Coarse-Grained Topology Estimation via Graph Sampling

Maciej Kurant *
Commun. Systems Group
ETH Zurich
maciej.kurant@gmail.com

Minas Gjoka
CalIT2
UC Irvine
mgjoka@uci.edu

Yan Wang
CalIT2
UC Irvine
wang.yan@uci.edu

Zack W. Almquist
Sociology Dept, CalIT2
UC Irvine
almquist@uci.edu

Carter T. Butts
Sociology Dept, CalIT2, IMBS
UC Irvine
buttsc@uci.edu

Athina Markopoulou
EECS, CalIT2, CPCC
UC Irvine
athina@uci.edu

ABSTRACT

In many online networks, nodes are partitioned into *categories* (*e.g.,* countries or universities in OSNs), which naturally defines a weighted *category graph i.e.,* a coarse-grained version of the underlying network. In this paper, we show how to efficiently estimate the category graph from a probability sample of nodes. We prove consistency of our estimators and evaluate their efficiency via simulation. We also apply our methodology to a sample of Facebook users to obtain a number of category graphs, such as the college friendship graph and the country friendship graph. We share and visualize the resulting data at `www.geosocialmap.com`.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Measurement techniques

General Terms

Measurement, Algorithms

Keywords

Online Social Networks, Coarse-Grained Topology, Estimators, Induced Subgraph Sampling, Star sampling, Facebook.

1. INTRODUCTION

Many large online networks, such as online social networks (OSNs) and the World Wide Web (WWW), are currently studied via sampling techniques. Sampling becomes necessary due to the sheer size of these networks and/or access limitations, which make it infeasible to collect (and, in some cases, to analyze) these networks in their entirety.

Most principled graph sampling methods to date have focused on collecting a probability sample of nodes [4,5,9,13, 22,24,26]. Based on such a sample, one can efficiently estimate many local graph properties, such as node attribute frequency, degree distributions, or clustering coefficients [11, 24]. However, these features reveal little about the global properties of the underlying graph.

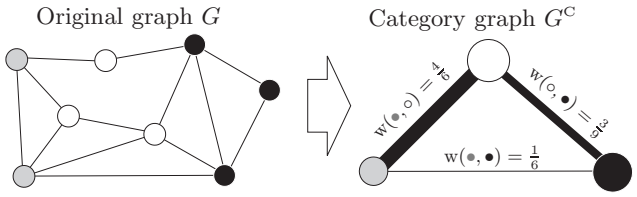

Figure 1: Illustration. Three node categories in the original graph G define a three-node weighted category graph G^{C}.

In this paper, we show how a particular aspect of global network structure, namely *coarse-grained topology*, can be efficiently estimated from a probability sample of nodes. Specifically, we note that nodes in many online graphs belong to *categories*, explicitly declared by users or clearly determined by observable characteristics. For example, in Facebook, users can officially declare the college or workplace with which they are affiliated, or a country/city in which they live. Similarly, in the WWW, all nodes can be categorized by their domain names, and the users of Internet radio sites like `Last.FM` may be grouped on the basis of listening behavior. This potentially allows us to build and study *category graphs*, in which each node corresponds to a category and edge weights reflect the frequency of edges between category members in the original graph.

For example, in Fig. 1, nodes in graph G belong in one of three categories (white, gray, and black), which defines the three-node weighted category graph G^{C}. The edge weight $\mathrm{w}(\circ,\bullet)$ in G^{C} is the probability that a black and a white node, randomly chosen from G, are connected in G.

The contribution of this paper lies in developing and evaluating several efficient estimators for two properties of the category graph, namely the size of the categories and the edge weights. These estimators take as input a (uniform or non-uniform) probability sample of nodes, measured via one of two strategies: *induced subgraph sampling* [11], in which

* Maciej Kurant was with CalIT2 at UC Irvine when this work was conducted.

WOSN'12, August 17, 2012, Helsinki, Finland.

we have information regarding only the sampled nodes; and *star sampling* [11], in which we also have category information about the neighbors of sampled nodes. We show that our estimators have good asymptotic properties (consistency, and hence asymptotic unbiasedness) and we evaluate their efficiency via simulation. Finally, as a practical illustration of our approach, we estimate several Facebook category graphs, such as the college-to-college and country-to-country friendship graphs.[1] These results are made available along with a highly-customizable, web-based visualization service at `www.geosocialmap.com`.

Our methodology has potential applications not only for descriptive and visualization purposes, but also for model-based analysis. For instance, given the estimated category edge weights, one can create models and test hypotheses on how category features (rank and type of a university, language and religion of a country, geographical distance) affect the inter-category interaction rates.

2. RELATED WORK

Node sampling in graphs. In some cases, it is possible to sample nodes uniformly and independently [5,11]. However, most state-of-the-art node sampling techniques use variants of random walks (RW), such as the classic RW [5,8,19,22,25], metropolized RW (MHRW) [5,22,26], multiple dependent RW [24], multigraph RW [4], RW with jumps [9,18], and weighted RW [13]. Based on the resulting (uniform or non-uniform) sample of nodes, there exist principled methods to estimate local graph properties (degree distribution, assortativity and clustering coefficient) [2,5,6,17,22,24].

Category graphs The use of partitions to produce reduced-form versions of larger networks has an extensive history in the social network literature, primarily under the label of block modeling [30]. There, categories correspond to positions, our category graph to the reduced graph or block image, and our edge weights to block densities or mixing rates. Given the full knowledge, one can easily create a category graph (see Section 3.2 and [1]). In contrast, our contribution lies in *estimating the category graph from a sample of nodes.*

Community Structure Sampling A related line of research is in subsampling a large fully known graph in order to significantly shrink its size while keeping the resulting subgraph "similar" to the original one [18,20]. In particular, [20] considered the community structure as a measure of similarity that should be preserved. Our work is fundamentally different, because here (i) the original graph is not known, (ii) the nodes are labeled, and (iii) the sampled subgraph may be arbitrary different from the original graph.

3. NOTATION AND PROBLEM STATEMENT

3.1 Basic graph G

We consider an undirected graph $G = (V, E)$, with $N = |V|$ nodes and $|E|$ edges. Denote by $\deg(v)$ the degree of node $v \in V$, and by $\mathrm{vol}(A) = \sum_{v \in A} \deg(v)$ the volume of a set of nodes $A \subseteq V$. We will often use

$$f_A = \frac{|A|}{|V|} \quad \text{and} \quad f_A^{\mathrm{vol}} = \frac{\mathrm{vol}(A)}{\mathrm{vol}(V)} \tag{1}$$

to denote the relative size of A in terms of number of nodes and volume, respectively.

3.2 Category graph G^{C}

The nodes V are partitioned into a set $\mathcal{C}$ of *categories*, *i.e.,* that $\bigcup_{C \in \mathcal{C}} = V$. We are interested in the *category graph* $G^{\mathrm{C}} = (\mathcal{C}, E^{\mathrm{C}})$, with node set given by the categories of G. For two different categories $A, B \in \mathcal{C}$, $A \neq B$, denote by $E_{A,B} \subset E$ the corresponding edge-cut in G, *i.e.,*

$$E_{A,B} = \{\{u, v\} \in E : \ u \in A \text{ and } v \in B\}.$$

If $|E_{A,B}| > 0$ then we draw an edge $\{A, B\}$ between A and B in G^{C}. We show an example of a category graph in Fig. 1.

The way we defined category graph G^{C} so far, prevents self-loops, but potentially allows for edge weights. The *weight* $\mathrm{w}(A, B)$ of edge $\{A, B\}$ can be defined in a number of ways. For instance, one could trivially set it always equal to 1. In some settings, *e.g.,* statistical modeling, the number of inter-category edges, $\mathrm{w}(A, B) = |E_{A,B}|$, is a good choice. For many purposes, however, it is useful to have a notion of edge weight that adjusts for category size, *e.g.,*

$$\mathrm{w}(A, B) \ = \ \frac{|E_{A,B}|}{|A| \cdot |B|}. \tag{2}$$

This definition has an intuitive interpretation. Because $|A| \cdot |B|$ is the size of the maximum possible edge-cut from A to B, $\mathrm{w}(A, B)$ is equal to the probability that a uniformly selected member of A is connected to a uniformly selected member of B. We give an example of these weights $\mathrm{w}(A, B)$ in Fig. 1.

3.3 Goal: Estimate G^{C} through sampling

Given the full knowledge of graph G, it is trivial to construct the category graph. In many cases, however, the knowledge of the full graph G is not available, rendering exact computation of Eq.(2) infeasible. For instance, downloading the entire Facebook social graph via HTML scraping would require downloading and processing about 115 terabytes of uncompressed HTML data [6], which is rather prohibitive in practice (but can be easily done from inside Facebook [1]).

In contrast, it is often possible to collect a sample $S \subseteq V$ of nodes of G. The challenge, then, and the main goal of this paper is to estimate the category graph G^{C} (*i.e.,* its nodes $\mathcal{C}$ and edge weights Eq.(2)) based on the sample S.

3.4 Sampling techniques

We consider only the sampling techniques where nodes are sampled with replacement (we permit S to contain multiple copies of the same node)[2], as follows.

Uniform Independence Sampling (UIS) samples nodes with equal probabilities.

Weighted Independence Sampling (WIS) samples v with probability proportional to a known weight $\mathrm{w}(v)$.

Simple Random Walk (RW) [19] selects the next-hop node v uniformly at random among the neighbors of the current node u. On a connected and aperiodic graph, RW samples node v with probability linearly proportional to its degree $\deg(v)$.

[1] Just after the submission of this paper, Facebook released a similar study of the "friendship ties between countries" [1] that drew significant attention.

[2] The without-replacement approaches (*i.e.,* BFS, DFS) are much more difficult to analyze [15,16].

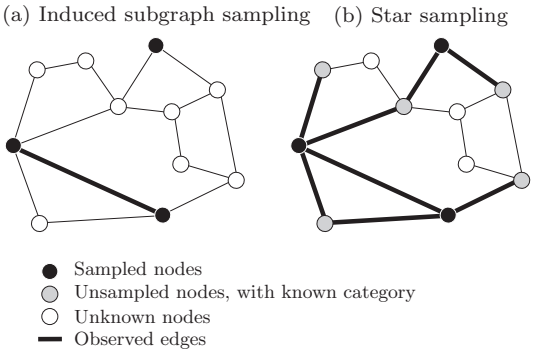

Figure 2: Observed categories and edges, under two studied scenarios.

Weighted Random Walk (WRW) is RW on a weighted graph [3]. In our simulations and implementation, we use "Stratified WRW," or S-WRW [13], *i.e.,* a version of WRW that increases the sampling efficiency by over-sampling graph regions relevant to the measurement objective and under-sampling the irrelevant ones.

3.5 Observed categories and edges

When sampling a node v, we obviously learn its category. However, in some cases we can also learn the categories of v's neighbors [11]. We therefore distinguish two graph sampling designs: *induced sampling* and *star sampling* (see Fig. 2), as follows.

Induced Subgraph Sampling learns the categories of the sampled nodes only, as shown in Fig. 2(a).

Star Sampling [5,11,13] reveals the categories (but not necessarily the degree) of *all* neighbors of a sampled node $u \in S$, as shown in Fig. 2(b).

4. ESTIMATION

In this section, we provide design-based estimators for category sizes and category graph edge weights, given a uniform independence (UIS) sample from the node set. All estimators shown in this section and in Section 4.2 are consistent; proofs are provided in our technical report [14].

4.1 Uniform Sampling (UIS)

4.1.1 Induced subgraph sampling

The size $|A|$ of category A can be estimated by multiplying by N the fraction of nodes sampled in A, *i.e.,*

$$|\widehat{A}| \;=\; N \cdot \frac{|S_A|}{|S|}, \tag{3}$$

where $S_A = \{v \in S : v \in A\}$ is a multiset containing all samples from category A.

To estimate the *edge weights*, recall from Eq.(2) that $\mathrm{w}(A,B)$ is obtained by dividing the number of edges between A and B by the maximal possible number of such edges. Analogously, under induced subgraph sampling, we *observe* $\sum_{a\in S_A}\sum_{b\in S_B} 1_{\{\{a,b\}\in E\}}$ edges between A and B, out of the maximal number $|S_A|\cdot|S_B|$ we could possibly observe, leading to the estimator

$$\widehat{\mathrm{w}}(A,B) \;=\; \frac{\sum_{a\in S_A}\sum_{b\in S_B} 1_{\{\{a,b\}\in E\}}}{|S_A|\cdot|S_B|}. \tag{4}$$

Note that when S contains the same node multiple times, we count any corresponding sampled edges multiple times as well.

4.1.2 Star sampling

Although not obvious at first blush, star sampling gives us an alternative way to estimate category sizes. Denote by

$$k_A \;=\; \frac{1}{|A|}\sum_{v\in A}\deg(v) \quad\text{and}\quad k_V \;=\; \frac{1}{|V|}\sum_{v\in V}\deg(v)$$

the average node degree in category A and in the entire graph, G, respectively. Because $\mathrm{vol}(A) = |A|\cdot k_A$, we can re-write the relative volume f_A^{vol} of category A (see Eq.(1)) as

$$f_A^{\mathrm{vol}} \;=\; \frac{\mathrm{vol}(A)}{\mathrm{vol}(V)} \;=\; \frac{|A|\cdot k_A}{|V|\cdot k_V} \;=\; \frac{|A|\cdot k_A}{N\cdot k_V}.$$

This allows us to estimate the size $|A|$ of category A as

$$|\widehat{A}| \;=\; N\cdot \widehat{f}_A^{\mathrm{vol}} \cdot \frac{\widehat{k}_V}{\widehat{k}_A}. \tag{5}$$

This formula may seem less attractive than Eq.(3), because we now have to estimate three different numbers. However, k_V and k_A can be easily estimated, respectively by

$$\widehat{k}_V = \frac{\sum_{v\in S}\deg(v)}{|S|} \quad\text{and}\quad \widehat{k}_A = \frac{\sum_{v\in S_A}\deg(v)}{|S_A|}. \tag{6}$$

Moreover, we have proposed in [13] an efficient star-based estimator of f_A^{vol}, *i.e.,*

$$\widehat{f}_A^{\mathrm{vol}} \;=\; \frac{1}{\mathrm{vol}(S)}\sum_{s\in S}\sum_{v\in\mathcal{N}(s)} 1_{\{v\in A\}}. \tag{7}$$

By plugging Eq.(6) and Eq.(7) into Eq.(5), we obtain a complex yet powerful star-based estimator of size $|A|$.

To estimate the *edge weights* under star sampling, note that on sampling node $a \in A$ we observe the set $E_{a,B} \subset E$ of all edges between a and category $B \neq A$. So we observe $|E_{a,B}|$ edges out of a potential $|B|$ edges between a and B. If we consider all nodes S_A we sampled from A, we observe $\sum_{a\in S_A}|E_{a,B}|$ out of a potential $|S_A|\cdot|B|$ edges. The same applies to nodes S_B sampled in B and their neighbors in A. Consequently, we can estimate the category graph edge weight $\mathrm{w}(A,B)$ by dividing the total number of edges we observed between A and B by our estimate of the maximal number we could potentially observe, *i.e.,*

$$\widehat{\mathrm{w}}(A,B) \;=\; \frac{\sum_{a\in S_A}|E_{a,B}| \;+\; \sum_{b\in S_B}|E_{b,A}|}{|S_A|\cdot|\widehat{B}| \;+\; |S_B|\cdot|\widehat{A}|}. \tag{8}$$

Note that because we usually do not know the real sizes of A and B, Eq.(8) uses their estimators $|\widehat{A}|$ and $|\widehat{B}|$. We can employ either Eq.(3) or Eq.(5), as needed.

4.2 Non-Uniform Sampling (WIS)

The estimators derived in Section 4.1 hold under UIS, where every node $v \in V$ is sampled with the same probability. Such a sampling design is rarely feasible in practice. A

more common scenario is *non-uniform* probability sampling, where every node $v \in V$ is sampled with probability proportional to a known weight $\mathrm{w}(v)$. Indeed, this is the case for WIS, RW, S-WRW and other principled walk-based sampling methods, provided that samples have adequately converged [5]. Non-uniform samples are by definition biased towards nodes of higher weight (typically degree), which may dramatically distort the estimation results if used without correcting for sampling probabilities [6].

Fortunately, a weighted sample can be unbiased using the Hansen-Hurwitz estimator [7] as shown *e.g.,* in [22,25,29]. Indeed, let every node $v \in V$ carry a value $x(v)$. We can estimate the population total $x_{\mathrm{tot}} = \sum_v x(v)$ by

$$\hat{x}_{\mathrm{tot}} = \frac{1}{n} \sum_{v \in S} \frac{x(v)}{\pi(v)}, \tag{9}$$

where $\pi(v)$ is the sampling probability of node v.

In practice, we usually know $\pi(v)$, and thus $\hat{x}_{\mathrm{tot}}$, only up to a constant, *i.e.,* we know the (non-normalized) weights $\mathrm{w}(v)$, $\mathrm{w}(v) \sim \pi(v)$. Fortunately, we can often address this problem by estimating the ratio of two totals, which makes the unknown constants cancel out. We will use this approach below.

4.2.1 Induced subgraph sampling

Following Eq.(9), we can estimate $|S_A|$ by setting $x(v) \equiv 1_{\{v \in A\}}$. This yields

$$|\widehat{S}_A| \;=\; \frac{1}{n} \sum_{v \in S} \frac{1_{\{v \in A\}}}{\pi(v)} \;=\; \frac{1}{n} \sum_{v \in S_A} \frac{1}{\pi(v)}.$$

Analogously, $|\widehat{S}| = \frac{1}{n} \sum_{v \in S} \frac{1}{\pi(v)}$. Consequently, we can rewrite Eq.(3) as

$$|\widehat{A}| = N \frac{\sum_{v \in S_A} \frac{1}{\pi(v)}}{\sum_{v \in S} \frac{1}{\pi(v)}} = N \frac{\sum_{v \in S_A} \frac{1}{\mathrm{w}(v)}}{\sum_{v \in S} \frac{1}{\mathrm{w}(v)}} = N \frac{\mathrm{w}_{\text{-}1}(S_A)}{\mathrm{w}_{\text{-}1}(S)} \tag{10}$$

where $\mathrm{w}_{\text{-}1}(X) = \sum_{v \in X} \frac{1}{\mathrm{w}(v)}$ is the 're-weighted size' of multiset $X \subset V$.

Now, to estimate the *edge weights*, note that in the numerator of Eq.(4), we have a sum over node *pairs*, rather than single nodes. In this case, Hansen-Hurwitz estimator divides every component by the product of weights of the two involved nodes [11], which yields

$$\widehat{\mathrm{w}}(A,B) \;=\; \frac{\sum_{a \in S_A} \sum_{b \in S_B} \frac{1_{\{\{a,b\} \in E\}}}{\mathrm{w}(a) \cdot \mathrm{w}(b)}}{\mathrm{w}_{\text{-}1}(S_A) \cdot \mathrm{w}_{\text{-}1}(S_B)}. \tag{11}$$

4.2.2 Star sampling

As in Section 4.1.2, we estimate the size of a category A using Eq.(5), *i.e.,*

$$|\widehat{A}| \;=\; N \cdot \widehat{f}_A^{\mathrm{vol}} \cdot \frac{\widehat{k}_V}{\widehat{k}_A}. \tag{12}$$

However, now, the terms $\widehat{f}_A^{\mathrm{vol}}$, $\widehat{k}_V$ and $\widehat{k}_A$ must be calculated taking into account the sampling weights. Indeed, the weighted version of $\widehat{f}_A^{\mathrm{vol}}$ is (after [13])

$$\widehat{f}_A^{\mathrm{vol}} \;=\; \frac{1}{\sum_{s \in S} \frac{\deg(s)}{\mathrm{w}(s)}} \cdot \sum_{s \in S} \left(\frac{1}{\mathrm{w}(s)} \sum_{v \in \mathcal{N}(s)} 1_{\{v \in A\}} \right). \tag{13}$$

Similarly, the estimators Eq.(6) of k_V and k_A can be rewritten respectively by

$$\widehat{k}_V = \frac{\sum_{v \in S} \frac{\deg(v)}{\mathrm{w}(v)}}{\mathrm{w}_{\text{-}1}(S)} \quad \text{and} \quad \widehat{k}_A = \frac{\sum_{v \in S_A} \frac{\deg(v)}{\mathrm{w}(v)}}{\mathrm{w}_{\text{-}1}(S_A)}. \tag{14}$$

Finally, Eq.(8) becomes

$$\widehat{\mathrm{w}}(A,B) \;=\; \frac{\sum_{a \in S_A} \frac{|E_{a,B}|}{\mathrm{w}(a)} \;+\; \sum_{b \in S_B} \frac{|E_{b,A}|}{\mathrm{w}(b)}}{\mathrm{w}_{\text{-}1}(S_A) \cdot |\widehat{B}| \;+\; \mathrm{w}_{\text{-}1}(S_B) \cdot |\widehat{A}|}. \tag{15}$$

Again, we have two size estimators Eq.(10) and Eq.(12) to choose from to plug into $|\widehat{A}|$ and $|\widehat{B}|$.

4.3 Sampling via crawling

In many online networks the only feasible sampling approach is via crawling [5]. Such techniques result in non-uniform sampling probabilities, and, consequently, sampling weights. For example, under RW the sampling weights converge asymptotically to $\mathrm{w}(v) = \deg(v)$ [19]. Using these weights in conjunction with the WIS estimators above allows for consistent estimation of coarse-grained topology from random walk samples.

4.4 Population size (N)

In our estimation of category sizes, the population size $N = |V|$ is required. In some cases N is known (e.g., in an OSN context, it may be published by the service provider), but in general this is not the case. Fortunately, where N is not available, we can turn to estimation [10,12,23]. For instance, [10] proposes an approach based on a "reversed coupon collector" problem, which can be used with both uniform and non-uniform sampling, and [12] significantly improves over [10].

Finally, we note that N is only necessary where absolute values of category sizes are required. Specifically, all edge weights and category sizes can be estimated up to a constant of proportionality without knowing the size of the total population. Thus, if we are interested in ratios of category sizes and/or edge weights (*e.g.,* the relative weight of the A, B connection versus the A, C connection in G^C), then N can be ignored (and replaced by an arbitrary constant in the above equations).

5. SIMULATION RESULTS

5.1 Objective and performance metrics

In this section, we evaluate our methodology. We use the Normalized Root Mean Square Error (NRMSE) to assess the error of our estimators:

$$\mathrm{NRMSE}(\widehat{x}) = \frac{\sqrt{\mathbb{E}\left[(\widehat{x} - x)^2\right]}}{x}, \tag{16}$$

where x is the real value and $\widehat{x}$ is the estimate.

5.2 Datasets

We consider two fully known OSN topologies:[3] Texas [27] (36K nodes, 1590K edges), and New Orleans [28] (63K nodes, 816K edges). We define as categories the 50 largest

[3]For the sake of space, an extensive study of our methodology on both synthetic and (other) real-life graphs can be found in [14].

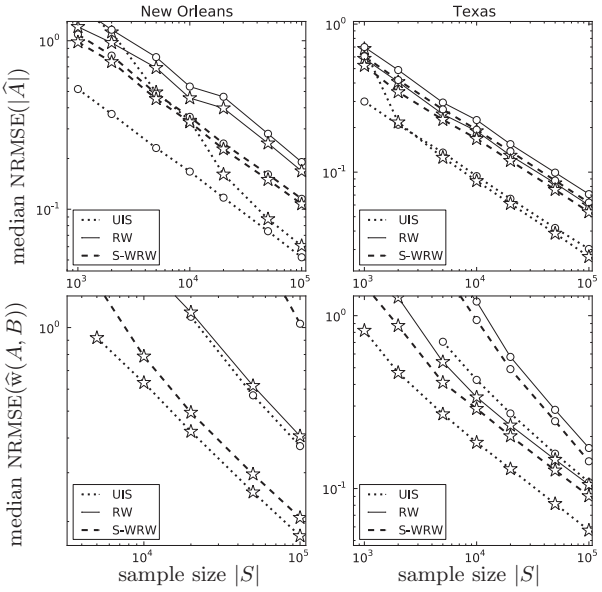

Figure 3: Simulations on real-life graphs. We estimate category sizes (top) and category edge weights (bottom), using induced subgraph sampling (circles) and star sampling (stars).

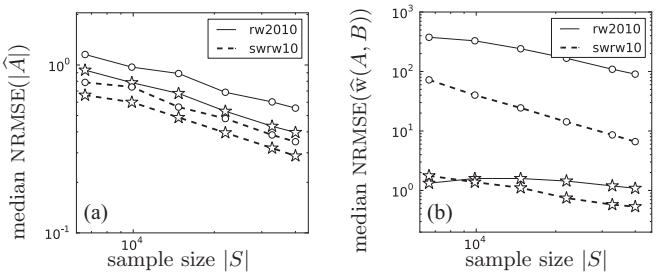

Figure 4: Results for 100 college networks: category size estimation (a), and edge weight estimation (b).

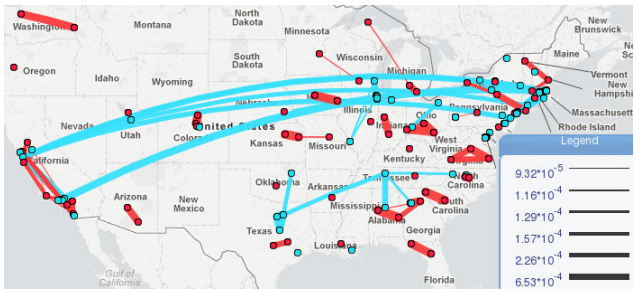

Figure 5: 100 strongest edges of the college-to-college friendship graph for top US colleges, as ranked by the "US News World Report '09".

communities in the graph, found by a standard community finding algorithm based on eigenvalues [21]. All the remaining smaller communities (if any) are then grouped together as the 51^{st} category. Next, we sample the resulting graphs by three different sampling methods: UIS, RW and S-WRW. Under S-WRW [13], we use equal category weights for all categories.

5.3 Results

We show the results in Fig. 3. First of all, in all cases the error approaches 0 with the sample size, which confirms that our estimators are consistent (asymptotically unbiased), as we show in [14].

The efficiency of the *category size* estimators (Fig. 3, top), Eq.(3) and Eq.(5), is comparable. Under UIS, induced estimators slightly outperform the star-based ones; under RW and S-WRW the latter usually perform better. This is because both RW and S-WRW visit high-degree nodes more often, and thus their star samples inherently collect and exploit more information about neighbor categories, which translates to a better performance.

While there is no clear winner in the category size estimation, in the *category edge weight* estimation, star sampling consistently and significantly outperforms induced sampling. Indeed, in Fig. 3(bottom), the induced estimators often need 5-10 times more samples to achieve the same accuracy as star estimators.

Finally, among the sampling techniques, UIS clearly performs best. Not surprisingly, direct independence sampling should be preferred whenever available. In the more practical scenarios, however, we are limited to exploration-based techniques. In our simulations, S-WRW is consistently better than RW. Note that because all categories (and thus nodes) are relevant, this advantage of S-WRW is purely due to stratification [13].

6. FACEBOOK CATEGORY GRAPHS

In this section, we use the estimators developed in this paper to infer several category graphs from Facebook.

6.1 Data sets

About 3.5% of Facebook users openly declare the college/university they attend. In our previous work [13], we collected samples of Facebook users using two methods: (i) RW (1M users), and (ii) S-WRW (tuned to oversample college students, 1M users). We discovered more than 10K colleges.

These datasets were collected using HTML scraping, which allowed us to collect for each user v not only v's category, but also the list of v's friends together with their categories. In other words, we collected a star sample of Facebook users, with no additional cost. By discarding the information about v's nodes, we can also use the induced subgraph estimators, for comparison.

6.2 Results

We present our results in Fig. 4. To calculate NRMSE we use as ground truth the average of estimation over all samples we collected.

In the estimation of Facebook *category sizes* (Fig. 4(a)), S-WRW outperforms RW, and the star version is better than induced. This is in agreement with our observations made in Section 5.

The estimation of *category edge weights* in Facebook, shown in Fig. 4(b), also confirms the observations in the simulations of Section 5. Indeed, all star estimators dramatically outperform their induced counterparts.

6.3 Geosocial visualization

Finally, we developed a highly customizable, web-based tool for visualization of our Facebook category graphs and made it available at `www.geosocialmap.com`. For example, in Fig. 5, we present a "college-to-college friendship graph" inferred from our Facebook dataset, with top 133 US colleges grouped by their (private/public) type. We observe that physical distance is a major factor for public colleges (red), but seemingly less so for private ones (blue). This and other datasets (*e.g.,* "country-to-country friendship graph", "North American friendship map") are available at `www.geosocialmap.com`.

7. CONCLUSION

In this paper, we derived a number of category graph estimators for (uniform and non-uniform) probability samples of nodes. We evaluated their performance in simulations and on Facebook samples. We showed (in [14]) that they all converge to their true values for reasonable sample sizes. Based on our evaluation, we also provide recommendations, summarized as follows. When estimating *category sizes*, there is no universal choice between induced and star sampling. For example, the performance of the star estimator improves (i) in dense graphs, (ii) in graphs with homogeneous node degree distribution, (iii) in graphs with weaker community structure, and (iv) under sampling techniques that oversample high degree nodes. However, a heterogeneous, highly skewed node degree distribution (very common in many real-life graphs) may strongly reduce or completely eliminate this gain. In contrast, when estimating the category *edge weights*, the star estimators are a clear winner; the induced subgraph estimators often need 5-10 times more samples to achieve the same accuracy. Finally, the sampling techniques strongly affect estimator efficiency. They can be ordered from best to worst as follows: UIS, S-WRW, and RW.

We applied our methodology to samples of Facebook users and we estimated potentially interesting category graphs, such as the global friendship map, or the friendship network of US colleges. We visualized and made publicly available these weighted topologies at `www.geosocialmap.com`.

8. ACKNOWLEDGMENTS

This work was supported by the following grants: Swiss SNF grant PBELP2-130871, NSF CDI Award 1028394, AFOSR Award FA9550-10-1-0310.

9. REFERENCES

[1] Mapping Global Friendship Ties: http://tinyurl.com/TiesFB.

[2] N. Ahmed, J. Neville, and R. Kompella. Reconsidering the Foundations of Network Sampling. In *Proc. of WIN*, 2010.

[3] D. Aldous and J. A. Fill. *Reversible Markov Chains and Random Walks on Graphs*. In preparation.

[4] M. Gjoka, C. T. Butts, M. Kurant, and A. Markopoulou. Multigraph Sampling of Online Social Networks. *IEEE JSAC on Measurement of Internet Topologies*, 2011.

[5] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In *Proc. of IEEE INFOCOM*, 2010.

[6] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Practical Recommendations on Crawling Online Social Networks. *IEEE JSAC on Measurement of Internet Topologies*, 2011.

[7] M. Hansen and W. Hurwitz. On the Theory of Sampling from Finite Populations. *Annals of Math. Statistics*, 1943.

[8] D. D. Heckathorn. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems*, 44:174–199, 1997.

[9] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near-uniform URL sampling. In *Proc. of WWW*, 2000.

[10] L. Katzir, E. Liberty, and O. Somekh. Estimating Sizes of Social Networks via Biased Sampling. In *WWW*, 2011.

[11] E. D. Kolaczyk. *Statistical Analysis of Network Data*, volume 69 of *Springer Series in Statistics*. 2009.

[12] M. Kurant, C. T. Butts, and A. Markopoulou. Graph Size Estimation. in preparation.

[13] M. Kurant, M. Gjoka, C. T. Butts, and A. Markopoulou. Walking on a Graph with a Magnifying Glass: Stratified Sampling via Weighted Random Walks. In *Proc. of Sigmetrics*, 2011.

[14] M. Kurant, M. Gjoka, Y. Wang, Z. W. Almquist, C. T. Butts, and A. Markopoulou. Coarse-Grained Topology Estimation via Graph Sampling. *arXiv:1105.5488*, 2011.

[15] M. Kurant, A. Markopoulou, and P. Thiran. On the bias of BFS (Breadth First Search). In *Proc. of ITC*, 2010.

[16] M. Kurant, A. Markopoulou, and P. Thiran. Towards Unbiased BFS Sampling. *IEEE JSAC on Measurement of Internet Topologies*, 2011.

[17] S. H. Lee, P.-J. Kim, and H. Jeong. Statistical properties of sampled networks. *Phys. Review E*, 73:16102, 2006.

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

[19] L. Lovász. Random walks on graphs: A survey. *Combinatorics, Paul Erdos is Eighty*, 2(1):1–46, 1993.

[20] A. Maiya and T. Berger-Wolf. Sampling community structure. In *WWW*, 2010.

[21] M. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Review E*, 2006.

[22] A. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach. Respondent-driven sampling for characterizing unstructured overlays. In *Proc. of IEEE INFOCOM Mini-conference*, 2009.

[23] R. Rejaie, M. Torkjazi, M. Valafar, and W. Willinger. Sizing up online social networks. *IEEE Network*, 24(5):32–37, 2010.

[24] B. Ribeiro and D. Towsley. Estimating and sampling graphs with multidimensional random walks. In *Proc.of IMC*, 2010.

[25] M. Salganik and D. D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34(1):193–240, 2004.

[26] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. In *Proc. of IMC*, 2006.

[27] A. Traud, P. Mucha, and M. Porter. Social Structure of Facebook Networks. *arXiv:1102.2166*, 2011.

[28] B. Viswanath, A. Mislove, M. Cha, and K. Gummadi. On the evolution of user interaction in facebook. In *Proc. of WOSN*, 2009.

[29] E. Volz and D. D. Heckathorn. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, 2008.

[30] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge Univ. Press, 1994.