Network Completion and Survey Sampling

Steve Hanneke
Machine Learning Department
Carnegie Mellon University
shanneke@cs.cmu.edu

Eric P. Xing
Machine Learning Department
Carnegie Mellon Univserity
epxing@cs.cmu.edu

Abstract

We study the problem of learning the topology of an undirected network by observing a random subsample. Specifically, the sample is chosen by randomly selecting a fixed number of vertices, and for each we are allowed to observe all edges it is incident with. We analyze a general formalization of learning from such samples, and derive confidence bounds on the number of differences between the true and learned topologies, as a function of the number of observed mistakes and the algorithm's bias. In addition to this general analysis, we also analyze a variant of the problem under a stochastic block model assumption.

1 Introduction

One of the most difficult challenges currently facing network analysis is the difficulty of gathering complete network data. However, there are currently very few techniques for working with incomplete network data. In particular, we would like to be able to observe a partial sample of a network, and based on that sample, infer what the rest of the network looks like. We call this the network completion task.

In particular, in this paper we look at the network completion task, given access to random survey samples. By a *random survey*, we mean that we choose a vertex in the network uniformly at random, and we are able to observe the edges that vertex is incident with. Thus, a random survey reveals the local neighborhood (or *ego network*) of a single randomly selected vertex. We assume the network is represented as an undirected graph, with n vertices, and that the random samples are performed without replacement. Thus, after m random surveys, we can observe all of the edges among the m surveyed vertices, along with any edges between those m vertices and any of the $n-m$ unsurveyed vertices. However, we cannot observe the edges that occur between any two unsurveyed vertices. Thus, there are precisely $\binom{n-m}{2}$ vertex pairs for which we do not know for sure whether they are adjacent or not. We measure the performance of a network completion algorithm based on how well it predicts the existence or nonexistence of edges between these pairs.

There has been a significant amount of work studying various sampling models, including survey sampling, in the social networks literature. For example, (Frank, 2005) provides an excellent overview and entry-point to the relevant classic literature. These methods have proven quite useful for analyzing social network data sets collected in various ways that best suit the particular social experiment. However, to our knowledge there has been no work studying the general problem of learning the network topology from survey samples, while providing formal statistical guarantees on the number of mistakes in the learned topology.

There are two main challenges in learning the network topology from survey samples. The first is that the vertex *pairs* present in the observable sample are not chosen uniformly, as would typically be required in order to apply most known results from the learning theory literature[1], so that special care is needed to describe confidence bounds on the number of mistakes. We address this issue by deriving confidence bounds specifically designed for learning from survey samples, in a style analogous to the PAC-MDL bounds of (Blum

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

[1] As the size of the graph grows, the assumption of sampling uniformly at random essentially becomes the usual i.i.d. assumption of inductive learning. Thus, much of this work can be viewed as handling a certain type of non-i.i.d. sampling method.

& Langford, 2003). The second difficulty is the exponential number of possible graphs; as is typically the case in learning, this issue requires any learning algorithm that provides nontrivial guarantees on the number of mistakes it makes for a given topology to have a fairly strong learning bias.

In addition to the general confidence bounds mentioned above (which hold for *any* network topology), we also analyze a special case in which the network is assumed to be generated from a stochastic block model. In this case, we propose a natural algorithm for estimating the network topology based on survey samples, and analyze its estimation quality in terms of the differences between the estimated and true probability of an edge existing between any particular pair of vertices.

The rest of the paper is organized as follows. In Section 2, we introduce the notation that will be used throughout the paper. This is followed in Section 3 with a derivation of confidence bounds on the number of mistakes made by an algorithm, as a function of an explicit learning bias or "prior." Continuing in Section 4, we describe and analyze an algorithm for a special case where the network is assumed to be generated from a stochastic block model. We conclude with some general observations in Section 5.

2 Notation

To formalize the setting, we assume there is a true undirected unweighted graph $G = (V, E)$ on n distinguishable vertices, for which E is unknown to the learner. However, the learner does know V (and thus also n). Let Γ_n denote the set of all graphs on the n vertices; so $|\Gamma_n| = 2^{\binom{n}{2}}$. The *ego network* of a vertex v is a partition of the $n-1$ other vertices into 2 disjoint sets: namely, those adjacent to v and those not adjacent to v. By a *survey* on a vertex v, we mean that the ego network of v is revealed to the learner. In other words, if we survey v, then we learn exactly which other vertices are adjacent to v and which are not.

The task we consider is that of learning the entire graph topology based on information obtained by surveying m vertices, selected uniformly at random from V. This is therefore a *transductive* learning task.

Let $\hat{G} \in \Gamma_n$ represent some observed graph, and $G' \in \Gamma_n$ be a reference graph; say $\hat{G} = (V, \hat{E})$ and $G' = (V, E')$. Define $T(\hat{G}, G') = |\hat{E} \vartriangle E'|$, where $\vartriangle$ denotes the symmetric difference. If $G = G'$, this plays a role analogous to the "true error rate" in inductive learning. However, we cannot directly measure $T(\hat{G}, G)$ from observables if $m < n$.

For any set $S \subset V$ of m vertices from V, define $S \times V = \{\{s, v\} : s \in S, v \in V\}$, and let $\hat{T}_S(\hat{G}, G') = |(S \times V) \cap (\hat{E} \vartriangle E')|$. If $G = G'$, this plays a role analogous to the "training error rate" in inductive learning. We can always directly measure $\hat{T}_S(\hat{G}, G)$ after surveying all $v \in S$.

Let $G_0 = (V, \varnothing)$ denote the *empty graph*. Define

$$F_{T,n,m}(t) = \max_{G'=(V,E') \in \Gamma_n : |E'| = T} \mathcal{P}r_S\{\hat{T}_S(G', G_0) \leq t\},$$

where $S \subset V$ is a set of size m selected uniformly at random (without replacement). This is analogous to the probability over the random selection of the training set that the training error is at most t when the true error is T. Essentially, G' here represents the "mistakes graph" of edges in $\hat{E} \vartriangle E$ when $T(\hat{G}, G) = T$, except that since we do not know G, we must maximize over all such mistakes graphs to be sure the bound derived below will always apply.

Let $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ denote the nonnegative integers, and define

$$T^{(m)}_{max}(t, \delta) = \max \left\{ T \middle| T \in \mathbb{N}_0, T \leq \binom{n}{2}, F_{T,n,m}(t) \geq \delta \right\},$$

where dependence on n is implicit for notational simplicity. This is analogous to the largest possible true error rate such that there is still at least a δ probability of observing training error of t or less.

We formalize the notion of a *learning bias* by a "prior," or distribution on the set of all graphs. Formally, let $p : \Gamma_n \to [0, 1]$ be an arbitrary function such that $\sum_{\hat{G} \in \Gamma_n} p(\hat{G}) \leq 1$. For instance, in the social networks context, it may make sense to give a larger $p(\hat{G})$ value to graphs $\hat{G}$ that often have links between people that are living in close geographic proximity, or have similar demographic or personality traits, etc. We could also define more complex $p(\cdot)$ distributions, for example through a combination of vertex-specific attributes along with global properties of the network, as prescribed by certain models of real-world networks (e.g., (Leskovec et al., 2005; Wasserman & Robins, 2005)).

3 Confidence Bounds for Learning From Survey Samples

Almost by definition of T_{max}, we get the following bound.

Lemma 1.

$$\forall \eta \in [0, 1], \forall \hat{G} \in \Gamma_n,$$
$$\mathcal{P}r_S\{T(\hat{G}, G) > T^{(m)}_{max}(\hat{T}_S(\hat{G}, G), \eta)\} \leq \eta.$$

For completeness, a formal proof of Lemma 1 is included in the appendix. By substituting $\delta p(\hat{G})$ for η, for $\delta \in [0,1]$, we obtain the following. $\forall \hat{G} \in \Gamma_n$,

$$\mathcal{P}r_S\{T(\hat{G},G) > T^{(m)}_{max}(\hat{T}_S(\hat{G},G), \delta p(\hat{G}))\} \leq \delta p(\hat{G}). \tag{1}$$

Applying the union bound, this implies

$$\begin{aligned}\mathcal{P}r_S\{\exists \hat{G} \in \Gamma_n : T(\hat{G},G) &> T^{(m)}_{max}(\hat{T}_S(\hat{G},G), \delta p(\hat{G}))\} \\ &\leq \sum_{\hat{G}\in\Gamma_n} \delta p(\hat{G}) \leq \delta.\end{aligned}$$

Finally, negating both sides, we have the following bound holding simultaneously for all $\hat{G} \in \Gamma_n$.

Theorem 1. *For any $G \in \Gamma_n$ and $m \in \{0,1,\ldots,n\}$, with probability $\geq 1-\delta$ over the draw of S (uniformly at random from V without replacement) of size m,*

$$\forall \hat{G} \in \Gamma_n,\ T(\hat{G},G) \leq T^{(m)}_{max}(\hat{T}_S(\hat{G},G), \delta p(\hat{G})).$$

3.1 Relaxations of the Bound

The only nontrivial part of calculating this bound is the maximization in $F_{T,n,m}(t)$. For the special case of $F_{T,n,m}(0)$, corresponding to zero training mistakes, one can show that $F_{T,n,m}(0) = \binom{n-x}{m}/\binom{n}{m}$, where x is an integer such that $\binom{x-1}{2} < T \leq \binom{x}{2}$. However, in general it seems an exact explicit formula for $F_{T,n,m}(t)$ without any maximization required may be difficult to obtain. We may therefore wish to obtain upper bounds on $F_{T,n,m}(t)$ (implying upper bounds on $T^{(m)}_{max}$ as well). We derive some such bounds below.

Theorem 2.

$$F_{T,n,m}(t) \leq e^{-\tilde{x}m/n},$$

where $\tilde{x}$ is the smallest nonnegative integer x satisfying

$$2(T-t) \leq x(x-1) + (n-x)\min\{t+x, 2t\}. \tag{2}$$

Before proving Theorem 2, as an example of how the bound on $T^{(m)}_{max}$ implied by this behaves, suppose we choose a hypothesis network $\hat{G} = (V, \hat{E})$ that is *consistent* with the observations: that is, $\hat{T}_S(\hat{G},G) = 0$. Then we have the following result.

Corollary 1. *For $0 \leq m \leq n \in \{2,3,\ldots\}$, with probability $\geq 1-\delta$ over the draw of S (uniformly at random from V without replacement) of size m, $\forall \hat{G} \in \Gamma_n$,*

$$\begin{aligned}\hat{T}_S(\hat{G},G) = 0 \Rightarrow T(\hat{G},G) &\leq T^{(m)}_{max}(0, \delta p(\hat{G})) \\ &\leq \frac{1}{2}\left(\frac{n}{m}\ln\frac{1}{\delta p(\hat{G})}\right)^2.\end{aligned}$$

Proof of Corollary 1. Let $T = T^{(m)}_{max}(0, \delta p(\hat{G}))$. Then

$$\delta p(\hat{G}) \leq F_{T,n,m}(0) \leq e^{-\tilde{x}m/n} \leq e^{-\sqrt{2T}m/n}.$$

This implies

$$T \leq \frac{1}{2}\left(\frac{n}{m}\ln\frac{1}{\delta p(\hat{G})}\right)^2.$$

□

Given a fairly strong prior $p(\cdot)$, this can be a rapidly decreasing function of the number of samples (see the example in Section 3.2).

To prove Theorem 2, the following lemma will be useful.

Lemma 2.

$$F_{T,n,1}(t) \leq 1 - \frac{\hat{x}}{n},$$

where $\hat{x}$ is the smallest nonnegative integer x satisfying

$$2T \leq x(x-1) + (n-x)\min\{t+x, 2t\} \tag{3}$$

Proof of Lemma 2. The maximizing graph in the definition of $F_{T,n,1}(t)$ maximizes the number of vertices having degree at most t. Call this graph $\hat{G}$. Say there are x vertices in $\hat{G}$ having degree $> t$. Then $F_{T,n,1}(t) = 1 - \frac{x}{n}$. The total degree is $2T$, so the sum of degrees of the x vertices with degree $> t$ is at least $2T - t(n-x)$. However, since this is a simple graph, the total degree of these x vertices is at most $x(x-1) + (n-x)\min\{x,t\}$. Therefore, $2T - t(n-x) \leq x(x-1) + (n-x)\min\{x,t\}$. This means $\hat{x} \leq x$, which implies $F_{T,n,1}(t) \leq 1 - \frac{\hat{x}}{n}$, as claimed. □

We are now ready for the proof of Theorem 2.

Proof of Theorem 2.

$$\begin{aligned}F_{T,n,m}(t) &\leq \prod_{i=0}^{m-1} F_{T-t,n-i,1}(t) \\ &\leq [F_{T-t,n,1}(t)]^m \leq \left(1 - \frac{\tilde{x}}{n}\right)^m \leq e^{-\tilde{x}m/n}.\end{aligned}$$

□

For nonzero values of $\hat{T}_S(\hat{G},G)$, the bound on $T^{(m)}_{max}(\hat{T}_S(\hat{G},G), \delta p(\hat{G}))$ implied by Theorem 2 may behave in ways more complex than Corollary 1. However, we can still solve for $\tilde{x}$ explicitly, in various ranges depending on which term in the min dominates, as follows.
When $0 \leq t < \frac{T-t}{n}$,

$$\tilde{x} = \left\lceil \max\left\{\tfrac{1}{2} + t + \tfrac{1}{2}\sqrt{(1+2t)^2 + 8(T-(n+1)t)}, \tfrac{2T-(n+2)t}{n-t-1}\right\}\right\rceil.$$

When

$$\left\lceil \frac{2T-(n+2)t}{n-t-1} \right\rceil \leq \frac{1}{2}+t-\frac{1}{2}\sqrt{(1+2t)^2+8(T-(n+1)t)},$$

and

$$\frac{T-t}{n} \leq t < \min\left\{ \frac{2(T-t)}{n}, n-\frac{1}{2}-\frac{1}{2}\sqrt{(2n-1)^2-(8(T-t)+1)} \right\},$$

or when

$$n-\frac{1}{2}-\frac{1}{2}\sqrt{(2n-1)^2-(8(T-t)+1)} \leq t < 2\frac{T-t}{n},$$

we have

$$\tilde{x} = \left\lceil \tfrac{2T-(n+2)t}{n-t-1} \right\rceil.$$

If

$$\left\lceil \frac{2T-(n+2)t}{n-t-1} \right\rceil > \frac{1}{2}+t-\frac{1}{2}\sqrt{(1+2t)^2+8(T-(n+1)t)}$$

and

$$\frac{T-t}{n} \leq t < \min\left\{ \frac{2(T-t)}{n}, n-\frac{1}{2}-\frac{1}{2}\sqrt{(2n-1)^2-(8(T-t)+1)} \right\}, \quad (4)$$

then

$$\tilde{x} = \left\lceil \tfrac{1}{2}+t+\tfrac{1}{2}\sqrt{(1+2t)^2+8(T-(n+1)t)} \right\rceil.$$

In the other cases (i.e., $t \geq 2\frac{T-t}{n}$), we have $\tilde{x} = 0$.

We can also calculate bounds of intermediate tightness, at the cost of a more complex description. The following is one such example. Its proof is included in Appendix B.

Theorem 3.

$$F_{T,n,m}(t) \leq \left(1-\frac{\hat{x}_0^{(1)}}{n}\right)\prod_{i=2}^{m}\left(1-\frac{1}{n}\min_{y\in\{0,1,\ldots,t\}}\hat{x}_y^{(i)}\right),$$

where $\hat{x}_y^{(i)}$ is the smallest nonnegative integer x satisfying

$$2(T-y) \leq x(x-1)+(n-i+1-x)\min\{t-y+x, 2(t-y)\}. \quad (5)$$

3.2 A Simulated Example

As an example application of this bounding technique, we present the results of a simulated network learning problem in Figure 1. The simulated network is generated as follows. First, we generate 1000 points uniformly at random in $[0,1]^2$. For each point, we create a corresponding vertex in the network, and we connect any two vertices with an edge if and only if the corresponding points are within Euclidean distance 0.1. This generates a graph where approximately 1% of the pairs of vertices are adjacent. In the learning problem, the prior value $p(\hat{G})$ for a graph $\hat{G} = (V, \hat{E})$ is uniform on those graphs such that there exists a threshold θ such that any two vertices in $\hat{G}$ are adjacent if and only if the corresponding points are within distance θ, and it is zero elsewhere. Thus, there are precisely $1+\binom{n}{2}$ graphs with nonzero $p(\hat{G})$ value, and for these graphs $p(\hat{G}) = \left(1+\binom{n}{2}\right)^{-1}$. The learning algorithm simply outputs the graph corresponding to the sparsest of these $1+\binom{n}{2}$ that is consistent with the observed pairs. Since the true graph is among these, we can use Corollary 1, which implies a bound on the fraction of pairs for which the prediction is incorrect of $\binom{n}{2}^{-1}\frac{1}{2}\left(\frac{n}{m}\ln\frac{1+\binom{n}{2}}{\delta}\right)^2$. The plotted values use $\delta = 0.1$, and are averaged over ten repeated runs. The true fraction of pairs for which the algorithm predicts incorrectly is less than 0.0012, even for $m = 1$. Note that the bound can be rather loose for small m values, but becomes increasingly informative as m increases.

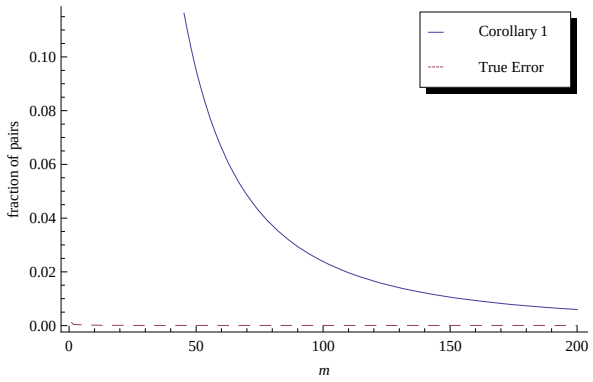

Figure 1: The true fraction of pairs that are incorrect and the bound on the fraction of pairs that are incorrect.

4 Learning with a Block Model Assumption

In this section, we provide an analysis of a particular algorithm, under a generative model assumption. As we will see, survey sampling is particularly well suited to the needs of this estimation problem. These results are entirely distinct from those in the previous section, except that they also involve learning from survey samples.

The particular modeling assumption we make here is a *stochastic block model* assumption. That is, every vertex $i \in \{1, 2, \ldots, n\}$ belongs to a *group* $g_i \in \mathcal{G}$, where $|\mathcal{G}| \leq n$. We assume that the g_i values are *unknown*, except for the m surveyed vertices. That is, for a random survey in this setting, the learner is informed of which other vertices that vertex is linked to *and* which group it is in.

Additionally, there is a *known* symmetric function $f(\cdot, \cdot)$ such that, for every i and j, $f(i, j) \in \{0, 1\}$; this will indicate the possibility for interaction between i and j (e.g., f could be a function of known features of the vertices, such as geographic proximity). We make the further assumption that for $g, h \in \mathcal{G}$, there is a value $p_{gh} \in [0, 1]$, such that for any i and j, the probability there is a link between i and j is precisely $p_{g_i g_j} f(i, j)$, and that these "link existence" random variables for the set of (i, j) pairs are independent.

As before, our task is to predict which of the unknown vertices are linked, based on information provided by m random surveys. However, given that edge existence is random, we may also be interested in estimating the probability $p_{ij} = p_{g_i g_j} f(i, j)$. We suggest the strategy outlined in Figure 2 to get an estimate $\hat{p}_{ij}$ of the probability that i and j are linked.

Let $\delta \in (0, 1)$, $\bar{f} = \min_{i,g} \frac{1}{n} \sum_{j: g_j = g} f(i, j)$, and $\bar{m} = m\bar{f} - \sqrt{2m\bar{f} \ln \frac{4n|\mathcal{G}|}{\delta}}$. The following theorem might be thought of as a coarse bound on the convergence of $\hat{p}_{ij}$ to p_{ij}.

Theorem 4. *Let $\hat{p}_{ij}$ be defined as in Figure 2, and let $m \in \{1, 2, \ldots, n\}$ be the number of random surveys. With probability $\geq 1 - \delta$, for all unsurveyed $i, j \in \{1, 2, \ldots, n\}$,*

$$|\hat{p}_{ij} - p_{ij}| \leq 9\sqrt{\frac{\ln(8n|\mathcal{G}|/\delta)}{2\bar{m}}}.$$

Proof of Theorem 4. For each $g, h \in \mathcal{G}$, let $m_{gh} = |Q_{gh}|$ denote the number of pairs (i, j) of surveyed vertices such that $g_i = g$ and $g_j = h$. Given the sample vertices, we have by Hoeffding's inequality that with probability $\geq 1 - \delta/2$,

$$\forall g, h \in \mathcal{G}, |\hat{p}_{gh} - p_{gh}| \leq \sqrt{\frac{1}{2m_{gh}} \ln \frac{4|\mathcal{G}|^2}{\delta}}.$$

Again by Hoeffding's inequality, with probability $\geq 1 - \delta/4$, for every $i \in \{1, 2, \ldots, n\}$ and $g \in \mathcal{G}$, if m_{ig} is the number of surveyed vertices j (with $j \neq i$) that have group g and $f(i, j) = 1$, and $\hat{p}_{ig}$ is the fraction of these to which i is linked, then

$$|\hat{p}_{ig} - p_{g_i g}| \leq \sqrt{\frac{1}{2m_{ig}} \ln \frac{8n|\mathcal{G}|}{\delta}}.$$

Thus, with probability $\geq 1 - \frac{3}{4}\delta$, every $i \in \{1, 2, \ldots, n\}$ and $g \in \mathcal{G}$ has

$$|\hat{p}_{ig} - \hat{p}_{g_i g}| \leq \sqrt{\frac{1}{2m_{ig}} \ln \frac{8n|\mathcal{G}|}{\delta}} + \sqrt{\frac{1}{2m_{g_i g}} \ln \frac{4|\mathcal{G}|^2}{\delta}}.$$

Let us suppose that this event occurs. Let $\tilde{m} = \min_{i \in V, g \in \mathcal{G}} m_{ig}$. Clearly we have every $m_{gh} \geq \tilde{m}$ and every $m_{ig} \geq \tilde{m}$. Now let $i, j \in \{1, 2, \ldots, n\}$ be unsurveyed vertices. Then

$$\begin{aligned}
& |\hat{p}_{ij} - p_{g_i g_j} f(i, j)| = |\hat{p}_{\hat{g}_i \hat{g}_j} - p_{g_i g_j}| f(i, j) \\
\leq\ & |\hat{p}_{\hat{g}_i \hat{g}_j} - p_{g_i g_j}| \leq |\hat{p}_{\hat{g}_i \hat{g}_j} - \hat{p}_{g_i, g_j}| + |\hat{p}_{g_i g_j} - p_{g_i g_j}| \\
\leq\ & |\hat{p}_{\hat{g}_i \hat{g}_j} - \hat{p}_{i \hat{g}_j}| + |\hat{p}_{i \hat{g}_j} - \hat{p}_{g_i \hat{g}_j}| \\
& + |\hat{p}_{\hat{g}_j g_i} - \hat{p}_{j g_i}| + |\hat{p}_{j g_i} - \hat{p}_{g_j g_i}| + |\hat{p}_{g_i g_j} - p_{g_i g_j}| \\
\leq\ & 4\sqrt{\frac{1}{2\tilde{m}} \ln \frac{8n|\mathcal{G}|}{\delta}} + 5\sqrt{\frac{1}{2\tilde{m}} \ln \frac{4|\mathcal{G}|^2}{\delta}} \\
\leq\ & 9\sqrt{\frac{\ln(8n|\mathcal{G}|/\delta)}{2\tilde{m}}}.
\end{aligned}$$

All that remains is to lower bound $\tilde{m}$. Note that for each i and g, $\mathbb{E}[m_{ig}] \geq \bar{f}m$. By a Chernoff and union bound, for any $\epsilon \in (0, 1)$, with probability $\geq 1 - n|\mathcal{G}|e^{-m\bar{f}\epsilon^2/2}$, for every $i \in \{1, 2, \ldots, n\}$ and $g \in \mathcal{G}$, $m_{ig} \geq \bar{f}m(1 - \epsilon)$. In particular, by taking $\epsilon = \sqrt{\frac{2\ln(4n|\mathcal{G}|/\delta)}{m\bar{f}}}$, we have that with probability $\geq 1 - \delta/4$, $\tilde{m} \geq \bar{m}$. A union bound to combine this with the results proven above completes the proof. □

After running this procedure, we must still decide how to predict the existence of a link using the $\hat{p}_{ij}$ values. The simplest strategy would be to predict an edge between pairs with $\hat{p}_{ij} \geq 1/2$. However, one problem for network completion algorithms is determining the right loss function. Because most networks are quite sparse, using a simple "number of mispredicted pairs" loss often results in the optimal strategy being "always say 'no edge'." However, this isn't always satisfactory. In many situations, we are willing to tolerate a reasonable number of false discoveries in order to find a few correct discoveries of unknown existing edges. So the need arises to trade off the probability of false discovery with the probability of missed discovery. We can take this preference into account in our network completion strategy, simply by altering this threshold for how high $\hat{p}_{ij}$ must be before we predict that there is an edge. An appropriate value of the threshold to maximize the true discovery rate while constraining the false discovery rate can be calculated using Theorem 4.

5 Conclusions

The problem of learning the topology of a network from survey samples has an interesting and subtle

Let Q_{gh} be the set of pairs (i,j) of surveyed vertices having $g_i = g$, $g_j = h$, and $f(i,j) = 1$
Let $\hat{p}_{gh}$ be the fraction of pairs in Q_{gh} that are linked in the network
For each unsurveyed i, let Q_{ig} be the set of surveyed j having $f(i,j) = 1$ and $g_j = g$
and let $\hat{p}_{ig}$ be the fraction of vertices $j \in Q_{ig}$ such that i and j are linked in the network
Let $\hat{g}_i = \arg\min_{g \in \mathcal{G}} \max_{h \in \mathcal{G}} |\hat{p}_{gh} - \hat{p}_{ih}|$
For each pair (i,j) of unsurveyed vertices, let $\hat{p}_{ij} = \hat{p}_{\hat{g}_i \hat{g}_j} f(i,j)$

Figure 2: A method for estimating the probability of edge existence, given a stochastic block model assumption and survey samples.

structure, which we have explored to some extent in this paper. In the first perspective we examined the problem from, we made essentially no assumptions other than the sampling method, and were able to derive general confidence bounds on the number of mistakes, in the style of PAC-MDL bounds. The main challenge was to account for the fact that the observable pairs of vertices are not chosen uniformly at random, as would be required for most of the known results in the learning theory literature to apply. The bounds we derived have several noteworthy properties. They indicate that, as usual, a strong prior is necessary in order to make nontrivial guarantees on the number of mistakes. Given such a strong prior, we can compare the rate of decrease of the bounds to some other rates we might imagine. For instance, in order to reduce this problem to a problem with uniform sampling of vertex pairs, we could simply retain only one of the observed pairs from each survey sample. In the simple zero training mistakes scenario, this would yield a bound on the fraction of predictions that are mistakes, decreasing as $\Theta(m^{-1})$ for a given hypothesis; comparing this to the $\Theta(m^{-2})$ bound proven above for using the full survey sample shows improvement. At the other extreme, perhaps the fastest rate we might conceive of for *any* type of sampling might be on the order of k^{-1}, where k is the number of vertex pairs we have observed in the sample. In our case, $k = \binom{m}{2} + m(n-m)$. The explicit bounds we derive seem not to achieve this $\Theta((mn)^{-1})$ rate, indicating that each observed pair carries less information under the non-uniform sampling compared to independent samples.

In the second perspective, we studied the convergence of a specific estimator of the probability any given edge exists, under a stochastic block model generative model assumption. The type of estimation we describe is particularly well suited to survey sampling, as it allows us to estimate the group memberships of the unsurveyed vertices based on how they interact with the surveyed vertices (whose group memberships are known). As they are a first attempt at this type of analysis, the rates we derive for this problem are admittedly coarse, and there may be room for further progress.

A Proof of Lemma 1

Proof. Let $\hat{G} = (V, \hat{E})$, and for $t \in \mathbb{R}$ let $F_{\hat{G},G}(t) = \mathcal{P}r_S\{\hat{T}_S(\hat{G}, G) \leq t\}$. Let $G' = (V, E \triangle \hat{E}) = (V, E')$. Then $\hat{T}_S(\hat{G}, G) = |(S \times V) \cap E'| = \hat{T}_S(G', G_0)$, and thus

$$\begin{aligned} & F_{\hat{G},G}(t) = \mathcal{P}r_S\{\hat{T}_S(G', G_0) \leq t\} \\ \leq & \max_{G''=(V,E'') \in \Gamma_n : |E''| = |E'|} \mathcal{P}r_S\{\hat{T}_S(G'', G_0) \leq t\} \\ = & F_{T(\hat{G},G),n,m}(t). \end{aligned}$$

Given $\eta \in [0,1]$, we have that

$$\begin{aligned} \eta \geq & \mathcal{P}r_S\{F_{\hat{G},G}(\hat{T}_S(\hat{G}, G)) < \eta\} \\ \geq & \mathcal{P}r_S\{F_{T(\hat{G},G),n,m}(\hat{T}_S(\hat{G}, G)) < \eta\} \\ \geq & \mathcal{P}r_S\Bigg\{T(\hat{G}, G) > \max\Bigg\{T \Bigg| T \in \mathbb{N}_0, T \leq \binom{n}{2}, \\ & F_{T,n,m}(\hat{T}_S(\hat{G}, G)) \geq \eta\Bigg\}\Bigg\} \\ = & \mathcal{P}r_s\{T(\hat{G}, G) > T^{(m)}_{max}(\hat{T}_S(\hat{G}, G), \eta)\}. \quad \square \end{aligned}$$

B Proof of Theorem 3

Proof. Let $\hat{G}$ be a maximizing graph in the definition of $F_{T,n,m}(t)$. Define $f^{(t)}_{T,n,i}(y) = \mathcal{P}r_S\{\hat{T}_S(\hat{G}, G_0) = y\}$, where $S \subset V$ is a set of size i selected uniformly at random. The theorem follows immediately from Lemma 2 if $m = 1$. Suppose $m > 1$.

$$\begin{aligned} F_{T,n,m}(t) \leq & \sum_{y=0}^{t} F_{T-y,n-m+1,1}(t-y) f^{(t)}_{T,n,m-1}(y) \\ \leq & \sum_{y=0}^{t} \left(1 - \frac{\hat{x}^{(m)}_y}{n}\right) f^{(t)}_{T,n,m-1}(y), \end{aligned} \tag{6}$$

where (6) follows from Lemma 2. Clearly, $\left(1 - \frac{\hat{x}_y^{(m)}}{n}\right) \leq \left(1 - \frac{1}{n}\min_{y\in\{0,1,\ldots,t\}} \hat{x}_y^{(m)}\right)$, so that (6) is at most

$$\sum_{y=0}^{t}\left(1 - \frac{1}{n}\min_{y'\in\{0,1,\ldots,t\}} \hat{x}_{y'}^{(m)}\right) f_{T,n,m-1}^{(t)}(y)$$
$$\leq \left(1 - \frac{1}{n}\min_{y'\in\{0,1,\ldots,t\}} \hat{x}_{y'}^{(m)}\right) F_{T,n,m-1}(t)$$
$$\leq \left(1 - \frac{\hat{x}_0^{(1)}}{n}\right)\prod_{i=2}^{m}\left(1 - \frac{1}{n}\min_{y\in\{0,1,\ldots,t\}} \hat{x}_y^{(i)}\right).$$

The final inequality follows by induction on m (with base case $m = 2$), and Lemma 2. □

Acknowledgments

This material is based upon work supported by an NSF CAREER Award to EPX under grant DBI-0546594, and NSF grant IIS-0713379. EPX is also supported by an Alfred P. Sloan Research Fellowship.

References

Blum, A., & Langford, J. (2003). PAC-MDL bounds. *16th Annual Conference on Learning Theory.*

Frank, O. (2005). Network sampling and model fitting. *Models and Methods in Social Network Analysis* (pp. 31–56). Cambridge University Press.

Leskovec, J., Chakrabarti, D., Kleinberg, J., & Faloutsos, C. (2005). Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. *European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD).*

Wasserman, S., & Robins, G. (2005). An introduction to random graphs, dependence graphs, and p*. *Models and Methods in Social Network Analysis* (pp. 148–161). Cambridge University Press.