Network Completion and Survey Sampling

Steve Hanneke and Eric P. Xing
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213 USA
{shanneke,epxing}@cs.cmu.edu

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. This task is quite challenging, and there have even been rumors that it is technically impossible. In this abstract, we argue that this is not the case, though a strong learning bias seems to be required.

In particular, 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 can not 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.

2 Agnostic Learning

In this setting, we have some *learning bias*, in the form of a distribution $p(E)$ over edge configurations in the graph. We then wish to upper bound, with high probability, the fraction of mistakes among the pairs of unsurveyed vertices. This is analogous to the standard Occam bounds from the learning theory literature. Let Γ_n be the set of all edge configurations; so $|\Gamma_n| = 2^{\binom{n}{2}}$. Also, let $V = \{1, 2, \ldots, n\}$ and let E^* denote the edge configuration in the *true* network. For any $\hat{E} \in \Gamma_n, E' \in \Gamma_n$, let $T(\hat{E}, E') = |\hat{E} \triangle E'|$, where $\triangle$ denotes the symmetric difference. For any set S of m vertices, let $\hat{T}_S(\hat{E}, E') = |(S \times V) \cap (\hat{E} \triangle E')|$. Let $E_0 = \varnothing$ denote the *empty graph*. Define

$$F_{T,n,m}(t) = \max_{E:|E|=T} \mathcal{P}r_S\{\hat{T}_S(E, E_0) \leq t\},$$

where $S \subset V$ is a set of size m selected uniformly at random. Now define

$$T^{(m)}_{max}(t, \delta) = \max\{T | T \in \mathbb{Z}, F_{T,n,m}(t) \geq \delta\},$$

where dependence on n is implicit for notational simplicity. Using these definitions, we have the following bound on the number of mistaken predictions between unsurveyed vertices, as a function of the number of observed mistakes.

Theorem 2.1. $\forall \delta \in (0,1)$, $\mathcal{P}r_S\{\forall \hat{E} \in \Gamma_n,\ T(\hat{E}, E^*) \leq T^{(m)}_{max}(\hat{T}_S(\hat{E}, E^*), \delta p(\hat{E}))\} \geq 1 - \delta$.

For a somewhat more explicit (though looser) bound, we can use the observation that

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

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

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

The proof of this (and some bounds of intermediate tightness) are available upon request. For an example of how this bound behaves, suppose we choose a hypothesis network $\hat{E}$ that is *consistent* with the observations. Then, for $m < n/2$, with probability $1-\delta$, the fraction of pairs of unsurveyed vertices on which $\hat{E}$ is mistaken is at most $\left(\frac{2}{m}\ln\frac{1}{\delta p(\hat{E})}\right)^2$.

3 Learning with a Block Model Assumption

We can also look at learning models with more assumptions, to ease learning. In this section, we consider using survey samples as before, but this time every vertex $i \in \{1,2,\ldots,n\}$ belongs to a *group* $g_i \in G$, where G is a finite set. We assume that the g_i are *unknown*, except for the m surveyed vertices. That is, for a random survey in this setting, we ask the vertex which other vertices it is linked to *and* which group it is in.

Additionally, there is a *known* 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 $g,h \in G$, 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 $p_{g_i g_j} f(i,j)$.

As before, our task is to predict which of the unknown vertices are linked, based on information provided by m random surveys. We suggest the following strategy to get an estimate $\hat{p}_{ij}$ of the probability that i and j are linked.

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 $\hat{g}_i = \arg\min_{g\in G} \max_{h\in G} |\hat{p}_{gh} - \hat{p}_{ih}|$
For each pair (i,j) of unsurveyed vertices, let $\hat{p}_{ij} = \hat{p}_{\hat{g}_i \hat{g}_j} f(i,j)$

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 as follows. Suppose we wish to constrain the false discovery probability below $\rho \in (0,1)$. We can predict an edge between unsurveyed vertices i and j whenever $1 - \hat{p}_{ij} + \beta \leq \rho$, for β defined below.

Let $\delta \in (0,1)$, $\bar{f} = \min_{i,g} \frac{1}{n-1} \sum_{j:g_j=g} f(i,j)$, and $\bar{m} = m\bar{f} - \sqrt{2m\bar{f}\ln\frac{6n|G|}{\delta}}$. The following theorem follows from simple bound arguments for Binomials. The proof can be made available upon request.

Theorem 3.1. *With probability $\geq 1-\delta$, for all $i,j \in \{1,2,\ldots,n\}$,*

$$|\hat{p}_{ij} - p_{g_i g_j}| \leq 9\sqrt{\frac{\ln(12n|G|/\delta)}{2\bar{m}}}.$$

Thus, we can define $\beta = 9\sqrt{\frac{\ln(12n|G|/\delta)}{2\bar{m}}}$. This guarantees that, with probability $1-\delta$, the false discovery probability for an edge existence prediction is at most ρ. Furthermore, it guarantees, with probability $1-\delta$, that the probability that there is a link between i and j when we predict there is not one is at most $f(i,j)p_{g_i g_j} \leq f(i,j)(1-\rho+2\beta)$.