Sampling Bias in User Attribute Estimation of OSNs[*]

Hosung Park
Department of Computer Science
KAIST, Daejeon, Korea
hosung@an.kaist.ac.kr

Sue Moon
Department of Computer Science
KAIST, Daejeon, Korea
sbmoon@kaist.edu

ABSTRACT

Recent work on unbiased sampling of OSNs has focused on estimation of the network characteristics such as degree distributions and clustering coefficients. In this work we shift the focus to node attributes. We show that existing sampling methods produce biased outputs and need modifications to alleviate the bias.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*

Keywords

Social networks; Sampling methods; User attributes

1. INTRODUCTION

With the growing size of online social networks (OSNs), efforts to derive a representative sample from OSNs has focused on accurate estimation of the topological features, such as degree distributions [3]. Not only structural features, but also nodal attributes, such as user profiles, tags, interests, and preferences, are important in market research and public opinion surveys. We expand the focus of sampling methods to the user attribute estimation of OSNs and evaluate the sampling bias of the existing methods.

2. SAMPLING METHODS

We consider the following seven sampling methods. The ultimate goal of all sampling methods in this work is to estimate $\hat{x_k}$, the number of nodes with attribute k.

Uniform Random Sampling (RS) selects a set of nodes S from all nodes in the network uniformly at random.

Snowball Sampling (SN) is equivalent to breadth-first search in our work.

Random Walk (RW) selects the next node uniformly at random from the neighbors of the current node. The transition probability of moving from x to y is $P_{RW}(x,y) = \frac{1}{degree(x)}$; it is dependent only on the degree of x. It is well known that RW is biased towards high-degree nodes.

Metropolis-Hastings Random Walk (MHRW) modifies the transition probabilities according to a target stationary distribution. Our target stationary distribution $f(x)$ is the uniform sampling, where $f(x) = \frac{1}{N}$ and N is the number of nodes in the network. Then the Metropolis-Hastings method builds a modified transition probability $P_{MH}(x,y)$ as follows:

$$P_{MH}(x,y) = \begin{cases} \frac{1}{degree(x)}\min(1, \frac{degree(x)}{degree(y)}) & \text{if } x \neq y, \\ 1 - \sum\limits_{x \neq y} P_{MH}(x,y) & \text{if } x = y \end{cases}$$

Unique-Sample MHRW (Uniq-MHRW): RW-based methods can visit a node multiple times, and thus can be construed as sampling with replacements. Uniq-MHRW removes multiple occurrences of a node in the random-walk sequence and returns only unique nodes.

Expansion Sampling (XSN) chooses the next node based on the degree to which a node $v \in N(S)$ contributes to the expansion factor $X(S)$, where $N(S)$ = neighborbood of S and $X(S) = \frac{|N(S)|}{|S|}$ [4]. XSN produces subgraphs representative of community structures in the original network.

Re-Weighted Random Walk (RWRW) differs from the six methods above in that it corrects for the bias after the RW sample is chosen. It re-weighs $\hat{x_k}$ using the Hansen-Hurwitz estimator. The estimator of the population total $t = \sum\limits_{i=1}^{N} y_i$ is $\hat{t} = \frac{1}{n}\sum\limits_{i=1}^{n}\frac{y_i}{p_i}$, where y_i is the attribute value of node i and p_i is its selection probability. We can estimate the proportion of X_k, a set of nodes with attribute k, by $\hat{\theta}(X_k) = \frac{\sum_{u \in X_k} 1/degree(u)}{\sum_{u \in S} 1/degree(u)}$ for RW samples.

Thinning (keeping only one every k samples) is applied to the family of RW-based methods to address correlation between consecutive samples.

3. SAMPLING BIAS OF USER ATTRIBUTES

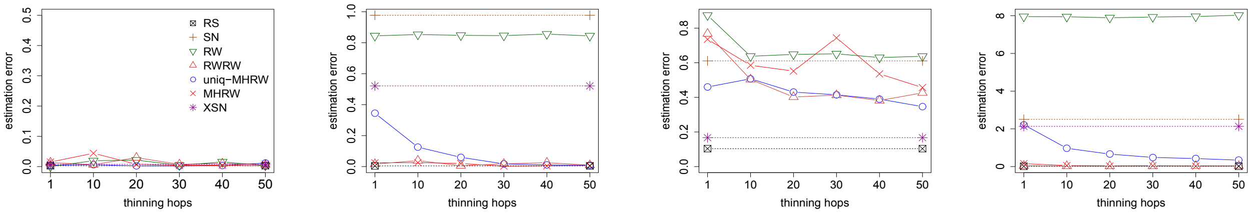
3.1 Network Topologies and User Attributes

We use 3 synthetic networks and 1 real one in our experiment: an Erdős-Rényi random graph (ER), a Barabási-Albert scale-free network (BA), a Watts-Strogatz small-world network (WS) and the Epinion network[1](EP). All have similar numbers of nodes and edges to those of EP. We make all the networks connected and undirected for the purpose of this work.

[*] This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No.2012033242)

[1] http://www.trustlet.org/wiki/Extended_Epinions_dataset

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.



(a) Scatter attributes (b) BFS attributes (c) Louvain attributes (d) Real attributes

Figure 1: Estimation errors of user attributes on the EP network (sampling rate=0.2)

	#nodes	#edges	clustering coeff.	power-law alpha	Scatter CI of att.1/att.2	BFS CI of att.1/att.2	Louvain #comm. / $\overline{CI}$	Epinion $\overline{CI}$
ER	100749	584829	0.0001	-	0.0027 / 0.0014	0.1691 / 0.1698	18 / 0.2255	-
BA	100751	503740	0.0006	2.499	0.0032 / -0.0034	0.2890 / 0.3089	26 / 0.2448	-
WS	100751	503755	0.4842	-	0.00001 / 0.00009	0.5405 / 0.5396	211 / 0.8990	-
EP	100751	584829	0.0934	1.760	0.0045 / -0.0091	0.8863 / 0.8824	2458 / 0.6884	0.5063

Table 1: Characteristics of the networks used in evaluation and their user attributes.

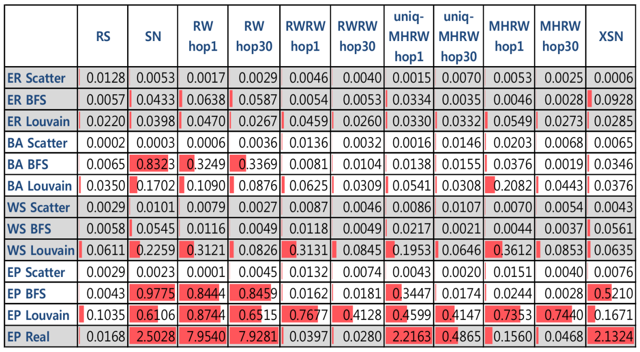
We assign user attributes to the nodes with the following three types of schemes. The Scatter scheme selects a node uniformly at random and assigns an attribute to the node. In the BFS scheme, we deploy user attributes tracking a breadth-first search from a random seed node. We deploy two attributes having 50% of the population each in the Scatter and BFS schemes. The Louvain scheme first divides a network into communities with the Louvain community detection method [1], and then assigns an attribute per community. That is, in the Louvain scheme, there are as many attributes as the number of communities. In case of EP network we also use 170, 940 real Epinion user attributes.

We summarize the network characteristics and user attributes in Table 1. Coleman Index (CI) indicates the intensity of homophily [2]. $CI = 0$, if attributes are randomly deployed regardless of others. We calculate $\overline{CI}$ only with the top 50 attributes.

3.2 Estimation of User Attributes

In our evaluation of the seven sampling methods, we use as a measure of bias, $\epsilon = |\frac{x_k - \hat{x_k}}{x_k}|$, where ϵ is the relative error of the estimated $\hat{x_k}$ against x_k and $\hat{x_k} = \frac{|X_k \cap S|}{\text{sampling rate}}$. In RWRW $\hat{x_k}$ is estimated differently: $\hat{x_k} = N \times \hat{\theta}(X_k)$. Figures 1 and 2 represent $\bar{\epsilon}$ with the sampling rate of 0.2. We calculate $\bar{\epsilon}$ with only 50 attributes as in $\overline{CI}$. The more realistic network topology (power-law and clustered) and user attributes deployment (homophily) are, the more erroneous estimation we obtain.

RS shows the best performance but is not applicable on real OSNs because the whole user-id space is not available to the public. SN and RW are highly biased methods in estimating user attributes. RWRW and MHRW show good performance except Louvain attributes. We conjecture that high biases from Louvain attributes are due to the limited extent of samples which only covers particular region of the network. The other limitaion of RWRW and MHRW is existence of duplicate elements in the samples. Only 8.7% and 44.5% of elements are unique for MHRW and RWRW respectively in the EP network. In contrast, XSN performs well in the Louvain scheme, but it performs badly for the real EP attributes. Uniq-MHRW with thinning can be a preferable sampling method as thinning lowers relative error. However, thinning brings about sampling overhead due to slow node coverage speed of MHRW. Uniq-MHRW with thinning by 50 hops requires 4.15M walks to sample 50% of unique nodes for EP network which has only 100k nodes.



	RS	SN	RW hop1	RW hop30	RWRW hop1	RWRW hop30	uniq-MHRW hop1	uniq-MHRW hop30	MHRW hop1	MHRW hop30	XSN
ER Scatter	0.0128	0.0053	0.0017	0.0029	0.0046	0.0040	0.0015	0.0070	0.0053	0.0025	0.0006
ER BFS	0.0057	0.0433	0.0638	0.0587	0.0054	0.0053	0.0334	0.0035	0.0046	0.0028	0.0928
ER Louvain	0.0220	0.0398	0.0470	0.0267	0.0459	0.0260	0.0330	0.0332	0.0549	0.0273	0.0285
BA Scatter	0.0002	0.0003	0.0006	0.0036	0.0136	0.0032	0.0016	0.0146	0.0203	0.0068	0.0065
BA BFS	0.0065	0.8323	0.3249	0.3369	0.0081	0.0104	0.0138	0.0155	0.0376	0.0019	0.0346
BA Louvain	0.0350	0.1702	0.1090	0.0876	0.0625	0.0309	0.0541	0.0308	0.2082	0.0443	0.0376
WS Scatter	0.0029	0.0101	0.0079	0.0027	0.0087	0.0046	0.0086	0.0107	0.0070	0.0054	0.0043
WS BFS	0.0058	0.0545	0.0116	0.0049	0.0118	0.0049	0.0217	0.0021	0.0044	0.0037	0.0561
WS Louvain	0.0611	0.2259	0.3121	0.0826	0.3131	0.0845	0.1953	0.0646	0.3612	0.0853	0.0635
EP Scatter	0.0029	0.0023	0.0001	0.0045	0.0132	0.0074	0.0043	0.0020	0.0151	0.0040	0.0076
EP BFS	0.0043	0.9775	0.8444	0.8459	0.0162	0.0181	0.3447	0.0174	0.0244	0.0028	0.5210
EP Louvain	0.1035	0.6106	0.8744	0.6515	0.7677	0.4128	0.4599	0.4147	0.7353	0.7440	0.1671
EP Real	0.0168	2.5028	7.9540	7.9281	0.0397	0.0280	2.2163	0.4865	0.1560	0.0468	2.1324

Figure 2: Relative error of estimated user attributes

Our preliminary results demonstrate the weaknesses of existing sampling methods. We plan to develop a sampling algorithm that produces unbiased estimates of user attributes.

4. REFERENCES

[1] V. D. Blondel et al. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[2] S. Currarini et al. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045, 2009.

[3] M. Gjoka et al. Walking in facebook: A case study of unbiased sampling of osns. In *IEEE INFOCOM 2010*.

[4] A. S. Maiya and T. Y. Berger-Wolf. Sampling community structure. In *WWW 2010*.