Exponential Random Graph Models

Karl Potisepp, Raivo Kolde
University of Tartu, Spring 2010
Why model social networks?

- Social behavior is complex - by using stochastic models we can introduce variability not easily modeled in detail.
- Statistical models allow inferences about the occurrence of certain network substructures.
- The more complex the network data structure, the more useful properly formulated models can be in achieving efficient representation.
- With good locally-specified models, it may be possible to traverse the micro-macro gap between localized social processes and global network patterns.
- **Simple answer:** modeling helps to understand how social networks are formed.
The logic behind exponential graph models I

- *observed network* - network data the researcher has collected and is interested in modeling. It is regarded as one realization from a set of possible networks with similar important characteristics.
- In general we do not know what process generated this network.
- Our goal in formulating a model is to propose a plausible and theoretically principled hypothesis for this process.
The logic behind exponential graph models II

- Taking a look at the structural characteristics in the graph helps to shape the form of the model.
- A *parameter* is an index representing the presence of some tendency in a graph.
- Example: since reciprocity of ties is a commonly observed feature in friendship networks, a good model is likely to imply that networks with reciprocation are more common than networks without reciprocation.

Reciprocation:

\[
\Pr(A \to B | B \to A) > \Pr(\neg(A \to B) | B \to A)
\]
We represent networks as graphs of nodes and edges. The range of possible networks and their probability of occurrence is represented by a probability distribution on the set of all possible graphs with this number of nodes (i.e. all adjacency matrices). At the outset we do not know which parameter values to use in assigning probabilities to graphs in the distribution. Our goal is to find the best values for parameters using the observed network as guide.
Once we have defined a probability distribution, we can draw graphs at random from the distribution according to their assigned probabilities, and compare the sampled graphs with the observed network on any other characteristic of interest. If the sampled graphs turn out to be similar to the observed network, we can then hypothesize that the modeled structural effects could explain how the network came to be.
Example - friendship in a classroom

- Observed network - a network for which we have measured friendship relations.
- We examine the observed friendship structure in the classroom in the context of all possible network structures in the classroom.
- We define a model and a way to compare networks generated using our model with the observed network.
- If the two networks are similar [enough], we have chosen the correct parameters.
- Knowing the correct parameters is good.
Constructing a model

1. Each network tie is regarded as a random variable.
2. A dependence hypothesis is proposed, defining contingencies among the network variables.
3. The dependence hypothesis implies a particular form to the model.
4. Simplification of parameters through homogeneity or other constraints.
5. Estimate and interpret model parameters.
General form of the model

$$\Pr(Y = y) = \frac{1}{k} \exp \{ \Sigma_A \eta_A g_A(y) \}$$

- summation over all configurations $A$.
- $\eta_A$ is the parameter corresponding to $A$ (and is non-zero only if all pairs of variables in $A$ are assumed to be conditionally dependent).
- $g_A(y)$ is the network statistic corresponding to $A$.
- $k$ is a normalizing quantity which ensures a proper probability distribution.

A configuration is a set of possible edges among a subset of the nodes in the graph.
Meaning of the equation

All exponential random graph models are in the form of the equation which describes a general probability distribution of graphs on \( n \) nodes.

The probability of observing any particular graph \( y \) in this distribution is given by the equation.

This probability is dependent both on the statistics \( g_A(y) \) in the network \( y \) and on the various non-zero parameters \( \eta_A \) for all configurations \( A \) in the model.
How it works
How it really works - Bernoulli graphs I

- We assume all edges are independent.
- Since the only configurations relevant to the model are those in which all possible ties are conditionally dependent on each other, the only possible configurations in this case are single edges.
- Based on the general form:
  \[
  \Pr(Y = y) = \frac{1}{k} \exp\{\sum_{i,j} \eta_{i,j} y_{i,j}\}
  \]
- There is a parameter \( \eta \) for each edge - so we impose a homogeneity assumption \( \eta_{i,j} = \Theta \) for all \( i \) and \( j \).
How it really works - Bernoulli graphs II

The resulting equation:

$$\Pr(Y = y) = \frac{1}{k} \exp\{\Theta L(y)\},$$

where $L(y) = \sum_{i,j} y_{i,j}$ is the number of arcs in the graph $y$.

- The parameter $\Theta$ is called the edge or density parameter.
Dependence structures with node-level variables

- In social networks, ties sometimes form due to the attributes of the nodes ([witty example]).
- We assume a vector $X$ of binary attribute variables $X_i = 1$, if actor $i$ has the attribute and otherwise $X_i = 0$.
- We can investigate a similarity or homophily hypothesis as a basis for social selection by looking at the distribution of ties given the distribution of attributes.
- In other words, our interest is in the probability of graph $y$ given the observations of attributes $x$, that is, $\Pr(X=x \mid Y=y)$. 
Estimation

To tweak the parameters and measure the accuracy of a model, it is easiest to use Markov Chain Monte Carlo maximum likelihood estimation.

Generating graphs from an exponential model is relatively straightforward (for example using the Metropolis-Hastings algorithm).

Estimation is based on the following central approach:

1. Simulate a distribution of graphs from a starting set of parameter values.
2. Refine parameter values by comparing the distribution of graphs against the observed graph.
3. Repeat the process until parameter estimates stabilize.
Thank you.