36-780: Social Network Modeling

Latent space and stochastic block models

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Outline

- Latent Space Models
  - Motivations and a simple formulation
  - An illustration
  - Some generalizations
- Stochastic Block Models
  - Motivations and a simple formulation
  - An illustration
  - Some generalizations
- HW04 is posted (reading and some R examples)

Latent Space Models (LSMs)

- A fairly recent invention (Hoff et al 2002).
- Mathematically similar to multidimensional scaling (MDS): find a lower-dimensional spatial representation of (complex) network data
  - Replace adjacency matrix (of order n^2) with a set of latent node attributes (of order n) – e.g. locations in Euclidean space
- Because distance is inherently symmetric, LSMs tend to work better for symmetric relations
  - One can add "fixed effects" to model directionality

LSMs – a basic model

- For adjacency matrix Y = [Y_{ij}] ,

$$\text{logit}\, P[Y_{ij} = 1] = \beta^{\mathsf{T}} X_{ij} - \text{dist}(Z_i, Z_j)$$

- Y_{ij} indep given Z's; the Z_i, Z_j are positions in a low-dimensional Euclidean space R^d, typically d=2 or 3
  - d=1 imposes constraints;
  - d >> 3 introduces computational and data problems
- X_{ij} can encode edge covariates (as in QAP or ERGM), or dyadic-independence network statistics

LSM - interpretation

$$\text{logit}\, P[Y_{ij} = 1] = \beta^{\mathsf{T}} X_{ij} - \text{dist}(Z_i, Z_j)$$

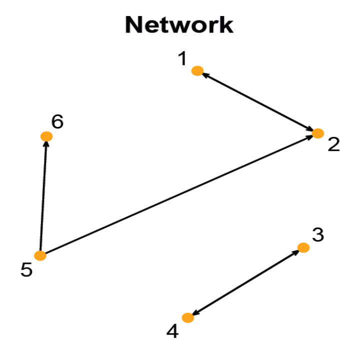
- The closer Z_i to Z_j, the more likely a tie (both ways)
  - A kind of "latent homophily" model
    - Like a latent proximity/unfolding model for choice
  - Z_i's can account for dependence between links not modeled by βX_{ij}
  - Z_i's can also generate a greater preponderance of triangles, stars, etc., than a p_1 model would
- X_{ij} can make $P[Y_{ij}=1] \neq P[Y_{ji}=1]$ (like p_1 model), so we can get directed graphs

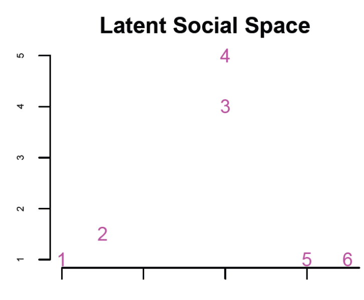
LSM – a generative view

- Account for affinity between actors that may not be group based

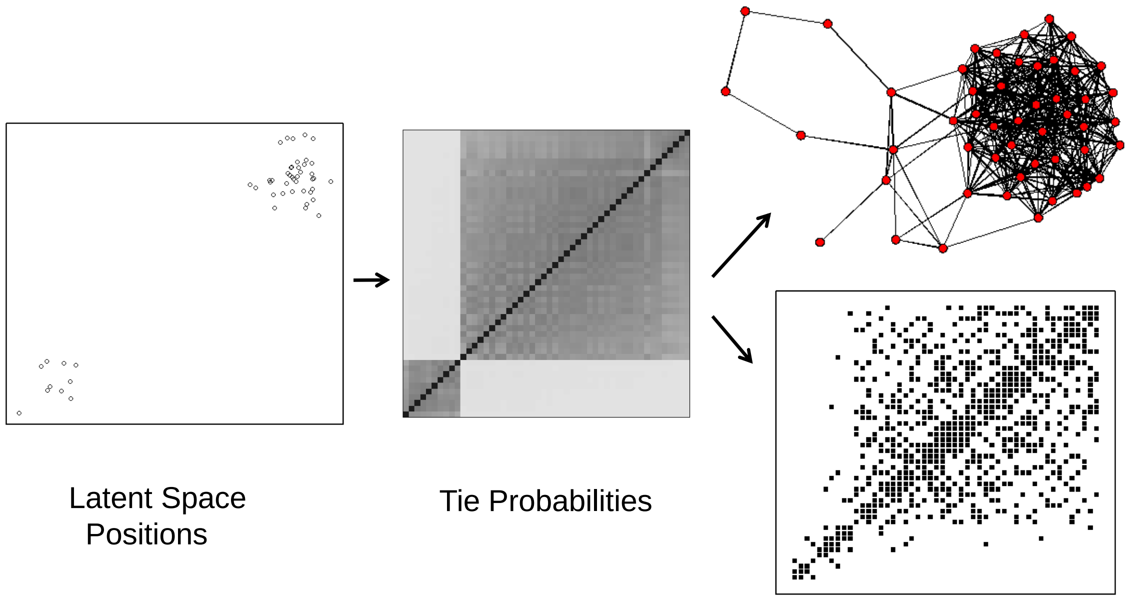
$$\text{logit}\, P[Y_{ij} = 1] = \beta^{\mathsf{T}} X_{ij} - \text{dist}\,(Z_i, Z_j)$$

- X_{ij} = edge covariates
- Z_i = latent space position for actor i





LSM – a larger generative example



Latent Space Positions

Tie Probabilities

Realized Network

What else to set up the LSM?

- Distribution for the Z_i's $\in R^d$
  - Convenient simple case:

$$Z_i \stackrel{iid}{\sim} N_d(\vec{\mu}, \Sigma)$$

  - Can incorporate model-based clustering as well:

$$Z_i \stackrel{iid}{\sim} \sum_{k=1}^{K} N_d(\vec{\mu}_k, \Sigma_k)$$

    (e.g. Fraley & Raftery, 2002, JASA)
- Any other tractable distribution for Z is fine also.

LSMs - Estimation

- Maximum likelihood (E-M) is plausible
  - **E-step** – impute Z's; **M-step** – estimate β's, μ, Σ lather, rinse, repeat…
- Bayesian MCMC fairly workable also
- Identifiability issues
  - The "closeness" dist(Zi, Z_j) will not be affected by the orientation of the ensemble of Z's in the latent space
  - Hoff (2002) imposes a "procrustean" constraint based on preliminary estimation of the Z_i's
  - Sweet (2013) uses a simple orientation heuristic on the first d latent locations Z_1, …. Z_d.

LSM generalizations – valued edges

- The logit LSM is clearly an instance of a generalized linear model (glm):

$$
\begin{aligned}
Y_{ij} &\sim p(y|E[Y_{ij}],\ldots) \\
E[Y_{ij}] &= g^{-1}(\eta_{ij}) \\
\eta_{ij} &= \beta^{\mathsf{T}} X_{ij} - |Z_i - Z_j|
\end{aligned}
$$

for example:

Data Y_{ij}	Distribution $p(y\|u)$	Link $g(u)$	
$Y_{ij} \in \{0,1\}$	Bernoulli	$\log u/(1-u)$	[logit]
$Y_{ij} \in \{0,1\}$	Bernoulli	$\Phi^{-1}(u)$	[probit]
$Y_{ij} \in \{0,1,2,\ldots\}$	Poisson	$\log u$	
$Y_{ij} \in \Re$	Normal	g(u) = u	

LSM generalizations – distance functions

- So far we have taken dist(Z_i, Z_j) = |Z_i – Z_j|, Euclidian distance
- Certainly d(Z_i, Z_i) in the linear predictor

$$\eta_{ij} = \beta^{\mathsf{T}} X_{ij} - d(Z_i, Z_j)$$

  need not be a distance, or even symmetric, e.g.
  - Hoff (2007) considers symmetric quadratic forms d(Z_i, Z_j) = $Z_i^{\mathsf{T}} \Lambda\, Z_j$ where Λ is a diagonal matrix
  - Why not d(Z_i, Z_j) = Zi^{T} A Z_j where A is nonsymmetric?
  - Other nonsymmetric s(Z_i, Z_j)?

Stochastic Block Models (SBMs)

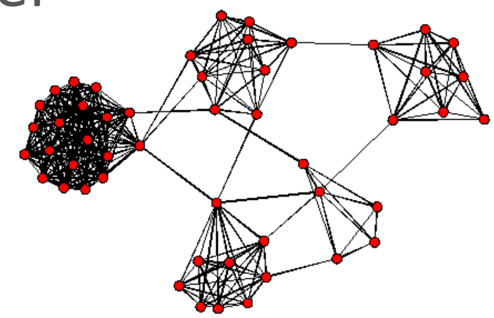
- The idea is to group the vertices into latent classes showing community structure
  - Typically all nodes in a cluster have similar affinities for other nodes in the network (*stochastic equivalence*)
- This is a much older method for identifying community structure in networks
  - Deterministic block-finding, like in the first few “sna” labs we did, goes back to Lorrain & White (1971)
  - Stochastic block models go back to Fienberg & Wasserman (1981), Holland, Laskey & Leinhardt (1983)

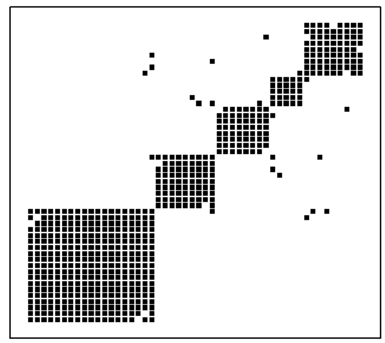
SBMs – a basic model

- B[k,ℓ] = P[tie from member of block k to member of block ℓ]

$$B = \begin{bmatrix} 0.99 & 0.05 & 0.01 & 0.01 & \cdots \\ 0.05 & 0.95 & 0.02 & 0.03 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}$$

- M_i = k if node i belongs to block k
- $p_k = P[M_i = k]$
- $P[Y_{ij} = 1] = B[M_i, M_j]$





SBM – edge covariates

$$\begin{aligned} \operatorname{logit} P[Y_{ij} = 1] &= \beta^{\mathsf{T}} X_{ij} + \operatorname{logit} B(M_i, M_j) \\ &= \beta^{\mathsf{T}} X_{ij} + \operatorname{logit} S_i^{\mathsf{T}} B S_j \end{aligned}$$

- Again, X_{ij} encodes edge covariates and/or dyadic-independence network statistics
- B is a K $\times$ K matrix, for a K-block model
- S_i is a 0/1 vector with all 0’s except for a 1 in position M_i
- M_i = k if node i is in block k.
- The M_i ’s (or S_i ’s) are the latent variables here

SBMs – Estimation

- Again E-M works
  - E step: impute M's
  - M step: estimate the entries of B, and p_k= P[M_i = k].

- Again, Bayesian MCMC also works

- Identifiability: any relabelling of the blocks gives the same data predictions
  - Latent class label-switching well-understood, e.g Jasra, Holmes & Stephens, *Statistical Science* (2005).

SBM generalizations – valued edges

- Again we can consider glm's

$$\begin{aligned} Y_{ij} &\sim p(y|E[Y_{ij}],\ldots) \\ E[Y_{ij}] &= g^{-1}(\eta_{ij}) \\ \eta_{ij} &= \beta^{\mathsf{T}} X_{ij} + S_i^{\mathsf{T}} B S_j \end{aligned}$$

(For the logit link function this leads to

$$\text{logit}\, P[Y_{ij}=1] = \beta^{\mathsf{T}} X_{ij} + S_i^{\mathsf{T}} B S_j$$

rather than

$$\text{logit}\, P[Y_{ij}=1] = \beta^{\mathsf{T}} X_{ij} + \text{logit}\, S_i^{\mathsf{T}} B S_j \ \ )$$

SBMs –unifications & generalizations

- Both SBMs and LSMs are conditionally independent dyad (CID) models
  - Conditional on the latent variables, these models have all the advantages of dyadic independence models (like p_1 or the beta model)
  - Averaging over the latent variables allows flexibility to model non-dyadic structure in the networks
- Hoff (2007) uses the $Z^T \Lambda Z$ structure to unify LSMs and SBMs in a general "eigenmodel" framework
- SBMs can be generalized to allow nodes to shift block membership for different ties (mixed membership SBMs, MMSBMs – more later!)

Summary

- Latent Space Models
  - Motivations and a simple formulation
  - An illustration
  - Some generalizations
- Stochastic Block Models
  - Motivations and a simple formulation
  - An illustration
  - Some generalizations
- HW04 is posted (reading and some R examples)