_t denotes the number of restarts up to time t. When the restarts occur with equal probability at every node, $N_t \sim \mathrm{Bin}(t, 1-\alpha)$, i. e., N_t has a binomial distribution with t trials and success probability $1-\alpha$. When the restart probabilities are unequal, the distribution of N_t is more involved. We note that

$$N_t/t \xrightarrow{a.s.} \sum_{i\in V} (1-\alpha_i)\pi_i(v), \tag{1.6}$$

where $\xrightarrow{a.s.}$ denotes convergence almost surely.

Both generalized Personalized PageRanks are probability distributions, i. e., their sum over $i \in V$ gives 1. When $v^T = e(j)$, where $e_i(j) = 1$ when $i = j$ and

$e_i(j) = 0$ when $i \neq j$, then both $\pi_i(v)$ and $\rho_i(v)$ can be interpreted as the relative importance of node i from the perspective of node j.

We see at least three applications of the generalized Personalized PageRank. The network sampling process introduced in [6] can be viewed as a particular case of PageRank with a node-dependent restart. We discuss this relation in more detail in Section 4. Secondly, the generalized Personalized PageRank can be applied as a proximity measure between nodes in semi-supervised machine learning [5, 16]. In this case, one may prefer to discount the effect of less informative nodes, e. g., nodes with very large degrees. Thirdly, the generalized Personalized PageRank can be applied for spam detection and control. It is known [11] that spam web pages are often designed to be highly ranked. By using the Location-of-Restart Personalized PageRank and penalizing the ranking of spam pages with small restart probability, one can push the spam pages from the top list produced by search engines.

Let us mention some other works generalizing the fixed probability of restart in PageRank. In [14] and [15] the authors consider the damping factor as a random variable distributed according to user behavior. These works generalize [9] where the random damping factor is chosen according to the uniform distribution. Also, there is a stream of works, starting from [8], that generalize the damping parameter to the damping function. In those works, the random walk restarts with probability as a function of the number of steps from the last restart.

In this paper, we investigate the two generalizations of Personalized PageRank. The paper is organised as follows. In Section 2, we investigate the Occupation-Time Personalized PageRank. In Section 3, we investigate the Location-of-Restart Personalized PageRank. In Section 4, we specify the results for some particular interesting cases. We close in Section 5 with a discussion of our results and suggestions for future research.

The conference version of this work is presented in [1].

2. Occupation-Time Personalized PageRank

The Occupation-Time Personalized PageRank can be calculated explicitly as follows:

Theorem 1 (Occupation-Time Personalized PageRank Formula).
The Occupation-Time Personalized PageRank $\pi(v)$ with restart vector v and node-

dependent restart equals

$$\pi(v) = \frac{1}{v^T[I - AP]^{-1}\underline{1}} v^T[I - AP]^{-1}, \tag{2.1}$$

with $P = D^{-1}W$ *the transition matrix of random walk on* G *without restarts.*

PROOF. By the defining equation for the stationary distribution of a Markov chain,

$$\pi(v)[AD^{-1}W + (I - A)\underline{1}v^T] = \pi(v), \tag{2.2}$$

so that

$$\pi(v)[I - A\ D^{-1}W] = \pi(v)(I - A)\underline{1}v^T, \tag{2.3}$$

and, since $\pi(v)\underline{1} = 1$,

$$\pi(v)[I - A\ D^{-1}W] = (1 - \pi(v)A\ \underline{1})v^T. \tag{2.4}$$

Since the matrix $AD^{-1}W$ is substochastic and hence $[I - AD^{-1}W]$ is invertible, we arrive at

$$\pi(v) = (1 - \pi(v)A\ \underline{1})v^T[I - A\ D^{-1}W]^{-1}. \tag{2.5}$$

Let us multiply the above equation from the right hand side by $A\underline{1}$ to obtain

$$\pi(v)A\underline{1} = (1 - \pi(v)A\ \underline{1})v^T[I - A\ D^{-1}W]^{-1}A\ \underline{1}. \tag{2.6}$$

This yields

$$\pi(v)A\underline{1} = \frac{v^T[I - AP]^{-1}A\underline{1}}{1 + v^T[I - AP]^{-1}A\underline{1}}, \tag{2.7}$$

so that

$$1 - \pi(v)A\ \underline{1} = \frac{1}{1 + v^T[I - AP]^{-1}A\underline{1}}. \tag{2.8}$$

Consequently, substituting (2.8) into (2.5), we arrive at

$$\pi(v) = \frac{1}{1 + v^T[I - AP]^{-1}A\underline{1}} v^T[I - AP]^{-1}. \tag{2.9}$$

Since $v^T\underline{1} = 1$, by the fact that v^T is a probability mass function, and since $A\underline{1} = AP\underline{1}$ because P is stochastic, we obtain

$$1 + v^T[I - AP]^{-1}A\underline{1} =$$
$$= 1 + v^T[I - AP]^{-1}AP\underline{1} = v^T[I + [I - AP]^{-1}AP]\underline{1} =$$
$$= v^T[I - AP]^{-1}[I - AP + AP]\underline{1} = v^T[I - AP]^{-1}\underline{1}, \quad (2.10)$$

from which the required equation (2.1) follows. □

By the renewal-reward theorem (see e. g., Theorem 2.2.1 in [22]), formula (2.1) admits the following probabilistic interpretation

$$\pi_i(v) = \frac{\mathbb{E}_v[\#\text{ visits to } i \text{ before restart}]}{\mathbb{E}_v[\#\text{ steps before restart}]}, \quad (2.11)$$

where $\mathbb{E}_v$ denotes expectation with respect to the Markov chain starting in distribution v.

Denote for brevity $\pi_i(j) = \pi_i(e_j^T)$, where e_j is the jth vector of the standard basis, so that $\pi_i(j)$ denotes the importance of node i from the perspective of node j. Similarly, $\pi_j(i)$ denotes the importance of node j from the perspective of i. We next provide a relation between these «direct» and «reverse» PageRanks in the case of *undirected* graphs.

THEOREM 2 (SYMMETRY FOR UNDIRECTED OCCUPATION-TIME PERSONALIZED PAGERANK). *When $W^T = W$ and $A > 0$, the following relation holds*

$$\frac{d_j}{\alpha_j K_j(A)}\pi_i(j) = \frac{d_i}{\alpha_i K_i(A)}\pi_j(i), \quad (2.12)$$

with $d_i = D_{ii}$ the degree of node i and

$$K_i(A) = \frac{1}{e_i^T[I - AP]^{-1}\underline{1}}. \quad (2.13)$$

PROOF. Note that the multiplier of (2.1) equals precisely $K_i(A)$. Thus, using a matrix geometric series expansion, we can rewrite equation (2.1) as

$$\pi_j(i) = K_i(A)e_i^T \sum_{k=0}^{\infty}(AD^{-1}W)^k e_j =$$
$$= K_i(A)e_i^T \sum_{k=0}^{\infty}(AD^{-1}W)^k D^{-1}A\ A^{-1}De_j =$$

$$
\begin{aligned}
&= K_i(A) e_i^T A\, D^{-1} \sum_{k=0}^{\infty} (WD^{-1}A)^k A^{-1} D e_j = \\
&= K_i(A) \frac{\alpha_i}{d_i} e_i^T \sum_{k=0}^{\infty} (WD^{-1}A)^k e_j \frac{d_j}{\alpha_j} = \\
&= \frac{K_i(A)}{K_j(A)} \frac{\alpha_i}{d_i} \frac{d_j}{\alpha_j} K_j(A) e_i^T [I - WD^{-1}A]^{-1} e_j = \\
&= \frac{K_i(A)}{K_j(A)} \frac{\alpha_i}{d_i} \frac{d_j}{\alpha_j} K_j(A) e_j^T [I - AD^{-1}W]^{-1} e_i, \qquad (2.14)
\end{aligned}
$$

which gives equation (2.12). □

We note that the term $(AD^{-1}W)^k$ can be interpreted as the contribution corresponding to all paths of length k, while $K_i(A)$ can be interpreted as the reciprocal of the expected time between two consecutive restarts if the restart distribution is concentrated on node i, i. e.,

$$K_i(A)^{-1} = \mathbb{E}_i[\#\text{ steps before restart}], \qquad (2.15)$$

see also (2.11). Thus, a probabilistic interpretation of (2.12) is that

$$\frac{d_j}{\alpha_j} \mathbb{E}_j[\#\text{ visits to } i \text{ before restart}] = \frac{d_i}{\alpha_i} \mathbb{E}_i[\#\text{ visits to } j \text{ before restart}]. \qquad (2.16)$$

Since

$$\mathbb{E}_i[\#\text{ visits to } j \text{ before restart}] = \sum_{k=1}^{\infty} \sum_{v_1,\ldots,v_k} \prod_{t=0}^{k-1} \frac{\alpha_{v_s}}{d_{v_s}}, \qquad (2.17)$$

where $v_0 = i$, we immediately see that the expression for

$$\mathbb{E}_j[\#\text{ visits to } i \text{ before restart}]$$

is identical to (2.17), except for the first factor of α_i/d_i, which is present in

$$\mathbb{E}_i[\#\text{ visits to } j \text{ before restart}],$$

but not in $\mathbb{E}_j[\#\text{ visits to } i \text{ before restart}]$, and the factor α_j/d_j, which is present in $\mathbb{E}_j[\#\text{ visits to } i \text{ before restart}]$, but not in $\mathbb{E}_i[\#\text{ visits to } j \text{ before restart}]$. This ex-

plains the factors d_i/α_i and d_j/α_j in (2.16) and gives an alternative probabilistic proof of Theorem 2.

3. Location-of-Restart Personalized PageRank

The Location-of-Restart Personalized PageRank can also be calculated explicitly:

THEOREM 3 (LOCATION-OF-RESTART PERSONALIZED PAGERANK FORMULA).
The Location-of-Restart Personalized PageRank $\rho(v)$ with restart vector v and node-dependent restart is equal to

$$\rho(v) = v^T[I - AP]^{-1}[I - A], \tag{3.1}$$

with $P = D^{-1}W$.

PROOF. This follows from the formula

$$\begin{aligned}\rho_i(v) &= \mathbb{E}_v[\#\text{ visits to } i \text{ before restart}]\mathbb{P}(\text{restart from } i)\\ &= \mathbb{E}_v[\#\text{ visits to } i \text{ before restart}](1-\alpha_i).\end{aligned} \tag{3.2}$$

Now we can use (2.17) and the analysis in the proof of Theorem 1 to complete the proof. □

We next state a formula connecting the Location-of-Restart Personalized PageRank with Occupation-Time Personalized PageRank:

THEOREM 4. *The Location-of-Restart Personalized PageRank can be expressed as*

$$\rho_i(v) = \frac{\pi_i(v)(1-\alpha_i)}{\sum_{j\in V}\pi_j(v)(1-\alpha_j)}. \tag{3.3}$$

PROOF. Since $\pi(v)$ is the stationary distribution of $\widetilde{P}$ (see (1.1)), it satisfies the equation

$$\pi(AP + [I - A]\underline{1}v^T) = \pi. \tag{3.4}$$

Rewriting this equation as

$$\pi[I - A]\underline{1}v^T = \pi[I - AP], \tag{3.5}$$

and postmultiplying by $[I - AP]^{-1}$, we obtain

$$\pi[I - A]\underline{1}v^T[I - AP]^{-1} = \pi \tag{3.6}$$

or

$$v^T[I - AP]^{-1} = \frac{\pi}{\sum\limits_{j\in V} \pi_j(v)(1 - \alpha_j)}. \tag{3.7}$$

Postmultiplying the above equation by $[I - A]$ and using (3.1) yields (3.3). Alternatively, by (3.2) and (2.11), we see that

$$\rho_i(v) = \frac{\pi_i(v)(1 - \alpha_i)}{\mathbb{E}_v[\#\text{ steps before restart}]}, \tag{3.8}$$

where the denominator serves to normalize so that $\sum\limits_{i\in V} \rho_i(v) = 1$. Therefore,

$$\mathbb{E}_v[\#\text{ steps before restart}] = \sum_{j\in V} \pi_j(v)(1 - \alpha_j) \tag{3.9}$$

and (3.3) follows. □

The Location-of-Restart Personalized PageRank admits an even more elegant relation between the «direct» and «reverse» PageRanks in the case of undirected graphs:

Theorem 5 (Symmetry for undirected Location-of-Restart Personalized PageRank). *When $W^T = W$ and $\alpha_i \in (0,1)$, the following relation holds*

$$\frac{1 - \alpha_j}{\alpha_j}\, d_j\, \rho_i(j) = \frac{1 - \alpha_i}{\alpha_i}\, d_i\, \rho_j(i). \tag{3.10}$$

Proof. This follows from a series of equivalent transformations (taking $v = e_i$ for $\rho_j(i)$)

$$\begin{aligned}
&\rho_j(i) = \\
&\quad = e_i^T[I - AP]^{-1}[I - A]e_j = e_i^T[I - AP]^{-1}e_j(1 - \alpha_j) = \\
&\quad = e_i^T[AD^{-1}(DA^{-1} - W)]^{-1}e_j(1 - \alpha_j) = e_i^T[(DA^{-1} - W)]^{-1}DA^{-1}e_j(1 - \alpha_j) =
\end{aligned}$$

$$= e_i^T [DA^{-1} - W]^{-1} e_j d_j \frac{1-\alpha_j}{\alpha_j} = e_i^T [(I - WD^{-1}A)DA^{-1}]^{-1} e_j d_j \frac{1-\alpha_j}{\alpha_j} =$$

$$= e_i^T AD^{-1} [I - WD^{-1}A]^{-1} e_j d_j \frac{1-\alpha_j}{\alpha_j} = \frac{\alpha_i}{d_i} e_i^T [I - WD^{-1}A]^{-1} e_j d_j \frac{1-\alpha_j}{\alpha_j} =$$

$$= \frac{\alpha_i}{d_i} \frac{\rho_i(j)}{1-\alpha_i} d_j \frac{1-\alpha_j}{\alpha_j}. \qquad \square \tag{3.11}$$

Alternatively, Theorem 5 follows directly from (3.2) and (2.16).

Interestingly, in (2.12), the whole graph topology has an effect on the relation between the «direct» and «reverse» Personalized PageRanks, whereas in the case of $\rho(v)$, see equation (3.10), only the local end-point information (i. e., α_i and d_i) have an effect on the relation between the «direct» and «reverse» PageRanks. We have no intuitive explanation of this distinction.

4. Interesting particular cases

In this section, we consider some interesting particular cases for the choice of restart probabilities and distributions.

4.1. Constant probability of restart

The case of constant restart probabilities (i. e., $\alpha_i = \alpha$ for every i) corresponds to the original or standard Personalized PageRank. We note that in this case the two generalizations coincide. For instance, we can recover a known formula [21] for the original Personalized PageRank with $A = \alpha I$ from equation (2.1). Specifically,

$$v^T [I - AP]^{-1} \underline{1} = v^T [I - \alpha P]^{-1} \underline{1} = v^T \sum_{k=0}^{\infty} \alpha^k P^k \underline{1} = \frac{1}{1-\alpha}, \tag{4.1}$$

and hence we retrieve the well-known formula

$$\pi(v) = (1-\alpha) v^T [I - \alpha P]^{-1}. \tag{4.2}$$

We also retrieve the following elegant result connecting «direct» and «reverse» original Personalized PageRanks on undirected graphs ($W^T = W$) obtained in [5]:

$$d_i \pi_j(i) = d_j \pi_i(j), \tag{4.3}$$

since in the original Personalized PageRank $\alpha_i \equiv \alpha$. Finally, we note that in the original Personalized PageRank, the expected time between restarts does not depend on the graph structure nor on the restart distribution and is given by

$$\mathbb{E}_v[\#\text{ steps before restart}] = \frac{1}{1-\alpha}, \tag{4.4}$$

which is just the mean of a geometrically distributed random variable with parameter $1-\alpha$.

4.2. Restart probabilities proportional to powers of degrees

Let us consider a particular case when the restart probabilities are proportional to powers of the degrees. Namely, let $\sigma \in \mathbb{R}$ and define

$$A = I - a\, D^{\sigma}, \tag{4.5}$$

with $ad_{\max}^{\sigma} < 1$. We first analyse $[I - AP]^{-1}$ with the help of a Laurent series expansion. Let $T(\varepsilon) = T_0 - \varepsilon T_1$ be a substochastic matrix for small values of ε and let T_0 be a stochastic matrix with associated stationary distribution ξ^T and deviation matrix $H = (I - T_0 + \underline{1}\xi^T)^{-1} - \underline{1}\xi^T$. Then, the following Laurent series expansion takes place (see Lemma 6.8 from [2])

$$[I - T(\varepsilon)]^{-1} = \frac{1}{\varepsilon} X_{-1} + X_0 + \varepsilon X_1 + \dots, \tag{4.6}$$

where the first two coefficients are given by

$$X_{-1} = \frac{1}{\pi^T T_1 \underline{1}} \underline{1}\xi^T, \tag{4.7}$$

and

$$X_0 = (I - X_{-1}T_1)H(I - T_1 X_{-1}). \tag{4.8}$$

Applying the above Laurent power series to $[I - AP]^{-1}$ with $T_0 = P$, $T_1 = D^{\sigma}P$ and $\varepsilon = a$, we obtain

$$\begin{aligned} [I - AP]^{-1} = [I - (P - aD^{\sigma}P)]^{-1} &= \frac{1}{a}\frac{1}{\pi^T T_1 \underline{1}} \underline{1}\xi^T + \mathrm{O}(1) = \\ &= \frac{1}{a}\frac{1}{\xi^T D^{\sigma} \underline{1}} \underline{1}\xi^T + \mathrm{O}(1). \end{aligned} \tag{4.9}$$

This yields asymptotic expressions for the generalized Personalized PageRanks for small a as

$$\pi_j = \xi_j + \mathrm{O}(a), \tag{4.10}$$

and

$$\rho_j = \frac{d_j^\sigma \xi_j}{\sum\limits_{i \in V} d_i^\sigma \xi_i} + \mathrm{O}(a). \tag{4.11}$$

In particular, if we assume that the graph is undirected ($W^T = W$), then $\xi_j = d_j / \sum\limits_{i \in V} d_i$ and we can further specify the above expressions as

$$\pi_j = \frac{d_j}{\sum\limits_{i \in V} d_i} + \mathrm{O}(a), \tag{4.12}$$

and

$$\rho_j = \frac{d_j^{1+\sigma}}{\sum\limits_{i \in V} d_i^{1+\sigma}} + \mathrm{O}(a). \tag{4.13}$$

We observe that using a positive or negative power σ of the degrees, we can significantly penalize or promote the score ρ for nodes with large degrees.

As a by-product of our computations, using (2.15), we have also obtained a nice asymptotic expression for the expected time between restarts in the case of undirected graphs:

$$\mathbb{E}_v[\#\text{ steps before restart}] = \frac{1}{a} \frac{\sum\limits_{i \in V} d_i}{\sum\limits_{i \in V} d_i^{1+\sigma}} + \mathrm{O}(1). \tag{4.14}$$

One interesting conclusion from the above expression is that when $\sigma > 0$ the highly skewed distribution of the degree in G can significantly shorten the time between restarts.

4.3. Random walk with jumps

In [6], the authors introduced a process with artificial jumps. It is suggested in [6] to add artificial edges with weights a/n between each two nodes of the graph. This process creates self-loops as well. Thus, the new modified graph is a combination of the original graph and a complete graph with self-loops. Let us demonstrate

that this is a particular case of the introduced generalized definition of Personalized PageRank. Specifically, we define the damping factors as

$$\alpha_i = \frac{d_i}{d_i + a}, \qquad i \in V, \tag{4.15}$$

and as the restart distribution we take the uniform distribution ($v = \underline{1}/n$). Indeed, it is easy to check that we retrieve the transition probabilities from [6]

$$p_{ij} = \begin{cases} \dfrac{a+n}{n(d_i+a)} & \text{when } i \text{ has an edge to } j, \\ \dfrac{a}{n(d_i+a)} & \text{when } i \text{ does not have an edge to } j. \end{cases} \tag{4.16}$$

As was shown in [6], the stationary distribution of the modified process, coinciding with the Occupation-Time Personalized PageRank, is given by

$$\pi_i = \pi_i(\underline{1}/n) = \frac{d_i + a}{2|E| + na}, \qquad i \in V. \tag{4.17}$$

In particular, from (1.6) we conclude that in the stationary regime

$$\mathbb{E}_v[\#\text{ steps before restart}] = \left(\sum_{j \in V} \left(1 - \frac{d_j}{d_j + a}\right) \frac{d_j + a}{2|E| + na} \right)^{-1} = \\ = \frac{2|E| + na}{na} = \frac{\bar{d} + a}{a},$$

where $\bar{d} = \frac{1}{n} \sum\limits_{i \in V} d_i$ is the average degree of the graph.

In the particular case of $\alpha_i = d_i/(d_i + a)$ and $v = \underline{1}/n$, the combination of (4.17) and (3.3) gives that $\pi_i(1 - \alpha_i)$ is independent of i, so that

$$\rho_i = 1/n. \tag{4.18}$$

This is quite surprising. Since $v^T = \underline{1}^T/n$, the nodes just after restart are distributed uniformly. However, it appears that the nodes just before restart are also uniformly distributed! Such an effect has also been observed in [7]. Algorithmically, this means that all pages receive the *same* generalized Personalized PageRank ρ, which,

for ranking purposes, is rather uninformative. On the other hand, this Personalized PageRank can be useful for sampling procedures. In fact, we can generalize (4.15) to

$$\alpha_i = \frac{d_i}{d_i + a_i}, \quad i \in V, \tag{4.19}$$

where now each node has its own parameter a_i. Now it is convenient to take as the restart distribution

$$v_i = \frac{a_i}{\sum\limits_{k \in V} a_k}. \tag{4.20}$$

Performing similar calculations as above, we arrive at

$$\pi_i(v) = \frac{d_i + a_i}{2|E| + \sum\limits_{k \in V} a_k}, \quad i \in V,$$

and

$$\rho_i(v) = \frac{a_i}{\sum\limits_{k \in V} a_k}, \quad i \in V. \tag{4.21}$$

Now in contrast with (4.18), the Location-of-Restart Personalized PageRank can be tuned to give any distribution that we like.

5. Discussion

We have proposed two generalizations of Personalized PageRank when the probability of restart depends on the node. Both generalizations coincide with the original Personalized PageRank when the probability of restart is the same for all nodes. However, in general, they show quite different behavior. In particular, the Location-of-Restart Personalized Pagerank appears to be more strongly affected by the value of the restart probabilities. We have further suggested several applications of the generalized Personalized PageRank in machine learning, sampling and information retrieval and analyzed some particularly interesting cases.

We feel that the analysis of the generalized Personalized PageRank on random graph models is a promising future research direction. We have already obtained some indications that the degree distribution can strongly affect the time between restarts. It would be highly interesting to analyze this effect in more detail on various random graph models (see e. g., [18] for an introduction into random graphs, and [12] for first results on directed configuration models).

Acknowledgements

The work of KA and MS was partially supported by the EU project Congas and Alcatel-Lucent Inria Joint Lab. The work of RvdH was supported in part by Netherlands Organisation for Scientific Research (NWO). This work was initiated during the 'Workshop on Modern Random Graphs and Applications' held at Yandex, Moscow, October 24–26, 2013. We thank Yandex, and in particular Andrei Raigorodskii, for bringing KA and RvdH together in such a wonderful setting.

Bibliography

1. ***K. Avrachenkov, R. W. van der Hofstad, M. Sokol,*** *Personalized PageRank with Node-dependent Restart*, Conference version in Proceeding of the 11th Workshop on Algorithms and Models for the Web Graph (WAW 2014).
2. ***K. Avrachenkov, J. Filar, P. Howlett,*** *Analytic perturbation theory and its applications*, SIAM Publisher, 2013.
3. ***K. Avrachenkov, V. Dobrynin, D. Nemirovsky, S. Pham, E. Smirnova,*** *Pagerank based clustering of hypertext document collections*, In Proceedings of ACM SIGIR 2008.
4. ***K. Avrachenkov, P. Gonçalves, A. Mishenin, M. Sokol,*** *Generalized optimization framework for graph-based semi-supervised learning*, In Proceedings of SIAM Conference on Data Mining (SDM 2012).
5. ***K. Avrachenkov, P. Gonçalves, M. Sokol,*** *On the Choice of Kernel and Labelled Data in Semi-supervised Learning Methods*, In Proceedings of WAW 2013, also in LNCS v. 8305, pp. 56–67, 2013.
6. ***K. Avrachenkov, B. Ribeiro, D. Towsley,*** *Improving random walk estimation accuracy with uniform restarts*, in Proceedings of WAW 2010, also Springer LNCS v. 6516, pp. 98–109, 2010.
7. ***K. Avrachenkov, N. Litvak, M. Sokol, D. Towsley,*** *Quick detection of nodes with large degrees*, *Internet Mathematics*, v. 10, pp. 1–19, 2013.
8. ***R. Baeza-Yates, P. Boldi, C. Castillo,*** *Generalizing pagerank*: *Damping functions for link-based ranking algorithms*, in Proceedings of ACM SIGIR'06, pp. 308–315, 2006.
9. ***P. Boldi,*** *TotalRank*: *Ranking without damping*, in Poster Proceedings of WWW2005, pp. 898–899, 2005.
10. ***S. Brin, L. Page, R. Motwami, T. Winograd,*** *The PageRank citation ranking*: *bringing order to the Web*, Stanford University Technical Report, 1998.
11. ***C. Castillo, D. Donato, A. Gionis, V. Murdock, F. Silvestri,*** *Know your neighbors*: *Web spam detection using the web topology*, In Proceedings of ACM SIGIR 2007, pp. 423–430, July 2007.

12. ***N. Chen, M. Olvera-Cravioto*** *Directed random graphs with given degree distributions*, Stochastic Systems, v. 3, pp. 147–186 (electronic), 2013.
13. ***P. Chen, H. Xie, S. Maslov, S. Redner,*** *Finding scientific gems with Google's PageRank algorithm*, *Journal of Informetrics*, v. 1(1), pp. 8–15, 2007.
14. ***P. G. Constantine, D. F. Gleich,*** *Using polynomial chaos to compute the influence of multiple random surfers in the PageRank model*, in Proceedings of WAW2007, LNCS v. 4863, pp. 82–95, 2007.
15. ***P. G. Constantine, D. F. Gleich,*** *Random alpha PageRank*, *Internet Mathematics*, v. 6(2), pp. 189–236, 2010.
16. ***F. Fouss, K. Francoisse, L. Yen, A. Pirotte, M. Saerens,*** *An experimental investigation of kernels on graphs for collaborative recommendation and semi-supervised classification*, Neural Networks, v. 31, pp. 53–72, 2012.
17. ***T. Haveliwala,*** *Topic-Sensitive PageRank*, in Proceedings of WWW 2002.
18. ***R. van der Hofstad,*** *Random Graphs and Complex Networks*, Lecture notes in preparation, Preprint (2014). Available from http://www.win.tue.nl/∼rhofstad/NotesRGCN.html.
19. ***X. Liu, J. Bollen, M. L. Nelson, H. van de Sompel,*** *Co-authorship networks in the digital library research community*, Information Processing & Management, v. 41, pp. 1462–1480, 2005.
20. ***P. Massa, P. Avesani,*** *Trust-aware recommender systems*, In Proceedings of the 2007 ACM conference on Recommender systems (RecSys '07), pp. 17–24, 2007.
21. ***C. D. Moler, K. A. Moler,*** *Numerical Computing with MATLAB*, SIAM, 2003.
22. ***H. C. Tijms,*** *A first course in stochastic models*, John Wiley and Sons, 2003.

Konstantin Avrachenkov

Inria Sophia Antipolis,
2004 Route des Lucioles,
06902 Sophia Antipolis,
France
konstantin.avratchenkov@inria.fr

Marina Sokol

Inria Sophia Antipolis,
2004 Route des Lucioles,
06902 Sophia Antipolis,
France
marina.m.sokol@gmail.com

Remco van der Hofstad

Department of Mathematics and
Computer Science,
Eindhoven University of Technology,
Postbus 513,
5600 MB Eindhoven,
The Netherlands
r.w.v.d.hofstad@tue.nl