Semantically Sampling in Heterogeneous Social Networks

Cheng-Lun Yang, Perng-Hwa Kung⋆, Chun-An Chen†, Shou-De Lin‡
School of Computer Science and Information Engineering, †Department of Electrical Engineering,
National Taiwan University
{r99944042, ⋆r00922048, ‡ sdlin}@csie.ntu.edu.tw,†andro0929@gmail.com

ABSTRACT

Online social networks sampling identifies a representative subnetwork that preserves certain graph property given heterogeneous semantics, with the full network not observed during sampling. This study presents a property, Relational Profile, to account for conditional dependency of node and relation type semantics in a network, and a sampling method to preserve the property. We show the proposed sampling method better preserves Relational Profile. Next, Relational Profile can design features to boost network prediction. Finally, our sampled network trains more accurate prediction models than other sampling baselines.

Categories and Subject Descriptors

H.2.4 [**Database Management**]: Database Applications—*Data Mining*

Keywords

Heterogeneous network; graph sampling; network prediction

1. INTRODUCTION

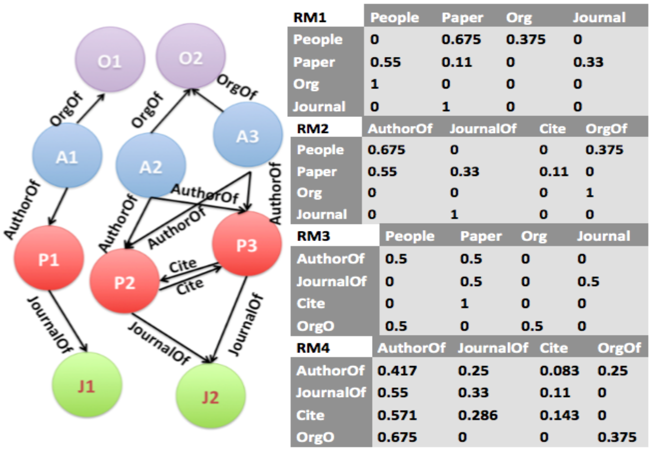
Online heterogeneous social networks are large-scale, with complex semantics having node and relation categories denoting various social roles. One obstacle to effectively utilizing such networks is that it is often impossible to fully access the whole networks. Therefore, we aim to sample a small network, G_s, that approximates the full network G under chosen properties. We propose Relational Profile (RP) that stores network's transitional probability between node and relation types, and a heuristic sampling strategy. Different sampling methods are tested for RP preservation, and we design RP features for node type and missing relation prediction. Results confirm RP's usability, and the superiority of our sampled networks.

Related Work: The two main objectives in network sampling are back-in-time and scale-down goals[3]. In the scale-down goal, explorative sampling adds nodes sequentially from the neighbors of sampled network in the partially visible network. Topology weighted or pure random walk sampling is used, depending on data[3][6], but they mainly focus on homogeneous network. Works on heterogeneous network examine inter/intra-link distribution[4], but transitional probabilities for both node/relation types and usage of semantic subnetworks have not been addressed.

2. METHODOLOGY

A directed heterogeneous graph G has node set V, relation set E, node and relation type labels NT and ET(see Figure 1's illustration). We define a Relational Profile as follows:

*A **Relational Profile** (RP) of a graph G consists four transitional probability matrices (RM) between two semantic types $RP = (RM1, RM2, RM3, RM4)$. Each r_{ij} in $RM1$ denotes $P(node\ type\ j|node\ type\ i)$, the probability of a relation with one end node type j given other end node type i. Similarly, $RM2$, $RM3$, and $RM4$ represent $P(edge|node)$, $P(node|edge)$, and $P(edge|edge)$ type distributions. RP terms come from counting each type transition (i,j)'s occurrence.*



RM1	People	Paper	Org	Journal
People	0	0.675	0.375	0
Paper	0.55	0.11	0	0.33
Org	1	0	0	0
Journal	0	1	0	0

RM2	AuthorOf	JournalOf	Cite	OrgOf
People	0.675	0	0	0.375
Paper	0.55	0.33	0.11	0
Org	0	0	0	1
Journal	0	1	0	0

RM3	People	Paper	Org	Journal
AuthorOf	0.5	0.5	0	0
JournalOf	0	0.5	0	0.5
Cite	0	1	0	0
OrgO	0.5	0	0.5	0

RM4	AuthorOf	JournalOf	Cite	OrgOf
AuthorOf	0.417	0.25	0.083	0.25
JournalOf	0.55	0.33	0.11	0
Cite	0.571	0.286	0.143	0
OrgO	0.675	0	0	0.375

Figure 1: Example of heterogeneous publication network and the corresponding Relational Profile with node types={Author, Paper, Org, Journal} and relation types = {AuthorOf, JournalOf, Cite, OrgOf}

To find ΔRP between graphs G_1 and G_2, we use average Root Mean Square Error (RMSE) of the four RM.

For explorative sampling, each node v is sequentially sampled among neighbors of the current G_s, $C_{G_s,k}$. We propose maximizing information gain due to $\mathcal{P}$ by selecting v from $\mathcal{D}(v; C_{G_s,k}) = E[\Delta_{\mathcal{P}}(G_s, G_s + v)|Obs(v, e)]$, a normalized distribution with observed G_s, v and its connecting edges e. Each step tries to optimize expected property change. We are now ready to introduce RP-Preserving Sampling:

RP-Preserving Sampling(RPS): With $\mathcal{P} = RP$, we incorporate semantic type knowledge by maximizing expected change of RP over all node types for adding node v to G_s,

$$\mathcal{D}(v; C_{G_s,k}) = E_t[\Delta RP(G_s, G_s + v)|Obs(v, e)]$$

$$= \sum_{t \in NT} P(type(v) = t|Obs(v, e))\Delta RP(G_s, G_s + v)$$

We estimate $P(type(v) = t|Obs(v, e))$ with:

$$\prod_{i \in N(v)} \left(\frac{RP(type(i)|type(v) = t)P(type(v) = t)}{Z}\right)$$

Z is the normalization constant and $N(v)$ is the neighboring nodes for v. We use Bayes' Rule and RP for estimating $P(type(v) = t|Obs(v, e))$, and assume conditional independence of node v to all its neighbors in joint type distribution.

3. EVALUATIONS

Evaluating Property Preservation: We take 3 real life social networks, using the largest connected component:

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

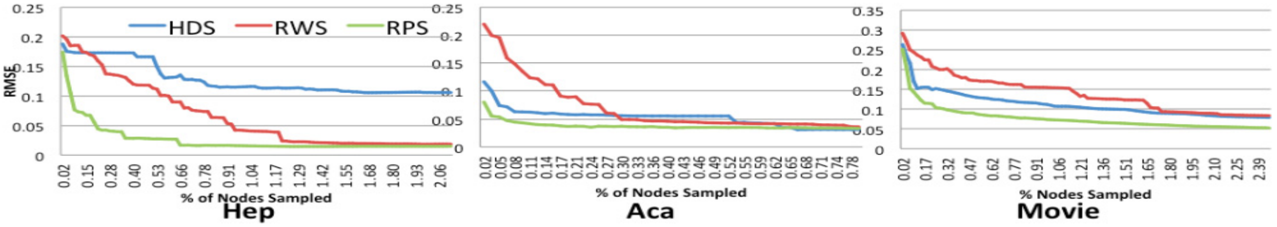


Figure 2: RMSE of RP versus % nodes sampled across 3 datasets, using different sampling methods

High Energy Physics (Hep) from 2003 KDDCup, Academic (Aca) from Taiwan Academic Archive, and Movie networks (Movie) from Internet Movie Database, with node numbers 41744, 63122, and 40520, respectively. Two baselines[6] are used: High Degree Sampling (HDS) for sampling v under $\mathcal{D}$ proportional to normalized degree; Random Walk Sampling (RWS) for randomly selecting among $C_{s,k}$. All experiments are averaged over 10 runs.

Results of RMSE are shown in Figure 2, where we sampled until stabilizing. RPS's RMSE drops the fastest, using only a small portion of sampled nodes (0.8% to 2.4% of total). We also examined node role approximation using semantically weighted PageRank[7] (LinkFusion), calculating the AUC for two weighted node lists for subnetwork and network. RPS is again consistently superior.

Evaluating Network Prediction Tasks: *Node Type Prediction* is a multi-class classification problem, where we predict $type(n)$ of a node n. Each node in G_s (node size 500) is a training instance and each testing instance randomly selects a node from the rest of the network. Testing size is set to 1000. *Missing Relation Prediction* predicts whether an arbitrary semantic relation exists in a graph. Each instance is a pair of nodes, with a relation between them labeled 1 and -1 otherwise. Training set uses G_s with negative instances downsampled to match number of positives. Testing instances are selected to have 10000 each label. Our model uses SVM with linear kernel via LIBLINEAR package[1].

Features for model training: We use f_{deg}(in/out degree, average neighbor in/out degree), f_{topo}(common neighbors, Jaccard Coefficient, Preferential Attachment, Adamic Adar, from [5]), and f_{nt}(node type distribution, $P(type(n) = t|n) = \frac{\#type(v)=t \forall v \in N(n)}{|N(n)|}$, with node n and its neighbors $N(n)$) as baselines. For the two Relational Profile features: **1.** $f_{RP_{node}}$, the expected type probability density for node n, estimated with RP using the earlier formulation:

$$f_{RP_{node}}(n): \; P(type(n) = t|n)$$
$$= \prod_{i \in N(n)} \frac{1}{Z} RP(type(i)|type(v) = t)P(type(v) = t)$$

2. $f_{RP_{path}}$ for each pair (s, t) of nodes with paths connecting s and t $Path(s, t)$ of length k (k=2 in this study):

$$f_{RP_{path}}(s, t, k) = \sum_{p \in Path(s,t)} \prod_{(p_1,p_2) \in p} P(type(p_2)|type(p_1))$$

Path type probability is estimated with bigram probabilities using $P(type(n_2)|type(n_1)) = RP(type(n_2)|type(n_1))$.

Results: First, compare different sampling methods. Feature sets $f_{RP_{node}}$ and f_{topo}+ $f_{RP_{path}}$ are used to train models for node type and missing relation predictions, respectively. Figure 3 shows Aca result (for brevity), with RPS the better performer. Accuracy reaches asymptote quickly with increasing size. Table 1 shows results with varying feature sets of node size 500. We find adding PR features universally improve prediction performance, with 5 to 10% in node type prediction and 2 to 6% in missing relation prediction.

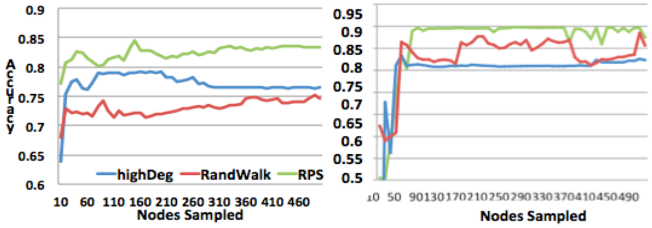


Figure 3: Prediction under different sampling methods in Aca dataset (Right: node type; Left: Missing Relation)

Task	Node Prediction		Relation Prediction	
	$f_{deg} + f_{nt}$	$f_{deg} + f_{nt} + f_{RP_{node}}$	f_{topo}	$f_{topo} + f_{RP_{path}}$
Hep	0.831	**0.884**	0.892	**0.908**
Aca	0.808	**0.876**	0.778	**0.825**
Movie	0.722	**0.828**	0.668	**0.729**

Table 1: Prediction using varying feature sets

4. CONCLUSIONS

We explore heterogeneous network sampling with explorative algorithms. RP and RP-preserving sampling are proposed that consider semantic type information. Experiments show RPS outperforms other sampling strategies in RP-preservation. RP as features is useful for prediction.

5. ACKNOWLEDGMENTS

This work was also supported by National Science Council, National Taiwan University and Intel Corporation under Grants NSC 100-2911-I-002-001, and 101R7501

References

[1] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2011.

[2] M. Kurant, M. Gjoka, Y. Wang, Z. W. Almquist, C. T. Butts, and A. Markopoulou. Coarse-grained topology estimation via graph sampling. In *WOSN'12*.

[3] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *SIGKDD*, 2006.

[4] J.-Y. Li and M.-Y. Yeh. On sampling type distribution from heterogeneous social networks. In *PAKDD*, 2011.

[5] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *JASIST*, 2007.

[6] A. S. Maiya and T. Y. Berger-Wolf. Benefits of bias: Towards better characterization of network sampling. *CoRR*, 2011.

[7] W. Xi, B. Zhang, Z. Chen, Y. Lu, S. Yan, W. ying Ma, and E. A. Fox. Link fusion: A unified link analysis framework for multi-type interrelated data objects. *WWW'04*.