Sampling Bias in User Attribute Estimation of OSNs

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

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

ABSTRACT
Recent work on unbiased sampling of OSNs has focused on estimation of 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.2.84 [**Database Management**]: Database Applications—*Data mining*

General Terms
Measurement,Experimentation

Keywords
Social networks, Sampling methods, User attribute

1. INTRODUCTION
With growing size of online social networks(OSNs), unbiased sampling of OSN [2] has been focused for accurate estimation of the interested features of OSNs. However little attention has been given to unbiased sampling on the user's attributes such as user profile, tag, interested topics and so on. Estimating user's attributes are more important in the market research and survey of public opinion like product preference surveys and political polls than estimating network characteristics. In this work, we show estimation bias of user attributes with the synthetic and real networks and various user attributes deployment schemes. We exhibit that homophily of user attributes and network characteristic of OSNs affect estimation bias in sampling user attributes.

2. SAMPLING METHODS
The sampling methods for this paper are described below.
Uniform Random Sampling (RS) : RS method selects a set of nodes N from all nodes in the network uniformly at random. Applying RS on real OSN requires whole user-id space which is hard to be attained by the public.
Snowball Sampling (SN) : We implement SN like Breadth-First-Search. SN starts from the seed node selecting the neighbor node which is not visited yet at each iteration.
Random Walk (RW) : 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 $P(x,y) = \frac{1}{degree(x)}$. It is well known that RW is biased towards high degree nodes.
Metropolis-Hastings Random Walk (MHRW) : MHRW provides a method to correct for the bias towards high degree nodes of RW. To collect unbiased uniform sample, we set target stationary distribution $\mu(x) = \frac{1}{N(v)}$ where $N(v)$ is the number of nodes in the network. Then Metropolis-Hastings method builds a modified transition oprobability $Q(x,y)$ as follows :

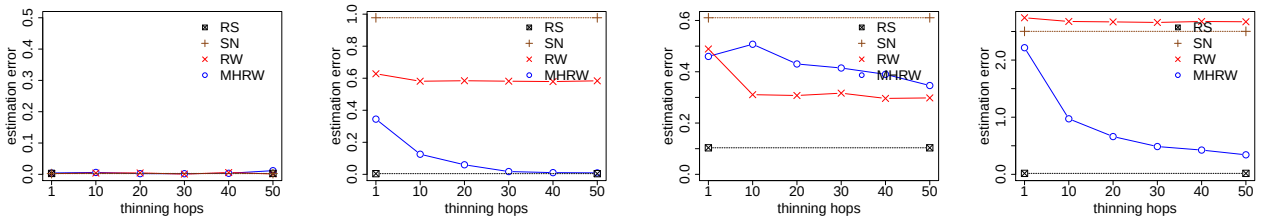
$$Q(x,y) = \begin{cases} \frac{1}{degree(x)} \min(1, \frac{degree(x)}{degree(y)}) & \text{if } x \neq y, \\ 1 - \sum_{x \neq y} Q(x,y) & \text{if } x = y \end{cases}$$

As we are interested in the estimation of user attributes, we sampled only nodes excluding edges in all sampling methods. Thinning(keeping only one every k samples) is applied to RW and MHRW samples to address correlation of consecutive samples.

3. SAMPLING BIAS OF USER ATTRIBUTES

3.1 Netowork Topology and User Attributes
We generate synthetic networks and deploy user attributes to the nodes in various schemes. Add to this we use real social network data and user attributes data.
Description of Used Networks : Four kinds of networks are prepared for the experiment; Erdős-Rényi random graph (ER), Barabási-Albert scale-free network (BA), Watts-Strogatz small-world network (WS) and Epinion network[1](EP). ER, BA and WS are synthetic networks and have similar number of nodes and edges to EP, the real social network data. We make all networks connected and undirected for the purpose of this work.
Deployment of User Attributes : Three schemes are chosen for the deployment of the synthetic user attributes.

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



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

Figure 1: Estimation errors of user attributes on 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. / mean CI	Epinion mean CI
ER	100749	584829	0.0001	-	-0.3308 / -0.3320	-0.1739 / -0.1733	18 / 0.1058	-
BA	100751	503740	0.0006	2.499	-0.3304 / -0.3363	-0.0490 / -0.0272	26 / 0.1260	-
WS	100751	503755	0.4842	-	-0.3333 / -0.3332	0.2527 / 0.2515	211 / 0.8165	-
EP	100751	584829	0.0934	1.760	-0.3292 / -0.3414	0.7848 / 0.7780	3458 / 0.5408	0.3313

Table 1: Characteristics of the prepared network and user attributes.

Scatter scheme selects a node uniformly at random and assigns an attribute to the node not allowing attribute overlapping (Scatter). In BFS scheme, user attributes are deployed tracking Breadth-First-Search from the random seed node allowing attribute overlapping for maintaning BFS structure of deployment (BFS). Louvain scheme first divides networks into communities with Louvain method for community detection, then assigns each attribute to each community members (Louvain). We deploy 170,940 real Epinion user attributes in addition to the above synthetic attributes for EP network (Epinion). We deploy two attributes having size of 50% of population for each attribute in Scatter and BFS schemes. The number of attributes of Louvain scheme is equal to the number of communities of the target network.

We depict characteristics of the prepared data in Table 1. The degree distributions of BA and EP network follow a power law which is observed in many real-world networks. EP network has both 'power-law' and 'clustered' characteristics which are distinguishing charateristics of realistic network. WS is well clustered but does not follow a power law. Coleman Index(CI) [1] indicates homophily of user attribute deployment which is the tendency of nodes to associate with similar others. CI is zero if attributes are randomly deployed regardless of others. Negative CI in Scatter attributes can be interpreted as associcating with different attributes because full random assignment make attributes of neighbor nodes alternate. We calculate mean CI with 50 most attributes in size if there are more than 50 attributes.

3.2 Estimation of User Attributes

We apply the sampling methods on the above network topologies and user attributes. Then we calculate relative error, $RE = |\frac{x-\hat{x}}{x}|$ of estimated number $\hat{x} = \frac{\#\text{ of attribute members}}{\text{sampling rate}}$ of each attribute from the sampled nodes. Figure 1, 2 represent relative error of the estimation with schemes mentioned above with sampling rate 0.2. The more realistic network topology (power-law and clustered) and user attributes deployment (homophily) are, the more erroneoues estimation we obtain. RS shows the best performance but it is hard to be used in the real OSN sampling. SN and RW are biased methods in estimating user attributes. MHRW with thinning can be a preferable sampling method as thinning lowers error. However, thinning brings about sampling overhead due to slow node coverage in MHRW. 1.89M walks are required to sample 50% of unique users in MHRW with 100k nodes EP network and thinning by 50 hops needs 4.15M walks.

	RS	SN	RW hop1	RW hop30	MHRW hop1	MHRW hop30
ER Scatter	0.0128	0.0052	0.0037	0.0039	0.0014	0.0069
ER BFS	0.0056	0.0432	0.0053	0.0053	0.0334	0.0034
ER Louvain	0.0219	0.0397	0.0359	0.0199	0.033	0.0331
BA Scatter	0.0002	0.0003	0.0023	0.0021	0.0015	0.0146
BA BFS	0.006	0.8323	0.0924	0.0878	0.0137	0.0154
BA Louvain	0.0349	0.1702	0.0408	0.0409	0.054	0.0307
WS Scatter	0.0028	0.0101	0.0053	0.0053	0.0086	0.0107
WS BFS	0.005	0.0544	0.0178	0.0058	0.0216	0.002
WS Louvain	0.061	0.2259	0.2153	0.0691	0.1952	0.0646
EP Scatter	0.0028	0.0022	0.0022	0.0007	0.0042	0.0019
EP BFS	0.0042	0.9777	0.6278	0.5814	0.3444	0.0174
EP Louvain	0.1034	0.6105	0.4887	0.3165	0.4598	0.4146
EP Real	0.0167	2.5028	2.7404	2.6584	2.2163	0.4865

Figure 2: Relative error of estimated user attributes on the all networks (sampling rate 0.2)

4. FUTURE WORK

More parameters, like overlapping ratio or attribute size distribution, should be considered in the investigation of estimating user attributes in OSNs. We also remain developing an algorithm complementing existing methods which can be utilized in the unbiased user attributes sampling problem as the ultimate goal.

5. REFERENCES

[1] S. Currarini, M. O. Jackson, and P. Pin. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045, 2009.

[2] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. IEEE, 2010.