Statistics of Big Data, Fall 2023-24

Class notes 8

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)

Network data modeling

A network or graph is a collection of N nodes and E arcs (directed or undirected) between then. The type of questions we want to ask about networks:

1. The nature of the network and connections in it, for example:
   - Is the network connected? What is the length of typical paths between connected nodes ("small world")?
   - Is the network reciprocal: If a node points to others, do they point back to it? If it points to many do many point back?
   - Clusters and high connectivity groups
   - Existence of "hubs" that are close to all
2. The connections between features or properties of the nodes or arcs and the nature of the network: what makes you "popular" etc.

It goes without saying that the answers to questions like this are critical and useful in many areas of science and business, increasingly so as networks (social and others) become central in our lives.

Erdos-Renyi-Gilbert model

This is the simplest and most classical analysis of networks, but still important and relevant. It generally deals with undirected graphs, though the main results also apply to directed.

Assume a graph with N nodes and we randomly generate E undirected edges between pairs of nodes. The formulation has two varieties:

- $G(N, E)$: A random draw from all possible graphs with exactly E edges, each with probability:

$$\frac{1}{\binom{\binom{N}{2}}{E}}.$$

- $G(N,p)$: Each of the $\binom{N}{2}$ arcs is selected with equal probability p, so the probability of a random graph with E edges is
$$p^E(1-p)^{\binom{N}{2}-E},$$
while the overall probability of seeing exactly E edges is:
$$\mathbb{P}(E) = \binom{\binom{N}{2}}{E} p^E(1-p)^{\binom{N}{2}-E}.$$

The main questions they asked about this graph is about the nature of connectivity in this "symmetric" setting. This leads to some powerful and famous results. Denote by λ the average rank for a node in this setting:

$$G(N,p):\ \lambda(N) = Np(N) \quad G(N,E):\ \lambda(N) = 2\frac{E(N)}{N},$$

Then Erdos-Renyi proved the following:

1. If $\lambda(N) < 1$ then as $N \to \infty$ the size of all connected components is $O(\log(N))$ with high probability, meaning the graph will be totally fragmented and most nodes can reach only very few others via the edges.

2. If $\lambda(N) \to 1$ then for large N there will (with high probability) be many components of size $O(N^{2/3})$, meaning the graph will still be highly fragmented, but each node can now reach a decent number of other nodes via the edges.

3. If $\lambda(N) \to c > 1$ then there will (with high probability) be one *huge* component that contains a positive percentage of the points (typically close to 100%), and all other components will be tiny with $O(\log(N))$ nodes. In practice, it means the graph is largely connected.

These results have been very influential, but they have some major simplifying assumption that limit their practical utility. They ignore phenomena that are important and prevalent in real graphs and networks:

- Some nodes are more central and connected than others (hubs)

- In directed graphs, pairs may have mutual relationship: if I have edge Y_{ij} from node i to node j, it is likely to affect (typically make more likely) the edge Y_{ji}.

etc. To build more useful models we have to get away from the completely random assumption and start considering these aspects.

The p_1 model of node properties and edge creation

Consider two nodes i, j and the four possible settings of the edges $Y_{ij} \in \{0,1\}$, $Y_{ji} \in \{0,1\}$, as a function of the parameters of the network and the nodes:

- θ: overall rate of connections (like in Erdos-Renyi)

- α_i : *Expansiveness*, measuring how friendly node i is

- β_i : *Popularity*, measuring how attractive node i is
- ρ : *Reciprocity*, measuring how likely $Y_{ij} = Y_{ji}$ is

We also have λ_{ij} a normalization factor. In this setting we can write the four probabilities as a function of the parameters:

$$\begin{aligned}
\log(\mathbb{P}_{ij}(0,0)) &= \lambda_{ij} \\
\log(\mathbb{P}_{ij}(1,0)) &= \lambda_{ij} + \alpha_i + \beta_j + \theta \\
\log(\mathbb{P}_{ij}(0,1)) &= \lambda_{ij} + \alpha_j + \beta_i + \theta \\
\log(\mathbb{P}_{ij}(1,1)) &= \lambda_{ij} + \alpha_i + \alpha_j + \beta_i + \beta_j + \rho + 2\theta
\end{aligned}$$

where λ_{ij} is such that the probabilities sum to 1.

Now note that if we choose $\rho = 0$, $\alpha_i = \beta_i = 0$, $\forall i$, then we get the Erdos-Renyi model with θ only (fixed probability).

To fit this model to data we would write the likelihood as a function of the parameters:

$$\mathcal{L}(\theta, \alpha, \beta, \rho) = C(\lambda) + \sum_{i,j=1, i\neq j}^{N} y_{ij}(\theta + \alpha_i + \beta_j) + \sum_{i<j} y_{ij} * y_{ji} * \rho,$$

if we ignore $C(\lambda) = \sum_{i,j} \lambda_{ij}$, then this is an exponential family log-likeilhood and we can find the MLE $\hat{\theta}, \hat{\alpha}_i, \hat{\beta}_i, \hat{\rho}$ with standard approaches. However this is not really accurate — the λ_{ij} are also unknown. However, they are not free parameters, rather complicated functions of the other parameters that violate the exponential family assumption:

$$\lambda_{ij} = -log\left(1 + \exp(\alpha_i + \beta_j + \theta) + \exp(\alpha_j + \beta_i + \theta) + \exp(\alpha_i + \alpha_j + \beta_i + \beta_j + \rho + 2\theta)\right).$$

Note also that α, β, θ are not identifiable in this setting, since we can take either all α, all β or θ and add and subtract constants that sum to zero with no change in the model.

To obtain a proper maximum likelihood solution we can call on more complex optimization approaches, specifically Markov Chain Monte Carlo (MCMC) that we may not have time to discuss in this course that seek a good combination of the parameters for the full likelihood. A more mainstream statistical approach is to use pseudo-likelihood. In this important family of approaches, we define a function that we can optimize and is "similar" to the likelihood but simplified. The main idea here is that if I am given Y_{ji} then the likelihood of Y_{ij} is simple and has a logistic form that does not depend on the $\lambda's$ which cancel out:

$$\begin{aligned}
\mathbb{P}(Y_{ij} = 1 | Y_{ji}) &= \frac{\mathbb{P}(Y_{ij} = 1, Y_{ji})}{\mathbb{P}(Y_{ij} = 1, Y_{ji}) + \mathbb{P}(Y_{ij} = 0, Y_{ji})} = \\
&\frac{\exp(\lambda_{ij} + \theta + \alpha_i + \beta_j + Y_{ji}(\theta + \alpha_j + \beta_i + \rho))}{\exp(\lambda_{ij} + \theta + \alpha_i + \beta_j + Y_{ji}(\theta + \alpha_j + \beta_i + \rho)) + \exp(\lambda_{ij} + Y_{ji}(\theta + \alpha_j + \beta_i))} = \\
&\frac{\exp(\theta + \alpha_i + \beta_j + Y_{ji}\rho)}{\exp(\theta + \alpha_i + \beta_j + Y_{ji}\rho) + 1},
\end{aligned}$$

a regular logistic regression:

$$\text{logit}\left(\mathbb{P}(Y_{ij} = 1 | Y_{ji})\right) = \theta + \alpha_i + \beta_j + Y_{ji}\rho,$$

with the linear constraints $\sum_i \alpha_i = \sum_j \beta_j = 0$ for identifiability, which are not a problem (also appear in regular logistic regression with intercept).

So now we have a standard logistic regression model as our maximum pseudo-likelihood solution, and we can also apply the regular logistic regression inference: significance on the parameters, F-tests for model selection, using of AIC and model selection criteria, etc. However, we should keep in mind that there are major problems here:

- We are not doing maximum likelihood, but maximum pseudo-likelihood, so by definition there is no guarantee that the theory on which ML inference is based is relevant. The help for the pstar function we are using even includes a warning:

  > Estimation of p^* models by maximum pseudo-likelihood is now known to be a dangerous practice. Use at your own risk.

- Even if we accept the PML approximation, note that we have $2N + 2$ parameters and N^2 observations (edges) in a naive view, but typically the number of actual edges is more likely $O(N)$, in which case the ML asymptotics, which are for number of parameters fixed, number of observations diverging, is not relevant anyway.

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:

- $S_k(y),\ k = 1, \ldots, N-1$ — 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 $u_1(y), \ldots, u_k(y)$ with corresponding parameters $\theta_1, \ldots, \theta_k$ 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.