Empirical Studies on Methods of Crawling Directed Networks

Junjie Tong, Haihong E , Meina Song and Junde Song
School of Computer, Beijing University of Posts and Telecommunications
Beijing, 100876, P. R. China

Abstract

Online Social Network has attracted lots of academies and industries to look into its characteristics, models and applications. There are many methods for crawling or sampling in networks, especially for the undirected networks. We focus on sampling the directed networks and intend to compare the efficiency, the accuracy and the stability between them. We consider the sampled nodes and links as a whole and separated from the original one. We evaluate experiments by deploying the snow ball method, the random walk method, DMHRW and MUSDSG with different sampling ratios on the datasets. The snow ball method and random walk method both have bias towards low outdegree nodes while the snow ball method tends to sample more hub nodes. DMHRW and MUSDSG can sample the network parallel but more complex than the snow ball and the random walk under the same sampling ratio. DMHRW will be the best choice of all while the computation capability and time are sufficient.

Keywords: *Sampling Method, Directed Networks, Measurements, Graph Sampling.*

1. Introduction

In recent years, the population of Online Social Networks (OSNs) has experienced an explosive increase. Twitter for example, has attracted more than 600 million individuals by August 2012 [1] counted by Twopcharts. The world-wide spreading of OSNs has motivated a large number of academies and researchers do studies and researches on the analysis and model on the structures and characteristics of OSNs. However, the complete dataset is typically not available for privacy and economic considerations at some extent. Therefore, a relative small but representative subset of the whole is desirable in order to study properties, characteristics and algorithms of these OSNs. How to get the relative small but representative subset of the whole accurately, efficiently and stably becomes an important problem.

Various graph sampling algorithms have been proposed for producing a representative subset of OSNs users. Currently, the algorithms for crawling OSNs can be roughly divided into two main categories: graph traversal based and random walk based. For the graph traversal based methods, each node in the connected component is visited only once, if we let the process run until completion. For the random walk based methods, they allow node re-visiting. BFS, in particular, is a basic technique that has been used extensively for sampling OSNs in the past studies [2, 3, 4]. And the comparison between the graph traversal based methods and the random walk based methods can be shown in Table 1. And we denote the graph traversal based methods as T and the random walk based methods as W in the method type column.

Table 1. Comparisons Between Main Methods

Method Name	*Method Type(T/W)*	*For Directed/Undirected Networks*	*Biased/Unbiase d*
BFS	T	Both	Towards high degree nodes [5], underestimate the level of symmetry [6]
Snow-ball [7]	T	Both	Underestimate the power-law coefficient [5]
Random Walk [8]	W	Both	Towards high degree nodes
Metro-Hastings RW [9]	W	Undirected	Unbiased
Re-Weighte d RW [8]	W	Undirected	Unbiased

There are many other random walk based methods in [8]. Although the random walk bases methods may be biased but the bias can be analyzed using classic results from Markov Chains and corrected by re-weighting the estimators [10].

Currently, there are a lot of works on new unbiased sampling method and bias analysis on existing sampling methods. The comparisons between BFS, Metro-Hastings RW and Re-Weighted RW in crawling undirected Facebook can be seen in [11, 12]. USDSG has been proposed for unbiased sampling in directed ONSs [13]. And the most widely used baseline sampling method is UNI which is usually called ground truth. This simple method is a textbook technique known as rejection sampling [14] and in general it allows to sample from any distribution of interest with some limitations [11].

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.

For the efficient crawling of OSNs, sometimes we have to adopt parallel processing which not only benefits the efficient but also overcomes the limitations such as capabilities on computation and storage and API requesting times etc. For example, the snow ball sampling and random walk sampling often choose several initiative nodes to start. For implementing MHRW and RWRW, we have to consider the convergence more seriously. And the Geweke diagnostic [15] and the Gelman-Rubin diagnostic [16] are widely used.

The focus of our work is how to crawling the directed networks accurately, efficiently and stably. We describe and implement several crawling methods on crawling directed networks not only limited to online social networks under practical experiment. And compare the stability, efficiency and bias between them. Our main contributions are the followings but not limited:

- Modifying and implementing the crawling algorithms in directed networks. Most of the current methods especially the unbiased methods are for crawling on undirected networks. And we implement the algorithms in various directed networks not only to the directed OSNs.
- Crawling without knowing the whole set. Most implementations of current sampling methods treat sampled nodes separately and focus mainly on the degree distribution but ignore the topology characteristics such as clustering coefficient and diameter.
- Comparisons between several crawling algorithms in various aspects which include efficiency, accuracy and stability.

2. Crawling Methodology

2.1 Scope and Assumptions

The directed networks can be modeled as a directed graph G=(V,E) , where V is a set of nodes (users) and E is a set of edges (denote relationships of some type). Let k_v^{in} and k_v^{out} be the indegree and outdegree of node v. Let k^{in} denotes the average indegree and r denote the sample ratio which is set before starting the crawling. N denotes the node number of the whole set and M denotes the edge number of the whole set. In this paper:

- We begin our crawling at some initial nodes without knowing the whole set so when we analyze the characteristics of the sampled networks we just consider the crawled nodes and corresponding connections between them.
- The initial nodes are selected in the largest SCC (strongly connect components) of each dataset for fast crawling and feasible experiments.
- We do not consider the missing links and implicit links. The datasets which we used are considered fully collected. We just consider the static snapshot of each dataset and do not consider its dynamics.

2.2 Sampling Methods

The crawling of the directed graph starts with one or several initial nodes and proceeds iteratively even in parallel. In every operation, we visit a node and discover all its neighbors. There are many ways, depending on the particular sampling method, in which we can proceed. In this section, we describe the crawling methods we implemented and compared in this paper.

1) Snow ball Sampling: Snow ball sampling typically selects one node as the seed node to start and processes iteratively. At each new iteration the neighbors of sampled nodes but not-yet-visited are explored. As BFS is widely used in graph traversal, snow ball sampling is an incomplete BFS to cover only some specific region of the graph. Sometimes, several seed nodes are selected for efficient crawling. The snow ball sampling method we adopted can be depicted as follows in Figure 1. And the above algorithm can be easily extended in choosing more than one seed node and processing parallel. The selection of the seed node will be described later in this section.
2) Random Walk: In the classic random walk [8], the next-hop node is chosen uniformly at random among the neighbors of the current node. And the classic random walk sampling is biased towards high degree nodes. The random walk smapling method we implemented as follows as in Figure 2.
3) Directed Metro-Hastings Random Walk (DMHRW): We modify the existing MHRW which is usually used in sampling undirected networks for crawling the directed networks. As the choice of the next-hop node depends on the ratio of indegree and outdegree of currentNode as depicted in Figure 3, it bias towards nodes with high indegree and low outdegree.

And While implementing this method, we use multiple parallel walks and we have to consider the covergence of the sampling process. We use the Geweke diagnostic to detect the convergence of the sampling process. The details of the Geweke diagnostic will be described later in this section.
And as shown between line 8 and 9 in algorithm 3, the algorithm will exit while current node numbers exceed some point. As in the experiment, we find out that sometimes it will not converge at last while the node number reaches the defined one before starting sampling.

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.

The algorithm can be divided into two parts as algorithm 3 and algorithm 4. And algorithm 4 in Figure 4 calls algorithm 3 and use the Geweke diagnositc after each iteration.

```
**Algorithm 1** : Snow-ball Sampling
**Input:** seed node $v$, sample ratio $r$, network size $N$
**Output:** Sampled network $\mathbf{H}$
$lastVisited \leftarrow NULL$
$nodeList \leftarrow NULL$
$lastVisited \leftarrow v$
$nodeList \leftarrow v$
**while** $len(nodeList) < r * N$ **do**
  **if** $lastVisited$ not NULL **then**
    $neighborList \leftarrow NULL$
    **for all** $v \in lastVisited$ **do**
      **for all** $w \in$ neighbors of $v$ **do**
        add edges between $v$ and $w$ to $\mathbf{H}$
        add $w$ to $neighborList$
        **if** $w \notin nodeList$ **then**
          add $w$ to $nodeList$
        **end if**
      **end for**
    **end for**
    $lastVisited \leftarrow NULL$
    $lastVisited \leftarrow neighborsList$
  **end if**
**end while**
**return** $\mathbf{H}$
```

Fig. 1 Algorithm 1.

```
**Algorithm 2** : Random Walk Sampling
**Input:** seed node $v$, sample ratio $r$, network size $N$
**Output:** Sampled network $\mathbf{H}$
$currentNode \leftarrow v$
$nodeList \leftarrow NULL$
$nodeList \leftarrow v$
**while** $len(nodeList) < r * N$ **do**
  $neighborList \leftarrow NULL$
  **for all** $w \in$ neighbors of $currentNode$ **do**
    add edges between $v$ and $w$ to $\mathbf{H}$
    add $w$ to $neighborList$
    **if** $w \notin nodeList$ **then**
      add $w$ to $nodeList$
    **end if**
  **end for**
  $w =$ randomly chose from $neighborList$
  $currentNode \leftarrow w$
**end while**
**return** $\mathbf{H}$
```

Fig. 2 Algorithm 2.

```
**Algorithm 3** : DMHRW-sub
**Input:** node id $currentNode$,
    node list $nodeList$, edge list $edgeList$
**Output:** $nodeList, edgeList$
$neighborList \leftarrow NULL$
**for all** $w \in$ neighbors of $currentNode$ **do**
  add edges between $currentNode$ and $w$ to $edgeList$
  add $w$ to $neighborList$
  **if** $w \notin nodeList$ **then**
    add $w$ to $nodeList$
  **end if**
**end for**
$u =$ randomly chose from $neighborList$
$m = (k^{in}_{currentNode} + 1)/(k^{out}_{currentNode} + 1)$
$p =$ randomly chose from $(0, 1)$
**if** $p < m$ **then**
  $currentNode \leftarrow u$
**end if**
**return** $nodeList, edgeList$
```

Fig. 3 Algorithm 3.

```
**Algorithm 4** : DMHRW
**Input:** sample ratio $r$, seed list $seedList$, network size $N$
**Output:** sampled network $\mathbf{H}$
$nLists \leftarrow NULL$
$eLists \leftarrow NULL$
**while** $len(nLists) < r * N$ or
Geweke-Diag $(nLists, eLists)$ not successful **do**
  **for all** $v \in seedList$ **do**
    $nLists[v], eLists[v] =$
    DHMRW-sub $(v, nLists[v], eLists[v])$
    Merge $nLists$ for eliminating duplicated nodes
    $nNum =$ number of nodes in merged $nLists$
    **if** $nNum > 1.2 * r * N$ **then**
      Break
    **end if**
  **end for**
**end while**
**for all** $edge \in eLists$ **do**
  **if** $edge \notin \mathbf{H}$ **then**
    add $edge$ to $\mathbf{H}$
  **end if**
**end for**
**return** $\mathbf{H}$
```

Fig. 4 Algorithm 4.

4) Modified Unbiased Sampling for Directed Social Graphs (MUSDSG): We have modified USDSG as the followings: using multiple parallel walks and the Geweke diagnostic to detect the sampling process, adding the neighbors of currentNode to sampled network and setting the upper bound of the sampled node number while the iteration processing is not converged. As we consider the topology of sampled network, we can compare topology in sampled network not just the degree of node and degree distribution which are already compared in [13]. The algorithm is depicted in Figure 5.

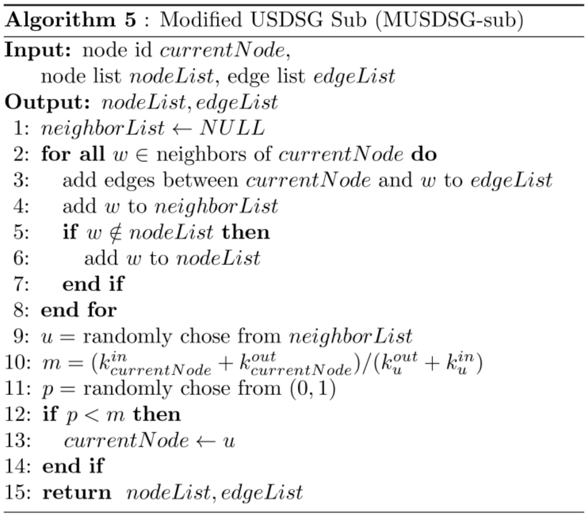
5) Modified Uniform Sampling (MUNI): The UNI [12] method allows us to obtain uniformly random users' ID by generating random IDs in certain space. This

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.

algorithm has limitations. First, the ID space must not be sparse for this operation to be efficient. Secondly, the operation which enables us to verify the user and retrieve the users whom he connects to can be easily implemented. As we have obtained the whole set and the IDs are integers without interval numbers, the ID space is not sparse and we can easily obtain the connections between any two users. We modify UNI as MUNI and add the connections between sampled nodes to form the sampled network. And the algorithm we implemented can be shown as follows in Figure 6.

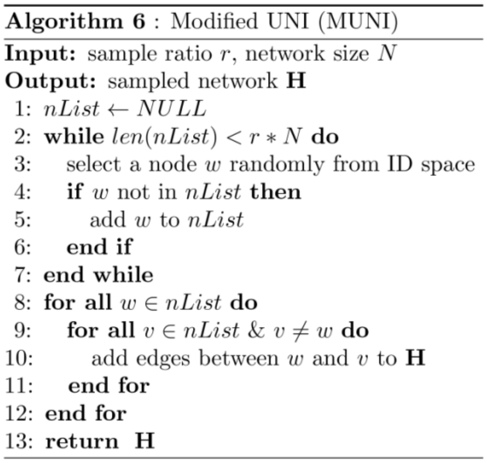
6) The Selection of the Initial or Seed Nodes: All the above sampling or crawling methods have to select one or more than one seed nodes. While we implement snow ball sampling and random walk sampling, we select one seed node randomly from the largest SCC separately. And while we implement DMHRW and MUSDSG, we select several seed nodes randomly from the largest SCC without multiple ones. Although the seed nodes in largest SCC will facilitate the crawling process, the diameter will be underestimated. As the small-world effect [17] leads to the small diameter, this will not affect much.

7) Detecting Convergence with Geweke Diagnostic: While we implement DMHRW and MUSDSG using multiple parallel walks, we have to detect convergence. The Geweke diagnostic detects the convergence of a single Markov chain. Let X be a single sequence of samples of our metric of interest. Geweke considers two subsequences of X, its beginning X_a (typically the first 10%), and its end X_b (typically the last 50%). Based on X_a and X_b, we compute the z-statistic: $z = (E(X_a) - E(X_b))/\sqrt{Var(X_a) + Var(X_b)}$. With increasing number of iterations, X_a and X_b move futher apart, which limits the correlation between them. And we declare rigid convergence when all the values of sequences fall in the [-1,1] interval.



```
Algorithm 5 : Modified USDSG Sub (MUSDSG-sub)
Input: node id $currentNode$,
    node list $nodeList$, edge list $edgeList$
Output: $nodeList, edgeList$
$neighborList \leftarrow NULL$
for all $w \in$ neighbors of $currentNode$ do
    add edges between $currentNode$ and $w$ to $edgeList$
    add $w$ to $neighborList$
    if $w \notin nodeList$ then
        add $w$ to $nodeList$
    end if
end for
$u =$ randomly chose from $neighborList$
$m = (k^{in}_{currentNode} + k^{out}_{currentNode})/(k^{out}_u + k^{in}_u)$
$p =$ randomly chose from $(0, 1)$
if $p < m$ then
    $currentNode \leftarrow u$
end if
return $nodeList, edgeList$
```

Fig. 5 Algorithm 5.



```
Algorithm 6 : Modified UNI (MUNI)
Input: sample ratio $r$, network size $N$
Output: sampled network $\mathbf{H}$
$nList \leftarrow NULL$
while $len(nList) < r * N$ do
    select a node $w$ randomly from ID space
    if $w$ not in $nList$ then
        add $w$ to $nList$
    end if
end while
for all $w \in nList$ do
    for all $v \in nList$ & $v \neq w$ do
        add edges between $w$ and $v$ to $\mathbf{H}$
    end for
end for
return $\mathbf{H}$
```

Fig. 6 Algorithm 6.

3. Datasets and Evaluation Methodology

This section contains two main parts. First, we describe our datasets and measure their topology characteristics. Second, we give details on how to compare or evaluate the different crawling methods.

3.1 The Datasets

We download four datasets from [18] including: Gnutella peer-to-peer network of August 2002 (gpn08 [19]), EU email communication network (eue [19]), Slashdot social network of November 2008 (slashdot01 [20]) and Slashdot social network of February 21 2009 (slashdot02 [20]).

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.

1) Explanations on Measured Characteristics: We have measured many topological characteristics of the networks in the datasets including the diameter, the correlation coefficient between indegree and outdegree r_0 [21], the coefficient of link reciprocity r_1[22, 23], the assortative coefficient r_2 [24], The average clustering coefficent in directed networks c. And r_2 falls in [-1,1]. An interesting observation is that essentially all social networks measured appear to be assortative, but other types of networks (information networks, technological networks, biological networks) appear to be disassortative [25]. And more discussions about clustering coefficent in directed networks are in [26].

2) Measurements on the Datasets: The characteristics are measured with NetworkX [27]. We use it to calculate the diameter and store the network. The characteristics of the datasets are shown in Table 2.

Table 2. Characteristics of the Networks in the Datasets

Network	gpn08	eue	slashdot01	slashdot02
N	6301	265214	77360	82168
M	20777	420045	905468	948464
k^{in}	3.297	1.5838	11.7046	11.543
d	9	14	12	13
effective diameter	6	5	5	5
c	0.015	0.4913	0.0228	0.0164
r_0	1.375	76.076	10.553	10.698
r_1	-0.00052	0.131	0.4391	0.427
r_2	0.194	-0.071	0.0072	0.0018

3.2 Evaluation Methodology

We compare the crawling methods described in section II in three aspects including efficiency, accuracy and stability. Before the description, we first define some parameters.

Each algorithm runs for Num times under certain sampling ratios R. To reach certain sampling ratio $r_i \in R$, each algorithm has to operate literately in t_j times in the jth ($j \leq Num$) sampling. And in num sampling times under certain sampling ratio r_i, the crawling process fails $fail_i$ times by using DMHRW and MUSDSG. While the sampled number exceeds certain upper bound without convergence, we consider it fails. And $c_j, d_j\ and\ k_j^{in}$ denote the average clustering coefficient, the diameter and the average indegree of the jth sampled network.

1) Efficiency: to evaluate the efficiency of the crawling method, we propose two parameters. One is average sampling time under certain sampling ratio defined as $t_{r_i} = \sum_{j=1}^{Num} t_j / Num$. And the other one is the successful ratio under certain sampling ratio which defined as $s_{r_i} = (Num - fail_i)/Num$. And the second one is only for DMHRW and MUSDSG.

2) Accuracy: to consider the accuracy of the crawling algorithm, we compare the node indegree and outdegree distribution, the average clustering coefficient, the diameter and the average node indegree of sampled networks by implementing the above algorithms and MUNI separately under certain sampling ratio.

3) Stability: we have to consider the expectation and standard deviation of the characteristics under certain sampling ratio during the separately Num sampling times by implementing the same algorithm. And we also consider the degree distribution seriously.

4. The Experiment and Analysis

In this section, we will compare snow ball sampling, random walk, DMHRW and MUSDSG in efficiency, accuracy and stability. And while we compare the characteristics of the sampled network, we consider MUNI as the ground truth. We evaluate the experiments of the different sampling methods on the datasets.

4.1 The Ground Truth: MUNI

First, we look into MUNI which is usually considered as the ground truth while comparing different crawling methods. We compare the characteristics under various sampling ratio of the four different directed networks to the whole set in two aspects: its stability and accuracy.

We run MUNI for 10 times under different sampling ratio for the four different datasets or directed networks. And Table 3 shows the expectation *E* and the stand deviation *D* of the characteristics under different sampling ratio for the four directed networks.

1) Accuracy: As shown in Table 3, we can easily compare the characteristics between the sampled networks and the whole network.

For the diameter, while the sample ratio becomes larger and larger, it gets closer and closer to the value of the whole network. And the different between the two is small. And it’s also the same as the average indegree.

But for the clustering coefficient, for the two online social networks (slashdot01 and slashdot02), the value of the

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.

sampled networks is nearly 200 times over the value of the whole network. It turns out that for the density directed network, MUNI is more likely towards the nodes with links to the connected nodes. And it is somehow like the proximity bias of link growth in Flickr [28] and FriendFeed [29].
As shown in Figure 7, the cumulative distribution of indegree and outdegree of the sampled network are totally fit for corresponding distributions of the whole network for slashdot02.

2) Stability: We run MUNI for 10 times under different sampling ratio for the four different dataset. And we can easily figure out from Table III that the standard deviations are small under different sampling ratio. And the standard deviations are often ten percent as the expectations.
The larger the sample ratio is, the closer the expectation is to the real value. But while the ratio is small, the MUNI sampling method may cause significant biases in characteristics such as the clustering coefficient of slashdot01 and slashdot02.

4.2 The Efficiency Comparison

The average sampling time under various sampling ratio are given in Figure 12. From that figure, we can easily figure out that under the same sample ratio, the snow ball method is the fastest but with extra store. And the random walk method is the slowest method. Both the snow ball method and the random walk method cannot be deployed parallel.

For the parallel methods: DMHRW and MUSDSG, DMHRW is usually faster than MUSDSG. But their successful ratios are different. We run DMHRW and MUSDSG for sampling gpn08 three times independently and show the results in Table 4. The successful ratio is low for both DMHRW and MUSDSG and the two methods cost much as parallel deployment. In Figure 8, we plot the convergence with Geweke Diagnostic in the only one successful time while we sampling gpn08 with the sample ratio is 0.4.

4.3 The Accuracy Comparison

From Figure 9, we can easily figure out that the cumulative indegree distributions are similar while using random walk method and snow ball sampling. And they are all close to the distribution of the whole network. But for the cumulative distributions of outdegree, significant bias towards the low outdegree nodes exists by using the random walk method. And from Figure 9, the bias reduces as the increasing of the sample ratio. And it gets closer and closer to the cumulative distribution of the whole network as shown in Figure 10.

After implementing DMHRW and MUSDSG, only one time successful in each one. And Figure 13 shows the CDD of indegree and outdegree of the sampled network in the successful deployment of DMHRW and MUSDSG. Significant bias towards low outdegree nodes exists in both of them.

From Table 5, the data are expectation and standard deviation about the diameter and clustering coefficient under different sample ratios in different datasets. The networks have smaller diameter and larger indegree by using snow ball method other than random walk method no matter to the sample ratio and the network. As the random walk method is likely towards the nodes with low indegree and less connections with other nodes, the diameter and the clustering coefficient is much smaller.

And from Table 4, although the successful ratio is low for both DMHRW and MUSDSG, the diameter and the clustering coefficient are closer to the whole network than snow ball method and random walk method.

4.3 The Stability Comparison

From Table 5, the standard deviation is only about ten percent of the corresponding expectation. For random walk method, the standard deviations decrease as the increasing of the sample ratio. And the deviations are smallest by using DMHRW.

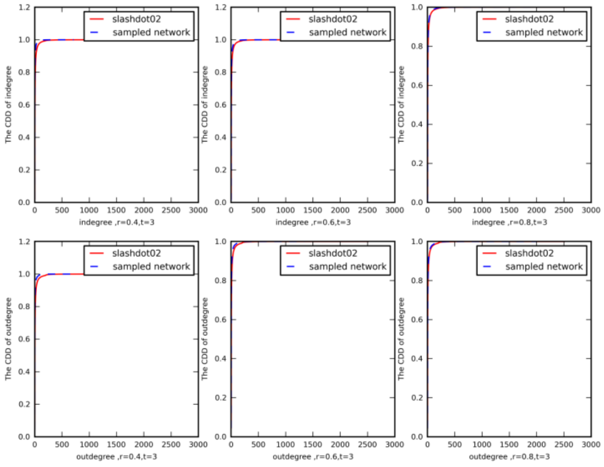


Fig. 7 The CDD of indegree and outdegree of slashdot02 by using MUNI

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.

Table 3. Expectation and Standard Deviation of the Characteristics (MUNI)

Network	r	$E(d)$	$D(d)$	$E(c)$	$D(c)$	$E(k^{in})$	$D(k^{in})$
gpn08	0.4	13.9	0.539	0.00845	0.00176	1.316	0.0449
	0.6	11	0.775	0.009	0.00105	1.979	0.0423
	0.8	10.1	0.7	0.0102	0.000374	2.62	0.0571
	1.0	9	-----	0.015	-----	3.297	-----
slashdot01	0.4	11.7	0.781	2.455	0.0172	5.265	0.157
	0.6	12.4	0.8	2.106	0.0146	7.437	0.132
	0.8	11.6	0.49	1.846	0.0108	9.595	0.130
	1.0	12	-----	0.0228	-----	11.7046	-----
slashdot02	0.4	12.2	0.748	2.302	0.01	5.21	0.110
	0.6	12.2	0.6	1.99	0.0135	7.25	0.135
	0.8	12.2	0.748	1.74	0.00567	9.43	0.0667
	1.0	13	-----	0.0164	-----	11.543	-----
eue	0.4	11.7	0.64	0.215	0.0113	0.616	0.0396
	0.6	13.7	2.002	0.308	0.0114	0.945	0.0241
	0.8	13.75	1.199	0.40	0.00654	1.25	0.0154
	1.0	14	-----	0.491	-----	1.5838	-----

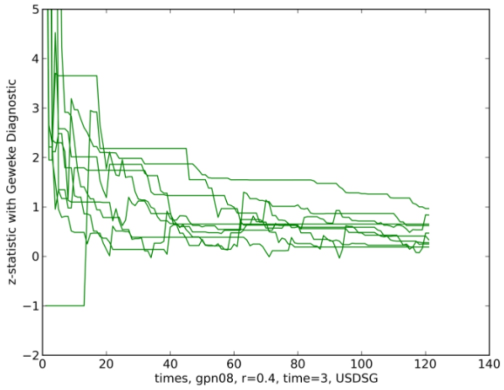


Fig. 8 The convergence with Geweke Diagnostic (MUSDSG)

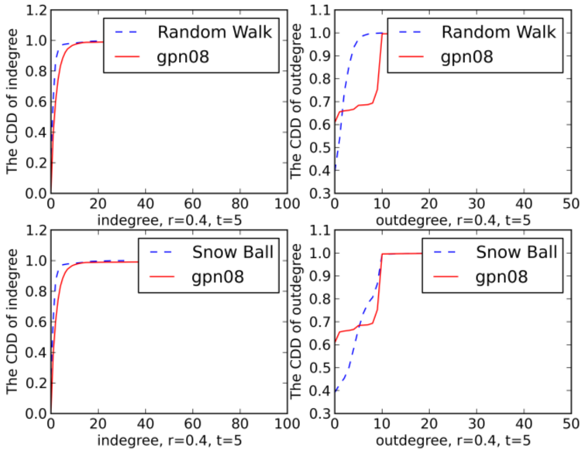


Fig. 9 The CDD of indegree and outdegree(r=0.4, t=5)

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.

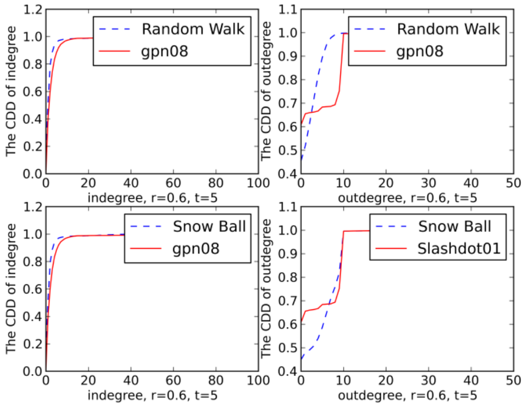


Fig. 10. The CDD of indegree and outdegree(r=0.6, t=5)

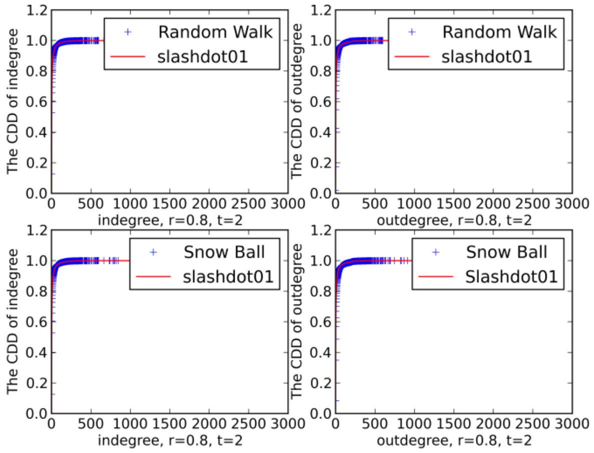


Fig. 11. The CDD of indegree and outdegree for slashdot02

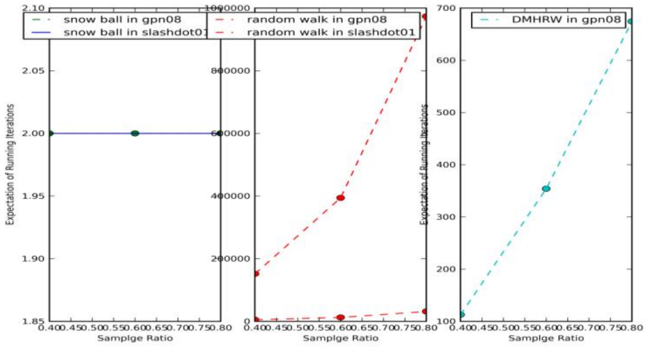


Fig 12. The sampling time under various sampling ratio

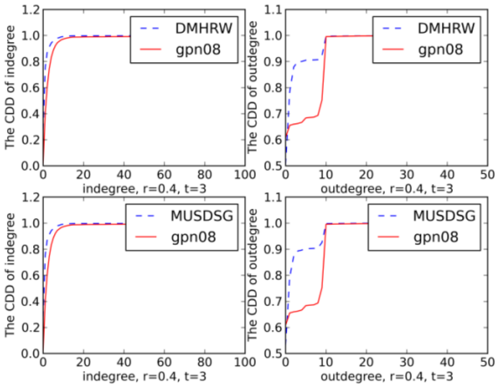


Fig. 13. The CDD of inderee and outdegree(DMHRW and MUSDSG)

Table 4. The successful ratio for DMHRW and MUSDSG in gpn08

r	**Sample Method**	S_r
0.4	DMHRW	33.3%
	MUSDSG	0
0.6	DMHRW	33.3%
	MUSDSG	0

5. Conclusions

In this paper, we have to explore into the sampling or crawling methods on directed networks. And we aim to give some insights into the methods in efficiency, accuracy and stability.

Despite the size of the original network and the difference of the sampling ratio, the MUNI performs well in accuracy and stability compared to different sampling methods. But it requires some conditions, such as getting the whole information of the user, the whole network is not sparse and etc.

The random walk method and snow ball method both have bias towards low outdegree nodes and increasing the sampling ratio can reduce the bias. As the increase of the sampling ratio, the sampling time of random walk increase much faster than

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.

the snow ball method while the snow ball method just needs several iterations.

Table 5. Expectation and Standard Deviation of the Characteristics under Different Sampling Ratio

Network	r	Sample Method	E(d)	D(d)	E(c)	D(c)	E(k^{in})	D(k^{in})
gpn08	0.4	Snow Ball	7	0	0.0224	0.0019	4.049	0.1824
		Random Walk	16	0.67	0.00466	0.00135	1.589	0.0187
		DMHRW	10	0	0.01	0.0019	1.59	0.046
		MUSDSG	9	0	0.0134	0.004	1.745	0.18
	0.6	Snow Ball	7	0	0.0209	0.0139	4.109	0.0278
		Random Walk	12	0	0.007	0.00125	2.138	0.0325
		DMHRW	9	0	0.01	0.001	2.37	0.0048
		MUSDSG	9	0	0.0132	0.0018	2.487	0.047
	0.8	Snow Ball	8	0	0.0134	0.00024	3.76	0.0113
		Random Walk	9	0	0.01	0.0006	2.923	0.0287
	1.0	-----	9	-----	0.015	-----	3.297	-----

MHRW and USDSG can be both deployed parallel and we modified both MHRW and USDSG for investigating into the unbiased sampling method in directed networks. Both the DMHRW and MUSDSG are more complex for computation and we have to detect the convergence while deploying the latter two. Although the successful ratio of DMHRW and MUSDSG is low, the characteristics of the sampled network are closer to the whole network than the random walk method and snow ball method.

For the lack of computation capability and the limits of the time, we suggest use the snow ball and the random walk methods. And sample the network with high sample ratio is much better for inducing the biases. If the computation capability and the time are enough, the DMHRW will be the better choice than MUSDSG with the characteristics closer to the whole network.

While comparing the sampling methods, we assume we know all the nodes and the links explicitly and the network is static. The network may have implicit links and nodes, and it is often dynamic [30]. Our future work will focus on completing the network using its growth mechanisms and the previous characteristics.

Appendix

Proposition I: For a given directed network $G=(V,E)$ with no loops and no self-loops, and convert it to an undirected network G'as the following way: if there is an edge between u and v and the edge (u,v) not in G', then add the edge (u,v) to G'. Then at last it forms an undirected network without loops and $d_{G'} \geq d_G$.

Proof 1: Suppose $d_{G'} < d_G$ and the length of the shortest path between u and v is equal to $d_{G'}$ inG'. As it has supposed that $d_{G'} < d_G$, then in G there is much shorter path ℓ between u and v. While converting G to G', the path is obviously added to G. And that means $d_{G'} < d_{G'}$, it is obviously not right. So our suppose is wrong and $d_{G'} \geq d_G$.

Acknowledgments

This work is supported by the National Kay project of Scientific and Technical Supporting Programs of China (Grant No. 2009BAH39B03); the National Natural Science Foundation of China (Grant No. 61072060); the National High Technology Research and Development Program of China Grant No. 2011AA100706); the Research Fund for the Doctoral Program of Higher Education (Grant No. 20110005120007); the Fundamental Research Funds for the Central Universities (2012RC0205); the Co-construction Program with Beijing Municipal Commission of Education; Engineering Research Center of Information Networks, Ministry of Education..

References

[1] http://twopcharts.com/twitter500million.php 2012-8-19.

[2] A. Mislove, M. Marcon, P. K. Gummadi and P. Druschel et al., Measurement and Analysis of Online Social Networks, IMC, 2007, pp. 29-42.

[3] Y. Ahn, S. Han, H. Kwak and S. Moon et al., Analysis of Topological Characteristics of Huge Online Social Networking Seervices, WWW, 2007, pp. 835-844.

[4] C. Wilson, B. Boe, A. Sala and K. P. Puttaswamy, User Interactions in Social Networks and Their Implications, EUROSYS, 2009, pp. 205-218.

[5] S. H. Lee, P. J. Kim, H. Jeong, Statistical Propertities of Sampled Netowrks, PHYS REV E, 2006, vol. 73, no. 1.

[6] L. Becchetti, C. Castillo, D. Donato and A. Fazzone, A Comparison of Sampling Techniques for Web Graph Characterization, LinkKDD, 2006.

[7] S. K. Thomson, Sampling, John Wiley & Sons, Inc., New York, 2002.

[8] L. Lovasz, Random Walks on Graphs: A Survey, Journal of Combinatorica, 1996.

[9] N. Metropolis, A. Rosenbluth, M. Rosenbluth and A. Teller et al., Equation of State Calculations by Fast Computing Machines, J. Chemical Physics, vol. 21, 2004, pp. 1087-1092.

[10] A. H. Rasti, M. Torkjazi, R. Rejaie and N. G. Duffield, Respondent-Driven Sampling for Characterizing Unstructured Overlays, INFOCOM, 2009, pp. 2701-2705.

[11] M. Gjoka, M. Kurant, C. Butts and A. P. Markopoulou, Walking in Facebook: A Case Study of Unbiased Sampling of OSNs, INFOCOM, 2010, pp. 2498-2506.

[12] M. Gjoka, M. Kurant, C. Butts and A. Markopoulou, Practical Recommendations on Crawling Online Social Networks, Journal on Selected Areas in Communications, vol. 29, no. 9, 2011, pp. 1872-1892.

[13] T. Y. Wang, Y. Chen, Z. B. Zhang and P. Sun et al., Unbiased Sampling in Directed Social Graph, ACM SIGCOMM, 2010, pp. 401-402.

[14] A. Leo-Garcia, Probability and Random Processes for Electrical Engінerring, 2nd ed, MA: Addison-Wesley Publishing Company, Inc., 1994.

[15] J. Geweke, Evaluating the Accuracy of Sampling-based Approaches to Calculating Posterior Moments, in Baysesian Statist. 4, 1992.

[16] A. Gelman and D. Rubin, Inference From Iterative Simulation Using Multiple Sequences, Statist. Sci., vol. 7, 1992.

[17] C. Karte, S. Milgram, Acquaintance Linking between Whie and Negro Populations: Application of the Samll World Problem, Journal of Personality and Social Psychology, vol. 15, no. 2, 1970, pp. 101-108.

[18] http://snap.stanford.edu/data/index.html, 2012-8-19

[19] J. Leskovec, J. Kleinberg and C. Faloutsos, Graph Evolution: Densification and Shrinking Diameters, ACM TKDD, vol. 1, no. 1, 2007, Articale 2.

[20] J. Leskovec, K. J. Lang, A. Dasgupta and M. W. Mahoney, Community Structure in Large Networks: Natural Cluster Size and the Absence of Large Well-Defined Clusters, Internet Mathematics, vol. 6, no. 1, 2009, pp. 29-123.

[21] R. Albert, A. L. Barabasi, Statistical Mechanics of Complex Networks, Reviews of Modern Physics, vol. 74, no. 1, 2002, pp. 47-97.

[22] D. Garlaschelli, M. I. Loffredo, Patterns of Link Reciprocity in Directed Networks, Physical Review Letters, vol. 93, no. 26, 2004, 268701.

[23] F. Zhao, T. Zhou, L. Zhang and M. H. Ma et al., Research Progress on Wikipedia, J. Univ. Electron. Sci. Technol., vol. 39, no. 3, 2010, pp. 321-334.

[24] M. E. J. Newman, Mixing Patterns in Networks, Phys. Rev. E, vol. 67, no. 2, 2003, 026126.

[25] M. E. J. Newman, The Structure and Function of Complex Network, SIAM Review, vol. 45, no. 2, 2003, pp. 167-256.

[26] G. Fagiolo, P. Martiri, Clustering in Complex Directed Networks, Physical Review E, vol. 76, no. 2, 2007, 026107.

[27] http://netowrkx.lanl.gov/#.

[28] A. Mislove, M. Marcon, P. K. Gummadi and P. Druschel et al., Growth of the Flickr Social Network, IMC, 2007, pp. 29-42.

[29] S. Garg, T. Gupta, N. Carlsson and Anirban Mahanti, Evolution of an Online Social Aggregation Network: An Empirical Study, IMC, 2009, pp. 315-321.

[30] J. Kunegis, D. Fay and C. Bauckhage, Netowrk Growh and the Spectral Evolution Model, CIKM, 2010, pp. 739-748.

Junjie Tong received his bachelor degree in computer science from China University of Mining and Technology in Beijing, and now is a Ph.D. candidate at the Department of Computer Science and Technology of Beijing University of Posts and Telecommunication. His research interests include CDN and complex networks.

Haihong E is a Ph.D., Lecturer at the Department of Computer Science and Technology of Beijing University of Posts and Telecommunication. Her research interests include service sciences and engineering, service network, and trusted service.

Meina Song is a Ph.D. Associate Professor at the Department of Computer Science and Technology of Beijing University of Posts and Telecommunication. Her research interests include service methodology, service system architecture, service sciences and engineering.

Junde Song is a Ph.D. Professor, Doctoral Advisor at the Department of Computer Science and Technology of Beijing University of Post & Telecom. His research interests include parallel computing, service sciences and engineering.

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.