Online Estimating the k Central Nodes of a Network

Yeon-sup Lim*, Daniel S. Menasché*, Bruno Ribeiro*, Don Towsley*, and Prithwish Basu†
*Department of Computer Science UMass Amherst, †Raytheon BBN Technologies

I. Introduction

Estimating the most influential nodes in a network is a fundamental problem in network analysis. Influential nodes may be important spreaders of diseases in biological networks, key actors in terrorist networks, or marketing targets in social networks. By identifying such central nodes, one can devise efficient strategies for prevention of diseases or crime and efficient marketing strategies.

The goal of this paper is to estimate the k most central nodes in a network through parsimonious sampling. Centrality determines the relative importance of a particular node within the network. Conventional measures of node centrality include degree, betweenness, and closeness [1].

For networks for which complete information is available, there are exact and approximation algorithms that can efficiently obtain the k most central nodes. However, when complete information is not available and networks need to be sampled online in the wild, these algorithms are no longer adequate for the task. In this paper we design and evaluate algorithms that identify the k most central nodes of a network in an online fashion, *i.e.,* by examining only a small fraction of the network nodes. We compare the performance of these algorithms on real-world data (two collaboration networks, an Internet autonomous system network and a social network) through simulation. We identify two sources of error: *(1)* sampling (collection) error and *(2)* identification error. Sampling error occurs when a node in the k most central nodes is not sampled by the sampling algorithm. Ideally a network crawler should collect the most central nodes as early as possible in the sampling process. We observe that in graphs with highly skewed degree distribution, random walks achieve low sampling errors early in the sampling process. Identification error occurs when a sampled top k node is not identified as such. Here, even if the most central nodes have been sampled (collected) we still need to be able to identify them.

This work was supported by the NSF under grant CNS-0721861 and CNS-1065133, ARO under grant MURI W911NF-08-1-0233, and ARL Cooperative Agreement W911NF-09-2-0053. DSM was also supported by a CAPES/Fulbright (Brazil) scholarship. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied of the NSF, ARO, ARL, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

For each of the three centrality metrics considered in this paper we pose the following questions

1) How do we estimate the top k most central nodes, and what is the impact of node degrees on our ability to identify central nodes?

2) What are the contributions of sampling (collection) error and identification error on the effectiveness of sampling and identifying the k top central nodes?

3) How can we take advantage of the ability to obtain node degrees while sampling nodes in a network?

We give the following (partial) answers to the questions above:

1) We observe that random walk sampling can effectively collect a significant fraction (more than 80%) of the k largest degree nodes in all the considered networks, once 5% of the nodes are sampled;

2) We discover that sampling error is larger in networks that have less skewed degrees whereas estimation error is larger in networks in which betweenness and closeness centrality are weakly correlated to degree centrality;

3) We observe that degree centrality acts as a good alias for other centrality metrics for a number of networks, and show that a simple random walk based algorithm outperforms complex strategies recently proposed in the literature in identifying the k most central nodes (for betweenness and closeness). Future work consists of devising methods to take advantage of collected information about node degrees, even when the latter is not a good alias for other centrality metrics.

II. Data Sets

A. Data Set Description

Our data sets are real world traces of representative graphs, which we classify into three categories: collaboration, device, and social networks. Let n and m be the number of nodes and edges in the network. In this paper we study four traces, referred to as CA-CondMat, CA-HepPh, AS-Snapshot and Email-Enron:

1) **Collaboration networks:** the CA-CondMat and CA-HepPh traces are collaboration networks from ArXiv Condensed Matter ($n = 23,133$ and $m = 186,936$) and ArXiv High Energy Physics ($n = 12,008$ and $m = 237,010$), respectively. The data covers papers published in the period January 1993 to April 2003 [2];

2) **Device networks:** the AS-Snapshot is a snapshot of the structure of the Internet at the level of autonomous systems,

978-1-4577-1051-3/11/$26.00 ©2011 IEEE

reconstructed from BGP tables posted by the University of Oregon Route Views Project ($n = 22,963$ and $m = 48,436$). This snapshot was created by Mark Newman from data of July 22, 2006 [3];

3) **Social networks:** the EMAIL-ENRON trace corresponds to an email communication network from Enron ($n = 36,692$ and $m = 367,662$) [4].

Note that each of the above networks consists of one connected component or one of the connected components is substantially larger than the others, in which case we restrict ourselves to the largest connected component.

B. Data Set Characteristics

In this section we report, for each of our four traces, 1) the skewness of degree distributions, and 2) the correlation between degree and other centrality metrics. As we will show in §IV, these two characteristics are important factors in determining the performance of sampling and identification strategies, respectively.

a) Skewness: Given the degree distributions, we seek a metric which correlates to how fast a random walk collects the top k nodes with highest degrees in such networks. To this goal, we consider the skewness s of the degree distribution (see Table I). In §IV we will show that the skewness is indeed correlated to how fast a random walk can collect the top k most central (degree, betweenness and closeness) nodes.

b) Correlation: In many networks, betweenness and closeness centralitity are known to be correlated with node degree [5]. To investigate the extent that this correlation occurs in our traces, Table I reports the Pearson correlation coefficient [6] between ranks by node degree and other centralities.

Note that the Pearson correlation coefficient characterizes the extent of linear relationship that exists between two ranks, over all nodes. Using all nodes, not only the top k nodes, prevents problems that arise when correlating sequences of elements composed of different elements. For a simple example, the Pearson correlation between sequences (1, 2, 3) and (4, 5, 6) equals 1, even though the two sequences are disjoint.

Table I shows that betweenness and closeness centrality have a positive correlation with degrees in all considered networks. Such correlation suggests that node degree can be used as an alias to identify the most central nodes according to other centrality metrics. In §IV we experimentally validate this observation through simulations, which indicate that this simple strategy in most cases outperforms more sophisticated strategies.

TABLE I
DEGREE DISTRIBUTION CHARACTERISTICS AND CORRELATION BETWEEN DEGREE, BETWEENNESS, AND CLOSENESS CENTRALITY

Network	r		Degree Skewness, s [7]
	Betweenness	Closeness	
AS-SNAPSHOT	0.7067	0.4757	43.6
EMAIL-ENRON	0.6951	0.4849	15.6
CA-CONDMAT	0.7142	0.6741	5.6
CA-HEPPH	0.6004	0.7020	4.8

III. SAMPLING AND IDENTIFICATION METHODS

A. Sampling Methods

In this paper we consider six sampling methods (see Table II): 1) random-walk sampling (RW) [8]; 2) random-walk sampling with uniform restarts (RWJ and RWU) [8]; 3) frontier sampling (FS) [9]; 4) expansion sampling (XS) [10]; 5) adjusted expansion sampling (AXS); and 6) randomized expansion sampling (RXS). The first four methods are described in [8], [9], [10]. For convenience, in what follows we briefly describe XS. Then we suggest a subtle adjustment to XS in §III-A2 and propose RXS in §III-A3.

1) Expansion sampling (XS): Given a graph $G(V, E)$, let $N(M)$ be the neighborhood of a given set of vertices $M \in V$. Starting from an arbitrary node, XS iteratively adds nodes from $N(M)$ to M, prioritizing those that have the largest number of neighbors in $V \setminus (N(M) \cup M)$.

It is often the case that, in order to retrieve the degrees of the nodes in $N(M)$, a crawler has to visit such nodes. Nevertheless, the original description of *XS* [10] does not account for the cost of visiting these nodes. In what follows we describe a subtle variation of expansion sampling, which allows us to 1) *fairly compare* XS against other methods, by charging XS for exploring $N(M)$ and 2) *use the information obtained* about $N(M)$ to estimate the most central nodes.

2) Adjusted Expansion Sampling (AXS): To adjust *XS* so as to account for the need to visit each node (in $N(M)$) before learning its degree, *AXS* maintains two sets: the set of *marked nodes* and the set of *sampled nodes*. The set of marked nodes, M, is equivalent to the one used by expansion sampling. The set of sampled nodes S equals M along with the nodes in $N(M)$ that have been visited by the crawler. Then, the top k most central nodes are identified among M, incurring cost $|S|$. Refer to [11] for the precise description of *AXS*.

3) Randomized Expansion Sampling (RXS): Next, we propose a randomized version of expansion sampling. Let $u(v), v \in S$ be the number of edges from v to $V \setminus S$. The aim of *RXS* is to avoid the cost of visiting all neighbors of the sampled set S before selecting the next node in $N(S)$ to be added to S. To this goal, *RXS* selects a node $v \in S$ with the largest $u(v)$ and proceed to sample one of its neighbors at random, adding it to S. The rationale behind *RXS* is to heuristically assume that the network has positive degree assortativity and that a neighbor w of the node with largest $u(v)$, $v \in S$, is likely to have the largest $u(w)$ among all vertices in $N(S)$. Refer to [11] for the precise description of *RXS*

B. Identification Methods

Given a sampled network, how do we select the top k nodes with largest centrality? The selection of the top k

TABLE II
TABLE OF NOTATION

method	parameter	$= RW$ when...
random walk [8]		
random walk fixed restart [8]	p, prob. restart	$p = 0$
random walk restart [8]	α, uniformity level	$\alpha = 0$
frontier sampling [9]	h, # of walkers	$h = 1$
(adjusted) expansion sampling (§III-A2)		
randomized expansion sampling (§III-A3)		

nodes with largest *degree* centrality is straightforward, once we assume that crawlers can sample node degrees when traversing the network. For *betweenness* and *closeness* centrality, we consider the following two approaches:

1) Degree Centrality as Alias to Betweenness and Closeness Centrality: Traversed nodes are sorted according to their queried degrees. The top k nodes in the sorted list are taken as the top k nodes with highest betweenness and closeness centrality.

2) Re-calculation of Centrality Metrics in the Sampled Network: Given the subgraph induced by the sampled nodes, closeness and betweenness centralities are computed for each of the nodes in the subgraph. Nodes are sorted based on the computed values and the top k are selected.

IV. EXPERIMENTAL RESULTS

In this section we report the results of our experiments on the impact of the system parameters on the different methods to find the top k nodes in the network. Our experimental goals are to show that random walk sampling coupled with degrees used as an alias to other centrality metrics, suffice in order to find the top k nodes in the four networks described in §II.

We performed fifty simulation runs for each network topology and set of parameters. We report the averages as well as the standard deviation of the performance metric of interest. When evaluating the performance of the sampling strategies, the metric of interest is the *fraction of top k nodes* that are contained in the sampled set. When evaluating the performance of the identification strategy, the metric of interest is the *overlap ratio* between the identified nodes and the actual top k nodes. Note that as k is fixed, both false positives and false negatives follow from our metrics.

A. Sampling Strategy: When Does a Random Walk Suffice?

In this section we compare *RW* against the other sampling strategies described in §III, where the goal is to collect the top 10 largest degree nodes. Results concerning other centrality metrics exhibit similar trends.

1) RWJ and RWU: Figure 1(a) and 1(b) compare *RW* against *RWJ* and *RWU*, respectively, for EMAIL-ENRON. (Figure 1(e) and (f) are for CA-CONDMAT). Recall that *RWJ* and *RWU* with parameters $p = 0$ and $\alpha = 0$, respectively, degenerate to *RW*. As shown in Figure 1, the performance of *RWJ* and *RWU* monotonically decreases with respect to p and α, i.e., *RW* outperforms the other strategies.

2) FS: Figure 1(c) and 1(g) compares *RW* against *FS*. *FS* leverages multiple random-walkers which is specially useful in order to sample graphs with multiple connected components. As shown in Figure 1, we observed no gains in using *FS* instead of *RW*, which might be explained in part by the fact that we focus only on the largest connected component of each network.

3) XS: Figure 1(d) and 1(h) compares *RW* against *XS* and its variants, *AXS* and *RXS*. Since *XS* utilize only part of the samples it collects in order to identify the top k nodes, its performance is the poorest among the four considered strategies. In EMAIL-ENRON, *RW* is the most efficient approach for collecting the 10 largest degree nodes. In CA-CONDMAT, however, *XS* and its variants outperform *RW*. In particular, *RXS* outperforms *AXS* even though *RXS* does not visit all the neighborhood of the marked nodes before deciding which node to subsequently mark (see §III-A1).

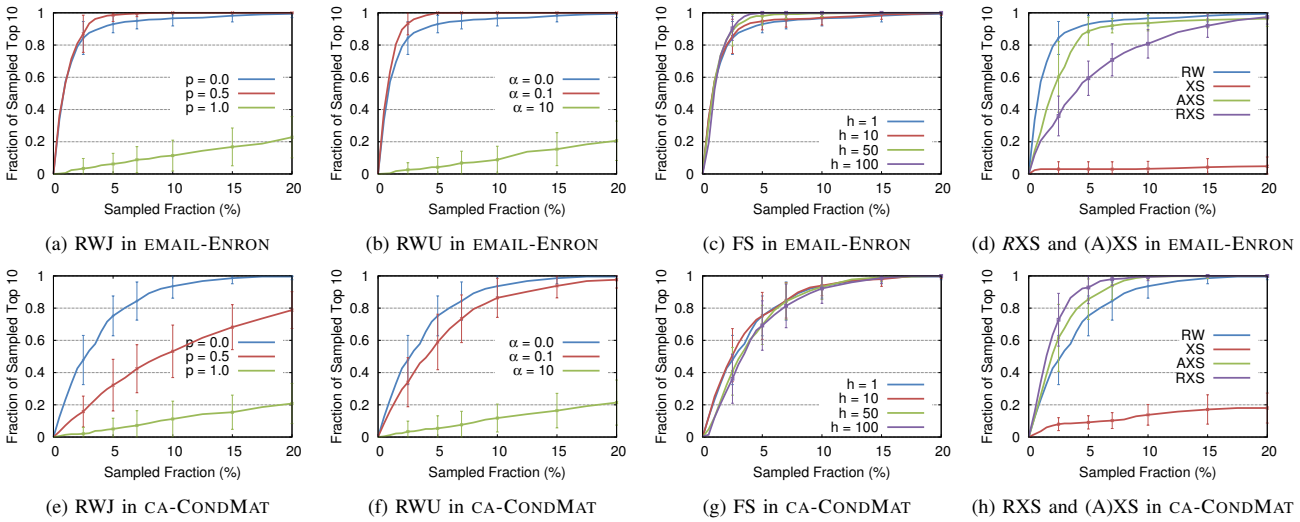
In order to compare the performance of the four sampling methods across the four considered traces, we present preliminary observations on the relationship between degree skewness and the performance of the different methods. Due to space constraints, here we present results for EMAIL-ENRON and CA-CONDMAT. Refer to [11] for the results of other traces.

The traces that present larger values of s, AS-SNAPSHOT and EMAIL-ENRON, are also those at which *RW* outperforms the other approaches. As s decreases (see Table I), the performance of *AXS* and *RXS* increases (see Figures 1(d) and 1(h)) and the one of *RW* decreases. The degree skewness, s, appears to be an important predictor of the performance of different sampling techniques. Future work consists of further investigating such relationship.

B. Identification Strategy: Degree as alias for closeness and betweeness Centralities?

We now focus on identifying the top k central nodes out of the sampled network. In this section we consider two strategies: $a)$ RW with degree as an alias to other metrics and $b)$ AXS with recomputation of centralities in the sampled network.

1) Betweenness Centrality: Figures 2(a)-(d) exhibit the overlap ratio of the top 10 most central betweenness nodes as a function of the sampled fraction of nodes. Note that the maximum overlap ratio that can be achieved at the identification step corresponds to the sampled fraction of top k nodes in the sample set (dotted lines in Figure 2). The degree alias strategy identifies more than 70% of the top k central nodes from a sample of no more than 5% of the network. The recalculation strategy, in contrast, only identifies 25% of the nodes with the same sample size. Interestingly, using the degree as an alias to



(a) RWJ in EMAIL-ENRON (b) RWU in EMAIL-ENRON (c) FS in EMAIL-ENRON (d) *R*XS and (A)XS in EMAIL-ENRON

(e) RWJ in CA-CONDMAT (f) RWU in CA-CONDMAT (g) FS in CA-CONDMAT (h) RXS and (A)XS in CA-CONDMAT

Fig. 1. Performance of RW against other sampling methods

betweenness centrality is not only more effective but also more efficient than calculating betweenness centrality values for the subgraph induced by the samples. Let n' and m' be the number of nodes and edges in the sampled network. The worst-case complexity of the former is $O(n' \log n')$ whereas of the later is $O(n'm' + n'^2 \log n')$ using state of the art algorithms [12].

Using degree as an alias to betweenness centrality yields better results in AS-SNAPSHOT and EMAIL-ENRON as opposed to CA-HEPPH and CA-CONDMAT. This suggest that skewness, s (see Table I) might affect not only the performance of the sampling strategy but also that of the identification strategy. Note that the degree alias strategy fails to estimate the correct top k nodes in CA-HEPPH. This might be explained, in part, by the fact that the rank correlation between node degree and betweenness in CA-HEPPH is the lowest, 0.6004. However, this observation requires further investigation, since as discussed in §II-B, rank correlation as reported in Table I accounts for the ranks of all nodes and not only for the ranks of top nodes.

2) Closeness Centrality: Figures 2(e)-(h) show the overlap ratio of the top 10 most central closeness nodes as a function of the sampled fraction of nodes. The degree alias strategy when used to find nodes with high closeness centrality yields poorer performance when compared to betweenness centrality. For small values of sampled fraction of nodes, the alias strategy still performs better than recomputing betweenness centrality at the graph induced by the samples.

Note that the recalculation strategy yields relatively consistent performance regardless of the trace, outperforming the degree alias strategy when the fraction of sampled nodes is large. However, the recalculation of closeness centrality yields the same worst-case time complexity as the one of betweenness [13]. Devising an improved identification strategy, which leverages information about node degrees collected during the sample phase, and tailored for closeness centrality, is subject for future work.

V. RELATED WORK

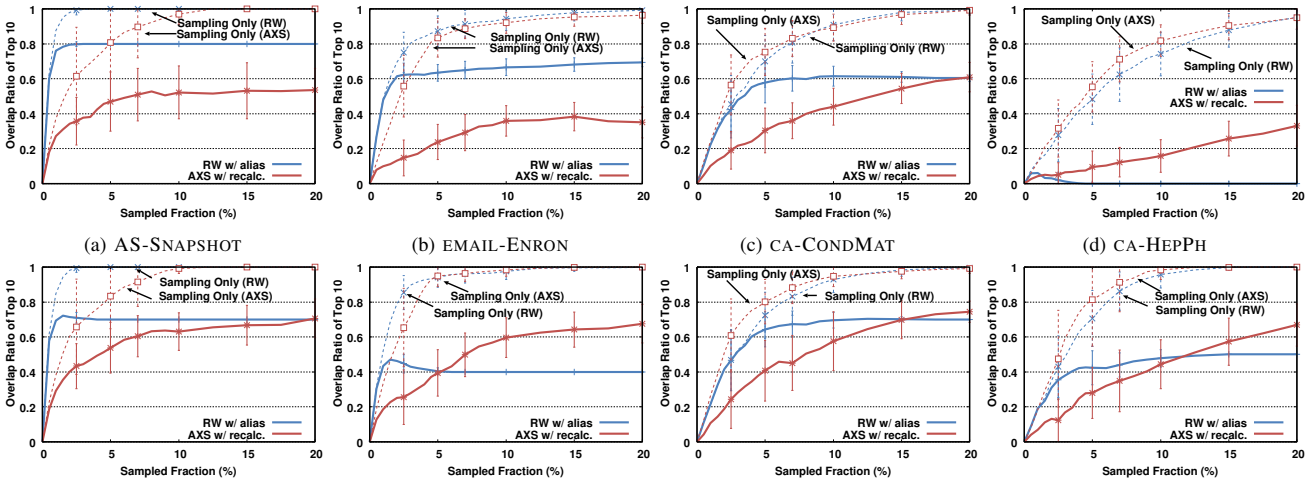
There is a vast literature on approximations and heuristics for estimating node centralities [12], [14], [15]. However, to the best of our knowledge, our work is the first to study the problem of estimating the top k most central nodes in a network, comparing multiple centrality metrics under different network topologies and accounting for both sampling and identification errors.

Chan *et al.* [14] introduce a framework to approximate node centralities in real-world networks that are known to exhibit modularity. Bader *et al.* [12] propose a betweenness centrality approximation algorithm based on an adaptive sampling technique. Our goal, in contrast, is to rank nodes according to their centralities rather than approximating the centrality values.

Okamoto *et al.* [15] propose an algorithm to rank the top k nodes with highest closeness centrality. They used the approximation algorithm presented in [13], which samples l vertices uniformly at random and computes the average of l shortest-path distances from each node to the l sampled nodes to estimate closeness centrality. In contrast, we show that the use of *random-walk sampling*, leveraging the fact that node degrees are visible while sampling the network, in many cases suffices to get the top k nodes with highest centrality.

Borgatti *et al.* [16] present the effect of sampling on centrality measures. Their results indicate that the accuracy of centrality measures degrades gracefully as the sample size decreases. In this paper, we extend such conclusion to the problem of estimating the top k most central nodes.

Maiya and Berger-Wolf [10] proposed Expansion Sampling (XS). The authors show that XS collects high central nodes



(a) AS-SNAPSHOT (b) EMAIL-ENRON (c) CA-CONDMAT (d) CA-HEPPH

(e) AS-SNAPSHOT (f) EMAIL-ENRON (g) CA-CONDMAT (h) CA-HEPPH

Fig. 2. Overlap ratio of top 10 central nodes and their estimates (dotted lines show the fraction of top 10 nodes in the sample set). The top is for betweenness centrality and the bottom is for closeness centrality.

sampling a small fraction of the network. In this paper, we adjust the costs incurred by XS and show that simpler strategies, such as random walk sampling, when leveraging degree information collected during the sampling process, yield estimates that are as accurate as the ones obtained through XS or other competing methods.

Avrachenkov *et al.* [17] present the advantages of the use of random walks to estimate page ranks, and study multiple variations over the standard random walk strategy. In a related paper [18], the authors discuss the use of random walks to estimate the top k nodes with highest page rank. In this paper, in contrast, we are interested in degree centrality, as well as other centrality metrics. In addition, while [17] and [18] focus on the web graph, in this paper we consider a more diverse set of topologies.

VI. CONCLUSION

A well known way to find the most central nodes in a network consists of coupling random walk sampling (or one of its variants) with a method to identify the most central nodes in the subgraph induced by the samples. Although it is commonly assumed that degree information is collected during the sampling step, in previous works this information has not been used at the identification step [10], [18]. In this paper, we showed that using degree information at the identification step in a very naive way, namely setting the degree as an alias to other centrality metrics, yields promising results.

REFERENCES

[1] Wikipedia, "Centrality — Wikipedia, the free encyclopedia," [Online; accessed 8-May-2011].

[2] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM KDD*, vol. 1, March 2007.

[3] Network Data, http://www-personal.umich.edu/~mejn/netdata/.

[4] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *ACM SIGKDD*. New York, NY, USA: ACM, 2005, pp. 177–187.

[5] M. J. Newman, "A measure of betweenness centrality based on random walks," *Social Networks*, vol. 27, no. 1, pp. 39 – 54, 2005.

[6] Wikipedia, "Spearman's rank correlation coefficient — Wikipedia, the free encyclopedia," [Online; accessed 8-May-2011].

[7] ——, "Skewness — Wikipedia, the free encyclopedia," [Online; accessed 8-May-2011].

[8] K. Avrachenkov, B. F. Ribeiro, and D. F. Towsley, "Improving random walk estimation accuracy with uniform restarts," in *WAW'10*, 2010.

[9] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," in *IMC*. ACM, 2010, pp. 390–403.

[10] A. S. Maiya and T. Y. Berger-Wolf, "Online sampling of high centrality individuals in social networks," in *PAKDD'10*, 2010.

[11] Y. sup Lim, D. S. Menasche, B. Ribeiro, D. Towsley, and P. Basu, "Online estimation of top k nodes of a network," in *Technical Report*, 2011. [Online]. Available: http://gaia.cs.umass.edu/networks/papers/nsw-2011.pdf

[12] D. Bader, S. Kintali, K. Madduri, and M. Mihail, "Approximating betweenness centrality," in *Algorithms and Models for the Web-Graph*. Springer Berlin, 2007, vol. 4863, pp. 124–137.

[13] D. Eppstein and J. Wang, "Fast approximation of centrality," in *SODA*, Philadelphia, PA, USA, 2001, pp. 228–229.

[14] S. Y. Chan, I. X. Leung, and P. Liò, "Fast centrality approximation in modular networks," in *ACM workshop complex networks*, 2009.

[15] K. Okamoto, W. Chen, and X.-Y. Li, "Ranking of closeness centrality for large-scale social networks," in *FAW'08*, 2008.

[16] S. P. Borgatti, K. M. Carley, and D. Krackhardt, "On the robustness of centrality measures under conditions of imperfect data," *Social Networks*, vol. 28, no. 2, pp. 124 – 136, 2006.

[17] K. Avrachenkov, N. Litvak, D. Nemirovsky, and N. Osipova, "Monte carlo methods in pagerank computation: When one iteration is sufficient," *SIAM J. Numer. Anal.*, vol. 45, pp. 890–904, February 2007.

[18] K. Avrachenkov, N. Litvak, D. Nemirovsky, E. Smirnova, and M. Sokol, "Monte Carlo Methods for Top-k Personalized PageRank Lists and Name Disambiguation," INRIA, Tech Report RR-7367, 09 2010.