STAT 538
Lecture 10
Problems in network modeling

©Marina Meilă
mmp@stat.washington.edu

1 Some definitions

The definitions and notations are the same as in Lecture 3.1.

2 Network problems

- **Connectivity** A large network is almost surely disconnected. We are interested in large subsets of nodes that are disconnected from the rest. Moreover, if between two large subsets of nodes only a few edges exist, for all interesting purposes these two sets are still disconnected. So "connectivity" really means: can we cut the graph into two parts of comparable sizes, that have very few edges crossing between them?

  This is exactly what the Normalized Cut criterion measures.

- **Finding communities** Amounts to graph clustering. Studied in computer science, social sciences, statistics, mathematics. Area where most statistical models have been developed.

  Finding communities is conceptually similar to finding "connected components" above. The quality measure for a community is called **conductance** and is related to the Normalized Cut.

  $$\phi(S) \;=\; \frac{Cut(S, V \setminus S)}{\min(Vol(S), Vol(V \setminus S))} \tag{1}$$

  Recent studies on real network show that community sizes do not grow in proportion to graph size! Hence realistic models have $K \to \infty$ when $n \to \infty$.

- **Centrality, prestige, and authority** The goal is to give each node a score that represents its prestige or social importance. For example

- authority of sources of information (like in PageRank or HITS) on the internet
- impact in citation networks
- influence, i.e. capacity of influencing others, or of attracting followers, in social networks

Various scores have been developed to quantify the above

(well understood measures)

- node degree (number of neighbors)
- eigenvector centrality

(not so well understood, may behave in unpredictable ways)

- closeness centrality

$$C_C(i) = \frac{n-1}{\sum_j d(i,j)} \tag{2}$$

- betweeness centrality

$$C_B(i) = \sum_{j,k} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \tag{3}$$

where $d(i,j)$ is the graph distance and $\sigma_{jk}(i)$ is the number of shortest paths between j,k that pass through i.

- ...and many more

- **Semisupervised learning** We want to estimate a function $y(i)$, $i \in V$ on the graph. For some nodes $i \in S$, y is observed; in other words these nodes are labeled, while the remaining nodes in $V \setminus V$ are unlabeled. This problem is similar to supervised learning, with the difference that we know for which future data we need to predict y.

- **Visualization**

3 Models for networks

- Erdos-Renyi (the null model)
- $p1$ and $p2$ models (GLM models)

- SBM (Stochastic Block Model)
- ERGM

- Latent space model
- Mixed membership SBM
- Multiplicative attributes model
- Graphons

SBM, MMSBM, and MAGM models communities explicitly.

4 Mixed membership SBM (MMSBM)

This is a generative hierarchical model, that can represent directed as well as undirected graphs.

- Parameters K number of clusters, α prior parameter for membership distribution, $B = [B_{kk'}]_{k,k'=1}^{K}$ matrix of edge probabilities, conditioned on cluster membership.
- for each node $i \in V$
  - draw its mixed membership, a distribution π_i over clusters $1 : K$

$$\pi_i \sim \text{Dirichlet}(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}) \tag{4}$$

- for each pair of nodes $i, j \in V$
  - draw current membership of i, $z_{i \to j} \sim \pi_i$
  - draw current membership of j, $z_{i \leftarrow j} \sim \pi_j$ (the notation $z_{i \leftarrow j}$ emphasises that the membership is in the cluster only for this particular edge, and another membership will be sampled for the next edge incident to j).
  - sample edge $Y_{ij} \sim \text{Bernoulli}(B_{z_{i \to j}, z_{i \leftarrow j}})$

– This sampling can generalize easily to directed or weighted graphs

The likelihood

$$p(Y, \pi_{1:n}, z_{i\to j}, z_{i\leftarrow j}|\alpha, B) \;=\; \prod_{i,j} P(Y_{ij}|z_{i\leftarrow j}, z_{i\to j}, B)P(z_{i\to j}|\pi_i)P(z_{i\leftarrow j}|\pi_j)\prod_{i\in V} p(\pi_i|\alpha) \tag{5}$$

Estimation of this model is done by MCMC Gibbs sampling.

5 Latent Space Model

- Observed graph $Y = [y_{ij}]_{i,j\in V}$, and optionally pairwise covariates $x_{ij} \in \mathbb{R}^d$
- Assumed each node i has unobserved location $z_i \in \mathbb{R}^k$ for some k fixed
- Wanted estimate node locations in **latent space**
- **Model**

$$\log \text{odds}\, y_{ij}|x_{ij}, z_i, z_j, \alpha, \beta \;=\; \eta_{ij} \;=\; \alpha + \beta^T x_{ij} - d_{ij} \quad \text{with } d_{ij} = ||z_i - z_j|| \tag{6}$$

Alternate model (that can also handle directed graphs) "Projection model"

$$\eta_{ij} \;=\; \alpha + \beta^T x_{ij} + |z_i| \cos(z_i, z_j) \tag{7}$$

Estimation (assume no x for simplicity)

1. Estimate α, $[d_{ij}]$ by convex minimization Note first that the log-likelihood

   $$l \;=\; \sum_{i<j} y_{ij}\eta_{ij} - \ln(1 + e^{\eta_{ij}}) \tag{8}$$

   is concave in η_{ij} (trivially) and since $\eta = \alpha \mathbf{1}\mathbf{1}^T - D$ (with $D = [d_{ij}]$), l is concave in (α, D).

2. Alternatively, use the **graph distance** (a.k.a. **shortest path distance**) to obtain initial estimates of $[d_{ij}]$.

3. Embedding (by MDS) Finding a set of k-dimensional z_i variables that achive the estimated distances is non-trivial (read "Hard"). Multidimensional scaling finds a minimum distortion embedding by a SVD-like method. For $[d_{ij}]$ graph distances this problem has been intensely studied in computer science.

4. Relax Local maximization of l w.r.t Z starting from the provious Z's. This estimate is denoted Z^{ML}.

5. Bayesian estimation Define priors for $\alpha, (\beta), Z$. Run Metropolis-Hastings starting from Z^{ML}, with e.g Gaussian proposal distribution around Z.

6. Include Procrustean transform The positions z are identifiable up to translation and rotation/reflection. Hence, in the MH algorithm, once a new Z^{t+1} is accepted, it is transformed by

$$Z^{t+1} \leftarrow \text{Procrustes}_{Z^t} Z^{t+1} \quad \text{with} \quad \text{Procrustes}_Z Z_0 = Z_0 Z^T (Z Z_0 Z_0^T Z^T)^{-1} Z \tag{9}$$

   (Note that Procrustes is a projection operator which find the element in the equivalence class of Z that is closest to Z_0 in Frobenius norm.)

   For the "cosine" version of the model, the projection operator is w.r.t the equivalence class of rotation/reflection and scaling.

6 Multiplicative Attributes Graph Model (MAGM)

The MAGM assumes that each node has a vector of binary attributes $a(i) \in \{0,1\}^K$, $i \in V$. Edges are sampled independently, with probability

$$P_{ij} = \prod_{k=1}^{K} \Theta^{(k)}_{a_k(i)a_k(j)} \tag{10}$$

where $\Theta^{(k)}$ is a 2×2 matrix of parameters associated to attribute k. The probability of sampling $a_k = 1$ is μ_k (independent of other i's). Hence, a MAGM is defined by the parameters (n, K, Θ, μ).

If one interprets $a_k(i) = 1$ as "i belongs to community K", the MAGM allows one to represent "overlapping communities". The simplified MAGM assumes that all K atrributes share the same parameters μ and Θ. For simplicity, one denotes $\alpha = \theta_{11}, \beta = \theta_{10} = \theta_{01}, \gamma = \theta_{00}$. A few possibilities offered by this parameterization are

- $\alpha, \gamma > \beta$: homophily, nodes prefer to attach to similar nodes
- $\alpha, \gamma < \beta$: heterophily, nodes prefer to attach to dissimilar nodes
- $\alpha > \beta > \gamma$: core-periphery, nodes prefer to attach to the core nodes

Comparing this parametrization with large real networks suggests that these are modeled well by the third model, with $\alpha \approx 0.99 > \beta \approx 0.5 + \epsilon > \gamma \approx \epsilon$ (where $\epsilon < 0.1$).

Since the probability of an edge given $i, a(i)$

$$Pr[Y_{ij} = 1 | a(i)] \; = \; (\mu\alpha + (1-\mu)\beta)^{|a(i)|}(\mu\beta + (1-\mu)\gamma)^{K-|a(i)|} \tag{11}$$

depends exponentially on K, a natural choice is to set

$$\textbf{Assumption 1 } K = \rho \log n. \tag{12}$$

This model can be seen as an averaged, unnormalized (and non-Bayesian) version of the MMSBM. The MAGM was extended to a generative model that incorporates distributions over visible and latent node attributes ($a(i)$ being the latent ones), called LMMG (Laten Multigroup Membership Graph model). The estimation is done by alternate (approximate) maximization of the likelihood (in a way reminiscent of EM); LMMG can potentially be estimated in the Bayesian framework in a way similar to the MMSBM.

By removing the normalization, and chosing appropriate Θ values, the probability of edge ij is guaranteed to increase when the two nodes i, j belong the the intersection of more commmunities.

Under the simple model, with Assumption 1, it can be shown that

- The diameter of the graph is bounded by a constant if $n \to \infty$ if $(\mu\beta + (1-\mu)\gamma)^{\rho} > \frac{1}{2}$
- The graph contains a unique **giant component** with high probability as $n \to \infty$ iff

  $$(\mu\alpha + (1-\mu)\beta)^{\rho\mu}(\mu\beta + (1-\mu)\gamma)^{\rho(1-\mu)} \geq \frac{1}{2} \tag{13}$$

  A giant component is a connected component of $\mathcal{G}$ which contains at least a constant fraction of the nodes (intuitively, “almost all” the nodes).