Improving Random Walk Estimation Accuracy with Uniform Restarts

Konstantin Avrachenkov, Bruno Ribeiro, Don Towsley

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Improving Random Walk Estimation Accuracy with Uniform Restarts

Konstantin Avrachenkov — Bruno Ribeiro — Don Towsley

N° 7394

September 2010

Thème COM

ISSN 0249-6399 ISRN INRIA/RR--7394--FR+ENG

INRIA
SOPHIA ANTIPOLIS

Improving Random Walk Estimation Accuracy with Uniform Restarts

Konstantin Avrachenkov* , Bruno Ribeiro† , Don Towsley‡

Thème COM — Systèmes communicants
Projet Maestro

Rapport de recherche n° 7394 — September 2010 — 17 pages

Abstract: This work proposes and studies the properties of a hybrid sampling scheme that mixes independent uniform node sampling and random walk (RW)-based crawling. We show that our sampling method combines the strengths of both uniform and RW sampling while minimizing their drawbacks. In particular, our method increases the spectral gap of the random walk, and hence, accelerates convergence to the stationary distribution. The proposed method resembles PageRank but unlike PageRank preserves time-reversibility. Applying our hybrid RW to the problem of estimating degree distributions of graphs shows promising results.

Key-words: Sampling, Random Walk, Spectral Gap, PageRank, Online Social Network

* INRIA Sophia Antipolis-Méditerranée, France, k.avrachenkov@sophia.inria.fr
† Dept. of Computer Science, University of Massachusetts Amherst, Amherst, MA, ribeiro@cs.umass.edu
‡ Dept. of Computer Science, University of Massachusetts Amherst, Amherst, MA, towsley@cs.umass.edu

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

L'amélioration de l'estimation par la marche aléatoire avec le redémarrage uniforme

Résumé : Ce travail propose et étudie les propriétés d'un méthode d'échantillonnage hybride qui mélange l'échantillonnage de noeuds uniforme et la marche alátoire. Nous montrons que notre méthode d'échantillonnage combine les forces des deux échantillonnages tout en minimisant leurs inconvénients. En particulier, notre méthode permet d'augmenter l'écart spectral de la marche aléatoire et donc d'accélèrer la convergence vers la distribution stationnaire. La méthode proposée ressemble PageRank, mais en revanche elle conserve la proprieté de réversibilité. L'applications de notre méthode au problème de l'estimation de la distribution du degré montre des résultats prometteurs.

Mots-clés : Echantillonnage, Marche Alátoire, Ecart Spectral, PageRank, Réseaux Sociaux

1 Introduction

Many networks, including on-line social networks (OSNs) and peer-to-peer (P2P) networks, exist for which it is impossible to obtain a complete picture of the network. This leaves researchers with the need to develop sampling techniques for characterizing and searching large networks. Sampling methods can be classified as based on independent uniform sampling or crawling. These two classes of sampling methods have their advantages and drawbacks. *Our work proposes and studies the properties of a hybrid sampling scheme that mixes independent uniform node sampling and random walk (RW)-based crawling. We show that our sampling method combines the strengths of both uniform and RW sampling while minimizing their drawbacks.*

Within the class of uniform sampling methods, uniform node sampling is widely popular and has the advantage of sampling disconnected graphs. In an online social network (OSN) where users are associated with unique numeric IDs, uniform node sampling is performed by querying randomly generated IDs. In a P2P network like Bittorrent, uniform node sampling is performed by querying a tracker server [14]. In practice, however, these samples are expensive (resource-wise) operations (the ID space in an OSN, such as Facebook and MySpace, is large and sparse and tracker queries can be rate-limited [14]). For instance, in MySpace we expect only 10% the IDs to belong to valid users [10], i.e., only one in every ten queries successfully finds a valid MySpace account.

Within crawl-based sampling methods, random walk (RW) sampling is among the most popular methods [5, 11, 12, 18, 20, 23]. Let $G = (V, E)$ be an undirected, non-bipartite graph with n nodes. RW sampling is preferred because it requires few resources and, *when G is connected*, can be shown to produce asymptotically unbiased estimates of

$$f(G) = \sum_{\forall v \in V} h(v). \tag{1}$$

Moreover, *when G is connected*, a RW visits all nodes in G in $O(n^3)$ [17] steps w.h.p., a useful property when searching unstructured networks (such as P2P networks).

Note that the above formal RW guarantees require G to be connected. In the real-world, however, networks may consist of several disconnected components, e.g. Twitter [22] and Livejournal [20], to cite two known examples. Moreover, the performance of such methods are closely tied to the difference between the largest and the second largest eigenvalues of the associated RW transition probability matrix. This difference is also known as the *spectral gap* which we denote as δ. More precisely, let $(X_t : t = 0, 1, 2, \dots)$ be a discrete-time Markov chain associated with a random walk over G with transition probability matrix $\mathbf{P} = [p_{ij}]$, $\forall i, j \in V$, where $p_{ij} = 1/d_i$ and d_i is the degree of node $i \in V$. The eigenvalues of $\mathbf{P}$ are $1 = \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n \geq -1$. The spectral gap is defined as $\delta = (1 - \lambda_2)$. When G is disconnected $\delta = 0$. Typically, even a connected real complex network has small δ, explained by the clustered structure of these networks [15], and as a consequence the RW tends to get "trapped" inside subgraphs.

Unfortunately the accuracy of estimates of $f(G)$ (eq. (1)) obtained with a RW is inversely proportional to the spectral gap δ (revisited later in this section). *The main contribution*

of this work is to increase the spectral gap, δ, of a RW by combining it with independent uniform node sampling. Although node sampling can be very expensive, when combined with RW sampling, the resulting algorithm can significantly reduce the estimation error and search time at a negligible increase in overhead. The idea is as follows, add auxiliary edges among all the nodes with weight $\alpha/n > 0$. The hybrid sampling scheme, which we call *RWuR*, corresponds to a random walk on this modified graph. Our results show a significant increase in the spectral gap δ when our hybrid scheme is used (even when α is small). Note that when $\alpha = 0$ we RWuR is a regular RW and in the limit when $\alpha \to \infty$ RWuR performs independent uniform sampling.

In what follows we revisit the connection between the spectral gap, δ, and estimation errors; connecting δ with the Mean Squared Error (MSE) of the RW estimates.

Spectral Gap and Estimation Error

For now we assume G is connected. Suppose we wish to estimate $f(G)$ of eq. (1) from B sampled nodes obtained by a stationary RW, $(X_0, X_1, \ldots, X_{B-1})$. A widely used estimator of $f(G)$ is [20, 23]

$$\hat{f} = \frac{1}{B} \sum_{t=0}^{B-1} h'(X_t) \ , \ X_t \in V \tag{2}$$

where $h'(v) = h(v)/\pi_v$, $v \in V$, and π_v is the stationary distribution of the random walk. We note that if the graph is undirected, the RW is time reversible and its stationary distribution is given by

$$\pi_i = \frac{d_i}{2|E|}, \quad \forall i \in V. \tag{3}$$

To simplify the notation we drop the dependence of f on G. The MSE of $\hat{f}$ is given by $E[(\hat{f} - f)^2]$. Now we explore the connection between the spectral gap, δ, and the MSE of $\hat{f}$. In a stationary RW, $\hat{f}$ is an unbiased estimator of f [23]. Thus, the Mean Squared Error (MSE) is also the variance of $\hat{f}$.

Let $\mathrm{var}_\pi(\hat{f})$ denote the variance of $\hat{f}$ in a stationary RW and

$$E[(\hat{f} - f)^2] = \mathrm{var}(\hat{f}) + \mathrm{bias}(\hat{f}),$$

where $\mathrm{bias}(\hat{f}) = (E[\hat{f}] - f)^2$. The assymptotic ratio between $(\mathrm{var}(\hat{f}) + \mathrm{bias}(\hat{f}))$ and $\mathrm{var}_\pi(\hat{f})$ is a function of the spectral gap δ [1, Chapter 4.1]

$$\sup_f \lim_{B\to\infty} \frac{\mathrm{var}(\hat{f}) + \mathrm{bias}(\hat{f})}{\mathrm{var}_\pi(\hat{f})} = \frac{1+\lambda_2}{1-\lambda_2} = \frac{2-\delta}{\delta}\,.$$

Note that δ also determines the mixing time of the RW [21], which means that when $\delta \ll 1$ (i.e., $\lambda_2 \approx 1$) it takes many steps for the random walk to converge to the stationary distribution (a potential source of bias in $\hat{f}$ when the RW does not start in steady state).

We now turn our attention to finding the relationship between $\mathrm{var}_\pi(\hat{f})$ and δ. Consider h from eq. (1). Assume that $\sum_{\forall v \in V} h(v) = 0$ (this assumption is used to simplify our calculations) and $\|h'\|_2^2 \triangleq \sum_{\forall v \in V} (h'(v))^2 \pi_v > 0$. Let

$$\sigma^2 \triangleq \lim_{B \to \infty} B \left(\mathrm{var}(\hat{f}) + \mathrm{bias}(\hat{f}) \right).$$

From the inequality [1, Chapter 4, Proposition 29]

$$\frac{\sigma^2}{B} \left(1 - \frac{2}{\delta B} \right) \leq \mathrm{var}_\pi(\hat{f}) \leq \frac{2\|h'\|_2^2}{\delta B} \left(1 + \frac{\delta}{2B} \right) \tag{4}$$

we have a relationship between the MSE of $\hat{g}$ and the spectral gap δ. Eq. (4) shows that the MSE upper bound decreases as δ increases. One important class of $f(G)$ functions is the class that measures θ_k, the fraction of nodes with degree k,

$$f_k(G) \triangleq \sum_{\forall v \in V} h_k(v) \triangleq \sum_{\forall v \in V} \mathbf{1}(d_v = k)/n, ,$$

where $\mathbf{1}(x = y) = 1$ if $x = y$ and $\mathbf{1}(x = y) = 0$, otherwise. Note that $\|h_k\|_2^2 = \theta_k^2 / \Pi_k$, where $\Pi_k = \sum_{\forall v \in V} \pi_v \, \mathbf{1}(d_v = k)$, which yields

$$\mathrm{var}_\pi(\hat{f}_k) \leq \frac{2\theta_k^2}{\Pi_k \delta B} \left(1 + \frac{\delta}{2B} \right) . \tag{5}$$

Thus, eq. (5) shows that the error in estimating the fraction of nodes with degree k is uper bounded by the inverse of: (1) the spectral gap δ and (2) the probability that the RW finds a node with degree k, Π_k. In Section 2 we see that increasing parameter α of our hybrid RW increases δ but also decreases Π_k when k is larger than the average degree. This tradeoff can be seen in the experiments of Section 3, where the MSE of the fraction of high degree nodes, at first, decreases as we increase α until a certain (unknown optimal) point where a further increase in α increases the MSE. As future work we will investigate this optimal value of α.

2 Reducing mixing time by restart

As we have observed in the previous section, many complex networks have small spectral gaps and hence random walks on such networks can have a negative impact on the accuracy (variance) and bias of the estimates.

In this work we are interested in methods for accelerating the rate of convergence to the stationary distribution based on the addition of auxiliary transitions. Thus, the natural first method to investigate is PageRank [8]. Namely, a random walk follows some outgoing link with probability c and with probability $1 - c$ it jumps to an arbitrary node of the network

chosen according to the uniform distribution. Then, the modified random walk is described by the following transition matrix

$$\tilde{P} = cP + (1-c)\frac{1}{n}\underline{1}\underline{1}^T, \tag{6}$$

where $\underline{1}$ is a vector of ones with an appropriate dimension. Then, the stationary distribution of the modified random walk $\tilde{\pi}(c)$ is a unique solution of the following equations

$$\tilde{\pi}(c)\tilde{P} = \tilde{\pi}(c), \quad \tilde{\pi}(c)\underline{1} = 1.$$

It is known [13] that the second largest eigenvalue of matrix $\tilde{P}$ is equal to c. Thus, by choosing c not close to one, we can significantly increase the spectral grap δ.

However, as was observed for example in [16], PageRank's steady state distribution can be just weakly correlated with node degree. Furthermore, there are cases when ranking of nodes according to PageRank is sensitive to the change of parameter c [7]. Thus, the stationary distribution of the modified random walk can be significantly distorted. To mitigate the latter problem, we suggest a variation where we connect all the nodes in the graph with a weight α/n. The difference with PageRank is that in our variation the uniform restart occurs not with a fixed probability but with a probability depending on the node degree. Specifically, the transition probability in our modification is given by

$$\hat{p}_{ij} = \begin{cases} \frac{\alpha/n+1}{d_i+\alpha}, & \text{if } i \text{ has a link to } j, \\ \frac{\alpha/n}{d_i+\alpha}, & \text{if } i \text{ does not have a link to } j. \end{cases} \tag{7}$$

The advantage of such a modification is that the new random walk is also reversible with the following stationary distribution

$$\hat{\pi}_i(\alpha) = \frac{d_i + \alpha}{2|E| + n\alpha} \qquad \forall i \in V, \tag{8}$$

from which the original stationary distribution (3) can easily be retrieved.

Let us now show that this second method also improves algebraic connectivity. First, we consider a regular graph where all nodes have degree d.

Theorem 2.1 *Let $G = (V, E)$ be an undirected regular graph with degree d. Let $\hat{\lambda}_k(\alpha)$ and λ_k be the eigenvalues of the Markov chains associated with the modified and original random walks on G, respectively. Then, all the eigenvalues corresponding to the modified random walk except the unit eigenvalue are scaled as follows:*

$$\hat{\lambda}_k(\alpha) = \frac{d}{d+\alpha}\lambda_k, \quad k = 2, ..., n. \tag{9}$$

Proof. The transition probabilities of the modified random walk (7) can be written in the following matrix form

$$\hat{P} = \frac{d}{d+\alpha}\left(P + \frac{\alpha}{dn}\underline{1}\underline{1}^T\right),$$

where P is the transition matrix corresponding to the original random walk.

Since $\underline{1}$ is an eigenvector corresponding to the unit eigenvalue, we can apply Brauer's Theorem [6]. Brauer's Theorem says that if λ is an eigenvalue and x the corresponding eigenvector of matrix A then $\lambda + v^T x$ is an eigenvalue of matrix $A + xv^T$ for any vector v and the other eigenvalues of $A + xv^T$ coincide with the eigenvalues of A. Thus, matrix $P + \frac{\alpha}{dn}\underline{1}\underline{1}^T$ has the following eigenvalues: $1+\alpha/d, \lambda_2, ..., \lambda_n$. Since the eigenvalues of a matrix multiplied by a scalar are the eigenvalues of that matrix multiplied by the same scalar, the eigenvalues of matrix $\hat{P}$ are: $1, d/(d+\alpha)\lambda_2, ..., d/(d+\alpha)\lambda_n$. □

If we expand $d/(d+\alpha)$ in (9) into a power series with respect to α we can rewrite (9) as follows:

$$\hat{\lambda}_2(\alpha) = \left(1 - \frac{\alpha}{d}\right)\lambda_2 + \mathrm{o}(\alpha).$$

Thus, for small values of α the spectral gap can be approximated as follows:

$$\delta \approx \frac{\alpha}{d}\lambda_2. \tag{10}$$

Let us now consider the case of a general undirected graph.

Theorem 2.2 *Let $G = (V, E)$ be a general undirected graph. Then, the second largest eigenvalue $\hat{\lambda}_2(\alpha)$ of the modified random walk has the following connection with the second largest eigenvalue λ_2 of the original random walk*

$$\hat{\lambda}_2(\alpha) = \left(1 - \frac{\sum_{k=1}^n \frac{1}{d_k} u_{2k} v_{2k}}{\sum_{k=1}^n u_{2k} v_{2k}}\alpha\right)\lambda_2 + \frac{\sum_{k=1}^n \frac{1}{d_k} u_{2k} \sum_{j=1}^n v_{2j}}{n \sum_{k=1}^n u_{2k} v_{2k}}\alpha + o(\alpha), \tag{11}$$

where u_2 and v_2 are respectively left and right Fiedler eigenvectors of the original graph.

Proof. Let us analyze the equation

$$\hat{P}(\alpha)\hat{v}_2(\alpha) = \hat{\lambda}_2(\alpha)\hat{v}_2(\alpha) \tag{12}$$

with the help of perturbation theory techniques [2, 4]. We expand $\hat{P}(\alpha)$ as a power series with respect to α. Namely, we write

$$\hat{P}(\alpha) = P + \alpha\Gamma^{(1)} + \alpha^2\Gamma^{(2)} + ... \,, \tag{13}$$

where in particular the coefficient of the first order term is given by

$$\Gamma^{(1)} = \mathrm{diag}\left(\frac{1}{d_k}\right)\left(\frac{1}{n}\underline{1}\underline{1}^T - P\right), \tag{14}$$

where $\mathrm{diag}\left(\frac{1}{d_k}\right)$ is a diagonal matrix with the elements $\frac{1}{d_k}$ on the diagonal. We also expand $\hat{\lambda}_2(\alpha)$ and $\hat{v}_2(\alpha)$ as power series

$$\hat{\lambda}_2(\alpha) = \lambda^{(0)} + \alpha\lambda^{(1)} + \alpha^2\lambda^{(2)} + ... \tag{15}$$

and

$$\hat{v}_2(\alpha) = v^{(0)} + \alpha v^{(1)} + \alpha^2 v^{(2)} + \dots \,. \tag{16}$$

Next we substitute $\hat{P}(\alpha)$, $\hat{\lambda}_2(\alpha)$ and $\hat{v}_2(\alpha)$ in the form of power series (13), (15) and (16) into equation (12). Thus, we have

$$(P + \alpha\Gamma^{(1)} + \alpha^2\Gamma^{(2)} + ...)(v^{(0)} + \alpha v^{(1)} + \alpha^2 v^{(2)} + ...) =$$

$$(\lambda^{(0)} + \alpha\lambda^{(1)} + \alpha^2\lambda^{(2)} + ...)(v^{(0)} + \alpha v^{(1)} + \alpha^2 v^{(2)} + ...).$$

Equating terms with the same powers of α yields

$$Pv^{(0)} = \lambda^{(0)} v^{(0)}. \tag{17}$$

Since we are interested in the second largest eigenvalue of $\hat{P}(\alpha)$, we conclude that $v^{(0)} = v_2$ and $\lambda^{(0)} = \lambda_2$, where v_2 is the eigenvector of P corresponding to the second largest eigenvalue λ_2 of P. Then, collecting terms with α, we get

$$\Gamma^{(1)} v_2 + Pv^{(1)} = \lambda^{(1)} v_2 + \lambda_2 v^{(1)}. \tag{18}$$

Premultiplication of equation (18) by the left eigenvector u_2^T corresponding to the second largest eigenvalue λ_2 ($u_2^T P = \lambda_2 u_2^T$) leads to

$$u_2^T \Gamma^{(1)} v_2 = \lambda_2^{(1)} u_2^T v_2.$$

or

$$\lambda^{(1)} = \frac{u_2^T \Gamma^{(1)} v_2}{u_2^T v_2} \lambda_2$$

Let us consider in more detail the expression $u_2^T \Gamma^{(1)} v_2$.

$$u_2^T \Gamma^{(1)} v_2 = \frac{1}{n} u_2^T \mathrm{diag}\left(\frac{1}{d_k}\right) \underline{1}\, \underline{1}^T v_2 - u_2^T \mathrm{diag}\left(\frac{1}{d_k}\right) P v_2$$

$$= \frac{1}{n} \sum_{k=1}^{n} \frac{1}{d_k} u_{2k} \sum_{j=1}^{n} v_{2j} - \sum_{k=1}^{n} \frac{1}{d_k} u_{2k} v_{2k} \lambda_2$$

In the latter equality we use the fact that $Pv_2 = \lambda_2 v_2$. Thus, we obtain formula (11). □

Similar to the above analysis we can derive any number of terms in the power series expressions for $\lambda(\alpha)$ and $v(\alpha)$. For example, equating the terms with α^2 in the perturbed equation yields

$$\lambda^{(2)} = \frac{1}{u_2^T v_2} \left(u_2^T \Gamma_2 v_2 \lambda_2 - u_2^T (\Gamma_1 P + \lambda^{(1)} I) v^{(1)} \right),$$

where $v^{(1)}$ can be found from equation (18). Namely, $v^{(1)}$ is a solution of the equation

$$(P - \lambda_2 I) v^{(1)} = (\Gamma_1 P + \lambda^{(1)} I) v_2,$$

in the orthogonal complement to the subspaces corresponding to the eigenvalues 1 and λ_2. We can obtain a solution with the help of the group reduced resolvent

$$H^{\{1,\lambda_2\}} = \frac{1}{2\pi i}\int_\Gamma \frac{1}{\zeta}(P-\zeta I)^{-1}d\zeta,$$

where Γ is a contour in the complex plane enclosing all the eigenvalues of P except 1 and λ_2 [2]. Then, we have

$$v^{(1)} = H^{\{1,\lambda_2\}}(\Gamma_1 P + \lambda^{(1)} I)v_2.$$

Even though Theorem 2.2 provides a connection between the second largest eigenvalues of the original and modified graphs, we cannot readily deduce from expression (11) if the spectral gap actually decreases. Therefore, next we analyse a typical case where we can obtain more insight from formula (11).

We note that v_2 and u_2^T are Fiedler vectors. The Fiedler vectors indicate principal clusters of the original graph. Let us represent the transition matrix for the original graph in the following form

$$P = P^{(0)} + \varepsilon C = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix} + \varepsilon C, \tag{19}$$

where P_1, P_2 represent the transitions inside the principal clusters and εC represents transitions between the principal clusters. We choose the blocks P_1 and P_2 to be transition matrices, which means that some elements of εC corresponding to the blocks P_1 and P_2 are negative. Of course, all elements of the sum are non-negative.

Now we are ready to state the next result.

Theorem 2.3 *Given that the original graph has two principal Fiedler components with the same average node degree $\bar{d}$, the following connection between the eigenvalues of the modified and original graphs take place*

$$\hat{\lambda}_2(\alpha) = \left(1 - \frac{\alpha}{\bar{d}}\right)\lambda_2 + o(\alpha) + O(\varepsilon, |E_1|^{-1}, |E_2|^{-1}). \tag{20}$$

Proof. The proof is postponed to the Appendix. □

In particular, we conclude from (20) that for small values of α the value of the spectral gap can be approximated as follows:

$$\delta \approx \frac{\alpha}{\bar{d}}\lambda_2. \tag{21}$$

Comparing (10) and (21), it is curious to observe that in the case of the general undirected graph the parameter d in (10) is replaced by the average node degree $\bar{d}$.

The expression (21) provides simple guidelines for the choice of α. For instance, if the original graph has spectral gap very close to zero and we would like the spectral gap of the modified graph to be approximately equal to 0.1 we choose $\alpha = 0.1\bar{d}$.

3 Numerical Results

In the following preliminary experiments we use one real-world graph and one random graph. The real-world graph has $5,204,176$ nodes and $77,402,652$ edges and was collected in a nearly complete crawl of the Livejournal social blog network [19]. The Livejournal graph has average degree 14.6 and a giant strongly connected component with $5,189,809$ nodes and the remaining nodes form a number of small connected components. The random graph is created by connecting, with one edge, two Barabási-Albert graphs [3], G_1 and G_2, with average degrees 2 and 10, respectively. We call the former the BA_2 graph. A RW over the BA_2 graph resembles the RW transition probability matrix described in eq. (19), where P_i, $i = 1, 2$, are the transition probability matrices of the RW on each Barabási-Albert graph (G_1 and G_2, respectively) and ϵ is small. Different from the example shown in Section 2, the average degrees of G_1 and G_2 are different.

Our goal is to compare Random Walks (RWs) against Random Walks with uniform Restarts (RWuRs) in estimating the degree distribution of the graph, i.e., we seek to estimate Θ_k, the fraction of nodes with degree greater than k. Let $\hat{\Theta}_k$ be the estimated value of Θ_k.We use

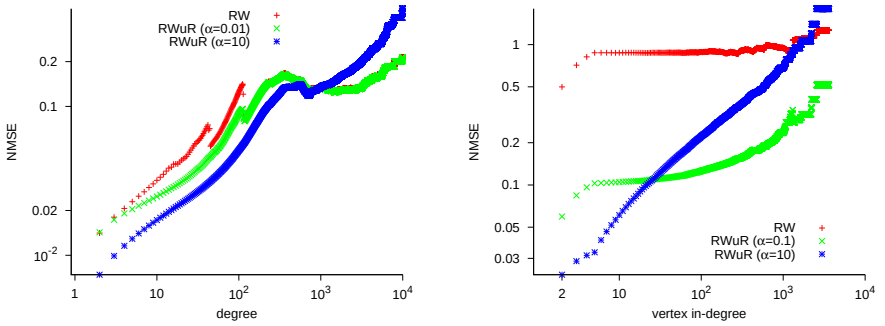
$$\text{NMSE}_k = E\left[(\hat{\Theta}_k - \Theta_k)^2\right]/\Theta_k \tag{22}$$

to measure the estimation accuracy.

Parameters: Our experiments have the following parameters. The sampling budget B, which is used in both RW and RWuR. The sampling budget of a RW determines the number of steps. The sampling budget of RWuR does not directly determine the number of steps. This is because there is a sampling budget penalty, c, associated with each restart. For instance, a RWuR that performs m uniform restarts walks $\lfloor B - mc \rfloor$ steps and gathers $\lfloor B - mc \rfloor + m$ observations. In all our experiments we use $B = n/100$ (i.e., the budget is 1% of the total number of nodes in the graph).

Initial RW states: *All experiments initialize RW in steady state while RWuR is initialized from uniformly sampled nodes (i.e., RWuR does not start in steady state). This initialization favors RW over our RWuR algorithm. Still, as seen next, our RWuR algorithm outperforms RW in all scenarios (for a given choice of α).*

Our first experiment is based on an undirected version of the Livejournal graph. Figure 1(a) shows the empirical NMSE, eq. (22), on the Livejournal graph obtained from $20,000$ runs, RWuR restart weights $\alpha = 0.01, 10$, and RWuR restart penalty $c = 10$. We choose $c = 10$ to match the 1/10 hit ratio of MySpace's uniform node sampling [10]. We observe that estimates obtained with RWuR $\alpha = 0.01$ are more precise than the estimates obtained with RW (particularly for small degree nodes and with almost no difference for high degree nodes). Thus, RWuR is able to reduce the NMSE even when restarts are rare (i.e., α is small). We perform the same experiment with increased restart weight $\alpha = 10$ (Figure 1(a)) and observe that increasing α also increases the accuracy of RWuR for estimating the head of the distribution but decreases the accuracy at estimating its tail than both RW and RWuR with $\alpha = 0.01$. Note that as we increase α RWuR gets closer to perfoming independent uniform node sampling.



(a) Livejournal (b) BA_2

Figure 1: Estimation error of RW and RWuR with varying α (larger is worst).

We also perform the same experiment over the BA_2 graph (with the only difference that the smallest restart weight is now $\alpha = 0.1$). Note that the BA_2 graph has a clear RW bottleneck (the edge that connects the two otherwise disconnected components). Figure 1(b) shows the empirical NMSE. Unsurprisingly, the improvement in NMSE obtained by RWuR against RWs is even more pronounced than the improvement observed in our Livejournal experiments. Note that here, similar to the Livejournal experiment, increasing the restart weight from $\alpha = 0.1$ to $\alpha = 10$ also increases the accuracy of RWuR for estimating the head of the distribution but decreases its accuracy at estimating the tail.

The above empirical observations, on the relationship between the NMSE of the degree distribution tail and α, prompt us to revisit the analysis performed at the end of Section 1. Putting together eqs. (5) and (22) yield

$$\text{NMSE}_k \propto \frac{1}{\Pi_k \delta}, \tag{23}$$

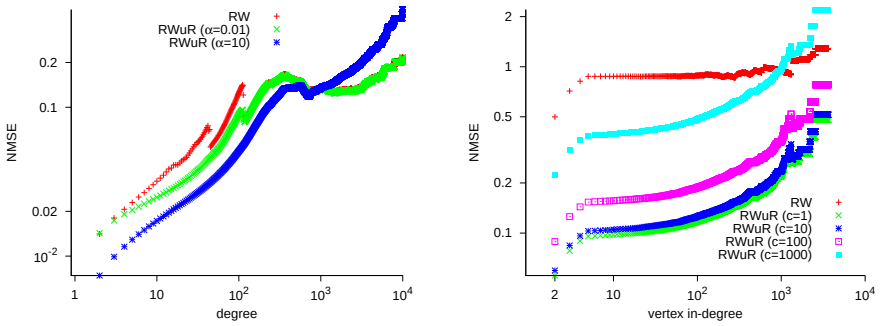
where Π_k is as defined at the end of Section 1. From Section 2 we know that

$$\Pi_k = \sum_{\forall v \in V} \mathbf{1}(d_v = k) \frac{k + \alpha}{2|E| + n\alpha}.$$

Thus, when $k > \bar{d}$ ($\bar{d}$ is the average degree), Π_k decreases with α which implies (eq. (23)) that the NMSE increases with α. Similarly, when $k < \bar{d}$ the NMSE decreases with α. Now let's look at the spectral gap δ. Section 2 shows that $\delta = 1 - d/(d + \alpha)\lambda_2$ for a d-regular graph and $\delta \approx \alpha\lambda_2/\bar{d}$ for graphs with two quasi-disconnected components that have same average degree, assuming α to be small. These results indicate that δ increases with α. As one increases α, when $k > \bar{d}$ the tradeoff between Π_k increasing and δ decreasing the

NMSE can explain the behavior observed in our numerical results. Similarly when $k < \bar{d}$, the NMSE monotonically decreases with α, which can also be observed in our numerical results.

Effect of the restart cost c: The uniform restarts required in our RWuR algorithm can be expensive. In MySpace [10] one needs to perform, in average, 10 queries in order to obtain an user that can be used to restart the RWuR. In the following experiments we compare RW and RWuR with different restart costs c. Figure 2(a) shows the empirical NMSE of RW and RWuR ($c = 1, 100$; $\alpha = 0.01$) on the Livejournal graph. The curves in Figure 2(a) are all on top of each other for degrees higher than 120. We notice no loss in NMSE when increasing the restart cost. This is expected as restarts are rare when $\alpha = 0.01$. Figure 2(b) shows the empirical NMSE of RW and RWuR ($c = 1, 10, 100, 1000$; $\alpha = 0.1$) on the BA_2 graph. Observe that RWuR outperforms RW when $c = 1, 10, 100$; the exception happens when $c = 1000$. Thus, in a graph with a strong RW bottleneck, RWuR is the estimation method of choice even if the cost of restart is high (e.g., $c = 100$).



(a) Livejournal ($\alpha = 0.01$)

(b) BA_2 ($\alpha = 0.1$)

Figure 2: Estimation error of RWuR with varying restart cost c (larger is worst).

4 Conclusions and Future Work

Our work proposed and studied the properties of a hybrid sampling scheme (RWuR) that mixes independent uniform node sampling and random walk (RW)-based crawling. Our sampling method combines the strengths of both uniform and RW sampling while minimizing their drawbacks. RWuR can be used in any OSN and P2P networks that allow uniform node sampling (usually at a premium cost), such as MySpace, Facebook, and Bittorrent. We have formally shown under two scenarios that, when compared to a regular RW, RWuR

has larger spectral gap, consequently reducing the mixing time. We also observe that RWuR has a positive impact on reducing the estimation error when compared to regular RWs. As part of our future work we plan to investigate the use of RWuRs to reduce the time to search for a file in a P2P network or a user in an OSN.

Acknowledgements

This research was sponsored in part by the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and do not represent the official policies, either expressed or implied, of the U.S. Army Research Laboratory or the U.S. Government. The U.S Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. This research was also sponsored in part by the European Commission through the ECODE project (INFSO-ICT-223936) of the European Seventh Framework Programme (FP7).

References

[1] Aldous, D., Fill, J.A.: Reversible Markov Chains and Random Walks on Graphs. Unpublished manuscript (1995)

[2] Avrachenkov, K.: Analytic perturbation theory and its applications. PhD Thesis, University of South Australia (1999)

[3] Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science 286, 509–512 (1999)

[4] Baumg artel, H.: Analytic perturbation theory for matrices and operators. Birkhauser (1985)

[5] Bisnik, N., Abouzeid, A.A.: Optimizing random walk search algorithms in p2p networks. Computer Networks 51(6), 1499–1514 (2007)

[6] Brauer, A.: Limits for the characteristic roots of a matrix, iv: Applications to stochastic matrices. Duke Math. J. 19, 75–91 (1952)

[7] Bressan, M., Peserico, E.: Choose the damping, choose the ranking? Proceedings of WAW 2009, LNCS 5427, 76–89 (2009)

[8] Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30, 107–117 (1998)

[9] Delebecque, F.: A reduction process for perturbed markov chains. SIAM Journal on Applied Mathematics 43, 325–350 (1983)

[10] Gauvin, W., Ribeiro, B., Liu, B., Towsley, D., Wang, J.: Measurement and gender-specific analysis of user publishing characteristics on myspace. IEEE Network special issue on Online Social Networks (2010)

[11] Gkantsidis, C., Mihail, M.: Hybrid search schemes for unstructured peer-to-peer networks. In: In Proceedings of IEEE INFOCOM. pp. 1526–1537 (2005)

[12] Gkantsidis, C., Mihail, M., Saberi, A.: Random walks in peer-to-peer networks: algorithms and evaluation. Perform. Eval. 63(3), 241–263 (March 2006)

[13] Haveliwala, T., Kamvar, S.: The second eigenvalue of the Google matrix. Tech. Rep. Available at `http://ilpubs.stanford.edu:8090/582/`, Stanford (2003)

[14] Konrath, M.A., Barcellos, M.P., Mansilha, R.B.: Attacking a swarm with a band of liars: evaluating the impact of attacks on bittorrent. In: Proc. of the IEEE International Conference on Peer-to-Peer Computing. pp. 37–44 (2007)

[15] Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Statistical properties of community structure in large social and information networks. In: Proc. of the WWW. pp. 695–704 (2008)

[16] Litvak, N., Scheinhardt, W., Volkovich, Y., Zwart, B.: Characterization of tail dependence for in-degree and pagerank. Proceedings of WAW 2009, LNCS 5427, 90–103 (2009)

[17] Lovász, L.: Random walks on graphs: a survey. Combinatorics 2, 1–46 (1993)

[18] Lv, Q., Cao, P., Cohen, E., Li, K., r r, S.S.: Search and replication in unstructured peer-to-peer networks. In: Proc. of the 16th international conference on Supercomputing. pp. 84–95 (2002)

[19] Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and Analysis of Online Social Networks. In: Proc. of the IMC (October 2007)

[20] Ribeiro, B., Towsley, D.: Estimating and sampling graphs with multidimensional random walks. In: Proc. of the ACM SIGCOMM IMC (Oct 2010)

[21] Sinclair, A.: Improved bounds for mixing rates of Markov chains and multicommodity flow. Combinatorics, Probability and Computing 1, 351–370 (1992)

[22] Twitter. `http://twitter.com` (2010)

[23] Volz, E., Heckathorn, D.D.: Probability based estimation theory for Respondent-Driven Sampling. Journal of Official Statistics (2008)

Appendix: Proof of Theorem 2.3

By construction, the matrix $P^{(0)}$ has two left eigenvectors corresponding to the eigenvalue one

$$u_{1,1} = \left[\begin{array}{cc} \pi_1 & 0 \end{array} \right], \quad u_{1,2} = \left[\begin{array}{cc} 0 & \pi_2 \end{array} \right],$$

and two right eigenvectors corresponding to the eigenvalue one

$$v_{1,1} = \left[\begin{array}{c} \underline{1} \\ 0 \end{array} \right], \quad v_{1,2} = \left[\begin{array}{c} 0 \\ \underline{1} \end{array} \right],$$

with $\pi_k P_k = \pi_k$. Let us introduce the following matrices

$$U_1 = \left[\begin{array}{c} u_{1,1} \\ u_{1,2} \end{array} \right], \quad V_1 = \left[\begin{array}{cc} v_{1,1} & v_{1,2} \end{array} \right].$$

Since the left and right eigenvectors of matrix $P^{(0)}$ form a bi-orthogonal system, we have

$$U_1 V_1 = I. \tag{24}$$

Let us also introduce the aggregated transition generator

$$M = U_1 C V_1. \tag{25}$$

Next we use the perturbation technique to determine a form of the Fiedler vectors. The perturbed eigenvalue problem can be written as follows:

$$(P^{(0)} + \varepsilon C)(v^{(0)} + \varepsilon v^{(1)} + ...) = (1 - \varepsilon \lambda^{(1)} + ...)(v^{(0)} + \varepsilon v^{(1)} + ...) \tag{26}$$

We set $\lambda^{(0)} = 1$, since the eigenvalue one of $P^{(0)}$ splits into two eigenvalues [2, 9]. The first perturbed eigenvalue remains equal to one and corresponds to the stationary distribution, the second perturbed eigenvalue corresponds to the Fiedler vectors. We are interested in this second eigenvalue. By equating the terms with the zero power of ε in (26) we obtain

$$P^{(0)} v^{(0)} = v^{(0)}.$$

We conclude from the above equation that $v^{(0)}$ is given by some linear combination of the right eigenvectors of $P^{(0)}$ corresponding to the eigenvalue one. Thus, we have

$$v^{(0)} = V_1 d, \tag{27}$$

with $d \in R^{2 \times 1}$. Equating the terms with the power one of ε we obtain

$$P^{(0)} v^{(1)} + C v^{(0)} = -\lambda^{(1)} + v^{(1)}.$$

Premultiplying the above equation by U_1 and noticing that $U_1(P^{(0)} - I) = 0$, we get

$$U_1 C v^{(0)} = -\lambda^{(1)} U_1 v^{(0)}.$$

Substituting (27) into the above equation, we obtain the eigenvalue problem for the aggregated generator

$$Md = -\lambda^{(1)} d.$$

Denote the entries of the aggregated generator as follows:

$$M = \begin{bmatrix} -\mu_1 & \mu_1 \\ \mu_2 & -\mu_2 \end{bmatrix}.$$

The above eigenvalue problem for the aggregated generator has two eigenvalues 0 and $\mu_1+\mu_2$. The zero eigenvalue corresponds to the perturbed stationary distribution. Thus we take the second eigenvalue $\lambda^{(1)} = \mu_1 + \mu_2$ corresponding to the Fiedler vectors. As vector d we can take $d = [\mu_1 \ -\mu_2]^T$ which leads to

$$v^{(0)} = V_1 d = \begin{bmatrix} \mu_1 \underline{1} \\ -\mu_2 \underline{1} \end{bmatrix}. \tag{28}$$

Similar considerations can be performed for the perturbed eigenvalue problem corresponding to the left Fiedler vector

$$(u^{(0)} + \varepsilon u^{(1)} + ...)(P^{(0)} + \varepsilon C) = (1 - \varepsilon\lambda^{(1)} + ...)(u^{(0)} + \varepsilon u^{(1)} + ...)$$

which results in the left eigenvector problem for the aggregated generator

$$eM = -\lambda^{(1)} e.$$

and, consequently, with $e = [\underline{1} \ -\underline{1}]$ we have

$$u^{(0)} = eU_1 = \begin{bmatrix} \pi_1 & -\pi_2 \end{bmatrix}. \tag{29}$$

Let us now substitute the expressions for the Fiedler vectors (28) and (29) into (11). Towards this end, we consider separately the two terms in (11)

$$a = \left(1 - \frac{\sum_{k=1}^n \frac{1}{d_k} u_{2k} v_{2k}}{\sum_{k=1}^n u_{2k} v_{2k}} \alpha\right) \lambda_2, \quad \text{and} \quad b = \frac{\sum_{k=1}^n \frac{1}{d_k} u_{2k} \sum_{j=1}^n v_{2j}}{n \sum_{k=1}^n u_{2k} v_{2k}} \alpha.$$

Let us first consider term a.

$$a = \left(1 - \frac{\alpha}{\mu_1 + \mu_2} \left(\mu_1 \sum_{k \in C_1} \frac{1}{d_k} \pi_{1,k} + \mu_2 \sum_{k \in C_2} \frac{1}{d_k} \pi_{2,k}\right)\right) \lambda_2 + \mathrm{O}(\varepsilon),$$

where the summations are taken over the two principal Fiedler clusters C_1 and C_2. Next we use the formula for the stationary distribution of the random walk.

$$a = \left(1 - \frac{\alpha}{\mu_1 + \mu_2} \left(\mu_1 \sum_{k \in C_1} \frac{1}{d_k} \frac{d_{1,k}}{2|E_1|} + \mu_2 \sum_{k \in C_2} \frac{1}{d_k} \frac{d_{2,k}}{2|E_2|}\right)\right) \lambda_2 + \mathrm{O}(\varepsilon).$$

We note that $d_{i,k} = d_k$ except for the nodes which connect the principal Fiedler clusters. Therefore, we can write

$$a = \left(1 - \frac{\alpha}{\mu_1 + \mu_2}\left(\mu_1 \frac{|C_1|}{2|E_1|} + \mu_2 \frac{|C_2|}{2|E_2|}\right)\right)\lambda_2 + \mathrm{O}(\varepsilon, |E_1|^{-1}, |E_2|^{-1})$$

$$= \left(1 - \frac{\alpha}{\bar{d}}\right)\lambda_2 + \mathrm{O}(\varepsilon, |E_1|^{-1}, |E_2|^{-1}),$$

where the last equality follows from the assumption that the principal Fiedler clusters have the same average node degree $\bar{d} = 2|E_1|/|C_1| = 2|E_2|/|C_2|$. Next we consider term b. Let us analyze the sum $\sum_{k=1}^{n} \frac{1}{d_k} u_{2k}$.

$$\sum_{k=1}^{n} \frac{1}{d_k} u_{2k} = \sum_{k \in C_1} \frac{1}{d_k} \pi_{1,k} + \sum_{k \in C_2} \frac{1}{d_k} \pi_{2,k} + \mathrm{O}(\varepsilon)$$

$$= \sum_{k \in C_1} \frac{1}{d_k} \frac{d_{1,k}}{2|E_1|} + \sum_{k \in C_2} \frac{1}{d_k} \frac{d_{2,k}}{2|E_2|} + \mathrm{O}(\varepsilon) = \frac{|C_1|}{2|E_1|} + \frac{|C_2|}{2|E_2|} + \mathrm{O}(\varepsilon, |E_1|^{-1}, |E_2|^{-1})$$

$$= \mathrm{O}(\varepsilon, |E_1|^{-1}, |E_2|^{-1})$$

In the latter equality we again use the fact that the principal Fiedler clusters have the same average node degree $\bar{d} = 2|E_1|/|C_1| = 2|E_2|/|C_2|$. Thus, only term a provides significant contribution to the final formula (20).

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399