Random Walk-based Graphical Sampling in Unbalanced Heterogeneous Bipartite Social Graphs

Yusheng Xie Zhengzhang Chen Lu Liu Ankit Agrawal Alok Choudhary
Northwestern University, Evanston, IL USA
{yxi389,zzc472,llg183,ankitag,choudhar}@eecs.northwestern.edu

ABSTRACT

We investigate sampling techniques in unbalanced heterogeneous bipartite graphs (UHBGs), which have wide applications in real world web-scale social networks. We propose random walked-based link sampling and stratified sampling for UHBGs and show that they have advantages over generic random walk samplers. In addition, each sampler's node degree distribution parameter estimator statistic is analytically derived to be used as a quality indicator. In the experiments, we apply the two sampling techniques, with a baseline node sampling method, to both synthetic and real Facebook data. The experimental results show that random walk-based stratified sampler has significant advantage over node sampler and link sampler on UHBGs.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Random walk, Network sampling, Heterogeneous bipartite graphs, Social networks

1. INTRODUCTION

A bipartite graph is a graph whose vertices can be divided into two disjoint sets U and W such that every edge connects a vertex in U to one in W. More formally, a bipartite graph G is defined as $G = (U \cup W, E)$ where $U = \{u_i | 1 \leq i \leq |U|\}$, $W = \{w_j | 1 \leq j \leq |W|\}$, and $E \in U \times W$ [9]. G is called a heterogenous bipartite graph (HBG) when its vertices from U and W model physically distinct categories [3].

Many popular online activities from social networks can be naturally modeled as HBGs [3]. In the study of employment and labor market dynamics, the individual users on LinkedIn can be considered as U, the *user nodes*, and the companies or employers as W, the *wall nodes*. Another example, from celebrities' or public brands' point of view, would be the fan engagement on social channels like Facebook. In this case, W consists of the public pages such Justin Bieber[1], Samsung Mobile USA[2], etc.; and U consists of individual Facebook users who "like" any such public pages. On websites like Youtube, the uploaded videos comprise W; and the users who view or comment on the videos comprise U.

The overwhelming popularity of online social networks has enriched web data with evolving interactions and communities both at mega-scale and in real-time. Social media is producing massive amounts of data with volume, velocity, and variety at an unprecedented scale. A special class of bipartite graphs, *unbalanced* heterogeneous bipartite graphs (UHBGs), is emerging from web-scale data. For example, Facebook may have over 1 billion active users, but only about 2,500 official public pages are registered on Facebook[3]. Now investigating Facebook user's interest distribution among public figures and brands would entail building a huge UHBG, where U contains a few thousand wall nodes and W contains hundreds of millions of user nodes.

One way to overcome the difficulty in processing ever-growing UHBGs is to use sampling. But generic sampling methods can be inefficient for UHBGs because of the large node degree difference among the nodes on the two sides of the UHBG. On the other hand, most UHBGs constructed from social data have known node degree distributions, which could be used to improve current generic sampling methods if properly incorporated. Being aware of the large node degree difference among the nodes in UHBG, the sampler can better allocate resources by avoiding redundant visits to nodes with high degrees.

1.1 Our contributions

In this paper, we develop two sampling techniques, link sampling and stratified sampling for UHBGs, based on random walk by incorporating the node degree distribution information. Instead of simply taking uniformly random links, our random walk-based link sampling deploys Metropolis-Hastings algorithm to uniformly sample nodes. The random walk-based stratified sampling further improves on link sampling by selecting the user nodes more efficiently. In addition, each sampler's node degree distribution parameter is analytically estimated and Maximum Likelihood Estimator (MLE) is used as a metric of the sampled network. We evaluate the performance of the two samplers with a baseline node sampling method on both synthetic and real datasets.

[1]http://www.facebook.com/JustinBieber
[2]http://www.facebook.com/SamsungMobileUSA
[3]According to http://voxsup.com

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2507822.

Algorithm 1: Random Walk-based Link Sampling

Input: $\widehat{\beta_W}$ and $\widehat{\beta_U}$, desired sampling densities
Output: W', sampled wall nodes; U', sampled user nodes;E', sampled edges

$w \leftarrow$ initial wall node
$\beta_W \leftarrow 0, \beta_U \leftarrow 0, W' \leftarrow \{w\}, U' \leftarrow \{\}, E' \leftarrow \{\}$
while $\beta_W \leq \widehat{\beta_W}$ *or* $\beta_U \leq \widehat{\beta_U}$ **do**
  $u \leftarrow$ a random neighbor of w
  $v \leftarrow$ a random wall node from W'
  if $\beta_W \leq \widehat{\beta_W}$ **then**
    $w \leftarrow$ a random neighbor of u
  else
    $w \leftarrow$ a random wall node from W'
  end
  if *True with probability* $P^{MH}_{w,v}$ **then**
    append (u, w) to E'; append u to U'
    append w to W'; increment β_U
    increment β_W when $\beta_W \leq \widehat{\beta_W}$
  end
end
return W', U', E'

Algorithm 2: Random Walk-based Stratified Sampling

Input: $\widehat{\beta_W}$ and $\widehat{\beta_U}$, desired sampling densities
Output: W', sampled wall nodes; U', sampled user nodes;E', sampled edges

$w_i \leftarrow$ initial wall node
$\beta_W \leftarrow 0, \beta_U \leftarrow 0, W' \leftarrow \{w_i\}, U' \leftarrow \{\}, E' \leftarrow \{\}$
while $\beta_W \leq \widehat{\beta_W}$ *or* $\beta_U \leq \widehat{\beta_U}$ **do**
  $w_k \leftarrow$ a random wall node from W'
  append w_i to W'
  $\beta_{U,i} \leftarrow 0$
  if $\beta_W \leq \widehat{\beta_W}$ **then**
    $w_i \leftarrow$ a random neighbor of u
    increment β_W with probability $P^{MH}_{w_i,w_k}$
  else
    $w_i \leftarrow$ a random wall node from W'
  end
  if *True with probability* $P^{MH}_{w_i,w_k}$ **then**
    while $\beta_{U,i} \cdot |w_i| \leq \beta_U \cdot \sum_{w_j \in W'} |w_j|$ **do**
      $u \leftarrow$ a random neighbor of w_i
      $s \leftarrow$ a random user node from U'
      if *True with probability* $P^{MH}_{u,s}$ **then**
        append (u, w) to E'; append u to U'
        increment $\beta_U, \beta_{U,i}$
      end
    end
  end
end
return W', U', E'

2. RELATED WORK

Bipartite graphs, especially heterogeneous bipartite graphs emerge as a central topic in many social studies [3]. Investigating online user's interest distribution among public brands and celebrities from UHBGs is an important topic and is a prerequisite for many popular applications such as online recommendation systems [9]. To produce real-time recommendations is often desired in most online activities such as video recommendation on Youtube. In order for collaborative filtering algorithms to efficiently serve the users, robust sampling techniques are very useful. But to the best of our knowledge, little work has been done on the sampling techniques for heterogeneous bipartite graphs.

[1] and [2] investigate novel samplers in the domain of large (social) graphs. The techniques described in [1] and [2] are directly based on random walk and the Metropolis-Hastings sampler [4] and are applied to general graphs. Powerful and general as they are, these methods are not the best fit for UHBGs. Because in UHBGs the degree of a wall node can be orders of magnitude higher that that of a user node, a generic random walker will be skewed and "trapped" by the wall nodes due to their high degrees. Directly applying Metropolis-Hastings algorithm would give each node an equal chance to be picked but this is done by controlling the traditional probability, which can prolong sampling time to achieve the same sampling density. In this work, we combine random walk and Metropolis-Hastings algorithm and modify the techniques to better suit the characteristics of UHBGs.

3. SAMPLING A SOCIAL NETWORK

3.1 Assumptions, evaluations, and notations

A sampled graph should, at least on a statistical level, preserve certain properties essential to a graph. A graph's node degree distribution is one of such properties. Previous works such as [7] and [6] describe power-law distributions that are ubiquitous in social networks. In our analysis, we assume that the node degrees in an UHBG follow discrete power-law distribution [7], $\mathbf{pow}(\gamma)$, whose probability mass function (pmf) is $p_{\mathbf{pow}}(k|\gamma) = \zeta^{-1}(\gamma) \cdot k^{-\gamma}$, $k = 1, 2, ...$, where γ is the distribution parameter and $\zeta(\gamma)$, the Riemann zeta function, is necessary to normalize $p_{\mathbf{pow}}(k|\gamma)$ to a proper pmf.

In the following analysis, X_i denotes a $\mathbf{pow}(\gamma_W)$ random variable for a wall node and Y_i denotes a $\mathbf{pow}(\gamma_U)$ random variable for a user node. β denotes the sampling fraction, which is calculated as the ratio of sampled nodes to all nodes in the original graph.

3.2 Random Walk Sampling

Random walk on graphs is a well studied topic [4]. A traditional random walker moves from current node w to the next node v by choosing v uniformly from the neighbors of w. That is, the random walker moves from w to v with probability:

$$P^{RW}_{w,v} = \begin{cases} \frac{1}{|w|}, & \text{if } v \text{ is a neighbor of } w \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

In addition to the basics, a practical random walker often has a reset probability, which, at each step, can send the walker back to the starting node with certain probability. To prevent the random walker from being stuck within a component or clique, which could happen if the walker starts very close to a major sink of the graph, a random walker usually executes several times on the same graph with different starting nodes.

A major drawback of random walk sampling is the bias it introduces into the sampled graph. Nodes with higher degrees are much more likely to be chosen because $P^{RW}_{w,v}$ makes the probability of a node to be visited proportional to its degree [1].

3.3 Random Walk-based Link Sampling

Generally speaking, link sampling (LS) *randomly* selects links between the heterogeneous nodes. The selected links and connecting nodes are kept in the sampled graph. Often in practice, *randomly* is technically interpreted as *uniformly random*. However, uniform LS is biased towards high-degree nodes in unbalanced graphs like UHBGs because it is implemented as a random walk algorithm. Such bias in the sampling process can be avoided by introducing proper transitional probabilities. Instead of using $P^{RW}_{w,v}$ from Equation 1, The Metropolis-Hastings algorithm [5] provides an alternative $P^{MH}_{w,v}$ in Equation 2.

Unlike other applications of the Metropolis-Hastings algorithm in random walk sampling [1], w and v in Equation 2 are not neighbors but homogeneous nodes on the same side in a UHBG. Based on this idea, Algorithm 1 implements a random walk-based LS, which has a transitional probability $P^{MH}_{w,v}$ at each step.

$$P^{MH}_{w,v} = \begin{cases} \min\left(\frac{1}{|w|}, \frac{1}{|v|}\right), & \text{if } w, v \in W' \text{ and } w \neq v \\ 1 - \sum_{x \neq v} P^{MH}_{w,x}, & \text{if } w = v \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Algorithm 1 is unique to UHBGs and brings several advantages over generic random walkers. First, it allows the two sides to be sampled *independently* with independent sampling densities. The number of nodes (and the degrees of the nodes) vary so much on two sides of UHBGs that it does not make sense to sample them using any homogeneous algorithms. Second, Algorithm 1 is still based on the Metropolis-Hastings algorithm and, within either side of the UHBG, can sample the nodes with even probability.

In addition to describing the LS sampling algorithm, we need to derive the Maximum Likelihood Estimator (MLE) formula for the distribution parameter in the node degree distribution as it is used to evaluate the sampling performance. By the symmetry of UHBGs, it suffices to analyze the node degree estimation from wall nodes. Suppose, for $1 < i < \beta_W|W|$, $X_i \sim \mathbf{pow}(\gamma_W)$ from the original graph. Then we are able to implicitly solve for $\widehat{\gamma_W}(LS)$, the MLE for γ_W from link sampling, as

$$\beta_W|W|\zeta'\left(\widehat{\gamma_W}(LS)\right) + \zeta\left(\widehat{\gamma_W}(LS)\right) \sum_{i=1}^{\beta_W|W|} \ln(X_i) = 0, \quad (3)$$

and $\widehat{\gamma_U}(LS)$ is derived in an analogous fashion. Because of the involvement of the Riemann zeta function, we cannot find satisfactory closed forms for the MLEs.

3.4 Random Walk-based Stratified Sampling

A practical problem with the random walk-based LS is that when it is applied to web-scale UHBGs, it actually cannot efficiently sample the user nodes due to the low acceptance rate of $P^{MH}_{w,v}$ in Equation 2. In UHBGs, wall nodes tend to have very high degrees, which diminishes the probability of acceptance, $\min\left(\frac{1}{|w|}, \frac{1}{|v|}\right)$.

A workaround of this problem is to use stratified sampling (SS) and sample the user nodes independently for each sampled wall node. Once a wall node w_k is chosen by probability $P^{MH}_{w_i,w_k}$, the SS sampler will again apply the Metropolis-Hastings algorithm to sample the user nodes linked to w_k. This scheme solves two problems. First, it improves the overall user node sampling efficiency. Second, it accommodates the degree differences among the wall nodes. That is, the number of user nodes linked to w_k visited by SS is proportional to the degree of w_k. Algorithm 2 incorporates the above ideas and implements random walk-based SS.

In addition to Algorithm 2, SS also has a different MLE formula from LS method. Select $X_i \sim \mathbf{pow}(\gamma_W)$ for $1 < i < \beta_W|W|$. Then for each X_i, user nodes linked to w_i are selected: $Y_{i,j} \sim \mathbf{pow}(\gamma_{U,i}) \cdot 1_{\{\text{linked with } W_i\}}$, for $1 < j < \beta_U|w_i|$, where γ_U is normalized to $\gamma_{U,i}$ due to the indicator function. We are able to implicitly solve for each MLE $\widehat{\gamma_{U,i}}$ as

$$\beta_U|w_i|\zeta'\left(\widehat{\gamma_{U,i}}\right) + \zeta\left(\widehat{\gamma_{U,i}}\right) \sum_{j=1}^{\beta_U|w_i|} \ln(Y_{j,i}) = 0. \quad (4)$$

We then propose $\widehat{\gamma_U}(SS) = \sum_{i=1}^{\beta_W|W|} \widehat{\gamma_{U,i}} \cdot |w_i|/|W|$ as the SS MLE for γ_U. In our estimator, parameters β_W and $\gamma_{i,U}$ control the stratification and allocation, respectively. The formula for $\widehat{\gamma_W}(SS)$ is the same as $\widehat{\gamma_W}(LS)$.

4. EXPERIMENTS

4.1 Baseline and Datasets

Node sampling (NS) is employed as a baseline sampler in our experiments. NS randomly selects a number of nodes. And only those links, both of whose end nodes are in the selection, are kept in the sampled graph.

Three datasets are used in our experiments:

SYN Synthesized data set that contains 2000 wall nodes, 1 million user nodes and 4 million edges. The dataset is generated using power law distributions with γ_W =1.830 and γ_U =1.295.

FB Public data collected from Facebook's Graph API from 2011 January to 2012 March. 5831 wall nodes are chosen to be the most popular public walls on Facebook in terms of page likes. 143 million user nodes, as well as 520 million edges, are included in this dataset. Part of this Facebook dataset has been publicly release [8]

LNK Public data collected from LinkedIn. This dataset contains 34 wall nodes, each of which corresponds to a public profile of an employer, 1.1 million associated user nodes, and 4.2 million edges.

Note that the three datasets are chosen to have different statistical characteristics. Table 1 summarizes important properties of the three datasets. Table 2 summarizes the estimations from the three mentioned samplers: Node Sampler (NS), Link Sampler (LS), and Stratified Sampler (SS). Table 2 also provides the sampling densities, β_W and β_U, on each dataset.

4.2 Performance of the samplers

Table 2 presents the three MLE estimates of γ_W and γ_U for the datasets SYN, FB, and LNK. The estimates obtained by using full data are regarded as ground truth. bold face cells indicate the best estimates. Overall, SS performs the best among all three methods. Both LS and SS outperform the baseline NS.

Table 1: SYN, FB1, and FB2 datasets

Metric	SYN	FB	LNK
Wall nodes	2,000	5,831	34
User nodes	1 million	143 million	1.1 million
Edges	4 million	520 million	4.2 million
Edges per wall	2,000	89,179	123,529
Edges per user	4.00	3.64	3.82
users per wall	500	24,524	32,353

Table 2: Parameter estimations

Parameter	SYN	FB	LNK
β_W	0.1000	0.0343	0.1321
β_U	0.05000	0.0006981	0.01352
True γ_W	1.830	1.706	2.383
$\widehat{\gamma_W}$ NS	1.533	2.259	1.552
$\widehat{\gamma_W}$ LS	1.601	1.394	2.898
$\widehat{\gamma_W}$ SS	**1.759**	**1.466**	**2.409**
True γ_U	1.295	2.289	10.75
$\widehat{\gamma_U}$ NS	1.508	3.501	12.08
$\widehat{\gamma_U}$ LS	1.579	3.029	7.205
$\widehat{\gamma_U}$ SS	**1.392**	**1.894**	**9.743**

Figure 1 visualizes the findings from Table 2. The "Full Data" curves in the mentioned figures are the actual *discrete* degree distribution from the entire bipartite graphs. The other three smooth curves are *continuous* power-law distributions generated from the parameters estimated by the three sampling methods.

5. CONCLUSION AND FUTURE WORK

In this work, we addressed the sampling problem in a class of unbalanced heterogeneous bipartite graphs (UHBGs) from social networks and proposed two random walk-based sampling techniques for the UHBGs. The experimental results showed that our techniques outperform the baseline approaches on both synthetic dataset and real world datasets. In the future, we plan to investigate the impact of each sampling technique on other statistical and topological properties of the graphs.

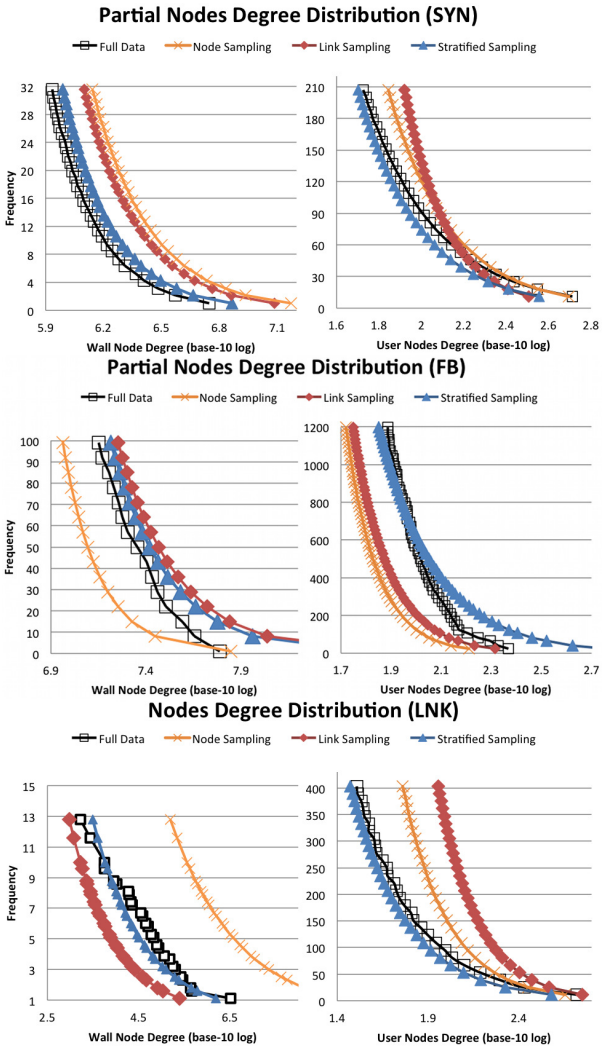

Figure 1: Estimation results from three samplers on all three datasets. Estimations are compared with the actual degree distributions.

6. ACKNOWLEDGMENTS

This work is supported in part by the following grants: NSF awards CCF-0833131, CNS-0830927, IIS-0905205, CCF-0938000, CCF-1029166, and OCI-1144061; DOE awards DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, DESC0005340, and DESC0007456; AFOSR award FA9550-12-1-0458.

7. REFERENCES

[1] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In *INFOCOM*, March 2010.

[2] J. Leskovec and C. Faloutsos. Sampling from large graphs. KDD '06, pages 631–636. ACM, 2006.

[3] L. Liu, J. Tang, J. Han, and S. Yang. Learning influence from heterogeneous social networks. *Data Min. Knowl. Discov.*, 25(3):511–544, 2012.

[4] L. Lovász. Random walks on graphs: A survey, 1993.

[5] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[6] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64:026118, Jul 2001.

[7] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *PNAS*, 99(Suppl 1):2566–2572, 2002.

[8] Y. Xie, Z. Chen, K. Zhang, Y. Cheng, A. Agrawal, W. keng Liao, and A. Choudhary. Detecting and tracking disease outbreaks in real-time through social media. In *IJCAI*, 2013.

[9] K. Zhang, Z. Chen, Y. Cheng, Y. Xie, D. Downey, A. Agrawal, W. keng Liao, and A. Choudhary. A probabilistic graphical model for brand reputation assessment in social networks. In *ASONAM '13*, 2013.