EURANDOM PREPRINT SERIES

2014-012

May, 2014

Personalized PageRank with node-dependent restart

K. Avrachenkov, R. van der Hofstad, M. Sokol
ISSN 1389-2355

Personalized PageRank with node-dependent restart

Konstantin Avrachenkov* Remco van der Hofstad† Marina Sokol‡

May 2014

Abstract

Personalized PageRank is an algorithm to classify the improtance of web pages on a user-dependent basis. We introduce two generalizations of Personalized PageRank with node-dependent restart. The first generalization is based on the proportion of visits to nodes before the restart, whereas the second generalization is based on the probability of visited node 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 the 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.

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 probability $\alpha \in (0, 1)$. Thus, in the Standard PageRank centrality measure [7], 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 [12] 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., [2, 3, 7, 10, 11, 12, 14, 15]).

In the (standard) Personalized PageRank, the random walker restarts with a given fixed probability $1-\alpha$ at each visited node. We suggest a generalization where a random walker restarts with probability $1-\alpha_i$ 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 of node j corresponds to jth Personalized PageRank and 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

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

where W is the (possibly non-symmetric) adjacency matrix, D is the diagonal matrix with diagonal entries $D_{ii} = \sum_{j=1}^{n} W_{ij}$, and $A = \mathrm{diag}(\alpha_1, \dots, \alpha_n)$ is the diagonal matrix of damping factors.

*Inria Sophia Antipolis, France, k.avrachenkov@sophia.inria.fr
†Eindhoven University of Technology, The Netherlands, r.w.v.d.hofstad@tue.nl
‡Inria Sophia Antipolis, France, marina.sokol@inria.fr

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 weakly connected and if some node does not have outgoing edges, we add artificial outgoing edges to all the other nodes.

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

Definition 1 (Occupation-time Personalized PageRank) *The* Occupation-Time Personalized PageRank *is given by*

$$\pi_j(v) = \lim_{t\to\infty} \mathbb{P}(X_t = j). \tag{2}$$

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

$$\pi_j(v) = \lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}_{\{X_s=v\}}. \tag{3}$$

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

Definition 2 (Location-of-Restart Personalized PageRank) *The* Location-of-Restart Personalized PageRank *is given by*

$$\rho_j(v) = \lim_{t\to\infty} \mathbb{P}(X_t = j \text{ just before restart}) = \lim_{t\to\infty} \mathbb{P}(X_t = j \mid \text{restart at time } t+1). \tag{4}$$

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

$$\rho_j(v) = \lim_{t\to\infty} \frac{1}{N_t}\sum_{s=1}^{t} \mathbb{1}_{\{X_t=j, X_{t+1} \text{ restarts}\}}, \tag{5}$$

where N_t denotes the number of restarts up to time t. When the restarts occur with equal probability for every node, we have that $N_t \sim \mathsf{Bin}(t, 1-\alpha)$, 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. In general, however,

$$N_t/t \xrightarrow{a.s.} \sum_{v\in V}(1-\alpha_v)\pi_j(v), \tag{6}$$

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

Both generalized Personalized PageRanks are probability distributions, i.e., their sum over $j \in V$ gives 1. When $v^T = e(i)$, where $e_j(i) = 1$ when $i = j$ and $e_j(i) = 0$ when $i \neq j$, then both $\pi_j(v)$ and $\rho_j(v)$ can be interpreted as the relative importance of node j from the perspective of node i.

We see at least three applications of the generalized Personalized PageRank. The network sampling process introduced in [5] 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 [4, 11]. In this case, one may prefer to discount the effect of less informative nodes, e.g., nodes with very large degrees. And thirdly, the generalized Personalized PageRank can be applied for spam detection and control. It is known [8] that spam web pages are often designed to be ranked highly. 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.

In this paper, we investigate these 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 **??**, we investigate monotonicity properties of the our Personalized PageRank measures with respect to α_j. 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.

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 node-dependent restart equals*

$$\pi(v) = \frac{1}{v^T[I-AP]^{-1}\underline{1}} v^T[I-AP]^{-1}, \tag{7}$$

with $P = D^{-1}W$ the transition matrix of random walk on G withour restarts.

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

$$\pi(v)[AD^{-1}W + (I-A)\underline{1}v^T] = \pi(v), \tag{8}$$

so that

$$\pi(v)[I - AD^{-1}W] = \pi(v)(I-A)\underline{1}v^T, \tag{9}$$

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

$$\pi(v)[I - AD^{-1}W] = (1 - \pi(v)A\underline{1})v^T. \tag{10}$$

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

$$\pi(v) = (1 - \pi(v)A\underline{1})v^T[I - AD^{-1}W]^{-1}. \tag{11}$$

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

$$\pi(v)A\underline{1} = (1 - \pi(v)A\underline{1})v^T[I - AD^{-1}W]^{-1}A\underline{1}. \tag{12}$$

This yields

$$\pi(v)A\underline{1} = \frac{v^T[I-AP]^{-1}A\underline{1}}{1 + v^T[I-AP]^{-1}A\underline{1}}, \tag{13}$$

and, consequently, since $A = \mathrm{diag}(\alpha_1, ..., \alpha_n)$ is a diagonal matrix, so that $A\underline{1} = (\alpha_1, ..., \alpha_n)^T$, and we arrive at

$$\pi(v) = \frac{1}{1 + v^T[I-AP]^{-1}A\underline{1}} v^T[I-AP]^{-1}. \tag{14}$$

Since $v^T\underline{1} = 1$, by the fact that v^T is a probability mass function, we obtain

$$1 + v^T[I-AP]^{-1}A\underline{1} = v^T[I-AP]^{-1}\underline{1}, \tag{15}$$

from which the required equation (7) follows. □

Formula (7) admits the following probabilistic interpretation in the form of renewal equation

$$\pi_j(v) = \frac{\mathbb{E}_v[\#\text{ visits to } j \text{ before restart}]}{\mathbb{E}_v[\#\text{ steps before restart}]}, \tag{16}$$

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

Denote for brevity $\pi_j(i) = \pi_j(e_i^T)$, where e_i is the ith vector of the standard basis, so that $\pi_j(i)$ denotes the importance of node i from the perspective of j. Similarly, $\pi_i(j)$ denotes the importance of node j from the perspective of i. We next prove 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*

$$\frac{d_i}{\alpha_i K_i(A)}\pi_j(i) = \frac{d_j}{\alpha_j K_j(A)}\pi_i(j), \tag{17}$$

with

$$K_i(A) = \frac{1}{e_i^T[I - AP]^{-1}\underline{1}}. \tag{18}$$

Proof. Note that the denominator of (7) equals precisely $K_i(A)$. Thus, using a matrix geometric series expansion, we can rewrite equation (7) as

$$\begin{aligned}
\pi_j(i) &= K_i(A)e_i^T\sum_{k=0}^{\infty}(AD^{-1}W)^k e_j \qquad (19)\\
&= K_i(A)e_i^T\sum_{k=0}^{\infty}(AD^{-1}W)^k D^{-1}AA^{-1}De_j\\
&= K_i(A)e_i^T AD^{-1}\sum_{k=0}^{\infty}(WD^{-1}A)^k A^{-1}De_j\\
&= K_i(A)\frac{\alpha_i}{d_i}e_i^T\sum_{k=0}^{\infty}(WD^{-1}A)^k e_j\frac{d_j}{\alpha_j}\\
&= \frac{K_i(A)}{K_j(A)}\frac{\alpha_i}{d_i}\frac{d_j}{\alpha_j}K_j(A)e_i^T[I - WD^{-1}A]^{-1}e_j\\
&= \frac{K_i(A)}{K_j(A)}\frac{\alpha_i}{d_i}\frac{d_j}{\alpha_j}K_j(A)e_j^T[I - AD^{-1}W]^{-1}e_i,
\end{aligned}$$

which gives equation (17). □

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.,

$$K_i(A)^{-1} = \mathbb{E}_i[\#\text{ steps before restart}], \tag{20}$$

see also (21). Thus, a probabilistic interpretation of (7) is that

$$\frac{d_i}{\alpha_i}\mathbb{E}_i[\#\text{ visits to } j \text{ before restart}] = \frac{d_j}{\alpha_j}\mathbb{E}_j[\#\text{ visits to } i \text{ before restart}]. \tag{21}$$

Since

$$\mathbb{E}_i[\#\text{ visits to } j \text{ before restart}] = \sum_{k=1}^{\infty}\sum_{v_1,\ldots,v_k}\prod_{t=0}^{k-1}\frac{\alpha_{v_s}}{d_{v_s}}, \tag{22}$$

where $v_0 = j$, we immediately see that the expression for $\mathbb{E}_j[\#\text{ visits to } i \text{ before restart}]$ is identical, except for the first factor of $\frac{\alpha_i}{d_i}$, which is present in $\mathbb{E}_i[\#\text{ visits to } j \text{ before restart}]$, but not in $\mathbb{E}_i[\#\text{ visits to } j \text{ before restart}]$, and the factor $\frac{\alpha_j}{d_j}$, which is present in $\mathbb{E}_j[\#\text{ visits to } i \text{ before restart}]$, but not in $\mathbb{E}_j[\#\text{ visits to } i \text{ before restart}]$. This explains the factors $\frac{d_i}{\alpha_i}$ and $\frac{d_j}{\alpha_j}$ in (21) 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 node-dependent restart is equal to*

$$\rho(v) = v^T[I - AP]^{-1}[I - A], \tag{23}$$

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

Proof. This follows from the formula

$$\begin{aligned}\rho_j(v) &= \mathbb{E}_v[\#\text{ visits to } j \text{ before restart}]\mathbb{P}(\text{restart from } j) \\ &= \mathbb{E}_v[\#\text{ visits to } j \text{ before restart}](1 - \alpha_j).\end{aligned} \tag{24}$$

Now we can use (22) and the analysis in the proof of Theorem 1 to complete the proof. □

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

Theorem 4 (Symmetry for undirected Location-of-Restart Personalized PageRank) *When $W^T = W$ and $\alpha_i \in (0, 1)$, the following relation holds*

$$\frac{1-\alpha_i}{\alpha_i}\ d_i\ \rho_j(i) = \frac{1-\alpha_j}{\alpha_j}\ d_j\ \rho_i(j). \tag{25}$$

Proof. This follows from a series of equivalent transformations

$$\begin{aligned}\rho_j(i) &= e_i^T[I - AP]^{-1}[I - A]e_j = e_i^T[I - AP]^{-1}e_j(1 - \alpha_j) \\ &= e_i^T[AD^{-1}(DA^{-1} - W)]^{-1}e_j(1 - \alpha_j) = e_i^T[DA^{-1} - W]^{-1}e_j d_j\frac{1-\alpha_j}{\alpha_j} \\ &= e_i^T[(I - WD^{-1}A)DA^{-1}]^{-1}e_j d_j\frac{1-\alpha_j}{\alpha_j} = e_i^T AD^{-1}[I - WD^{-1}A]^{-1}e_j d_j\frac{1-\alpha_j}{\alpha_j} \\ &= \frac{\alpha_i}{d_i}e_i^T[I - WD^{-1}A]^{-1}e_j d_j\frac{1-\alpha_j}{\alpha_j} \\ &= \frac{\alpha_i}{d_i}\frac{\rho_i(j)}{1-\alpha_i}d_j\frac{1-\alpha_j}{\alpha_j}.\end{aligned} \tag{26}$$

Alternatively, Theorem 4 follows directly from (24) and (21). □

Interestingly, in (17), 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 (25), 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., $\alpha_j = \alpha$ for every j) 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 [16] for the original Personalized PageRank with $A = \alpha I$ from equation (7). Specifically,

$$v^T[I - AP]^{-1}\underline{1} = \alpha v^T[I - \alpha P]^{-1}\underline{1} = v^T \sum_{k=0}^{\infty} \alpha^k P^k \underline{1} = \frac{1}{1-\alpha}, \tag{27}$$

and hence we retrieve the well-known formula

$$\pi(v) = (1 - \alpha)v^T[I - \alpha P]^{-1}. \tag{28}$$

We also retrieve the following elegant result connecting direct and "reverse" original Personalized PageRanks on undirected graphs ($W^T = W$) obtained in [4]:

$$d_i \pi_j(i) = d_j \pi_i(j), \tag{29}$$

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

$$\mathbb{E}_v[\text{time between consecutive restarts}] = \frac{1}{1-\alpha}, \tag{30}$$

which is just the mean of the geomatrically distributed random variable.

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

$$A = I - aD^\sigma, \tag{31}$$

with $ad^\sigma_{\max} < 1$. 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 [1])

$$[I - T(\varepsilon)]^{-1} = \frac{1}{\varepsilon} X_{-1} + X_0 + \varepsilon X_1 + \dots, \tag{32}$$

where the first two coefficients are given by

$$X_{-1} = \frac{1}{\pi^T T_1 \underline{1}} \underline{1}\xi^T, \tag{33}$$

and

$$X_0 = (I - X_{-1}T_1)H(I - T_1 X_{-1}). \tag{34}$$

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

$$[I - AP]^{-1} = [I - (P - aD^\sigma P)]^{-1} = \frac{1}{a}\frac{1}{\pi^T T_1 \underline{1}} \underline{1}\xi^T + \mathrm{O}(a) = \frac{1}{a}\frac{1}{\xi^T D^\sigma \underline{1}} \underline{1}\xi^T + \mathrm{O}(a). \tag{35}$$

This yields the following asymptotic expressions for the generlized Personalized PageRanks

$$\pi_j(a) = \xi_j + \mathrm{o}(a), \tag{36}$$

and

$$\rho_j(a) = \frac{d_j^\sigma \xi_j}{\sum_{i\in V} d_i^\sigma \xi_i} + \mathrm{o}(a). \tag{37}$$

In particular, if we assume that the graph is undirected ($W^T = W$), we can further specify the above expressions

$$\pi_j(a) = \frac{d_j}{\sum_i d_i} + \mathrm{o}(a), \tag{38}$$

and

$$\rho_j(a) = \frac{d_j^{1+\sigma}}{\sum_{i\in V} d_i^{1+\sigma}} + \mathrm{o}(a). \tag{39}$$

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

As a by-product of our computations, we have also obtain nice asymptotic expression for the expected time between restarts in the case of undirected graph:

$$\mathbb{E}_v[\text{time between consecutive restarts}] = \frac{1}{a}\frac{\sum_{i\in V} d_i}{\sum_{i\in V} d_i^{1+\sigma}} + \mathrm{O}(a). \tag{40}$$

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

4.3 Random walk with jumps

In [5], the authors introduced a process with artificial jumps. It is suggested in [5] to add artificial edges with weights a/n between each two nodes to 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 introduce generalized definition of Personalized PageRank. Specifically, we define the damping factors as

$$\alpha_i = \frac{d_i}{d_i + a}, \quad i \in V, \tag{41}$$

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 [5]

$$p_{ij} = \begin{cases} \frac{a+n}{n(d_i+a)} & \text{when } i \text{ has an edge to } j, \\ \frac{a}{n(d_i+a)} & \text{when } i \text{ does not have an edge to } j. \end{cases} \tag{42}$$

As was shown in [5], the stationary distribution of the modified process, coinciding with the Occupation-time Personalized PageRank, is given by

$$\pi_i = \pi_i(\underline{1}/n) = \frac{d_i + a}{2|E| + na}, \quad i \in V. \tag{43}$$

Since $\pi(v)$ is the stationary distribution of $\tilde{P}$ with $v = \underline{1}/n$ (see (1)), it satisfies the equation

$$\pi(AP + [I - A]\underline{1}v^T) = \pi. \tag{44}$$

Rewriting this equation as

$$\pi[I - A]\underline{1}v^T = \pi[I - AP], \tag{45}$$

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

$$\pi[I - A]\underline{1}v^T[I - AP]^{-1} = \pi \tag{46}$$

or

$$v^T[I - AP]^{-1} = \frac{\pi}{\sum_{i=1}^n \pi_i(1-\alpha_i)}. \tag{47}$$

This yields

$$\rho_j(v) = \frac{\pi_j(1-\alpha_j)}{\sum_{i=1}^n \pi_i(1-\alpha_i)}. \tag{48}$$

In our particular case of $\alpha_i = d_i/(d_i + a)$, the combination of (43) and (48) gives that $\pi_j(1 - \alpha_j)$ is independent of j, so that

$$\rho_j = 1/n. \tag{49}$$

This is quite surprising. Since $v^T = \frac{1}{n}\underline{1}^T$, the nodes just after restart are distributed uniformly. However, it appears that the nodes just before restart are also uniformly distributed! Such effect has also been observed in [6]. 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. Note that to compute it, we do not need to know how many nodes the graph has. The only thing that we need to be able to do is to retrieve the neighbors of nodes. In pratice, this can be very useful!

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 stronger 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 analized some particular interesting cases.

We feel that the analysis of the generalized Personalized PageRank on random graph model is a promising future research directions. We have already obtained some indications that the degree distribution can strongly affect the time between restarts. It would be highly interesting to analyse this effect in more detail on various random graph models (see e.g., [13] for a introduction into random graphs, and [9] 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.

References

[1] K. Avrachenkov, J. Filar and P. Howlett, *Analytic perturbation theory and its applications*, SIAM Pulisher, 2013.

[2] K. Avrachenkov, V. Dobrynin, D. Nemirovsky, S. Pham and E. Smirnova, "Pagerank based clustering of hypertext document collections", In Proceedings of ACM SIGIR 2008.

[3] K. Avrachenkov, P. Gonçalves, A. Mishenin and M. Sokol, "Generalized optimization framework for graph-based semi-supervised learning", In Proceedings of SIAM Conference on Data Mining (SDM 2012).

[4] K. Avrachenkov, P. Gonçalves and 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.

[5] K. Avrachenkov, B. Ribeiro and D. Towsley, "Improving random walk estimation accuracy with uniform restarts", in Proceedings of WAW 2010, also Springer LNCS v.6516, pp.98-109, 2010.

[6] K. Avrachenkov, N. Litvak, M. Sokol and D. Towsley, "Quick detection of nodes with large degrees", *Internet Mathematics*, v.10, pp.1-19, 2013.

[7] S. Brin, L. Page, R. Motwami and T. Winograd, "The PageRank citation ranking: bringing order to the Web", Stanford University Technical Report, 1998.

[8] C. Castillo, D. Donato, A. Gionis, V. Murdock and F. Silvestri, "Know your neighbors: Web spam detection using the web topology", In Proceedings of ACM SIGIR 2007, pp.423-430, July 2007.

[9] N. Chen and M. Olvera-Cravioto. "Directed random graphs with given degree distributions", Stochastic Systems, v.3, pp.147-186 (electronic), 2013.

[10] P. Chen, H. Xie, S. Maslov and S. Redner, "Finding scientific gems with Googles PageRank algorithm", *Journal of Informetrics*, v.1(1), pp.8-15, 2007.

[11] F. Fouss, K. Francoisse, L. Yen, A. Pirotte and M. Saerens, "An experimental investigation of kernels on graphs for collaborative recommendation and semi-supervised classification", *Neural Networks*, v.31, pp.53-72, 2012.

[12] T. Haveliwala, "Topic-Sensitive PageRank", in Proceedings of WWW 2002.

[13] R. van der Hofstad, *Random Graphs and Complex Networks*, Lecture notes in preparation, Preprint (2014). Avaliable from `http://www.win.tue.nl/~rhofstad/NotesRGCN.html`.

[14] X. Liu, J. Bollen, M.L. Nelson and H. van de Sompel, "Co-authorship networks in the digital library research community", *Information Processing & Management*, v.41, pp.1462-1480, 2005.

[15] P. Massa and P. Avesani, "Trust-aware recommender systems", In Proceedings of the 2007 ACM conference on Recommender systems (RecSys '07), pp.17-24, 2007.

[16] C.D. Moler and K.A. Moler, *Numerical Computing with MATLAB*, SIAM, 2003.