Scribed Notes 36-720 Statistical Network Models

Brendan McVeigh

September 14, 2016

Review

We define a graph $G = (V, E)$, where $E \subseteq V \times V$. Where V is the set of nodes and E is the set of edges between nodes. Graphs are mathematical abstractions while networks are things that exist in the real world.

Random Graph (review of properties)

For n nodes we define an adjacency matrix A where $A_{i,j} \underset{iid}{\sim} Bern(p)$

- Phase transition to giant component at $\lambda = p(n-1)$. If $\lambda > 1$, will have a large component. If $\lambda < 1$, then will have small chunks.
- Binomial degree distribution $\to$ Poisson as $n \to \infty$ and $p \to 0$.
- Diameter of giant compontent O(log(n)) (for $\lambda > 1$.
- $\hat{p}_{MLE} = \sum_{ij} \frac{A_{ij}}{2\binom{n}{2}}$, exponential family
- Rare (p^3) triangles, little transitivity. To get a triangle all three edges need to apear and they appear independently with probability p, $P(triangle \,|\, two\ edges) = p$.

These properties are not a good fit for most observed networks in the real world. Thus, we try to relax some assumptions "As much as a random graph as possible but trying to fix some of these problems" result is the block model.

Block Model

Terminolog

All nodes are divided into k blocks where a vector Z is the block assignemnts, so node i belongs to block z_i. Edges are still individually follow a Bernoulli distribution where

$$P(A_{ij}|Z_i = r, Z_j = s) = b_{rs}$$

where b is a $k \times k$ (symmetric for undirected graphs) matrix. Known as the "affinity matrix" although other names are also used in the literature.

Use n_r to denote the number of nodes in block r, $\left(n = \sum_{r=1}^{k} n_r\right)$

Density

Baseline $P(edge)$ if we don't know blocks

$$\begin{aligned}
P_{\text{effective}} &= P(A_{ij}) \\
&= \sum_{r=1}^{k}\sum_{s=1}^{k} P(A_{ij}=1, Z_i=r, Z_j=s) \\
&= \sum_{r,s} P(A_{ij}=1|Z_i=r, Z_j=s)P(Z_i=r, Z_j=s) \\
&= \sum_{r,s} b_{rs}P(Z_i=r, Z_j=s) \\
&= \sum_{r,s} b_{rs}\frac{n_r}{n}\frac{n_s}{n} \\
&= \sum_{r,s} b_{rs}\frac{n_r n_s}{n^2}
\end{aligned}$$

Triangle Formation

$$\begin{aligned}
&P(\textit{completing triangle}) \\
&= P(A_{ik}=1|A_{ij}=1, A_{jk}=1) \\
&= \sum_{r,s,q} P(A_{ik}=1, z_i=r, z_j=s, z_k=q|A_{ij}=1, A_{jk}=1) \\
&= \sum_{r,s,q} P(A_{ik}=1|z_i=r, z_j=s, z_k=q, A_{ij}=1, A_{jk}=1)P(z_i=r, z_j=s, z_k=q|A_{ij}=1, A_{jk}=1) \\
&= \sum_{r,s,q} b_{rq}P(z_i=r, z_j=s, z_k=q|A_{ij}=1, A_{jk}=1)
\end{aligned}$$

We can think of this as a weighted average of elements in the affinity matrix. If two nodes prefer links to nodes of the same type $b_{rr} > b_{rs}$ we will see more triangles than we would otherwise expect looking only at the overall density ($P_{\text{effective}}$). The opposite is also true, will get fewer than expected triangles if nodes prefer nodes of other types (blocks).
Overall density

$$\sum_{r,s}\frac{n_r n_s}{n^2} \qquad \text{vs.} \qquad P(A_{ik}=1|A_{ij}=1, A_{jk}=1) = \sum_{r,s,q} b_{rq}P(z_i=r, z_j=s, z_k=q|A_{ij}=1, A_{jk}=1)$$

Likelihoods

Block model

$$L(b) = \prod_{i<j}(b_{z_i z_j})^{A_{ij}}(1-b_{z_i z_j})^{1-A_{ij}}$$

$$\ell(b) = \sum_{i<j} A_{ij}\log(b_{z_i z_j}) + (1-A_{ij})\log(1-b_{z_i z_j})$$

If n_r is the number of nodes in block r, and e_{rs} is the number of edges between block r and block s. Then

$$\begin{aligned}
\ell(b) &= \sum_{r=1}^{k}\sum_{s=1}^{k} e_{rs}\log(b_{rs}) + (n_r n_s - e_{rs})\log(1-b_{rs}) \\
&= \sum_{r,s}^{k} n_r n_s \log(1-b_{rs}) + e_{rs}\text{logit}(b_{rs})
\end{aligned}$$

Random Graph	Block Modle
• All edges are independent	• All edges are independent given block assignments Z
• All edges appear with the same probability p	• All edges between two blockes, r, s appear with the same probability b_{rs}. Each individual block r looks like a random graph with $p = b_{rr}$
• distribtuion of all edges is Binomial(n, p)	• Edges within a block r follow a Binomial$((n_r^2 - n_r)/2, b_{rr})$ distribution, between blocks r, s is Binomial$(n_r n_s, b_{r,s})$
• distribution of edges is an exponential family distribution, number of edges is a sufficient statistic for p	• e_{rs} distribution of edges is an exponential family distribution, the number of edges between block r and block s is a sufficient statistic for b_{rs}
• Propportion of triangles is p^3, no transitivity	•

Table 1: Comparison between properties of random graphs and of block models

Since b_{rs} only appears in one term of sum the we can get and MLE for b_{rs}, $\hat{b}_{rs} = e_{rs}/(n_r n_s)$ since the distribution of e_{rs} is Binomial$(n_r n_s, b_{rs})$

Routes to Graphical Model

Given Partition

Given a partition into blocks, "most random" graph with observed edge densities. By most random we mean the maximum entroy distribution, the distribution closest to homogenous random graphs in kullback-leibler distance. One caveat is that we must be dividing nodes into blocks correctly in order for the model to be useful.

Graph Theory

Szemerédi regularity lemma: For any graph G of n nodes, and any ϵ, we can divide the nodes into $s(n, \epsilon)$ blocks, and the edge counts come within ϵ if expections for a k-block model.

Role Models

From sociology: "regular equivalence", two nodes are structureally equivalent when they have the same neighbors. Thus, Professor A and Professor B as shown in Figure 1 are not structurally equivalent but we want to say they are simiar. This leads to the concept of regular equivalence, two nodes are regularly equivalent if they have the same types of connections even if they are not exactly the same ones. The types of relationships or "role model" are shown in Figure 2.

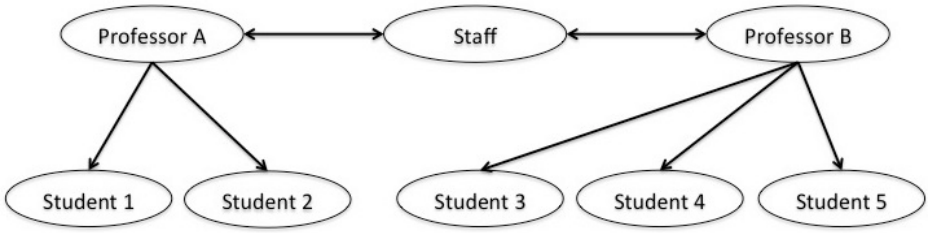

Figure 1:

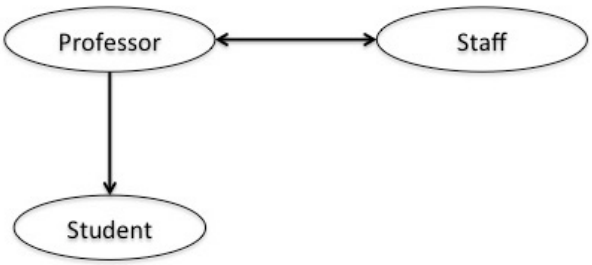

Figure 2: A directed graph with a loop