STA4513: Statistical Models of Networks Fall 2014

Lecture 1 — September 10th, 2014

Prof. Daniel M. Roy *Scribe: Alberto Camacho*

1 Overview

This lecture is an introduction to the typical setting in network data and how it is modeled. We introduce the *Latent Variable Models* and the *Stochastic Block Model*.

This lecture borrows heavily from Peter Hoff's 567 course at U. Washington, in particular, lectures 1, 2, and 17.

2 Typical Statistical Setting

In a typical scenario, we are given datasets X_1, $X_2, \ldots$, and we want to:

(a) Identify hidden structure/patterns.

(b) Predict future or missing data.

Examples of datasets: Data can be represented in matrix form. In the following examples, the rows refer to **objects** (users, days, ...) and columns refer to **attributes** of these objects.

1. $(height, weight, age)$ of students:

$$X = \begin{bmatrix} 171 & 65 & 21 \\ 182 & 70 & 19 \\ 170 & 60 & 20 \end{bmatrix}$$

2. Daily average $(temp, humid., rainfall)$ in Toronto:

$$X = \begin{bmatrix} 21 & 40 & 1 \\ 25 & 50 & 0 \\ 22 & 20 & 0 \end{bmatrix}$$

3. Movie ratings by users:

$$X = \begin{bmatrix} 5 & 5 & 5 & 22 \\ 5 & 1 & 2 & 55 \\ 5 & 2 & 5 & 23 \end{bmatrix}$$

We notice that data can have different *nature*. E.g., it can represent:

- a *set* of measurements
- a *sequence* of measurements in time (ordered!)
- an *array* of measurements, each associated with a pair of objects

3 Network Data

Relational Data are data that include:

- a set of object labels. We *assume* that different rows correspond to different objects.
- measurements associated with pairs of object labels.

Example: Symmetric Social Network The symmetric social network is a typical example of a relational dataset.

$$\{y_{ij} | 1 \leq i < j \leq n\} = \text{binary indicators of friendship between individuals}$$

$$y_{ij} = \begin{cases} 1 \text{ if } i, j \text{ friends} \\ 0 \text{ otherwise} \end{cases}$$

4 Components of Network Data

A network dataset consists of:

- **Nodes:** set or multiple sets of object labels: objects, egos, ...
- **Measurements** associated with individual/pairs/triplets of nodes

  dyadic variables (structural variables): measurements on pair of nodes (dyal)
  nodal variables (actor attributes): measured on nodes.

The typical setting of statistics is to have nodes and nodal variables. On the other hand, triadic variables are rare. We distinguish among *standard* and *relational* data.

Standard Data consists of nodes and nodal variables.

Relational Data consists of nodes, nodal variables and **dyadic** variables.

5 Types of Relations

Networks can be characterized according to different criteria:

1. According to symmetry of edges:

   undirected networks one measurement per pair $(y_{ij}, i < j)$
   directed networks two measurements (y_{ij}, y_{ji})

Example:

- undirected: facebook, distance between cities
- directed: emails sender/recipient

2. Binary vs. valued:

   Binary (or **dichotomous**:) A binary relation takes only two values.

   Valued: A valued relation takes more than two values (that can be ordinal or categorical).

3. More dichotomies, e.g. Bipartite Relations where groups of objects are disjoint (e.g. in movie ratings row and column objects are two partitions).

6 Structure of Course

Probabilistic Symmetries lead to **models**. Network data may manifest:

- uniform structure
- exchangeable (or *partial* exchangeable)/conditional independence
- **RCE** Row/Column Exchangeable
- graph limits: $G_1, G_2, \ldots \to G$. The limit of a sequence of graphs with $|G_n| \to \infty$

While there is a whole theory in dense graphs, there is less work in sparse graphs and stationary network sampling.

7 Conditionally Uniform Model

Consider a graph $X = (X_{ij})_{\{i,j\}\in[n]}$, $[n] = \{1, \ldots, n\}$, where: $X_{ij} = \begin{cases} 1 \text{ if edge between } i, j \\ 0 \text{ otherwise} \end{cases}$

Let $T_1, T_2, \ldots$ be a collection of statistics, where:

- $T_1 = \#$ of edges
- $T_2 = \#$ of triangles
- $\ldots$

Then, conditional on $T_1(X), \ldots, T_k(X)$, the graph X is uniformly distributed.

7.1 Erdős-Rényi Model

Important special case is the Erdős-Rényi model, when we consider the statistic T_1 alone. That is, we consider only the number of edges (or edge density) in the graph, so the graph is uniquely characterized by the probability p of having an edge between two nodes.

8 Latent Variable Models

The idea is that:

- Nodes have attributes (u_i).
- Edges are independent given these attributes. That is, given u_i, u_j, then $X_{ij} \perp\!\!\!\perp X \setminus X_{ij}$.

The aim is to look for good attributes (E.g., $X_{ij} \sim F_\theta(u_i, u_j)$, where u_i and θ are attributes) that allow us to make predictions over the edges.

9 Block Model

The Block Model consists of:

- A **partition** of the nodes into classes (classes/blocks/clusters/types). The partition is represented by a classification function $c : [n] \to [K]$, where K is the number of blocks.
- An **estimation** of the rate of ties X_{ij} between & within blocks. Between-group **density matrix**:

$$\Theta = \begin{bmatrix} \theta_{11} & \cdots & \theta_{1k} \\ \vdots & \ddots & \vdots \\ \theta_{k1} & \cdots & \theta_{kk} \end{bmatrix}$$

  where $\theta_{c_i,c_j} = Pr(X_{ij} = 1)$.

Edges are independent in this model and nodes within the same block are *stochastically equivalent.*

Definition 1. *Two nodes i, k are stochastically equivalent if, $\forall j$, $Pr(X_{ij} = 1) = Pr(X_{kj} = 1)$.*

In this case, $c_i = c_k$ and $\theta_{c_i,c_j} = \theta_{c_k,c_j}$.

Question: How would we identify stoc. equiv. nodes?

10 Stochastic Block Model

We want to estimate:

- class function (c_i)
- between-class rates

Under SBM:

$$\begin{aligned} Pr(X = x|c, \theta) &= \Pi_{i \neq j} \theta_{c_i, c_j}^{y_{ij}} (1 - \theta_{c_i, c_j})^{1 - y_{ij}} \\ &= \Pi_{k,l \in [K]} \theta_{kl}^{s_{kl}} (1 - \theta_{kl})^{n_{kl} - s_{kl}} \end{aligned}$$

where:

$$\begin{aligned} n_{kl} &= \#\text{ pairs } i, j \text{ s.t. } c_i = k,\ c_j = l \\ s_{kl} &= \#\text{ pairs } i, j \text{ s.t. } c_i = k,\ c_j = l \text{ and } X_{ij} = 1 \end{aligned}$$

11 Approximate MLE of $c_1, \ldots, c_n, \theta$

Estimate graph model with:

- EM
- Gibbs sampling/MLMC

We can order classes and (more easily, less costly) see relations & structure.

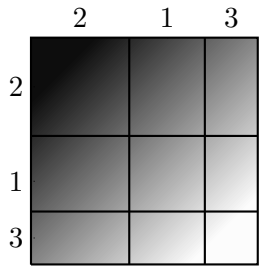

Figure 1: Block Model

Example: In the block model of figure 1, we see $\theta_{12} < \theta_{22}$.

The expected number of edges in block ij is $\binom{n_{ij}}{2}\theta_{ij}$

In principle, more classes imply that θ_{ij} is less noisy.

12 Novicki & Snijders (2001)

The Stochastic Block Model has been extended in a number of directions:

- covariates
- directed data (X_{ij}, X_{ji})
- valued data
- unknown K (number of blocks)
- hierarchical block models

Ordinal Data:

$$Z_{ij} = \theta_{c_i,c_j} + \epsilon_{ij}$$
$$X_{ij} = g(Z_{ij}) = \begin{cases} y_1 \text{ if } z_{ij} \in [c_1, c_2) \\ \vdots \\ y_m \text{ if } z_{ij} \in [c_m, c_{m+1}) \end{cases}$$

Directional Data:

$$p(X_{ij}, X_{ji}|\mu, \gamma, a_i, b_i) \propto e^{\mu_{ij}x_{ij} + \mu_{ji}x_{ji} + \gamma x_{ij}x_{ji}}$$
$$p_i \equiv \mu_{ij} = \mu + a_i + b_j$$
$$\text{[NS01] } \mu_{ij} = \theta_{c_i,c_j}$$

13 Matrix Form of Block Model

$\mathbb{E}X_{ij} = \theta_{c_i,c_j}$ (c_i, c_j unobserved)

Θ matrix can be expressed as $\theta_{c_i,c_j} = u_i^T \Theta u_j$, where:

- $u : K \times 1$ vector, all zero except $u[c_i] = 1$
- $\Theta : K \times K$ matrix of rates

13.1 Limitations of SBMs

stochastic equivalence nodes in the same class are stoc. equiv. In reality, don't expect any two nodes to act identically.

sender/receiver equivalence latent variable (c_i's) are the same regardless of direction of interaction.

14 Latent Factor Models

Generalize from $\theta_{c_i,c_j} = u_i^T \Theta u_j$.

$\mathbb{E}X_{ij} = u_i^T D v_j$, where:

- u_i is a vector of latent *factors* describing i as a sender.
- v_j is a vector of latent *factors* describing i as a receiver.
- D is a diagonal matrix.

Continuous, binary, ordinal data:

$$Z_{ij} = u_i^T D v_j + \epsilon_{ij}$$
$$X_{ij} = g(Z_{ij})$$

15 Understanding Latent Factors

We can write $Z = U^T DV + E$, where $U, V \in \mathbb{R}^{k \times n}, D \in \mathbb{R}^{k \times k}$.

Then, $Z_{ij} = \Sigma_{k \in [K]} d_k u_{ik} v_{jk} + \epsilon_{ij}$

Interpretation:

- $u_i \approx u_j \Rightarrow$ approx. stoc. equiv. as senders.
- $u_i \approx v_j, D > 0 \Rightarrow$ sender i approx. receiv. j.

Matrix Decomposition Interpretation:

1. every $m \times n$ matrix Z can be written $Z = UDV^T$, where
   - $D = diag(d_1, \ldots, d_{mn}$ $(d_1 > d_2 > \ldots)$.
   - U, V orthonormal
   - $U \in \mathbb{R}^{n \times m}, V \in \mathbb{R}^{m \times n}, D \in \mathbb{R}^{m \times m}$
2. if UDV^T is the SVD of Z, then
   $\hat{Z} = U_{[1:k]} D_{[1:k,1:k]} V_{[1:k]}^T$ least square rank-k approx of Z.
3. Z is symmetric
4. $Z = U \bigwedge U^T + \epsilon_{ij}$
   eigenmodel (Hoff) $\bigwedge = diag(\lambda_1, \ldots, \lambda_k)$.
5. $\lambda_r > 0$: homophily.
6. $\lambda_r < 0$: anti-homophily.