Online Sampling of High Centrality Individuals in Social Networks

Arun S. Maiya and Tanya Y. Berger-Wolf

Department of Computer Science
University of Illinois at Chicago
851 S. Morgan, Chicago, IL 60607, USA
{amaiya2,tanyabw}@uic.edu

Abstract. In this work, we investigate the use of online or "crawling" algorithms to sample large social networks in order to determine the most influential or important individuals within the network (by varying definitions of network centrality). We describe a novel sampling technique based on concepts from expander graphs. We empirically evaluate this method in addition to other online sampling strategies on several real-world social networks. We find that, by sampling nodes to maximize the expansion of the sample, we are able to approximate the set of most influential individuals across *multiple* measures of centrality.

1 Introduction and Motivation

Given a large or even massive social network, how can one efficiently identify the most important or influential individuals *without* complete access to the entire network at once? This scenario arises when the network is too large for conventional analysis to be computationally feasible. It may also arise in the context of mining data from a network whose complete structure is hidden from public view (such as a friendship network in Web-based social media) or has a highly distributed structure (such as the network of blogs or the Web itself). The efficient identification of influential individuals (by varying definitions of influence) in large social networks has many applications, from the prevention of computer worms to viral marketing. In this work, we investigate the use of online sampling in identifying such critical individuals.

Online Sampling of Centrality. One of the key tasks in social network analysis is determining the relative importance of individuals based on their positions in the structure of the network [1, 2]. This is referred to as the *centrality* of individuals, and there are many notions of what it means to be *central* to a network [2]. In a sampling approach to centrality approximation, a subset of the individuals in the network is sampled, and an induced subgraph consisting only of these individuals and the links among them is produced. The centrality computation, then, is performed on this induced subgraph *instead* of the entire network, with the centrality scores of the sample being used as approximations of the true centrality of sampled individuals. It is clear that, for this approach to

M.J. Zaki et al. (Eds.): PAKDD 2010, Part I, LNAI 6118, pp. 91–98, 2010.

be useful, two criteria must be met. First, the sampling method must produce an induced subgraph that is representative of centrality in the original network (e.g. the centrality ranking in the sample should be consistent with that of the original network). Ideally, the sampling method will quickly find high centrality nodes, and these nodes will also be ranked highly when computing centrality on the sample. A second criterion is that the sampling method must be an *online* algorithm, since the network may be too large for its global structure to be accessed in its entirety or the global view of the entire network may be limited. By *online*, we mean that the sampling is produced in an iterative, sequential manner, without a priori access to the entire input (*i.e.*, sampling via crawling); at each iteration the new addition to the sample is based on the properties of the nodes crawled thus far. The next node selected for inclusion in the sample is always chosen from the set of nodes connected directly to the current sample. This is also referred to as *snowball sampling* or *neighborhood sampling.* In this work, we systematically investigate the task of *online sampling* of centrality in large social networks. We show that, by sampling nodes to maximize the *expansion* of the sample, the set of most influential individuals can be approximated across *three* different centrality measures: betweenness, closeness, and eigenvector centrality. Remarkably, these sets of top ranked individuals can be approximated reasonably well with sample sizes as small as 1%.

2 Background and Related Work

Centrality in Networks. The idea that the structural position of an individual in a social network may be correlated with the relative influence or importance of that individual was first postulated by Bavelas in the 1940s in the context of organizational communication [3,1]. Since then, many notions of what it means to be important or *central* in a network have been proposed and applied with great success in a variety of different contexts (e.g. [4,5]). In this paper we focus on three widely-used measures: *betweenness centrality*, *closeness centrality*, and *eigenvector centrality.* The *betweenness* of a node is defined as the fraction of the overall shortest paths passing through a particular node [6,7,1]. The *closeness* of an individual in a network is a function of the inverse of the average distance to every other individual [1]. Finally, *eigenvector centrality* is a measure of prestige or popularity proposed by Bonacich [8]. When respresenting the entire network (or graph) as an adjacency matrix $\boldsymbol{A}$, the eigenvector centrality of individuals in the network is the eigenvector $\boldsymbol{x}$ of matrix $\boldsymbol{A}$ corresponding to the largest eigenvalue λ [8]. The PageRank measure [5] used by Google to rank search results is, in fact, simply a variant of eigenvector centrality and has been shown to be highly effective in approximating the prestige and authority of Web pages.

Related Work. Web crawlers, programs that traverse the Web and index pages for search engines in an automated manner, can be viewed as *online* sampling algorithms. Although the goal of Web crawlers is to collect and store the Web link graph for offline processing, it is highly advantageous for crawling algorithms to seek out high PageRank pages early in the crawl [9]. As a result, there have been several studies evaluating these algorithms in their ability to sample PageRank (e.g. [9,10,11,12]). In Section 4.2, we evaluate the performance of several of these algorithms in the context of undirected social networks and in their ability to sample alternative notions of network centrality. Finally, also related to this work are the existing studies on representative subgraph sampling such as [13,14,15].

3 Proposed Method

We employ an *online* sampling algorithm to sample individuals in the social network. Let $G = (V, E)$ be a *network* or *graph* where V is set of vertices (or nodes) and $E \subseteq V \times V$ is a set of edges (or links between the nodes). We begin by selecting a single individual $v \in V$ uniformly at random and specifying a desired sample size k where $k \ll V$. The sample $S \subset V$, then, is initialized to $\{v\}$. Each subsequent individual selected for inclusion in the sample is chosen from the current neighborhood $N(S)$, where $N(S) = \{w \in V - S : \exists v \in S \; s.t. \; (v, w) \in E\}$. Next, upon constructing the sample S, a centrality measure is computed on the induced subgraph of the sample, $G(S)$. The critical question is how to select individuals from $N(S)$ such that:

1. The ranking of individuals by centrality scores in $G(S)$ corresponds to the ranking of individuals in G (most importantly for the highest centrality individuals, as they are generally of greater interest).
2. The highest centrality individuals are quickly included in the sample.

Moreover, as an *online* algorithm, these selection decisions must be made solely on the basis of local information (i.e. information obtained only from those individuals already crawled). In the following sections, we describe several different approaches to online sampling.

Expansion Sampling. We now describe a novel sampling technique based on the concept of expansion in graphs [16]. We refer to this method as *expansion sampling* (XS). The *expansion* of a sample S is defined as $\frac{|N(S)|}{|S|}$. In this approach, we seek out the sample S of size k with the maximal expansion: $\text{argmax}_{S:\,|S|=k} \frac{|N(S)|}{|S|}$. We propose a simple algorithm that greedily selects nodes in order to maximize the expansion of the current sample. That is, the next node v selected for inclusion in the sample is chosen based on the expression: $\text{argmax}_{v \in N(S)} |N(\{v\}) - (N(S) \cup S)|$.

Web Crawlers. As mentioned previously, the process of web crawling is essentially an online sampling process. There are several crawlers explicitly designed to include high PageRank nodes into the sample more quickly. One such approach is referred to as *Backlink Count* (BLC) in which the next node selected for inclusion into the sample is the node with the most links to nodes already in the sample [11]. A second approach is referred as the *OPIC* algorithm [10]. In this approach, all individuals are assigned a default “cash” value. When a node is included in the sample, its cash is distributed to its neighbors in equal proportion. The next node selected for inclusion into the sample, then, is a function of the sum of cash a node has received from its neighbors. We evaluate both these methods not only in their ability to sample PageRank (or Eigenvector Centrality), but also other centrality measures.

BFS, DFS, and Random Walks. As a basis for comparison, we evaluate the performance of several basic approaches to online sampling. Specifically, we evaluate the breadth-first search (BFS), the depth-first search (DFS), and sampling based on a random walk (RW) of the social network.

4 Evaluation

4.1 Experimental Setup

Datasets. We evaluate four diverse, real-world social networks[1]. These include a co-authorship network (Cond-Mat [17]), an email network (Enron [18]), an online trust network (Epinions [19]), and an online social network (Slashdot [20]).

Sampling Methodology. As described earlier, our aim is to sample a *minute* fraction of the individuals in the original network in a way that the individuals ranked highly in the entire network are both present *and* ranked highly in the induced subgraph of the sample. We execute each sampling algorithm on each dataset and sample up to 5% of the nodes. Ten samples are produced by each algorithm on each dataset. During the sampling process, at one percent intervals (e.g. $1\%, 2\%, \ldots, 5\%$), we compute centrality and evaluation statistics on the induced subgraph of the sample and compare those with the original network (evaluation criteria are described in the next section). When evaluating eigenvector centrality, we employ the PageRank variant (PageRank is a variant of eigenvector centrality, as described in Section 2).

Measuring Sample Quality. We employ two evaluation criteria to measure the quality of samples. First, we perform a rank correlation between the centrality ranking in the sample and the true centrality ranking in the original network using Kendall's Tau [21]. This measure (which ranges in value from -1 to 1) evaluates the extent to which the relative ordering by centrality of all nodes in the sample is consistent with that of the original network. A value of 1 indicates the rankings are perfectly consistent, and a value of -1 indicates they are inversely consistent. However, in most cases, it is the set of *high* centrality individuals in the network that is of the most interest. A sample exhibiting a high rank correlation, but consisting only of *low* centrality individuals is not of interest in most cases. Therefore, we employ a second, more informative, evaluation criterion based on the Jaccard measure of set similarity [22], which is the size of

[1] For all networks, we extract the giant component as an undirected, unweighted graph.

the intersection of two sets divided by the size of the union. A Jaccard similarity of 1 indicates that all the elements are shared between the sets, and a score of 0 indicates that none of the elements are shared. We take the top k individuals in the sample centrality ranking and the top k individuals in the ranking of the original network and measure the Jaccard set similarity. This measure indicates how well each sampling algorithm is able to determine the *identity* of the top k highest centrality individuals. For our experiments, we use $k = 50$.

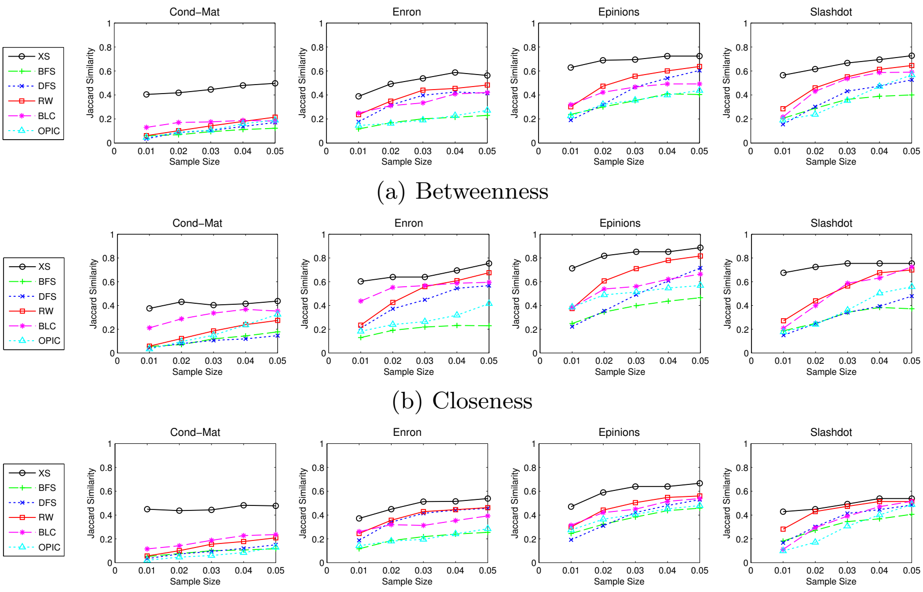
4.2 Experimental Results

Identifying High Betweenness Individuals. Figure 1a shows the extent to which each sampling algorithm is able to identify individuals with the highest betweenness scores in the network. On all datasets, the *expansion sampling* algorithm consistently (and significantly) outperforms all other approaches. Recall that, in *expansion sampling*, nodes selected for inclusion in the sample are those that maximize the expansion of the sample at each step: $\text{argmax}_{v \in N(S)} |N(\{v\}) - (N(S) \cup S)|$. By sampling nodes that contribute most to the expansion, the algorithm seeks out individuals with the most dissimilar neighborhood. These nodes, then, should be brokers or bridges *between* different neighborhoods or clusters of individuals, which captures the notion of betweenness.

The random walk sampling, in some cases, also identifies high betweenness fairly well (though, not as well as *expansion sampling*). Intuitively, high betweenness individuals will appear with high frequency on the paths between other individuals. Performing a random walk, then, should tend to traverse through high betweenness nodes, thereby, including them in the sample. Finally, we see that BFS and OPIC sampling perform particularly poorly when being used to find high betweenness individuals.

Identifying High Closeness Individuals. As shown in 1b, *expansion sampling* also outperforms other methods in its ability to sample high closeness individuals. Recall that closeness centrality is a measure of how close an individual is to all other individuals in the network. Why should *expansion sampling* also perform best on this centrality measure? Consider a sample S with high expansion (i.e. $\frac{|N(S)|}{|S|}$ is significantly large). This means that the sample S, as a whole, is one hop away from many other individuals in the network, which, by definition, is closeness. In addition, as with betweenness, BFS sampling again performs poorly in identifying high closeness individuals.

Identifying High PageRank Individuals. Figure 1c shows the performance of each sampling technique in sampling PageRank. Surprisingly, *expansion sampling* again outperforms other methods (although, not as dramatically as with betweenness and closeness). This result is quite unexpected, as the Web crawling algorithms have been specifically designed to sample high PageRank individuals quickly. Moreover, one would expect a random walk sampling strategy to perform much better (if not the best) because, by the very formulation of PageRank centrality, high PageRank individuals will tend to be visited more frequently by



(a) Betweenness

(b) Closeness

(c) PageRank

Fig. 1. Jaccard set similarity between Top 50 of original network and Top 50 of sample for each centrality measure on each dataset. For all datapoints, standard error is very low, and, for ease of illustration, the standard error bars are omitted. **Key:** *XS = Expansion Sampling, BFS = Breadth-First Search, DFS = Depth-First Search, RW = Random Walk, BLC = Backlink Count, OPIC = OPIC algorithm.*

a random walker. But, clearly, this property holds only in the limit and not for a small sample. Overall, once again, it is *expansion sampling* that identifies high PageRank individuals most effectively.

On the Concordance Across Centrality Measures. Thus far, we have shown that *expansion sampling* performs best in identifying influential individuals by *all three* centrality measures evaluated. A single sample produced by the *expansion sampling* algorithm, then, includes individuals ranking highly on multiple centrality measures. This begs the question: to what extent do these different centrality measures coincide or correspond with one another in real-world, social networks? That is, does *expansion sampling* simply find individuals that simultaneously rank highly on all three of the centrality measures? Upon closer inspection, we find this to *not* always be the case. Although we do find some agreement or overlap across the centrality measures (i.e. there exist individuals ranking highly on multiple measures), *expansion sampling* does, in fact, find individuals over and above this overlap. We compute the Jaccard similarity between the sets of the top 50 ranked individuals in the *entire* network by each centrality measure (i.e. the similarity between each of the *true*, *global* centrality rankings). For instance, in the Enron dataset, we find that the Jaccard similarity of the betweenness ranking and the closeness ranking (both of the entire Enron network) is 0.37. However, as shown in Figure 1b, the Jaccard similarity between the closeness ranking of the entire network and the closeness ranking of the sample produced by *expansion sampling* ranges from 0.6 to 0.75 (over and above 0.37). This indicates that the *expansion sampling* method does, in fact, find individuals ranked highly only on closeness and not on betweenness. Using similar analysis, *expansion sampling* also finds individuals ranking highly on betweenness and not on closeness. Although *expansion sampling* does indeed identify individuals ranking highly on multiple measures, remarkably, this method also finds individuals that each rank highly on *different* measures of centrality. This is striking, as one would not expect a *single*, biased sampling strategy to be able to find high ranked nodes across *multiple* (and diverse) measures of centrality. The *expansion sampling* method, however, does just this.

Consistency in Relative Ordering of Sample and Original Network. Finally, we show the Kendall's Tau rank correlation between the centrality ranking of the samples and the true centrality ranking in the original network (over all individuals in the sample). Due to space constraints, we only show the results for the Enron network in Figure 2. Although results vary slightly across measures and datasets, all sampling algorithms seem to exhibit a relatively strong consistency in the relative ordering of sampled individuals by centrality in comparison to the true centrality ranking. The key to effectively identifying the top ranked individuals, then, is finding and including them early in the sampling process. As discussed in Sections 4.2, 4.2, and 4.2, it is *expansion sampling* that is most effective in this regard.

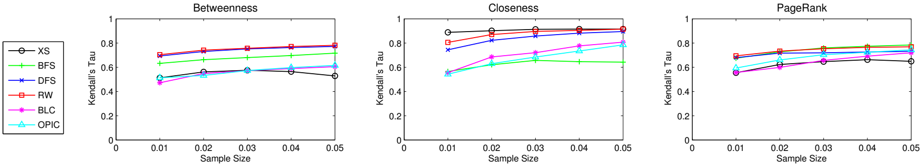


Fig. 2. [Enron Dataset] Kendall's Tau rank correlation between the sample centrality rankings and the true ranking in the original network for each centrality measure

5 Conclusion

In this work, we have studied the use of *online sampling* to identify the set of individuals exhibiting the highest centrality in large social networks. We showed that, by sampling nodes to maximize the *expansion* of the sample, the set of most influential individuals can be approximated across multiple centrality measures. For future work, we plan to investigate the effect of network and graph-theoretic properties on the performance of these and other sampling strategies.

References

1. Freeman, L.C.: Centrality in social networks. Social Networks 1, 215–239 (1979)
2. Wasserman, S., Faust, K., Iacobucci, D.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (November 1994)
3. Bavelas, A.: Communication patterns in task-oriented groups. J. Acoustical Soc. of Am. 22(6), 725–730 (1950)
4. Russo, T., Koesten, J.: Prestige, centrality, and learning: A social network analysis of an online class. Communication Education 54(3), 254–261 (2005)
5. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab (1998)
6. Anthonisse, J.: The rush in a graph. Mathematische Centrum, Amsterdam (1971)
7. Freeman, L.: A set of measures of centrality based on betweenness. Sociometry 40, 35–41 (1977)
8. Bonacich, P.: Power and centrality: A family of measures. American J. Sociology 92(5), 1170–1182 (1987)
9. Boldi, P., Santini, M., Vigna, S.: Paradoxical effects in pagerank incremental computations. In: Workshop on Web Graphs (2004)
10. Abiteboul, S., Preda, M., Cobena, G.: Adaptive on-line page importance computation. In: WWW (2003)
11. Cho, J., Molina, H.G., Page, L.: Efficient crawling through url ordering. Computer Networks and ISDN Systems 30(1-7), 161–172 (1998)
12. Najork, M.: Breadth-first search crawling yields high-quality pages. In: WWW 2001 (2001)
13. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. In: KDD 2005 (2005)
14. Krishnamurthy, V., Faloutsos, M., Chrobak, M., Cui, J., Lao, L., Percus, A.: Sampling large internet topologies for simulation purposes. Computer Networks 51(15), 4284–4302 (2007)
15. Hubler, C., Kriegel, H.P., Borgwardt, K., Ghahramani, Z.: Metropolis algorithms for representative subgraph sampling. In: ICDM 2008 (2008)
16. Hoory, S., Linial, N., Wigderson, A.: Expander graphs and their applications. Bull. Amer. Math. Soc. 43, 439–561 (2006)
17. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution: Densification and shrinking diameters. ACM TKDD 1(1), 2 (2007)
18. Shetty, J., Adibi, J.: Enron email dataset. Technical report (2004)
19. Richardson, M., Agrawal, R., Domingos, P.: Trust Management for the Semantic Web. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 351–368. Springer, Heidelberg (2003)
20. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Statistical properties of community structure in large social and information networks. In: WWW 2008 (2008)
21. Kendall, M., Gibbons, J.D.: Rank Correlation Methods, 5th edn. (September 1990)
22. Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin del la Société Vaudoise des Sciences Naturelles 37, 547–579 (1901)