Sampling Community Structure

Arun S. Maiya
Department of Computer Science
University of Illinois at Chicago
851 S. Morgan Street
Chicago, Illinois 60607
amaiya@cs.uic.edu

Tanya Y. Berger-Wolf
Department of Computer Science
University of Illinois at Chicago
851 S. Morgan Street
Chicago, Illinois 60607
tanyabw@cs.uic.edu

ABSTRACT

We propose a novel method, based on concepts from expander graphs, to sample communities in networks. We show that our sampling method, unlike previous techniques, produces subgraphs representative of community structure in the original network. These generated subgraphs may be viewed as *stratified* samples in that they consist of members from most or all communities in the network. Using samples produced by our method, we show that the problem of community detection may be recast into a case of statistical relational learning. We empirically evaluate our approach against several real-world datasets and demonstrate that our sampling method can effectively be used to infer and approximate community affiliation in the larger network.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications—*Data Mining*

General Terms

Algorithms; Experimentation

Keywords

sampling, social network analysis, community detection, complex networks, graphs, clustering

1. INTRODUCTION AND MOTIVATION

In this work, we present a method to produce subgraph samples from networks such that these samples are representative of *community structure*, a characteristic prevalent in many complex networks under study today, from online social networks to telecommunication call graphs to biological systems. With advances in technology, pervasive use of the Internet, and the proliferation of location-aware devices, there is an ever increasing availability of these social and biological network data. Many web-based services, from LinkedIn to Wikipedia, produce large amounts of data on interactions and associations among entities [7, 23]. In the same vein, location-aware devices such as mobile phones produce copious amounts of data on physical proximity between individuals (i.e. associations and interactions) [6]. In the domain of biology also, from neurons to proteins to food webs, there is now access to large networks of associations among various entities and a need to analyze and understand these data [3, 21].

Whether the beginning of network science is taken to be the birth of graph theory or the dawn of social network analysis, it is clear that the networks under analysis in the past were relatively small as compared to those of today. The networks of today can be so large that analysis of the network in its entirety can be intractable and impractical. How, then, should one proceed in analyzing and mining these networks? Traditional approaches include designing more efficient algorithms or leveraging computing power through parallelization or distributed computing. Unfortunately, these existing methods are not always easily available as an option. Another approach that has received very little attention is *sampling.*

1.1 Sampling Networks

Sampling is fundamental to statistics and employed when there is a need to study a population and direct analysis of the entire population is infeasible due to sheer size and inaccessibility. In these cases, random samples are taken, the samples are analyzed, and results are generalized to the population from which the samples were drawn. Can this approach be applied to networks?

Networks are normally represented as graphs with the vertices (or nodes) representing entities and the edges (or links) representing interactions or associations between the entities. In a social network, for instance, the nodes represent individuals and edges may represent associations such as social interactions, emails sent or received, physical proximity, or demographic similarity. Simply mimicking traditional statistical sampling approaches would entail taking a random sample of nodes. However, the induced subgraph on the sampled nodes may very well be a collection of disconnected isolated singletons, which is essentially useless for any meaningful analysis. The task, then, must be to sample a subgraph in such a way that the subgraph is representative of the original or global graph. The question of what it means for a sample to be *representative* of the original network must be addressed. Existing works consider such measures as similarity in degree distributions and clustering coefficients [20, 24]. We argue that the measure of representativeness should vary and depend on the analysis being performed. We further argue that existing subgraph sampling techniques and corresponding measures of representativeness are inadequate for a key analysis task in social and biological networks: representing and identifying community structure.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

1.2 Communities in Networks

A community in a network is a subgroup of relatively densely connected nodes [33]. The discovery of communities in networks is important as they often correspond to real social groups, functional groups, or similarity [15]. Many popular community detection algorithms considered to be accurate are also computationally expensive [8, 11]. Representative subgraph sampling, then, provides a potential solution for inferring and approximating global, latent properties such as these in large graphs. By sampling a representative subgraph, analysis can be performed on the sample instead of the larger network. Results could, then, be generalized to the larger population, which, in this case, is the original network. In this work, we focus on two specific tasks: representing global community structure in samples and using these samples to infer community affiliation in the larger network.

1.3 Contributions

In this paper, we propose a sampling algorithm capable of representing and inferring community structure in the original network. Specifically, our contributions in this paper are as follows:

- We propose a novel method for representative subgraph sampling based on concepts from expander graphs and show that our approach produces subgraphs representative of community structure in the original network.
- Using subgraph samples produced by our method, we show that the problem of community detection can be recast into a case of statistical relational learning, or more specifically, univariate collective inference. In doing so, we show that subgraph samples may be used to infer the community affiliation of nodes *not* present in the sample.
- We empirically demonstrate our sampling method can be used to represent global community structure and infer community affiliation on several real-world datasets.

To the best of our knowledge, this is the first work using sampling to make non-trivial inferences of latent properties such as community structure in the larger network and to apply statistical relational learning and collective inference to the problem of community detection. Before describing our sampling method and approach, we first discuss the existing related work in this area.

2. RELATED WORK

Leskovec and Faloutsos [24] authored what may be the first real study on representative sampling in real-world networks. They tested a number of different sampling methods to assess the ability of these algorithms to match various properties of the original network such as degree distribution, clustering coefficient, and the distribution of component sizes. More recently [20], an innovative subgraph sampling technique based on the Metropolis algorithm [29] was proposed and again tested to assess the degree of consistency with graph properties. The results reported in [20] indicate that this approach outperforms all previous approaches and would seem to be the current state-of-the-art for representative subgraph sampling. Although there is little other work in sampling representative subgraphs with the intent of matching properties of the original network, there are contributions involving sampling graphs for other purposes such as graph compression [1, 10, 14], visualization [31], sociology [12], and epidemiology [16]. Also related to our work is the vast body of research in community detection, a discussion of which is well beyond the scope of this paper but is excellently surveyed in [11]. Finally, in [28], under the assumption that a network sample already exists and contains nodes from a *single* community, a method is proposed to grow the sample to include all members of this single community in question.

Our contributions in this paper differ from existing work on several fronts. First, the work in [20, 24] only assessed the degree to which subgraph samples are representative of *explicit* graph properties, many of which are relatively easy to compute on the original network in the first place (e.g. the degree distribution). To date and to the best of our knowledge, there is little or no work on how to sample subgraphs for inference of *implicit* or *latent* properties in the original network. Second, sampling has not previously been applied to the problem of community detection, despite the vast amount of literature in this area. Our work is wholly different from work by Mehler et al. [28] in that they 1) did not propose how to produce a subgraph sample from a network and 2) only show how to determine members of a *single* community in the network. In contrast, our aim is to show how best to produce samples representative of *all* or *most* of the communities in the network and further show how these samples may be used to infer the community affiliation of nodes *not* present in the sample. We begin a discussion of our work with some preliminaries.

3. PRELIMINARIES

3.1 Notations and Definitions

Definition 1. $G = (V, E)$ is a *network* or *graph* where V is set of vertices (or nodes) and $E \subseteq V \times V$ is a set of edges (or links between the nodes). We will use the terms *network* and *graph* interchangeably.

Definition 2. S is a *sample* of nodes where $S \subset V$.

Definition 3. $G(S)$ is the *induced subgraph* of G based on the sample S. That is, $G(S) = (S, E_S)$ where $S \subset V$ is the vertex set and the edge set $E_S = (S \times S) \cap E$.

Definition 4. A graph (or subgraph) is *connected* if and only if there is a path of edges from v to w for every pair of nodes v and w in the graph.

Definition 5. $N(S)$ is the *neighborhood* of S. That is, $N(S) = \{w \in V - S : \exists v \in S \; s.t. \; (v, w) \in E\}$.

Definition 6. The *expansion factor*, $X(S)$, of a sample S is:

$$X(S) = \frac{|N(S)|}{|S|}$$

The terms *expansion ratio* and *expansion parameter* are synonyms for the expansion factor[1].

[1] In this paper, we focus our attention on vertex expansion (the number of nodes connected to a sample) rather than edge expansion (the number of edges emanating from a sample).

Definition 7. The *maximum expander set* of size k is a sample S of size k with the maximal expansion factor:

$$\underset{S:\,|S|=k}{\text{argmax}} \frac{|N(S)|}{|S|}$$

3.2 Problem Formulation

Having stated the necessary definitions of terms and notation, we are now ready to formulate the specific problem addressed in this paper. In this work, our primary goal is to sample subgraphs representative of *community structure*. As stated previously, a community is a set of relatively densely connected nodes in a network G [33]. Although there are many ways to represent community structure depending on various factors such as whether or not overlapping is allowed, in this paper, we represent community structure as a *partition*: a collection of disjoint subsets whose union is the vertex set V. Under this representation, each subset in the partition represents a community. The task of a community detection algorithm, then, is to identify a partition such that vertices within the same subset in the partition are more densely connected to each other than to vertices in other subsets.

As mentioned in Section 2, there is a large body of work on the identification of community structure in networks. Nevertheless, the problem of community detection is still considered a very much open problem for a number of different reasons. While there are many different community detection algorithms, they often produce different answers on the same graph [17]. Furthermore, many community detection algorithms are computationally expensive making them unscalable [11, 15]. At the same time, there is also some evidence to suggest that these more expensive algorithms tend to be more accurate than faster, less costly alternatives [8, 11]. These issues necessitate the very problem of representative subgraph sampling we address in this work.

Given a graph $G = (V, E)$, our goal is to sample a set of nodes, S, such that the sampled subgraph, $G(S)$, is representative of community structure in the larger network, G. What do we mean by "representative of community structure" in the larger network? First, we would like the sampled subgraph to contain nodes from all (or most) of the communities present in the larger network (i.e. a *stratified* sample of community structure). Second, if executing a community detection algorithm separately on both the sampled subgraph and the original network, we would like vertices grouped together in the subgraph to be also grouped together in the larger network. To measure this, we employ a measure of partition distance with the distance being low if groupings of vertices are consistent between the two partitions. This can be formally defined as follows:

Definition 8. A sample $S \subset V$ is a *community representative sample* if, given a graph $G = (V, E)$, a sample size k, a community detection algorithm A, and a measure of partition distance $\mathcal{D}[\cdot, \cdot]$, S is a sample of size k that satisfies the following two conditions:

- Condition 1: S minimizes $\mathcal{D}[P_S(G(S)), P_S(G)]$ where $P_S(G(S))$ is the partition of S produced by A on $G(S)$ and $P_S(G)$ is the partition of S produced by A on G.
- Condition 2: The number of non-empty intersections between S and each partition set of $P(G)$ is maximized, where $P(G)$ is the partition of V produced by A on G.

As will be shown later, such a sample can be used to infer the community affiliation of nodes *not* present in the sample. The approach we employ to produce these representative samples is rooted in work on expander graphs, which we describe in the next section.

4. PROPOSED METHOD

At its core, our method is based on the conjecture that samples with good expansion properties tend to be more representative of community structure in the original network than samples with worse expansion. This approach is derived from concepts on *expander graphs*. Expander graphs are highly connected graphs which, at the same time, are relatively sparse [19]. Formally, a graph is a (k, α)-expander if $|N(S)| \geq \alpha|S|$ for each $S \subset V$ where $|S| \leq k$. The expansion factor of an entire graph, then, is defined as:

$$\min_{S:|S|\leq k} \frac{|N(S)|}{|S|}$$

For our purposes, rather than finding the sample with the *minimum* expansion factor, we are interested in finding the sample with the *maximum* expansion factor:

$$\underset{S:\,|S|=k}{\text{argmax}} \frac{|N(S)|}{|S|}$$

Moreover, we are interested in the specific *sample* that produces the maximum expansion factor rather than the *value* of the expansion factor itself (hence, we employ argmax instead of max above). As defined in Definition 7, we refer to this sample as the *maximum expander set*. We refer to our sampling approach as *Expansion Sampling*. Intuitively, by including nodes in our sample that best contribute to the expansion factor, we are essentially sampling nodes which act as bridges to *new* clusters (i.e. communities). And, by including these "bridge" nodes (and surrounding nodes), we hope to produce a sample which 1) contains most or all of the communities from the network and 2) is a condensed representation of the overall community structure of the entire graph. We propose two methods to approximate the maximum expander set. The first is a greedy algorithm using snowball sampling. The second employs Markov Chain Monte Carlo simulation (MCMC). Before describing these methods, we first examine the relationship between expansion and community structure in greater depth.

4.1 Expansion and Community Structure

The approach we employ to sample community structure in networks is based upon finding samples with the best expansion in the network. The rationale is that better expansion equates to better community representativeness. Intuitively, as sampling progresses and nodes are added to the sample, the expansion factor of the sample changes. Nodes belonging to *new* communities not already represented in the sample should result in a relatively larger expansion than nodes belonging to *existing* communities already represented in the sample. By sampling to maximize expansion, we include more communities into the sample. To verify this, we first show the criteria for changes in expansion as sample size grows. It is clear that changes in the magnitude of expansion are a function of the number of *new* neighbors each node contributes to the sample (because the denominator of $\frac{|N(S)|}{|S|}$ only ever increases by one as nodes are added to the sample). It is also trivial to show the specific criteria for increases and decreases in the expansion as sample size grows.

Proposition 1. *There exist networks for which the expansion factor, $\frac{|N(S)|}{|S|}$, is non-monotonic as the sample size $|S|$ grows.*

Proof. Consider a current sample, S. Assume that the next node selected for inclusion in the sample is v and let $N(\{v\})$ be the neighbors of v. Then, the *new* sample is $S \cup \{v\}$. By Definition 6 in Section 3, the expansion of S is $\frac{|N(S)|}{|S|}$ and the expansion of $S \cup \{v\}$ is $\frac{|N(S\cup\{v\})|}{|S\cup\{v\}|}$.
The expansion will increase when:

$$\begin{aligned}
\frac{|N(S \cup \{v\})|}{|S \cup \{v\}|} &> \frac{|N(S)|}{|S|} \\
|N(S \cup \{v\})| &> |N(S)| \cdot \frac{|S \cup \{v\}|}{|S|} \\
|N(S \cup \{v\})| - |N(S)| &> |N(S)| \cdot \frac{(|S|+1)}{|S|} - |N(S)| \\
|N(\{v\}) - (N(S) \cup S)| &> \frac{|N(S)|}{|S|} \qquad (1)
\end{aligned}$$

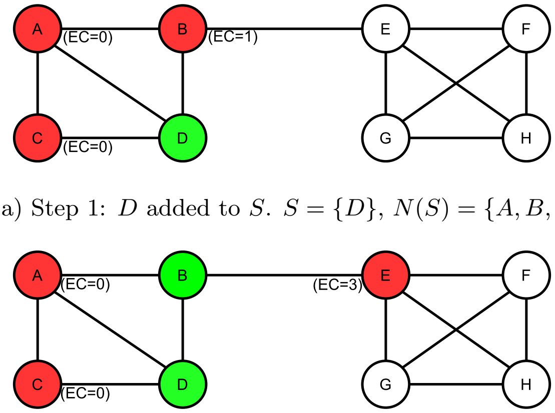
Conversely, expansion will decrease when:

$$|N(\{v\}) - (N(S) \cup S)| < \frac{|N(S)|}{|S|} \qquad (2)$$

The increase or decrease in expansion, then, is a function of the number of *new* neighbors contributed by v (i.e. $(|N(\{v\}) - (N(S) \cup S)|)$ *relative to the current expansion factor.* □

We now examine how networks exhibiting community structure affect changes in expansion. Recall from Section 3.2 that a community is a set of relatively densely connected nodes in a network. A network is said to exhibit community structure if it divides into sets of nodes with dense connections within sets and sparse connections between sets. From proof of Proposition 1, we see that expansion is lower when $|N(\{v\}) - (N(S) \cup S)|$ is sufficiently small. For a network exhibiting community structure, this, in fact, happens precisely when v is affiliated with an *existing* community included within S (as v will have few new neighbors if it is already densely connected to nodes in $N(S) \cup S$). Conversely, by similar reasoning, when v is affiliated with a *new* community (i.e. a community *not* already included in S), $|N(\{v\}) - (N(S) \cup S)|$ will be relatively larger resulting in larger expansion. By the intuitive definition of what it means to be a community in a network, the expansion of a sample is directly related to community structure. Figure 1 shows a simple example illustrating this connection.

We conclude this section with two final points. First, the extent to which expansion and community structure are related depends on the strength of community structure within a network and the "tightness" of communities. The connection between the two is stronger when the community structure of the network is stronger. A widely used measure for the "goodness" or the strength of a community in graph clustering and community detection is *conductance* [22], which is a function of the fraction of total edges emanating from a sample (lower values mean stronger communities):



(a) Step 1: D added to S. $S = \{D\}$, $N(S) = \{A, B, C\}$

(b) Step 2: B added to S. $S = \{D, B\}$, $N(S) = \{A, C, E\}$

Figure 1: **[Best viewed in color.]** A simple illustration showing two 4-node communities: green nodes are in S and red nodes are in $N(S)$. Numbers in parentheses beside each node $v \in N(S)$ show the value of $|N(\{v\}) - (N(S) \cup S)|$ or the *expansion contribution* of v (denoted as EC). Notice that nodes in new communities (and nodes that *lead* to new communities) have relatively better expansion contributions at each step.

$$\varphi(S) = \frac{\sum_{i \in S, j \in \overline{S}} a_{ij}}{\min(a(S), a(\overline{S}))}$$

where a_{ij} are entries of the adjacency matrix representing the graph and $a(S) = \sum_{i \in S} \sum_{j \in V} a_{ij}$, which is the total number of edges incident to S. It is easy to prove that when the conductance is sufficiently low (and clustering is sufficiently high), the aforementioned relation between expansion and community structure is stronger.

For the second and final point, we also note here that there is no direct correspondence between the degree centrality of a node (i.e. number of neighbors) and the extent to which that node contributes to the expansion. A node v with high degree ($|N(\{v\})|$) will have low expansion if $|N(\{v\}) - (N(S) \cup S)|$ is sufficiently small. By explicitly selecting nodes that best contribute specifically to the *expansion*, we incorporate more communities into the sample, thereby producing a sample representative of community structure in the larger network. In the following two sections, we describe two specific methods for finding samples with the best expansion in a network: a "snowball" sampling approach based on neighborhood dissimilarity (Section 4.2) and an MCMC-based method (Section 4.3).

4.2 Expansion Sampling: Snowball

The Snowball approach to *Expansion Sampling* (XSN) is shown in Algorithm 1. We use the term "snowball" for this technique because subsequent members of the sample, S, are selected from the current neighborhood set $N(S)$. In this way, the sample grows like a snowball. We select the first node of the sample uniformly at random from the entire graph, G. Subsequent elements of the sample are chosen based on the degree to which a node $v \in N(S)$ con-

Algorithm 1 Snowball Expansion Sampler (XSN)

```
Input:
    Graph $G = (V, E)$
    $k$, the sample size.
$S = \emptyset$                    // initialize sample to empty set
$v = random(V)$     // choose node from V at random
$S = S \cup \{v\}$
while $|S| \leq k$ do
    Select new node $v \in N(S)$ based on maximization of:
                $|N(\{v\}) - (N(S) \cup S)|$
    $S = S \cup \{v\}$
end while
```

tributes to the expansion factor of the currently constructed sample S, which is expressed as $|N(\{v\}) - (N(S) \cup S)|$. (Recall that this is precisely the expression discussed in Proposition 1.) New sample members may be chosen either deterministically or probabilistically. In the deterministic version, the new sample member, v, is selected using $\text{argmax}_{v \in N(S)} |N(\{v\}) - (N(S) \cup S)|$. Alternatively, a probabilistic approach may also be employed to account for occasional scenarios in which it is the *near-highest* expansion (rather than the absolute highest) that better leads to the most new communities. For instance, a new sample member, v, can be chosen with probability proportional to $\beta^{|N(\{v\}) - (N(S) \cup S)|}$ where β is some constant. We did not find significant performance differences between the two, so we only describe results for the deterministic version.

The Snowball approach to *Expansion Sampling* always produces *connected* samples. For many analysis tasks such as community inference, connected samples are, in fact, desirable. In addition to this greedy approach, we also propose a Monte Carlo-based approach for finding samples with high expansion, which we describe in the next section.

4.3 Expansion Sampling: MCMC

An alternative approach to *Expansion Sampling* is Markov Chain Monte Carlo simulation (MCMC), a standard technique to sample and evaluate probability distributions (see Chapter 29 in [26] for an excellent introduction to Monte Carlo methods). Our MCMC method to approximate the maximum expander set employs the well-known Metropolis algorithm[2] [29], which has recently been applied to subgraph sampling [20].

4.3.1 Overview

Given a graph $G = (V, E)$, the idea behind the *MCMC Expansion Sampler* (XMC) is to construct a Markov chain in which each state represents a subgraph of size k where $k \ll |V|$. A quality measure is chosen to determine the degree of representativeness of each subgraph sample. Starting with a randomly selected sample, we randomly perturb the sample by one node and accept or reject the new sample based on the change in quality score (this is the standard random-walk Metropolis-Hastings method [26]). Upon convergence to a stationary distribution, sampling the Markov chain is equivalent to sampling subgraphs with probability proportional to the quality scores of the subgraphs. By sampling the Markov chain and retaining the subgraph with the maximum quality score, a representative subgraph is obtained.

4.3.2 Quality Measure

Different quality measures may be used depending on the graph property of interest (e.g. distance between degree distributions which was employed in [20]). In this work, however, we are interested in a very specific property: the expansion factor. Notice that, given a sample S, the maximum possible expansion factor on *any* graph of $|V|$ nodes is: $\frac{|V-S|}{|S|}$. Therefore, a normalized quality measure to determine the relative expansion ability of samples can be defined as: $\frac{|N(S)|}{|V-S|}$, with higher quality scores equating to better expansion. We refer to this measure as the *expansion quality.* The *expansion quality* measures the degree to which a sample acheives the maximum possible expansion on any network of $|V|$ nodes. A score of 1 indicates that the sample "touches" or is one hop away from every other node in the network.

4.3.3 Acceptance Probability

After perturbing the current sample by one node, if the quality score has increased, then we accept the transition to the new subgraph sample (i.e. Markov state) with probability 1. If the quality score has decreased, we employ the acceptance probability proposed in [20]:

$$\left[\frac{quality(S_{new})}{quality(S_{current})}\right]^p$$

where $p = 10\frac{|E|}{|V|}\log_{10}|V|$ and $quality(\cdot)$ is the *expansion quality.* For networks with many nodes and many connected samples, there can be a large quantity of medium-quality samples and few high-quality samples [20]. In these cases, good samples must be rewarded to a greater extent in order to reach these high quality samples, and p, consequently, must be higher. Hence, this setting proposed for p is a function of both the number of nodes (which captures the size of the network) and the edge-to-node ratio (which captures the number of possible connected samples) [20].

5. EXPERIMENTAL EVALUATION

5.1 Datasets

We evaluate and compare our aforementioned sampling algorithms on several network datasets used extensively in the literature as standard testbeds for community detection. Summary statistics for each dataset are shown in Table 1, and the datasets are briefly described below. For the purposes of this paper, all networks are treated as undirected and unweighted graphs.

Dataset	**Vertices**	**Density**	**CC**
Net Science	379	0.0127	0.4306
C. elegans Metabolic	453	0.0449	0.1244
PGP	10680	0.0004	0.3780
HepTh	27400	0.0009	0.1196
HepPh	34401	0.0007	0.1457
Epinions	75877	0.0001	0.0657

Table 1: Vertex count, density, and clustering coefficient (CC) for each dataset

[2] The Metropolis algorithm was selected as one of the top 10 algorithms of the twentieth century [4].

Network Science Collaboration Network [30] is a collaboration network of researchers within the network science community. Nodes represent authors and edges exist between authors if coauthoring a paper.

C. elegans Metabolic Network [9] consists of the metabolic network of the C. elegans worm: nodes are substrates and edges are reactions among the substrates.

PGP Key-Signing Network [2] is a social network consisting of users of the Pretty Good Privacy (PGP) encryption system. Nodes represent users with digital keys and edges represent a key-signing between the users, also referred to as the "Web of Trust".

HepTh [13, 25] is a citation network between papers in Arxiv HEP-TH (high energy physics theory) from the e-print archive, arxiv.org.

HepPh [13,25] is another citation network from arxiv.org. These citations cover papers published in Arxiv HEP-PH (high energy physics phenomenology).

Epinions.com [32] is a trust-based online social network of the consumer review site, Epinions.com.

As shown in Table 1, some datasets we evaluate, such as the *Network Science* dataset, are smaller networks. Clearly, sampling is not required for networks this small. We include these and the other networks in our analysis due to the fact that these datasets are standard testbeds for community detection for which some community structure is known to exist.

5.2 Experimental Setup

We compare our Snowball Expansion Sampler (referred to hereafter as XSN) and our MCMC Expansion Sampler (referred to hereafter as XMC) against two state-of-the-art approaches described in Section 2:

- Metropolis Using Degree Distribution (MDD) [20]
- Metropolis Using Clustering Coefficient (MCC) [20]

In [24], it is implied that, if producing samples with clustering coefficients matching the larger network, these samples will be representative of community structure, and one of our aims is to investigate this claim. The authors of [20] report the MCC method as the best method for producing samples with consistent clustering coefficients. Also in [20], the MDD method is reported to be the best overall method for producing representative subgraphs samples, beating out all other existing methods in general performance. Hence, these two methods are chosen as a basis for comparison, as they represent the current state-of-the-art.

For each dataset described in Section 5.1, using each of these sampling approaches, we sample 15% of the nodes in the network and produce 25 samples from each sampling algorithm on each dataset (a sample size of 15% is chosen based on experimental findings in [24]). For the approaches based on the Metropolis algorithm, we perform 10,000 iterations to produce each sample. Next, we execute a community detection algorithm on each of the subgraph samples in addition to the original network. Finally, we compare the community structure of the samples with the community structure in the original network to evaluate each of the sampling algorithms. (The exact evaluation criteria for this comparison are described in Section 5.2.2.) Comparisons consider averages over all the 25 samples produced by each algorithm.

5.2.1 *Detecting Communities*

The question of which community detection algorithms to evaluate must be addressed, as the number of such algorithms are numerous. For a thorough evaluation, we evaluate multiple community detection algorithms including:

- Girvan-Newman algorithm (GN) [15]
- Newman's leading eigenvector method (NLE) [30]
- An algorithm based on greedy optimization (CNM) [5]

Of these three algorithms, only the CNM algorithm is executable on the larger networks we evaluate in our experiments. For instance, the Girvan-Newman algorithm, one of the most well-known and well-cited community detection algorithms proposed, is notorious for being computationally expensive and starts to become unusable when the size of the network exceeds 10,000 nodes. The algorithm iteratively removes edges with the highest edge betweenness to identify communities with the running time being $O(n^2m)$ where n is the number of vertices and m is the number of edges [11,15]. Interestingly, on networks for which all algorithms *do* run, the detected community structures are *not* identical. As mentioned in Section 3.2, these issues with scalability and inconsistency necessitate the very task of subgraph sampling we address in this work.

5.2.2 *Recognizing Good Samples of Communities*

We now describe the precise *evaluation criteria* we employ to assess how representative samples are of global community structure in the larger network. As mentioned in Section 3.2, one way to view community structure of a network is as a partition on the nodes with each subset representing a single community, and this is the representation employed in this paper. A measure of partition distance may be used to evaluate the degree to which a sample is representative of the community structure in the original network. We employ the measure of partition distance proposed by Gusfield [18], which is essentially the minimum number of elements that would need to be removed in order to make the partitions identical. Using this measure, we calculate the distance between the community structure of a sample and the community structure in the larger network. The partition distance will be low if nodes grouped together in the sample's community structure are also grouped together in the community structure of original network with lower distances corresponding to better representativeness (and high otherwise).

There is, however, an issue with *only* considering partition distance when measuring the degree to which samples are representative of community structure. Consider a network and two samples from the network. The first sample contains nodes from multiple communities. But, in the second sample, all the nodes belong to a single community. Let us assume that, when running a community detection algorithm on the samples, nodes are grouped together correctly in both samples. That is, nodes grouped together in the sample are also grouped together in the larger network. In this case, the partition distance for both samples will be low even though the first sample is clearly a *substantially* better representative of the overall community structure in the larger network. In fact, a sample containing a *single* community will always have a perfect partition distance of zero.

Moreover, lower (or better) partition distances are harder to achieve as more communities are incorporated into the sample. Observing, then, that two samples exhibit similar partition distances may not tell the whole story. Therefore, we must not only consider partition distance when evaluating the representativeness of samples, but also consider the *number* of communities represented in the sample. Specifically, we measure the fraction of total communities in the larger network represented in each sample, a number ranging in value from 0 to 1. By considering both measures, we obtain a better picture of the true representativeness of samples.

The fraction of communities in the sample ($FRAC$) is a normalized value ranging from 0 to 1 (higher values are better). For ease of illustration, the partition distance ($PART$), is also converted to an accuracy score ranging from 0 to 1 by normalizing and subtracting from one (also resulting in higher values being better). Finally, in addition to considering each measure individually, we compute a *composite score*, which is the harmonic mean (or F-score) of the two measures[3]:

$$\text{Composite} = \frac{2 \cdot \text{FRAC} \cdot \text{PART}}{\text{FRAC} + \text{PART}}$$

We are left with three performance indicators to consider:

- Fraction of Communities Represented in Sample
- Partition Accuracy
- Composite Score

The *composite score*, as with the other measures, also ranges in value from 0 to 1 with higher values being better.

5.3 Main Results

We now discuss our results, focusing on two areas of analysis:

- *Community Representativeness*: the degree to which samples are representative of community structure in the larger network
- *Community Affiliation Inference*: the degree to which samples can be used to infer community affiliation of nodes *not* present in the sample

5.3.1 Community Representativeness

As mentioned in Section 5.2, we consider three performance measures to evaluate the degree to which samples are representative of community structure in the larger network. Figure 2 shows results on each of these measures. As can be seen, the Expansion Sampling approaches (XSN and XMC) outperform other sampling methods. We also see that the Expansion Sampling methods not only have the highest *composite score*, but also dominate on each of the measures individually. This is striking, as higher partition accuracies are more difficult to achieve as more communities are incorporated into the sample (as discussed in Section 5.2.2).

In addition, we see that the XSN algorithm, in particular, dominates the remaining three Metropolis-based algorithms including the XMC method. We further find that the XMC method continues to improve the expansion of the sample to the very last iteration. (Recall that we execute all Metropolis-based algorithms for $10,000$ iterations.) The ability of the XMC method to find the sample with maximum expansion, then, might be improved by executing for more iterations or fine-tuning the MCMC parameters (e.g. the acceptance probability), resulting in performance more comparable to the XSN algorithm.

These results in Figure 2, then, show empirically what was illustrated theoretically in Section 4.1. By sampling to maximize expansion, we incorporate more communities from the network into samples, thereby producing samples representative of community structure in the larger network. In the next section, we show how these representative samples may be used to infer the community affiliation of nodes *not* included in the sample.

5.3.2 Inferring Community Affiliation from Samples

In the previous section, we assessed the extent to which samples are representative of community structure in the original network using partition accuracy and fraction of communities represented in the sample. A supplementary approach to assessing the representativeness of samples is population inference. If samples are truly representative of the larger population, analysis on the sample should generalize well to members of the population *not* present in the sample. In our case, we assess the degree to which samples can be used to infer the community affiliation of nodes *not* present in the sample. In other words, using a sample S and its induced subgraph $G(S)$, we attempt to infer the community affiliation for all nodes v such that $v \in V - S$. In order to do this, we recast the problem of community detection into a problem of statistical relational learning, or, more specifically, univariate collective inference. *Univariate Collective Inferencing* was formally defined in [27]. We restate their definition here directly:

Definition 9. Given a graph $G = (V, E, X)$ where $x_i \in X$ is the single attribute of vertex $v_i \in V$ and given known values $x_i \in X$ for some subset of vertices $S \subset V$, *univariate collective inferencing* is the process of simultaneously inferring the values $x_i \in X$ for the remaining vertices, $\overline{S} = V - S$, or inferring a probability distribution over those values for each vertex.

In our case, X is the set of community assignments to the vertices. For any two vertices $v_i, v_j \in V$, $x_i = x_j$ if and only if v_i and v_j are members of the same community. As before, we sample a set of nodes S and execute the community detection algorithm on the induced subgraph $G(S)$. Using these community assignments to S, we infer the community affiliation for the remaining vertices $\overline{S} = V - S$ with collective inferencing. As before, we also execute a community detection algorithm on the entire graph G, which leaves us with two sets of community assignments to the nodes $\overline{S}$: one resulting from collective inferencing and the other resulting from execution of the community detection algorithm on the entire, original network. For all nodes in $\overline{S}$, we compare these two sets of community assignments using the measures of partition accuracy described previously. In [27], a number of different collective inferencing schemes were tested and compared. Some of the best results were obtained from relaxation labeling combined with what the authors' refer to as a weighted majority relational model [27]. For this

[3] PART and FRAC can be viewed in terms of precision and recall, respectively.

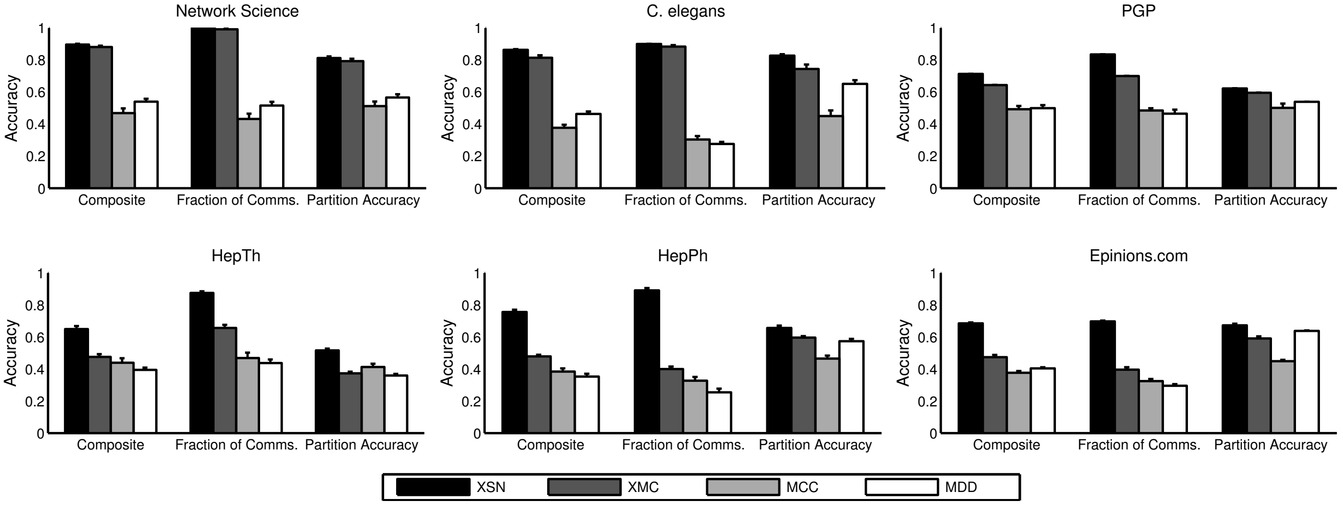


Figure 2: Representativeness of samples produced by each sampling algorithm on each dataset. Plots show the accuracy of each of the four sampling algorithms by various performance measures (standard error is also shown). The Expansion Sampling approaches (most notably XSN) outperform other methods on the three performance measures of community representativeness.

reason, it is this collective inferencing scheme[4] that we employ in our tests. For more information on this and other univariate collective inferencing schemes, one may refer to the original case study by Macskassey et al. [27].

Figure 3 shows the results. As shown, our XSN and XMC approaches outperform others on inference accuracy. These results again show that *Expansion Sampling* produces samples most representative of community structure in the larger network and points to an intriguing yet previously uninvestigated approach to community inference in complex networks.

It is also interesting to note that the collective inferencing scheme employed (relaxation labeling with a weighted majority relational model) produces a probability distribution over the possible community assignments for each node. The output of this scheme, then, is a *soft* clustering of the nodes, rather than a hard clustering produced by most community detection algorithms. This soft clustering holds the potential for additional knowledge discovery. For instance, these probability distributions may be useful in assessing the strength of community affiliation for specific individuals and identifying the individuals most and least representative of the communities to which they belong. The Zachary Karate network [34] is social network of a karate club consisting of thirty-four members. In this social network (for which ground truth is known to some degree), individual 3 is often misclassified by community detection algorithms (e.g. [15]). When executing our sample method on this network and examining community affiliation distributions produced by the relaxation labeling method employed, we see that the probability distribution over the community assignments for node 3 is the closest to a uniform distribution. There is, then, the most uncertainty associated with the community assignment of node 3 – information absent from traditional approaches to community detection (most noticeably in the three community detection algorithms we tested). This represents an interesting avenue for future research, as effective community detection is still an open problem.

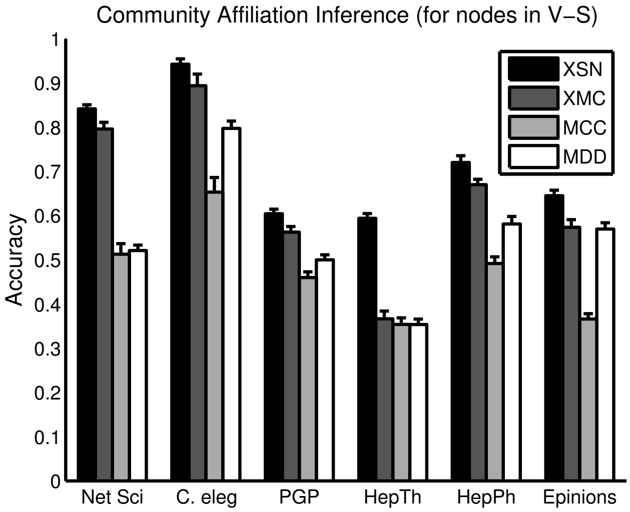


Figure 3: Inference accuracy of each sample on each dataset. The XSN and XMC approaches outperform the remaining sampling algorithms when being used to infer the community affiliation of nodes in the unsampled, original network.

5.4 Additional Findings

We now return to our earlier claim that the measure of representativeness for samples may have to vary and depend on the analysis being performed. In this section, we examine other graph-theoretic properties of samples. Specifically,

[4] For collective inferencing, we use the NetKit-SRL toolkit [27].

we look at the extent to which samples are representative of degree distributions and clustering coefficients in the larger network. For evaluating the similarity in degree distributions, we employ the Kolmogorov-Smirnov D-statistic also used in both [24] and [20]. For evaluating the similarity in clustering coefficients, we employ the distance measure used in [20] based on the L_1 norm. As we did with the partition distance measures, we convert these distance measures to accuracy scores by subtracting the distances from 1 and plot the performance measures in Figure 4 (higher bars are better).

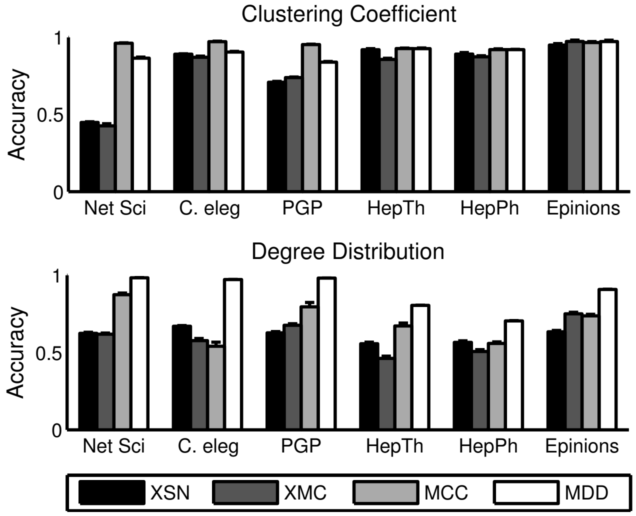


Figure 4: Degree Distribution and Clustering Coefficient of Samples. For each bar, the standard error is also shown.

Figure 4 shows the extent to which samples from each sampling method are representative of degree distributions and clustering coefficients in the original network. Surprisingly, the *Expansion Sampling* approaches seem to match the clustering coefficients and degree distributions better on larger datasets than smaller ones. Also, as one would predict, the MDD method performs best for degree distribution accuracy and the MCC method performs best for clustering coefficient performance (albeit, only slightly in some cases). We see that, although the *Expansion Sampling* algorithms outperform other methods in terms of community representativeness, they do not do as well in matching degree distributions and clustering coefficients of the larger network as other sampling methods. Similarly, the MDD and MCC methods that focus on producing samples consistent with degree distributions and clustering coefficients, respectively, do not do as well in representing community structure. There is, then, some evidence to indicate that samples most consistent with one property, such as degree distributions on the original network, may not be representative of other properties especially in terms of community representativeness. Additionally, the ability to match degree distribution and clustering coefficient seems to vary significantly across different networks of different sizes. More investigation, however, is required to draw firm conclusions. An open question is how best to capture multiple properties of the original network in a single sample, which we plan to address in future work.

6. CONCLUSION

We have proposed a novel approach based on concepts from expander graphs to produce subgraph samples representative of community structure in the original network. We referred to this approach as *Expansion Sampling* and described two different methods following this approach: a Snowball Expansion Sampler (XSN) and an MCMC Expansion Sampler (XMC). We empirically evaluated our approach against two state-of-the-art sampling methods on several real-world datasets. We showed that subgraph samples produced by our methods are more representative of community structure in the larger network than samples produced by existing methods. Further, with sample sizes of only 15% of the original network, we demonstrated that our sampling method outperforms the existing methods in terms of the ability to infer community affiliation of nodes in the larger network.

For future work, we plan to further investigate the capability of subgraph samples to infer community affiliation by assessing the relationship between sample size and inference accuracy in addition to examining how best to capture *multiple* graph properties (both latent and not) in a single subgraph sample. One approach to this is to couple *Expansion Sampling* with other methods known to sample different properties well. For instance, there is some evidence (e.g. [24]) to suggest that strategies based on breadth-first searches (BFS) may sample degree distributions well. In other experiments *we* have conducted, we found that samples produced by *Expansion Sampling* exhibited a significantly high *expansion quality* with relatively small sample sizes (smaller than 15% in many cases). We conjecture that, by tracking the expansion and switching to an alternate strategy (e.g. BFS-based sampling) when the *expansion quality* is reasonably high, the precision (i.e. partition accuracy) might be improved while simultaneously maintaining a high recall (i.e. fraction of communities represented in the sample). Finally, we also plan to investigate the interplay between subgraph sampling and various other network characteristics such as directed and weighted edges, overlapping communities, and soft clustering. Overall, our results indicate that representative sampling in networks may be a promising new approach to complex network analysis.

7. ACKNOWLEDGMENTS

We are grateful to David Kempe for useful discussions and to the anonymous reviewers for constructive comments. This research is supported in part by NSF grants IIS-0705822 and IIS-0747369 (CAREER).

8. REFERENCES

[1] M. Adler and M. Mitzenmacher. Towards compressing web graphs. In *Proc. of the IEEE DCC*, pages 203–212, 2000.

[2] M. Boguna, R. P. Satorras, A. D. Guilera, and A. Arenas. Models of social networks based on social distance attachment. *Physical Review E*, 70(5), 2004.

[3] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews. Neuroscience*, Feb 2009.

[4] B. A. Cipra. The best of the 20th century: Editors name top 10 algorithms. *SIAM News*, 33, 2000.

[5] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111+, Dec 2004.

[6] R. Clayford and T. Johnson. Operational parameters affecting use of anonymous cell phone tracking for generating traffic information. In *82nd TRB Annual Meeting*, Jan 2003.

[7] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *KDD '08*, pages 160–168, 2008.

[8] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *J. Stat. Mech.: Theory and Experiment*, 2005(09):P09008+, Sep 2005.

[9] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72(2):027104+, 2005.

[10] T. Feder and R. Motwani. Clique partitions, graph compression and speeding-up algorithms. In *J. Comp. and Sys. Sci.*, pages 123–133, 1991.

[11] S. Fortunato. Community detection in graphs. Jun 2009.

[12] O. Frank. *Network Sampling and Model Fitting.* Cambridge University Press, February 2005.

[13] J. Gehrke, P. Ginsparg, and J. Kleinberg. Overview of the 2003 kdd cup. *SIGKDD Explor. Newsl.*, 5(2):149–151, 2003.

[14] A. C. Gilbert and K. Levchenko. Compressing network graphs. In *Proc. LinkKDD workshop at the 10th ACM Conference on KDD*, Aug 2004.

[15] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proc Natl Acad Sci USA*, 99(12):7821–7826, Jun 2002.

[16] S. Goel, Matthew, and J. Salganik. Respondent-driven sampling as markov chain monte carlo. *Stat. in Medicine*, 2009.

[17] B. H. Good, Y.-A. de Montjoye, and A. Clauset. The performance of modularity maximization in practical contexts. *arXiv ePrints*, 0910.0165, 2009.

[18] D. Gusfield. Partition-distance: A problem and class of perfect graphs arising in clustering. *Info. Proc. Letters*, 82(3):159–164, May 2002.

[19] S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bull. Amer. Math. Soc.*, 43:439–561, 2006.

[20] C. Hubler, H.-P. Kriegel, K. Borgwardt, and Z. Ghahramani. Metropolis algorithms for representative subgraph sampling. In *ICDM '08*, pages 283–292, 2008.

[21] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, Oct 2000.

[22] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, May 2004.

[23] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD '08*, pages 462–470, 2008.

[24] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *KDD '06*, pages 631–636, New York, NY, USA, 2006.

[25] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05*, pages 177–187, 2005.

[26] D. J. C. Mackay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, Jun 2002.

[27] S. A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. *J. Mach. Learn. Res.*, 8:935–983, 2007.

[28] A. Mehler and S. Skiena. Expanding network communities from representative examples. *ACM TKDD*, 3(2):1–27, 2009.

[29] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–92, 1953.

[30] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3), 2006.

[31] D. Rafiei. Effectively visualizing large networks through sampling. In *Proc. VIS 05*, pages 375–382, 2005.

[32] M. Richardson, R. Agrawal, and P. Domingos. *Trust Management for the Semantic Web*, volume 2870. January 2003.

[33] S. Wasserman, K. Faust, and D. Iacobucci. *Social Network Analysis : Methods and Applications.* Cambridge University Press, Nov 1994.

[34] W. Zachary. An information flow model for conflict and fission in small groups. *J. Anthropological Research*, 33:452–473, 1977.