



## Efficient Crawling in Online Social Networks Using Metro-polis Hastings Random Walk Technique

<sup>1</sup>Kulvinder Singh, <sup>2</sup>Sanjeev Dhawan, <sup>3</sup>Kirti Saini

<sup>1,2</sup>Faculty of Computer Science & Engineering,

University Institute of Engineering and Technology,

Kurukshetra University, Kurukshetra-136119, Haryana, INDIA

<sup>3</sup>M.Tech. (Computer Engineering) Research Scholar

University Institute of Engineering and Technology,

Kurukshetra University, Kurukshetra-136119, Haryana, INDIA.

**Abstract:** The immense interest generated by OSNs (Online Social networks) [1] has given rise to a number of measurement and characterization studies that attempt to provide a first step towards their understanding. Unlike the Web, which is largely organized around content, online social networks are organized around users. Participating users join a network, publish their profile and any content, and create links to any other users with whom they associate. Sampling techniques are essential for practical estimation of OSN properties. While sampling can, in principle, allow precise population-level inference from a relatively small number of observations, this depends critically on the ability to draw a sample with known statistical properties. The lack of a sampling frame (i.e., a complete list of users, from which individuals can be directly sampled) for most OSNs makes principled sampling especially difficult. To elide this limitation, our work focuses on sampling methods that are based on crawling of friendship relations - a fundamental primitive in any OSN. As online social network is very wide and densely connected, so it is very difficult to crawl a network. Many crawling techniques are there having their own cons and pros. In this paper, MHRW is used for crawling a social network. MHRW (Metro-polis Hastings Random Walk) technique is a crawling technique which is an enhanced technique of RW (Random walk). In RW node is selected randomly and so it is biased towards higher degree nodes and this limitation is overcome in MHRW. In MHRW, node is selected on probability basis. The node which is having more number of edges connected to it have more probability of being chosen and the node with less number of edges have less probability. Node with higher probability is selected as next node for crawling.

Keywords: OSN, MHRW, RW, degree bias, nodes

### 1. Introduction

MHRW [2] is the node sampling method and is employed to get unbiased samples in undirected social graphs i.e keeping the node degree distribution of the original graph unchanged. The process of sampling a graph is usually starts from one single seed or multiple seeds. After the nodes have been sampled, the knowledge of the node's in edges and out edges can be used to choose the next node. The policy of choosing the next node depends on the design of sampling algorithms. Here MHRW is implemented. In MHRW, a proposal function is designed based on the probability distribution. By randomly accepting or refusing the proposal, the proposal function changes the transition probabilities which makes the samples converge to the probability distribution.

For implementation Java is used in front end and for database oracle is used as back end. Code is generated for implementing MHRW for selecting the nodes for crawling. In this a scale is created for putting the node in different range, a node with higher probability has the maximum chances of probability that maximum number of nodes are being chosen from them and the one with lower probability has chances of less nodes are selected. The scale is divided into 10 partitions for putting the value in different range. The starting range of scale contains maximum number of nodes range and as the scale reaches to the end the number of nodes becomes very less.

The technique is implemented for undirected network which is bidirectional. This means that if a user is in his/her friend list then the user also contains that friend in his list also. This means that if a node 1 is connected to node 2 then node 2 is also connected to node 1. For example if Kirti is in friend list of Priya then Priya must be in friend list of Kirti and they are connected to each other. The value of the node is computed by how densely it is connected to network. By applying formula if the value comes in between 0 and 1 then the node is added to the list otherwise it is rejected. The probability of selecting a node is maximum in the beginning of scale and as

it reaches to an end nodes selected are very less. And at the end, the nodes are collected in different partitions of the scale giving the probabilities on the basis of which nodes are selected for crawling. For implementing this technique Java is used with NetBeans IDE 7.3 and Oracle is used as the back end database. Different classes were introduced for nodes, edges, for selecting nodes and for database connectivity.

## II. Related Work

There were many techniques studied for crawling by researchers and discussed their cons and pros. In the literature survey different techniques were studied and one of the technique MHRW is selected for implementation as it produced unbiased results. The results of BFS are variant and they produced samples that are not only biased but also do not have any statistical proof which means that their results are not accurate and cannot be trusted. So it is the poorest method but it is good if the network is simple and small and is widely used techniques. FS samples edge uniformly, which indicates that the sample set is biased towards higher degree vertices. Unlike, MHRW which aims at obtaining an unbiased sample directly, Frontier Sampling (FS) corrects the bias by specific estimators. Junjie *et al.* [3] focused on sampling the directed networks and intended to compare the efficiency, the accuracy and the stability between them. It was considered, the sampled nodes and links as a whole and separated from the original one. Experiments were evaluated by deploying the snow ball method, the random walk method, DMHRW and MUSDSG with different sampling ratios on the datasets. Yuming Mai [4] surveyed about the use and performance of the Metropolis-Hasting Random Walk (MHRW) and other algorithms to obtain uniform samples from the facebook and other online social networks (OSNs). The survey was composed of two subsections: Firstly, he compared MHRW with one algorithm and the other was comparing MHRW with more than one algorithms and then described recent research efforts to improve the efficiency and accuracy of MHRW and their results. If RWRW is treated as an edge sampling method, RWRW uses exactly the same estimators as FS. The comparison between MHRW and RWRW are also suitable for comparing MHRW and FS. Firstly, MHRW has the “ready to use” merits, since vertices are sampled uniformly. However, FS is biased towards the high-degree vertices and requires re-weighting appropriately. Secondly, specific estimators should be built for estimating different graph properties. However, only estimators of degree distribution and global clustering coefficients are currently available and estimators for purely data-analytic procedures, such as hierarchical clustering or multi-dimensional scaling are impossible to be constructed. Thus MHRW is more simple and versatile than FS in practice. We can conclude that the BFS, RW and RJ do not converge to uniform sampling and are biased towards the high degree-vertices. Further improvements or variants of random walks include random walk with jumps, multiple dependent random walks, weighted random walks, or multigraph sampling [5]. Rasti *et al.* [6] made an attempt to sample unstructured point to point overlays. They compared Metropolized Random Walk with their own developed RDS. Initially both the sampling techniques were tested using various static and dynamic graphs where the distribution of the sampled property is known exactly. Then both techniques were evaluated using Gnutella overlay and finally compared the samples and true distribution using Kolmogorov-Smirnov (KS) basis. Rasti *et al.* claimed that RDS performed significantly better than MRW when the overlay structure represent highly skewed node degree and highly skewed node clustering coefficients combination. Wang *et al* [7] applied MHRW and get the first unbiased sample of directed social graph. Thesis by Alan [8], used novel measurement techniques to study online social networks at scale, and used the resulting insights to design innovative new information systems. RW sampling, bias can be quantified by Markov Chain analysis and corrected by appropriate re-weighting (RWRW). MHRW samples vertices uniformly; however, one design assumption of MHRW is that the social graph is well connected. If efficiency is required then RWRW is better and if simplicity is the requirement then MHRW is the better choice.

## III. Implementation

MHRW is one of the efficient techniques for crawling as it does not produced biased results for higher degree nodes. Our focus is to crawl a network more efficiently in less time. First of all we select the number of nodes for which we want to create a network for crawling by MHRW. Now for each node, the numbers of edges connected to it are calculated. As MHRW crawl network on probability basis. Diff is calculated by subtracting the maximum value from the minimum value and dividing the output with the 9.

$$\text{diff} = \frac{\text{max} - \text{min}}{9}$$

While d is calculated by the formula

$$d = \frac{\text{val} - \text{min}}{\text{diff}} \text{ where val is the random value selected to find its probability.}$$

Random. Math is used to compare the coming values. Random.math always contains value between 0 and 1.

$$\text{Random.math} < \frac{70 + y - i * 10}{100}$$

Enter no. of nodes to work for.

100

final friend count list is

[7, 2, 1, 14, 7, 10, 8, 8, 8, 8, 6, 7, 3, 4, 5, 7, 10, 11, 5, 4, 6, 3, 5, 9, 6, 8, 3, 9, 5, 8, 4, 7, 7, 10, 8, 11, 3, 13, 11, 5, 9, 7, 4, 4, 6, 8, 7, 8, 5, 3, 9, 8, 6, 9, 7, 5, 3, 10, 4, 5, 13, 8, 6, 10, 4, 9, 15, 8, 10, 6, 9, 7, 7, 8, 5, 3, 11, 8, 6, 7, 4, 3, 7, 7, 14, 10, 8, 4, 11, 6, 9, 5, 6, 12, 3, 8, 7, 7, 3]

min is 1 and max is 15

final break up list is

[4,67,86,38,61,95,6,17,18,34,36,39,58,64,69,77,87,90,7,8,9,10,24,26,28,30,35,41,46,48,51,52,54,62,66,68,71,74,78,88,92,97,1,5,11,12,16,21,25,32,33,42,45,47,53,55,63,70,72,73,79,80,83,84,85,91,94,98,99,14,15,19,20,23,29,31,40,43,44,49,56,59,60,65,75,81,89,93,2,13,22,27,37,50,57,76,82,96,100, ,3]

THE FINAL SAMPLE GENERATED BY MHRW IS

Node 4, Node 34, Node 69, Node 36, Node 87, Node 9, Node 74, Node 26, Node 45, Node 1, Node 11, Node 16, Node 47, Node 29, Node 60, Node 31

BUILD SUCCESSFUL (total time: 3 seconds)

So, this was the sample nodes generated for crawling 100 nodes in OSN by MHRW. This can be done for any number of nodes which helps in generating efficient samples for crawling OSN in less time.

#### IV. Results

The MHRW is implemented to select the node with higher probability first with maximum chances. Firstly, a scale is created for putting the node in different range, a node with higher probability has the maximum chances of probability that maximum number of nodes are being chosen from them and the one with lower probability has chances of less nodes are selected. The scale is divided into 10 partitions for putting the value in different range. The starting range of scale contains maximum number of nodes range and as the scale reaches to the end the number of nodes becomes very less. The value of the node is computed by how densely it is connected to network. By applying formula if the value comes in between 0 and 1 then it is selected otherwise it is rejected. The probability of selecting a node is maximum in the beginning of scale and as it reaches to an end nodes selected are very less. And at the end, the nodes are collected in different partitions of the scale giving the probabilities on the basis of which nodes are selected for crawling. The scale gives the probability of selecting nodes and so the range having maximum percentage of nodes will select highest number of nodes and as the percentage decreases the number of nodes to be selected will be less. The study gives the efficient way of crawling OSN by the MHRW technique.

#### V. Database Tables

The database tables which were used for storing the data are described as follows. Each table contains its required attributes. The screenshots of the tables are:

Table No. 1 City

| Column Name | Data Type      | Nullable | Default | Primary Key |
|-------------|----------------|----------|---------|-------------|
| CITYID      | NUMBER         | No       | -       | -           |
| CITYNAME    | VARCHAR2(4000) | Yes      | -       | -           |

Table No. 2 Institutions

| Column Name | Data Type | Nullable | Default | Primary Key |
|-------------|-----------|----------|---------|-------------|
| EDUID       | NUMBER    | Yes      | -       | -           |

|              |                |     |   |   |
|--------------|----------------|-----|---|---|
| EDUINSTITUTE | VARCHAR2(4000) | Yes | - | - |
|--------------|----------------|-----|---|---|

**Table. 3 Link**

| Column Name | Data Type    | Nullable | Default | Primary Key |
|-------------|--------------|----------|---------|-------------|
| TOPIC       | VARCHAR2(50) | No       | -       | 1           |
| SEQ         | NUMBER       | No       | -       | 2           |
| INFO        | VARCHAR2(80) | Yes      | -       | -           |

**Table No. 4 Node**

| Column Name     | Data Type      | Nullable | Default | Primary Key |
|-----------------|----------------|----------|---------|-------------|
| NODEID          | NUMBER         | Yes      | -       | -           |
| FNAME           | VARCHAR2(4000) | Yes      | -       | -           |
| INAME           | VARCHAR2(100)  | Yes      | -       | -           |
| HOMETOWNID      | NUMBER         | Yes      | -       | -           |
| HIGHSCHOOID     | NUMBER         | Yes      | -       | -           |
| CURRENTCITYID   | NUMBER         | Yes      | -       | -           |
| COLLEGEID       | NUMBER         | Yes      | -       | -           |
| FAVSPORTSTEAMID | NUMBER         | Yes      | -       | -           |
| ORGID           | NUMBER         | Yes      | -       | -           |
| POLITICALVIEW   | NUMBER         | Yes      | -       | -           |
| EMAILID         | VARCHAR2(4000) | Yes      | -       | -           |
| PSWD            | VARCHAR2(4000) | Yes      | -       | -           |

**Table No. 5 Organisations**

| Column Name | Data Type | Nullable | Default | Primary Key |
|-------------|-----------|----------|---------|-------------|
|-------------|-----------|----------|---------|-------------|

|               |                |     |   |   |
|---------------|----------------|-----|---|---|
| ORGANISATIONS | NUMBER         | No  | - | - |
| ORNAME        | VARCHAR2(4000) | Yes | - | - |

**Table No. 6 Sports team name**

| Column Name    | Data Type      | Nullable | Default | Primary Key |
|----------------|----------------|----------|---------|-------------|
| Teamid         | NUMBER         | No       | -       | -           |
| Sportsteamname | VARCHAR2(4000) | Yes      | -       | -           |

## VI. Conclusions and Future Scope

In online social networks, crawling the graph efficiently is important since the graphs are very large and highly dynamic. Often, crawling an entire connected component is not feasible, and one must resort to using samples of the graph. The MHRW is implemented to select the node with higher probability first with maximum chances. Firstly, a scale is created for putting the node in different range, a node with higher probability has the maximum chances of probability that maximum number of nodes are being chosen from them and the one with lower probability has chances of less nodes are selected. The number of nodes are taken out of which the nodes will be selected for crawling. The scale is divided into 10 partitions for putting the value in different range. The starting range of scale contains maximum number of nodes range and as the scale reaches to the end the number of nodes becomes very less. The value of the node is computed by how densely it is connected to network. By applying formula if the value comes in between 0 and 1 then it is selected otherwise it is rejected. The probability of selecting a node is maximum in the beginning of scale and as it reaches to an end, nodes selected are very less. And at the end, the nodes are collected in different partitions of the scale giving the probabilities on the basis of which nodes are selected for crawling. The study gives the efficient way of crawling OSN by the MHRW technique.

Further techniques will be introduced which will produce more efficient results and make the crawling process in OSN to be traversed in no time. Social networking analysis has now moved from being a suggestive metaphor to an analytic approach to a paradigm, with its own theoretical statements, methods, social network analysis software, and researchers. There are various measurement techniques for analyzing the efficiency, user behavior, searching of social networking sites. Online social networking is still very much in its infancy, yet it already forms the basis for some enormously popular applications.

## References

1. Dr. Sanjeev Dhawan, Dr. Kulwinder Singh and Kirti Saini, "Comparing Crawling Techniques in Social Network" IJRS, Vol. 3 Issue 7, pp. 216-223.
2. Tianyi Wang, Yang Chen, Zengbin Zhang, Tianyin Xu, Long Jin, Pan Hui, Beixing Deng, Xing Li, "Understanding Graph Sampling Algorithms for Social Network Analysis", 2011 31st International Conference on Distributed Computing Systems Workshops Rev. 41, 401402.
3. Junjie Tong, Haihong E, Meina Song and Junde Song, "Empirical Studies on Methods of Crawling Directed Networks", Vol. 10, Issue 1, No 1, January 2013.
4. Yuming Mai, "Uniform Sampling of the Facebook Social Network Using the Metropolis Hasting Random Walk" ACM Transactions on Applied Perception, Vol. 2, No. 3, Article 1, 2010
5. Minas Gjoka, Maciej Kurant, Carter T. Butts and Athina Markopoulou, "A Walk in Facebook: Uniform Sampling of Users in Online Social Networks", (2011)..
6. Rasti, A., Torkjazi, M., Rejaie, R., Duffield, N., Willinger, W., And Stutzbach, "Respondent-driven sampling for characterizing Unstructured overlays, INFOCOM 2009, IEEE, 27012705.
7. Tianyi Wang, Yang Chen, Zengbin Zhang, Tianyin Xu Long Jin, Pan Hui, Beixing Deng, Xing Li, "Understanding Graph Sampling Algorithms for Social Network Analysis", 2011 31st (IEEE) Rev. 41, 401402.
8. Alan E. Mislove, "Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems", 2009.