Sampling Heterogeneous Networks

Cheng-Lun Yang, Perng-Hwa Kung†, Cheng-Te Li, Chun-An Chen*, Shou-De Lin
School of Computer Science and Information Engineering, Department of Electrical Engineering*
National Taiwan University
Taipei City, Taiwan
{r99944042, r00922048†, d98944005, sdlin}@csie.ntu.edu.tw, *andro0929@gmail.com

***Abstract*—Online social networks are mainly characterized by large-scale and heterogeneous semantic relationships. Unfortunately, for online social network services such as Facebook or Twitter, it is very difficult to obtain the fully observed network without privilege to access the data internally. To address the above needs, social network sampling is a means that aims at identifying a representative subgraph that preserves certain properties of the network, given the information of any instance in the network is unknown before being sampled. This study tackles heterogeneous network sampling by considering the conditional dependency of node types and link types, where we design a property, Relational Profile, to account such characterization. We further propose a sampling method to preserve this property. Lastly, we propose to evaluate our model from three different angles. First, we show that the proposed sampling method can more faithfully preserve the Relational Profile. Second, we evaluate the usefulness of the Relational Profile showing such information is beneficial for link prediction tasks. Finally, we evaluate whether the networks sampled by our method can be used to train more accurate prediction models comparing to networks produced by other methods.**

I. Introduction

The widespread hype in social network sees an unprecedented information load with complex semantics for various applications. Network semantics are mainly classified into individual attributes and interaction relationships (or node and link types). For example, in publication network DBLP, entities can be authors, papers, etc. Furthermore, two entities can interact through different activities, such as a person can author a paper while a paper can cite another paper. Identifying compact representation and property statistics of such heterogeneous information in a social network can provide a better understanding to the already-complicated network structure.

Generally there are two mainstream strategies to scale down networks: summarization and sampling. Summarization condenses the graph for efficient visualization or processing[14], and assumes the full network is observed. However, many online social networks such as Facebook do not reveal the complete graph, making summarization infeasible. Social network sampling is more preferred for such scenario since it is designed to gradually observe a subgraph. For instance, the technique of crawling a social network service such as Facebook can be considered as a sampling process[4]. In the beginning the crawler has no information about any individual persons or the network structure. Through extracting the friends, friends of friends and so on given certain seed persons, the crawler can gradually obtain a sub-network through sampling. Note that when we say 'sample a node' here, we essentially mean to observe the type, attributes, degree, as well as all kinds of information about this node, since such information is considered as unknown before a node is sampled. So far, there have been a diverse studies on sampling homogeneous social networks[6][9]. Here, a sampling algorithm is designed to preserve certain network topology such as clustering coefficient, degree distribution, diameter, etc. These are well-known properties for a homogeneous social network where there is only one type of node and link. However, in the real world, people are connected through different types of relationships (e.g. friends, family, co-authors, etc), and it is more natural to represent them as a heterogeneous social network in which there are different types of vertices and edges in the network. Up to date, only very few works try to tackle the problem of sampling heterogeneous social networks, where we only identify a couple [3][10] empirically evaluating means to preserve the node or link type distribution.

The major aim of this paper is to design a sampling framework that extracts a network with gradually observed information. During the sampling process, the framework first evaluates the nearby neighbors, then selects suitable candidate nodes, and finally issues a request to access a particular candidate's information. This type of iterative subgraph expansion is generally considered as a strategy for explorative sampling [10][13]. In particular, we attempt to determine the best means to sample a network G_s that minimizes the difference between the sampled network and the full network $\Delta(G, G_s)$ trying to preserve a chosen property. As the set of outputs, we envision a compact sampled network and property statistics. The main challenge of sampling lies in that during the sampling the information of the unsampled part is considered unknown, therefore it is very difficult to know whether certain property (e.g. overall degree distribution) is preserved.

This paper investigates three main challenging aspects in heterogeneous network sampling with echoing contributions:
- **First**, we would like to investigate the type of property beneficial to preserve while sampling a heterogeneous social network. Our solution proposes a novel property, Relational Profile (RP), that captures both topological and semantic

†corresponding author

information in heterogeneous network.
- **Next**, Given the desired property (RP), we would like to design a sampling algorithm to preserve it. We propose a RP preserving sampling method that utilizes a probabilistic model for RP-aware matching in node selection, given limited network observation.
- **Finally**, we would like to design a systematic mechanism to evaluate the effectiveness of the proposed property as well as the sampling algorithm. Our solution involves designed experiments from different aspects for evaluation, and should provide proper example on how to evaluate sampling method where ground truth is not existent.

II. Related Work

Representative subgraph sampling selects a subgraph via a choosing strategy. Leskovec et al. [9] explored two sampling goals: back-in-time and scale-down. On scale-down goal, they found random walk based methods do best in graph approximation. Maiya and Wolf [13] discovered that with appropriate greedy expansion of graph properties, the subgraph better approximates the full graph under comparing measures such as degree, network reach, among others.

Two approaches in sampling are: network-wide or explorative sampling. The former can choose the next instance for sampling from the full network whenever possible and the latter can only access the current neighbors. **Network-wide sampling** has recently been applied to large-scale online social networks such as Facebook[1][4]. Some works concern with uniform sampling[5][16]. More recent works sample under the bias of properties such as degree distribution[3]. Hübler et al.[6] proposed an adapted Metropolis algorithm to approximate different topological properties. However, the above methods need to constantly access the entire network, which may be problematic when network is very large or only partially observable. **Explorative sampling** is largely characterized by methods that include Forest Fire [9], Random Walk[4][11], and Multiple Ego-Centric-Exploration Sampling [10]. These methods are variations to random surfer model, where there is a small number of egos, and new nodes are selected by exploring the proximity of the egos. However, surfer models do not consider type relations present in a heterogeneous network.

In **heterogeneous network sampling**, [3] proposes a multi-graph approach that first decomposes the heterogeneous network into a set of homogeneous ones for each type of links. Random walk sampling is run for each single type network, and networks are combined as output. This method suffers the drawback that cross-type relations between nodes and links are not considered. [10] proposed node type distribution for inter-relationship and intra-relationship that denote proportion of links that connect same node type. Several homogeneous sampling methods were compared, and Respondent Driven Sampling, a variant to explorative sampling, appeared to be the best for the two proposed goals. Our work's goals differ from theirs in that they did not propose a property tailored for heterogeneous social network, whereas we are preserving a the higher-order relationship of node and link types. Furthermore, we examine application scenarios to predictive model building. We summarize the related works in Table I.

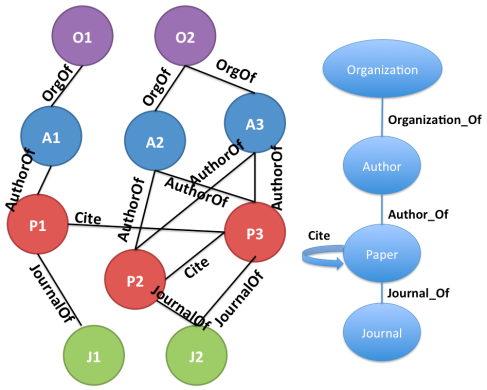


Fig. 1. A sample graph schema, using publication network as example. Left: a sample network, with node annotation (O: Organization; A:Author; P:Paper; J:Journal). Right: corresponding graph schema.

	Homogeneous Network	Heterogeneous Network
Network-wide Sampling	[1][6][8][9][15]	
Sampling by Exploration	[4][5][7][8][9][11][13][15][16]	[3][10]

TABLE I. Network Sampling Method Comparison

III. Definition of Relational Profile

A. Background and Terminology

Given a graph $G = (V, E)$, where V is a set of N vertices (entities) and E represents a set of M edges (relations), we define a **heterogeneous graph** as the following:

Definition 1: *a **heterogeneous graph G = (V, E)** is a graph, with each node n of type $NT(n)$. A triple (v_1, v_2, $ET(e)$) expresses edge e(nodes v_1 and v_2 related by relation $ET(e)$). Node and edge label sets are NT and ET, respectively. We use $N(v)$ for node v's neighbors.*

Figure 1 shows a toy graph schema and the corresponding network using publication network as example, where *paper* publishes in *journal*; *paper* cites other *papers*; *author* publishes *papers*; and *author* belongs to *organization*. To compare two different graphs, under property $\mathcal{P}$ (e.g. degree), we define operator $\Delta_{\mathcal{P}}$ to measure difference of the two graphs under $\mathcal{P}$ (i.e. $\Delta_{\mathcal{P}}(G_1, G_2)$ for graphs G_1, G_2).

We formalize the problem, property-preserving heterogeneous network sampling, as: given a heterogeneous network G, sampled network G_s, a network property $\mathcal{P}$, distance operator $\Delta_{\mathcal{P}}$ under property $\mathcal{P}$, a ***Property Preserving Social Network Sampling*** is a problem to sample a subgraph G_s of given size k such that $\Delta_{\mathcal{P}}(G_s, G)$ is minimized (e.g. sample to minimize difference in degree distribution).

B. Relational Profile: our proposed property to preserve

Previously, most works considered preserving topological properties or first order semantics (i.e. type distribution) only. Ideally, we believe a suitable property to represent a heterogeneous social network should capture both the topological and semantic information, which becomes the focus of our design:

Definition 2: *Given a heterogeneous graph G, we define the **Relational Profile of G,** $RP(G)$, as a ($|NT|$+$|ET|$) x ($|NT|$+$|ET|$) matrix which contains the transitional probabilities from one vertex/edge type to another, denoted as:*

$$RP = \begin{bmatrix} RM_{NN} & RM_{NE} \\ RM_{EN} & RM_{EE} \end{bmatrix} \quad (1)$$

	Author	Paper	Org	Journal	AuthorOf	JournalOf	Cite	OrgOf
Author	0	0.675	0.375	0	0.675	0	0	0.375
Paper	0.5	0.2	0	0.3	0.5	0.3	0.2	0
Org	1	0	0	0	0	0	0	1
Journal	0	1	0	0	0	1	0	0
Author_of	0.5	0.5	0	0	0.333	0.25	0.167	0.25
Journal_of	0	0.5	0	0.5	0.556	0.222	0.222	0
Cite	0	1	0	0	0.556	0.333	0.111	0
Org_of	0.5	0	0.5	0	0.714	0	0	0.286

Fig. 2. RP for network in Figure 1

Each Relational Matrix, or RM, is thus defined:
• *The (i,j)th element in RM_{NN}, or $RM_{NN}(i|j)$, is the probability of reaching a neighbor of type i from a randomly picked node of type j.*
• *The (i,j)th element in RM_{NE}, or $RM_{NE}(i|j)$ is the chance that edges connecting node of type j is of type i.*
• *The (i,j)th element in RM_{EN}, or $RM_{EN}(i|j)$, is the chance an edge's (of type j) endnodes contain a node of type i.*
• *Finally, The (i,j)th element in RM_{EE}, or $RM_{EE}(i|j)$, is the chance that for two randomly picked edges that share a node, given one edge is of type j, the other is of type i.*

Actual calculation of an entry of Relational Matrix $RM(i|j)$ for any particular RM in RP is derived from observed node type counts. See Figure 2 for example.

In order to measure similarity between networks with RP, the Root Mean Square Error (RMSE) for all entries is applied:

$$\Delta RP(G_1, G_2) = \sum_k RMSE(RMk_{G_1}, RMk_{G_2}) \quad (2)$$

$$RMSE(RMk_{G_1}, RMk_{G_2}) = \sqrt{\sum_{(i,j)\in NT;ET} \Delta RM_k(i|j)^2}$$

where RMk_G represents graph G's Relational Matrices. (i, j), shows all corresponding entries of the Relational Matrix.

IV. Sampling Methodology

A. Explorative Sampling Framework

Our proposed algorithm bases on **explorative sampling**, which samples the new node n_{new} at each step from the set of candidate nodes C_{G_s}, consisting unsampled one-step neighbors of the current sampled network. Under the goal $argmin_{G_s}\Delta_{\mathcal{P}}(G_s, G)$, all subsets k nodes need to be inspected, which is intractable as the search space grows exponentially with subset size. A relaxation heuristic is: $\forall v \in C_{G_s}$,

$$P(n_{new} = v) \propto Score(v, G_s, \mathcal{P}),\ v \in C_{G_s} \quad (3)$$

Equation (3) essentially maps node selection probability to informativeness of preserving property $\mathcal{P}$. We can incorporate any property in (3). For example, $Score(v) = Deg(v)$ can be a good choice to preserve high-degree nodes in sampling.

We summarize the process in Algorithm 1. Starting with empty G_s, a node is selected uniformly at random in G. Then, we calculate the score value of each one degree unvisited node neighbor using Equation(3), and sample node v based on the distribution normalized from such scores.

B. Property Preserving Sampling for Relational Profile

We propose **Relational Profile Preserving Sampling (or RPS)**, to preserve our Relational Profile. If the $RP(G)$ for original network and candidate nodes' types were revealed, problem would be easy: we greedily select at each step the node v whose $\Delta_{RP}(G_s + v, G)$ is the smallest (i.e. adding v best approximates $RP(G)$). Unfortunately, the ultimate $RP(G)$ is unknown. Coping with this challenge, we sample that brings the largest change to the *existing* RP (i.e. $RP(G_s)$). In other words, given $\Delta_{RP}(G_s + v, Gs)$ is known, we opt node v that causes the largest change. The intuition is, node/edge type pairs that are sampled less frequently usually provide a larger change to the RP value, and including less frequent pairs whenever they occur conceives more sufficient observation and better estimates the statistics.

Calculating a candidate node's change to RP requires knowing the node type, which lies the second challenge as the node type is unknown before sampled. We propose to predict the type distribution of each candidate node v, and use such distribution to generate the 'expected' change of RP given v's inclusion. Finally, the node that produces the largest 'expected change' is sampled. The probability type distribution of each candidate node can be estimated using the existing sampled RP. Referring back to Algorithm 1, RPS effectively represents $Score(v,\ G_s, \mathcal{P})$ using $E[\Delta_{RP}(G_s + v, G)|G_s]$. Now, we generate $E[\Delta_{RP}(G_s + v, G)|G_s]$: first of all, if the type distribution of v, $P(type(v) = t|G_s)$, is estimated, then $E[\Delta_{RP}(G_s + v, G)|G_s]$ can be obtained as:

$$\sum_{t\in NT} P(type(v) = t|G_s) * \Delta_{RP,type(v)=t}(G_s + v, G_s) \quad (4)$$

we can view it as the weighted sum of possible node types with respective RP change. Next, type of v can be estimated from its neighboring nodes, or mathematically, $P(type(v) = t|G_s) = P(type(v) = t|type(N_v))$, where type($N_v$) is the type information of the jointly observed neighbor edges and nodes of v. Using Bayes rule, we can transform $P(type(v) = t|type(N_v))$ into $P(type(N_v)|type(v) = t) * P(type(v) = t)$, since the $P(type(N_v))$ is already observed. As Naive Bayes assumes the type of each neighbor node/link given type(v) is independent, $P(type(N_v)|type(v) = t)$ is further decomposed as: $\prod_{i\in N_v} P(type(i)|type(v) = t)$, where each term is obtainable from the RP of G_s as the node i's type is observed. Summarizing, the chance of choosing a candidate node v to be sampled is proportional to $E[\Delta_{RP}(G_s + v, G)|G_s]$, and can be approximated by:

$$\sum_{t\in NT}\Big[\prod_{i\in N_v} P(type(i)|type(v) = t) * P(type(v) = t) * \Delta_{RP,type(v)=t}(G_s + v, G_s)\Big] \quad (5)$$

Existing sampling by exploration methods suffer from problems of easily overfitting local topology, where in graphs with predominant but sparsely connected graphs, such node pairs may be missed. Instead, RPS sampling prioritizes maximizing type information gain while sampling a new node.

Algorithm 1 Property Preserving Explorative Sampling Algorithm in Heterogeneous Network

Require: Heterogeneous graph G = (V, E), k = sample size, Property $\mathcal{P}$, Score function
v = random(V) # or given input
G_s = (V_s, E_s), V_s = {V_s, v}, E_s = edges(v)
while $|V_s| \leq$ k **do**
$C_{G_s} = \bigcup_{v \in V_s} \{n \in N_v \wedge n \notin V_s | v\}$
$P(n) \propto$ Score(v, G_s, $\mathcal{P}$)# P(n) represents the probability the node n is chosen as the next sampled node.
n_{new} = Select(P(n)) #sampling proportional to density function
V_s = {V_s+ n_{new}}
E_s = {E_s+edges(n_{new})}
end while
return sampled Graph G_s = (V_s, E_s)

V. Experiments

A. Evaluating Property Preservation

We first illustrate how RPS is better at approximating Relational Profile during sampling process.

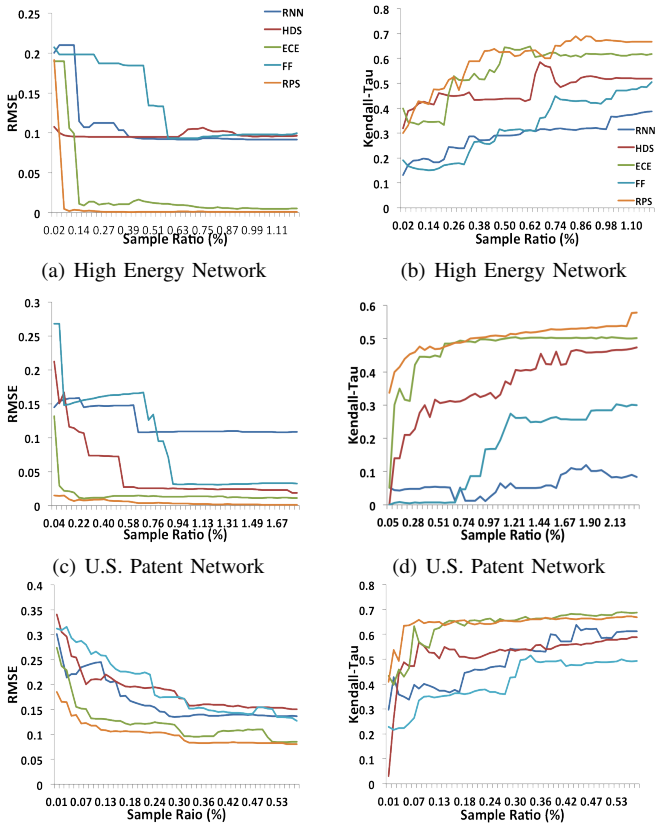
1) Datasets: We used 3 public social networks with distinct semantics. High Energy Physics Citation network contains Arxiv papers published from 1993 to 2003, and covers the citation relationships between authors, affiliating organizations, papers, and journals. Patent network extracts registered United States patents. Nodes involve patents, inventors, associated categories, and affiliations. DBLP publication network's individual nodes are denoted by author, paper, and conference. Table II shows the data statistics. We used the largest connected component in each dataset.

Dataset	# Node	# Edge	NT size	ET size
High Energy Physics	41,744	483,217	4	5
U.S. Patent Office	27,547	81,255	4	5
DBLP	86,535	244,176	3	3

TABLE II. Dataset Statistics

2) Baselines: Topology-rooted baselines include Random Nearest Neighbor Sampling (RNN), High Degree Sampling (HDS), Egocentric random walk with restart (ECE) and Forest Fire Sampling (FF). For Random Nearest Neighbor, we sample the next node uniformly at random from all one degree neighbors. For High Degree Sampling, the density function is proportional to candidate node's connecting degree to G_s, ensuring greedy expansion. Egocentric algorithm samples by keeping a random surfer and select from its neighbors. Forest Fire Sampling takes n_{seed} as initial surfer. For all surfers, the neighbors are burnt with probability p_f, which would become a new surfer with current one deleted. A burn back probability p_b burns nodes in G_s. We set restart probability for ECE = 0.1, $p_f = 0.25,\ p_b = 0.2$ as suggested in [9].

3) Evaluating RP Preservation and Weighted PageRank Scores: We first evaluate through computing RMSE of the Relational Profiles between the original graph and the sampled subgraph using Equation (2). The experiment runs all the algorithms 10 times. Results are shown in Figure 3, where the x-axis indicates the amount of nodes in the subgraph. From Figure 3, we observe the drop of RMSE with more nodes sampled, in particular, we see RPS significantly outperform



(a) High Energy Network (b) High Energy Network
(c) U.S. Patent Network (d) U.S. Patent Network
(e) DBLP Network (f) DBLP Network

Fig. 3. (a,c,e): Comparison of RMSE in RP; (b,d,f):Comparison of Kendall-Tau of LinkFusion scores across 3 datasets, using different sampling methods

all other baselines in all datasets in terms of Relational Profile preservation. Note RPS produces sharper drops in RMSE early, while others often bias toward locality.

We assign a weighted PageRank score to each node to compare node ranking between sampled and full networks with goal to preserve ranking order. Link Fusion [19] algorithm is used, where the parameters for modified PageRank have smoothing factors 0 and weights α_M, β_{MN} assigned with $RP(type = M|type = M), RP(type = N|type = M)$. We measure ranking similarity using Kendall-Tau, over all node types (n_t = nodes of type t):

$$\tau = \sum_{t \in NT} \frac{(concordant\ pairs_t) - (discordant\ pairs_t)}{|NT| * \binom{n_t}{2}} \quad (6)$$

Figure 3 shows the results (scaled to [0,1]). Our method competes well and beat baselines in approximating node importance. Generally, Kendall-Tau rises pretty quickly for Patent and DBLP networks. This is desirable behavior as very few nodes are required to obtain correct relative importance of the nodes while maintaining node/edge type diversity.

B. Introducing Prediction Tasks and RP-based features

Our next evaluation concerns two network prediction tasks: In **Node Type Prediction**, a node's type information is often not available in a network. We want a model that predicts the node's label, which is a multi-class classification problem, where each training instance takes a node n, with node type as class labels. In **Link Existence Prediction**, missing links often exists due to incomplete observation. Prediction of link is

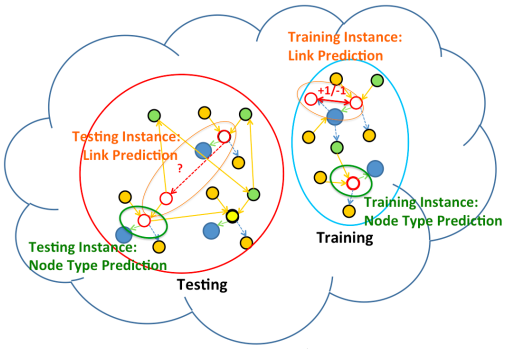


Fig. 4. Prediction Tasks Illustration. In Green: Node Type Prediction; In Orange: Edge Existence Prediction

a binary classification problem, where each training instance is based on node pair (n_s, n_t). Label is 1 if there is a link between the node pair and 0 otherwise.

1) Data Preparation: For data collection, we gradually sample 500 nodes as the training data. In **Node Type Prediction**, testing instances are chosen from unsampled portion of the network. Sampled network is used to train a prediction model. In **Link Prediction**, positive instances are edges in the considered portion of network. Negative examples are node pairs without edge connection, and are down-sampled for balanced class proportion. We set the total testing instances to be 20000. See Figure 4 for illustration. For the experiments, we use Logistic Regression as our model, implemented using LIBLINEAR [2] . For multi-class classification, one-versus-all classifiers are deployed.

2) Features: We compare set of topological features (referring to [17]) with Relational Profile-based features. Readers may see Table III for baseline features. If function f takes one variable, it denotes single node feature, whereas two variables denotes a pair of nodes (s, t) with starting and target nodes s and t are involved.

Feature Name	Formulation
Single Node Features	
Degree: $f_{deg}(n)$	$count(N(n))$
Neighboring degree: $f_{new\_deg}(n)$	$\sum_{z \in N(n)} \frac{deg(z)}{\lvert N(n) \rvert}$
Node Type: $f_{nt}(n)$	$\frac{\lvert \forall z \in N(n) \ s.t. \ type(z)=t \rvert}{\lvert N(n) \rvert}$
Two Node Features	
Common Neighbors: $f_{cn}(s,t)$	$\lvert N(s) \bigcap N(t) \rvert$
Jaccard Coefficient: $f_{jc}(s,t)$	$\frac{\lvert N(s) \bigcap N(t) \rvert}{\lvert N(s) \bigcup N(t) \rvert}$
Preferential Attachment: $f_{pa}(s,t)$	$\lvert N(s) \rvert * \lvert N(t) \rvert$
Adamic Adar: $f_{aa}(s,t)$	$\sum_{z \in N(s) \bigcap N(t)} \frac{1}{log \lvert N(z) \rvert}$
Relational Profile Features	
RP_{node} feature: $f_{RP_{node}}(n)$	Equation (8)
RP_{path} feature: $f_{RP_{path},l}(s,t)$	Equation (9)

TABLE III. FEATURES USED IN PREDICTION EXPERIMENTS

For the design of Relational Profile-based features, we first define node type scores. RP_{node} **feature** $f_{RP_{node}}(n)$: is expected type distribution given node n's observed neighbors, based on Equation (5), with distribution $P(type(n) = t|G_s)$ =

$$\prod_{i \in N(n)} RP(type(i|n)) * P(type(n) = t) \tag{7}$$

For each pair (s, t) of nodes, we define RP_{path} **feature**:

$$f_{RP_{path},l}(s,t) = \sum_{p \in T(s,t)} P(type(p)) \tag{8}$$

where $T(s, t)$ denotes the set of all paths with maximum length l from s to t. $P(type(p))$ is approximated with the sum of products of bigram probabilities, which calculates the collective effect of all possible meta-paths from s to t:

$$\frac{1}{Z} \sum_{p \in T(s,t)} \prod_{(n_1,n_2) \in p} P(type(n_2)|type(n_1)) \tag{9}$$

Where $(n_1, n_2) \in p$ denotes each edge (n_1, n_2) in the path, and $P(type(n_2)|type(n_1))$ is calculated using $RP(type(n_2)|type(n_1))$. Z is a normalization factor.

In short, Relational Profile-based features emphasize the essence of significance in type-based information propagation along the network topology.

C. Evaluate Usefulness of RP in Prediction Results

Node Type Prediction: We wish to test whether RP_{node} is a useful feature set for node type prediction. In this prediction task, for each testing node, its neighbors' information, including node type and connecting edges' type are used for feature generation. For the experiments, different combinations of features were tried, as shown in Figure 5(a)(c)(e). We first find that as more nodes as sampled, RP_{node} incorporated model's accuracy improves, and stabilizes quickly with small amount of nodes sampled. Looking at the the largest sample size (500 in this case), we see that using proposed combined feature set can increase prediction accuracy up to 4.4%. This shows the nature of our proposed RP_{node}: it emphasizes more on whether the type information for all neighboring nodes is consistent with global Relational Profile, and effectively acts as regularization, complementing topological information.

Link existence prediction: Since we are predicting sparse instances, we use Area Under ROC Curve (AUC) as the evaluating measure. For the features, $f_{baseline\_link}$ denotes the set of all baseline features (topological features). We compute single node topological features $f_{deg}(s, t)$ through $f_{deg}(s) + f_{deg}(t)$. For our proposed feature RP_{Path}, we set maximum path length to be 2. From Table IV, it shows that after 500 nodes are sampled, using RP_{path} features is possible to build a better link prediction model than without them. The implication of this observation is that RP_{path} differs from other topological features in how each neighbor is weighted: type information is further included.

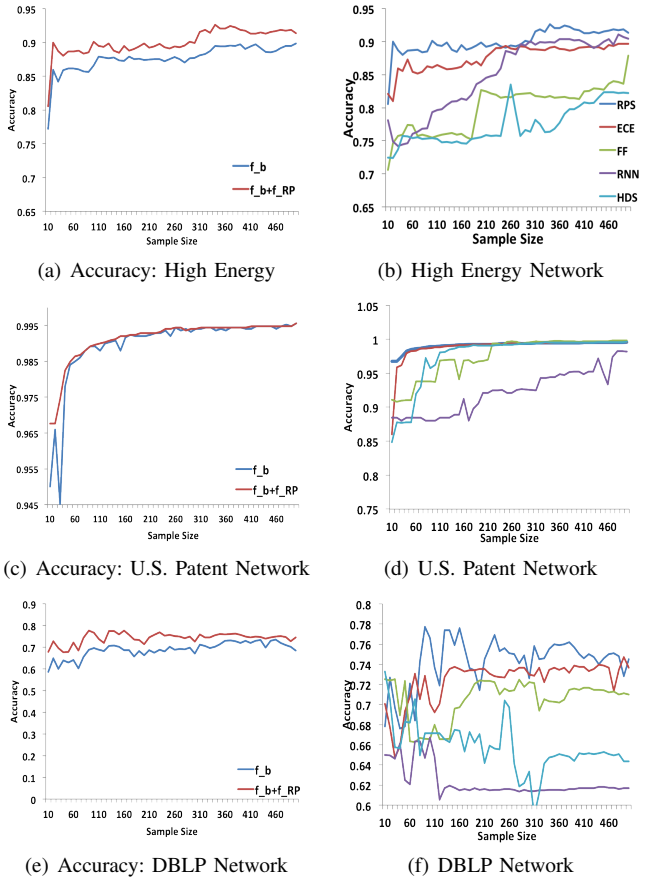
Without prior knowledge of schema, or type dependency information, pure topology features may face bias towards statistically significant populations and neglect types that have much lower ratios in comparison to frequently appearing types.

Features	HepTh	Patent	DBLP
All $f_{baseline\_link}$	0.929	0.864	0.714
All $f_{baseline\_link}$+$f_{RP_{path}}$	**0.950**	**0.867**	**0.753**

TABLE IV. PREDICTION ACCURACIES ON THREE DATASETS UNDER DIFFERENT FEATURE COMBINATIONS AFTER 500 NODES ARE SAMPLED

D. Comparing Different Sampling Methods

In this section, we focus on the quality of prediction performance between different sampling methods. For the feature sets, f_{deg} + f_{nei_deg} + f_{nt} + RP_{node} is used to train the prediction models for node type and edge existence prediction tasks, respectively.



(a) Accuracy: High Energy (b) High Energy Network

(c) Accuracy: U.S. Patent Network (d) U.S. Patent Network

(e) Accuracy: DBLP Network (f) DBLP Network

Fig. 5. (a,c,e): Node prediction accuracy for two feature sets f_b:$\{f_{deg}$+f_{nei_deg}+$f_{nt}\}$, f_b+$f_{RP_{node}}$:$\{f_{deg}$+f_{nei_deg}+f_{nt}+$f_{RP_{node}}\}$ (b,d,f): Node Type Prediction Tasks Results for Different Sampling Methods: Sample Size v.s. Accuracy. Feature set:$\{f_{deg} + f_{nei\_deg} + f_{nt} + f_{RP_{node}}\}$

First, for node type prediction, Figure 5 presents prediction accuracy under different subgraph size and sampling methods. Generally, using Relational Profile Sampling consistently outperforms other baselines, as shown. In addition, the accuracy of RPS reaches upper limit quickly and stays stabilized, which averts possible noise due to local network topology.

For link existence prediction, consider Figure 6, where we show results for the Patent Network for brevity. We see RPS outperforms other methods and reaches ceiling early. Also, RPS gives a consistent performance with varying size.

To summarize, RP is a meaningful property since minimizing its error can indeed achieve better prediction results, and RPS sampling is effective in that it generates a more representative subgraph for predictive model training.

VI. Conclusions

Heterogeneous social network is represented via a compact subgraph is crucial for effective analysis. We explore heterogeneous graph sampling based on generic explorative sampling algorithms. At the same time, Relational Profile-preserved graph sampling for heterogeneous network is proposed. We design a series of experiments to verify the validity and usefulness of not only the sampling algorithm but also the proposed property. The future work includes the speed-up of our sampling method as well as the design of a more general, Bayesian-inference driven prediction model for sampling.

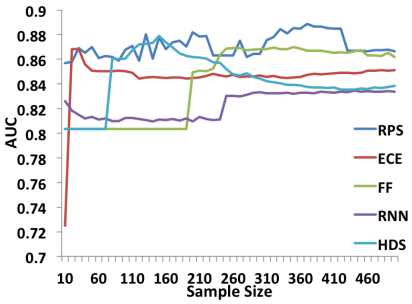


Fig. 6. Link Existence Prediction: Sample Size v.s. AUC on Patent Network

Acknowledgments

This work was supported by III Innovative and Prospective Technologies Project of the Institute for Information Industry, subsidized by the Ministry of Economy Affairs of the Republic of China, and Intel Corporation under grant NTU102R7501.

References

[1] Ahmed, N. and Berchmans, F. and Neville, J. and Kompella, R., "Time-based sampling of social network activity graphs", *Workshop on Mining and Learning with Graphs*, 2010

[2] Chang, C. and Lin, C., "LIBSVM: A library for support vector machines", *ACM Transactions on Intelligent Systems and Technology*, 2-27:127:27, 2011

[3] Gjoka, M. and Butts, C.T. and Kurant, M. and Markopoulou, A., "Multigraph Sampling of Online Social Networks", *IEEE Journal on Selected Areas in Communications*, 29(9):1893 -1905, October 2011

[4] Gjoka, M. and Butts, C.T. and Kurant, M. and Markopoulou, A.,"Walking in Facebook: A Case Study of Unbiased Sampling of OSNs",*INFOCOM, 2010 Proceedings IEEE*, p.p. 1-9,2010

[5] Henzinger, M. and Heydon, A. and Mitzenmacher, M. and Najork, M., "On near-uniform URL sampling", *WWWC*, pages 295-308, 2000

[6] Hübler, C. and Kriegel, H. and Borgwardt, K. and Ghahramani, Z., "Metropolis Algorithms for Representative Subgraph Sampling", *ICDM*, pages 282-293, 2008

[7] V. Krishnamurthy and M. Faloutsos and M. Chrobak and L. Lao and J.-H. Cui and A. G. Percus., "Reducing large internet topologies for faster simulations. In Networking", *IN IFIP NETWORKING*, 2005

[8] Kurant, M. and Gjoka, M. and Butts, C. and Markopoulou, A., "Walking on a graph with a magnifying glass: stratified sampling via weighted random walks," *SIGMETRICS*, 2011

[9] Leskovec, Jure and Faloutsos, Christos. "Sampling form large graphs", *SIGKDD*, pages 631-636, 2006

[10] Li, Jhao-Yin and Yeh, Mi-Yen, "On sampling type distribution from heterogeneous social networks", *PAKDD*, pages 111-122, 2011

[11] Lszl Lovsz, "Random Walks on Graphs: A Survey", 1993

[12] Lu, L. and Zhou, T., "Link Prediction in Complex Networks: A Survey", *CoRR*, abs/1010.0, 2010.

[13] Maiya, A. and Berger-Wolf, T., "Benefits of Bias: Towards Better Characterization of Network Sampling", *SIGKDD*, 2011

[14] Navlakha, S. and Rastogi, R. and Shrivastava, N., "Graph summarization with bounded error", *SIGMOD*, pages 419-432, 2008

[15] Ribeiro, Bruno and Towsley, Don, "Estimating and sampling graphs with multidimensional random walks", *SIGCOMM*, 2010

[16] Stutzbach, D. and Rejaie, R. and Duffield, N. and Sen, S. and Willinger, W., "On unbiased sampling for unstructured peer-to-peer networks", *IEEE/ACM Transactions on Networks*, Apr. 2009

[17] Yang, Y. and Chawla, N. and Sun, Y. and Han, J., "Predicting Links in Multi-relational and Heterogeneous Networks," ICDM, 2012

[18] Ye, Shaozhi and Lang, Juan and Wu, Felix,"Crawling Online Social Graphs",*APWeb*, pages 236-242, 2010

[19] Xi, W. and Zhang, B. and Chen, Z. and Lu. Y. and Yan, S. and Ma, W. and Fox, E., "Link Fusion: A Unified Link Analysis Framework for Multi-Type Interrelated Data Objects", *WWW*, 2004