Statistics of Big Data, Fall 2023-24

Class notes 9

Sources for today's material:
survey by Goldberg et al. on statistical modeling of network data (that appeared in 2010 in the Foundations and Trends in Machine Learning)

The more general ERG-p^* approach

So far the only complication we assumed is mutuality for directed graphs. Frank & Strauss considered a larger family of graphs, where the probability of an edge can depend on all edges that share a node with it (rather than only the opposite edge between the two nodes). They showed that for undirected graphs all these models can be written in general form:

$$\mathbb{P}(Y=y)=\exp\left(T(y)\tau+\sum_{k=1}^{N-1}S_k(y)\theta_k+\psi(\theta,\tau)\right).$$

This is a model with N parameters where the summary statistics are:

- $S_k(y),\ k=1,\dots,N-1$ — the number of k-stars in the graph, where a k-star is a node connected to k neighbors. 1-star is an arc, 2-star is a node with two arcs, etc.
- $T(y)$ — the number of triangles (3-clicks) in the graph

This model is very general, but not so good to work with: the N parameters are a large number, and the counts S_k are heavily dependent on each other, creating strong instability. We are also mostly interested in directed graphs (although the mapping between models for directed and indirected is usually simple).

As a practical approach inspired by this formulation, Wasserman & Patterson proposed the general Exponential Random Graph (ERG) model, also called p^*. Instead of specifying the statistics S_k, T above, they suggest a flexible framework where the user can define a set of statistics $u_1(y),\dots,u_k(y)$ with corresponding parameters $\theta_1,\dots,\theta_k$ and posit the model:

$$\mathbb{P}(Y=y)=\exp\left(\theta^T u(y)-\psi(\theta)\right),$$

where $\psi(\theta)$ is a normalization term (like the λ_{ij} above). The way to fit this model is with the same pseudo-likelihood approach, where modeling $\mathbb{P}(Y_{ij}=1|Y_{-ij})$ gives a simple logistic rergression in the parameters θ. The problems with this approach include the unreliability of PML and the strong dependence between summary statistics like number of k-stars. Some of the classical statistics that are included in $u(y)$:

1. **Edges**: The number of edges S_1
2. **Mutuality/reciprocity**: The number of directed pairs as in p_1
3. **Stransitivity**: the number of directed triangles in the graph

etc.

Latent space models

Now we switch to a different way of thinking about graphs. We assume the nodes have unobserved latent variables which are *locations* in some latent space, and they affect the affinity between nodes and their tendency to connect: closer nodes are more likely to connect. Formally, assume each node i has a latent (unobserved) location $Z_i \in \mathbb{R}^d$, and there is some distance (say Euclidean) on $\mathbb{R}^d$ such that $\mathbb{P}(Y_{ij} = 1)$ depends on $D(Z_i, Z, j)$.

We may also assumed that each edge has observed covariates X_{ij}, and nodes can have observed covariates too, in which case we assume things like $X_{ij} = X_i^T X_j$.

In this model we assume that given the latent variables the edges are independent, and the distribution has some parameters Θ:

$$\mathbb{P}\left(Y|Z, X; \Theta\right) = \prod_{i \neq j} \mathbb{P}(Y_{ij}|Z_i, Z_j, X_{ij}; \Theta),$$

and typically assume simply a logistic model for the node probabilities:

$$\text{logit}\left(\mathbb{P}(Y_{ij}|Z_i, Z_j, X_{ij}; \Theta)\right) = \alpha + \beta^T X_{ij} - D(Z_i, Z_j),$$

where $D(Z_i, Z_j) = \|Z_i - Z_j\|_2$ for example.

In the simple case that there are no observed features X_{ij} this takes the form:

$$\text{logit}\left(\mathbb{P}(Y_{ij}) = \alpha - D(Z_i, Z_j).\right.$$

Formally this ia missing data problem (since the Z_i are unobserved). Solving this maximum likelihood involves integrating over the unobserved variables, this is typically done by MCMC. It can lead to parameter estimates $\hat{\alpha}, \hat{\beta}$, but often we want to make use of these approaches for clustering or other ways to learn about structures in the latent space. For this we can infer the "likely" locations Z_i from the MCMC and apply clustering to them.

A more direct approach is to assume that there is a natural clustering model that generated the Z_i's and actually fit the parameters of this model as well as the parameters for $\mathbb{P}(y|Z)$ above. The simplest and most common approach is a Gaussian Mixture Model (GMM) assumption: $Z_i \sim \sum_k p_k N(\mu_k, \sigma_k^2 I)$, which assumes a collection of spherical Gaussians generated the latent locations. The number of Gaussians is the number of clusters, and we can think of this approach as combining the latent space modeling approach with the GMM approach for clustering. This is the approach of Handcock et al. (2002).

If we write out the complete log-likelihood for this we need to figure out whether the model we found fits the data well, and we have to consider both the GMM log-likelihood of the locations we inferred (which now also has parameters):

$$\sum_{i=1}^{n} \log(\sum_{k=1}^{K} p_k |\Sigma_k| {-1/2} \exp\left\{-0.5(z_i - \mu_k)^T \Sigma_k^{-1} (z_i - \mu_k)\right\},$$

as well as the regular log-likelihood for Y_{ij} given the latent locations. This gives the complete data log likelihood, but the latent locations are missing data, so that distribution has to be integrated over, using MCMC or EM...

To select between models we also have to penalize for the number of parameters as we always do. The theory behind the approximations that Handcock et al. (2002) employ is complex, but they come up with an approximate model selection measure based on BIC. We will not go into details, but accept that these are usable but not very reliable measures of how well the model fits and they can help us (together with visual and intuitive arguments) to find what models fit our data.

Scale-free networks

There is strong folklore that large "natural" networks (the internet, Facebook...) have typical properties:

1. Small number of nodes with a large or huge number of edges ("hubs")
2. Most nodes have very few edges (long tailed phenomenon)
3. The few-edges nodes are strongly clustered (communities)

It has been widely argued that for properties 1+2 a good fit is the scale-free model for the number of edges of each node, where

$$\mathbb{P}(X = k) \propto k^{-\gamma},$$

where $2 < \gamma \leq 3$: note that for $\gamma \leq 2$ there is no expectation, while for $\gamma \leq 3$ there is no variance. Hence we are assuming that these fat-tailed distributions in the scale-free network have an expectation but no variance.

Why is this called scale-free? Note that with this form:

$$\frac{\mathbb{P}(X = c \cdot k)}{\mathbb{P}(X = k)} = c^{-\gamma},$$

regardless of k, so the tail behavior is the same for small relative to medium, medium relative to large, etc.

Other "soft" properties of scale-free networks:

1. *Small world:* there are short paths from each node to each node going through hubs
2. *Robustness:* deleting nodes does not hurt connectivity or typical path lengthes
3. *Clustering:* formation of tight communities

Assuming we accept the fundamental importance of scale-free graphs (today widely disputed), we can ask what type of random processes can create such graphs? One important one is the **Preferential Attachment** model of Barabasi-Albert (1999). This simple model evolves as:

- Start with a set of m_0 nodes, randomly connected between them

- At stage N we add another node and connect it to $m < m_N$ existing nodes, with probability that is proportional to the number of connections each of them already has:

$$p_i \propto \frac{k_i}{\sum_j k_j}.$$

For this simple process they show:

- As the network grows we get $\mathbb{P}(k) \propto k^{-3}$, at the edge of the range for scale-free.
- The length of an average path is about $\frac{\log(N)}{\log\log(N)}$ when the network has N nodes $\Rightarrow$ a small world.

As mentioned above, in recent years there has been extensive skepticism about the usefulness of these models, and how well they fit real data.