The exchangeable graph model with applications to dynamic network analysis

Edo Airoldi

Computer Science Department &
Lewis-Sigler Institute for Integrative Genomics
Princeton University

Joint work with: David Blei, Kathleen Carley, Stephen Fienberg & Eric Xing

Overview

- Problem: how can we think quantitatively about social structure and social dynamics?
- Data:
  - Sampson's monastery data
  - National survey of adolescent health
  - Linked-In
- Disclaimer: do not think probability, statistical methodology or learning, rather think substantive

Key notions

- Complexity of observed connectivity is resolved in a <u>structure</u> of simple motifs and their evolution
- Mixed membership
- Dynamics
  - State-space models
  - Birth-death processes

The role of structure

- Structural hypotheses drive inference

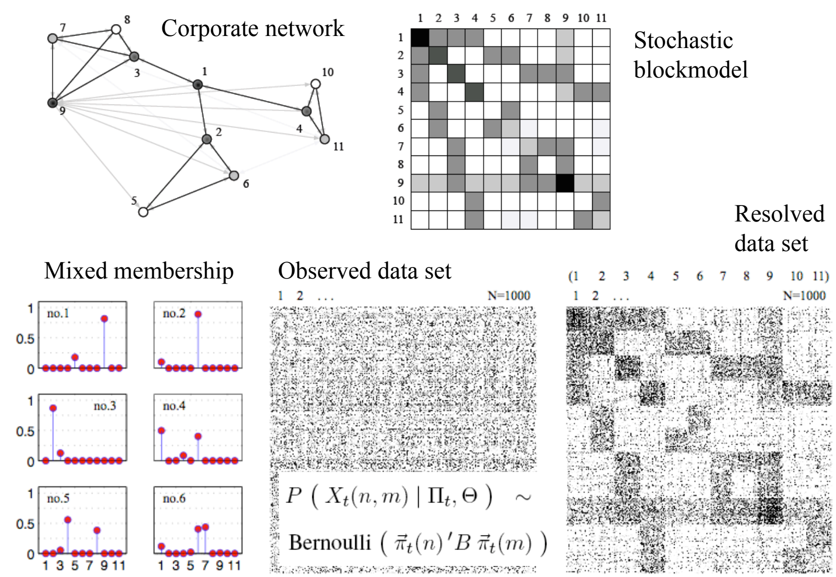
Domain Knowledge
↓
Structural Hypotheses
↓
Data

Probabilistic Generative Models

Bayesian Posterior Inference

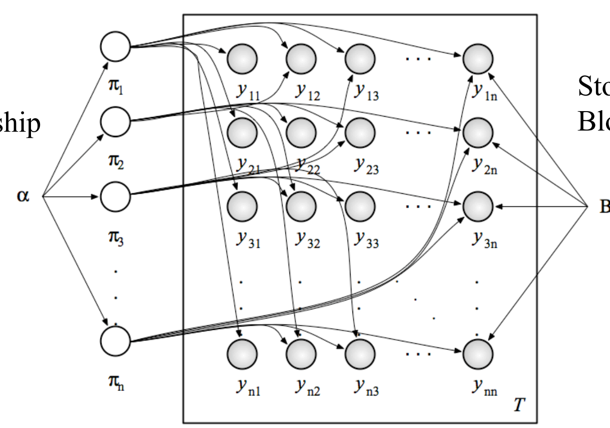
Agenda

- Static network analysis
- Methodological themes
- Dynamics of social failure
- The exchangeable edge model
- Concluding remarks



A projection onto *Π×B*

Mixed Membership

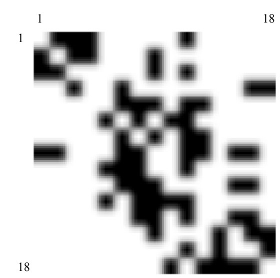


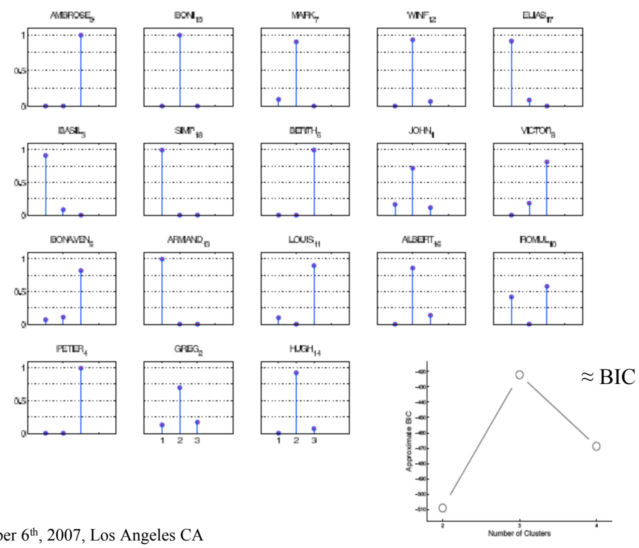
Stochastic Blockmodel

Sampson’s monastery data

- How many factions are there?
- How do factions relate to one another?
- Who belongs to which faction?

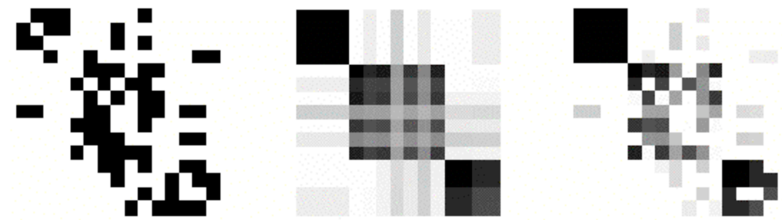
18 Novices
Edge = Like



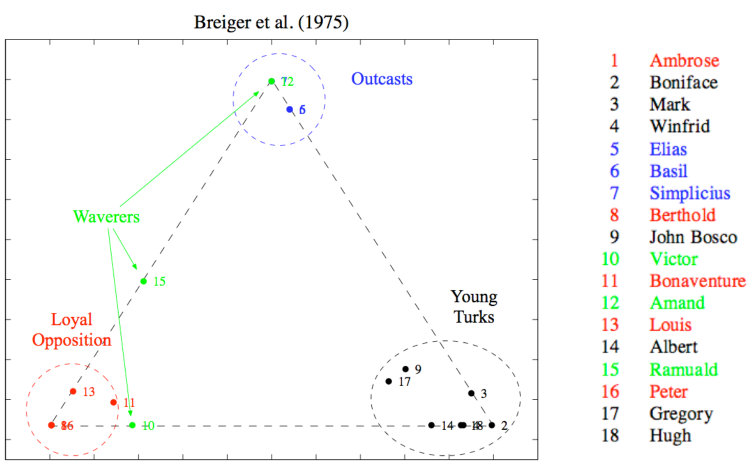


Recovering observed connectivity

- Two model variants (node-specific, relation-specific) provide increasing levels of definition



Original data | node-specific (summary) | relation-specific (de-noising)



National study on adolescents

- A friendship network among 69 students in grades 7-12

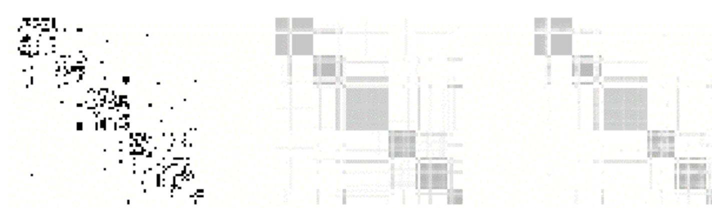


Fig. 8. Original matrix of friensdhip relations (left), and estimated relations obtained by thresholding the posterior expectations $\pi_p{}'B\,\pi_q|R$ (center), and $\phi_p{}'B\,\phi_q|R$ (right).

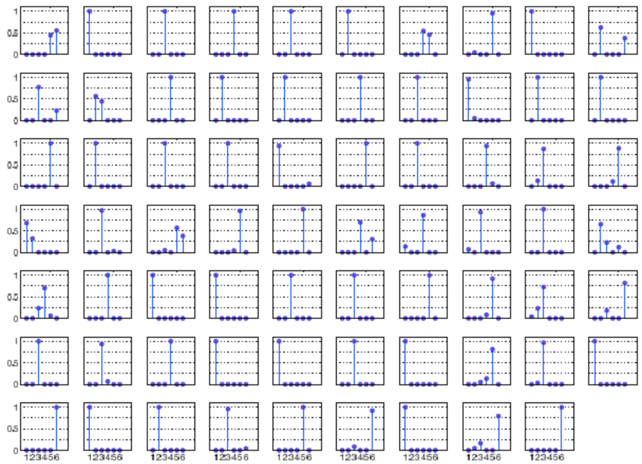


Fig. 7. The posterior mixed membership scores, π, for the 69 students in a school. Each panel correspond to a student; on the Y axis we measure the grade of membership, corresponding to the six grade levels from 7 to 12, on the X axis.

	MMSB Clusters						MSB Clusters						LSCM Clusters					
Grade	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
7	13	1	0	0	0	0	13	1	0	0	0	0	13	1	0	0	0	0
8	0	9	2	0	0	1	0	10	2	0	0	0	0	11	1	0	0	0
9	0	0	16	0	0	0	0	0	10	0	0	6	0	0	7	6	3	0
10	0	0	0	10	0	0	0	0	0	10	0	0	0	0	0	0	3	7
11	0	0	1	0	11	1	0	0	1	0	11	1	0	0	0	0	3	10
12	0	0	0	0	0	4	0	0	0	0	0	4	0	0	0	0	0	4

Table 1: Grade levels versus (highest) expected posterior membership for the 69 students, according to three alternative models. MMSB is the proposed mixed membership stochastic blockmodel, MSB is a simpler stochastic block mixture model (Doreian et al., 2007), and LSCM is the latent space cluster model (Handcock et al., 2007).

$$\hat{B} = \begin{bmatrix} 0.3235 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.3614 & 0.0002 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.2607 & 0.0 & 0.0 & 0.0002 \\ 0.0 & 0.0 & 0.0 & 0.3751 & 0.0009 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0002 & 0.3795 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.3719 \end{bmatrix}$$

Problem revisited

- Given: A collection of relational measurement on the same sets of objects (units of analysis)

  (square matrices, or unipartite graphs, with integer, real or multivariate edge weights)

- Find: (i) A pool of recurrent connectivity patterns among blocks of nodes —how many and what they look like, and (ii) A mapping of nodes to connectivity patterns — at the block level

  (PCA for relational data, with symmetry constraints)

Summary

- Observed connectivity structure is described in terms of two main sources of variability:

1. Stochastic blockmodel
   - Blocks and block-to-block connectivity patterns

     (the community structure, global, asymmetric)

2. Membership map
   - Nodes-to-blocks map

     (mixed membership, object-specific, symmetric)

Agenda

- Static network analysis
- Methodological themes
- Dynamics of social failure
- The exchangeable edge model
- Concluding remarks

- For each node $p \in \mathcal{N}$:
  - Draw a K dimensional mixed membership vector $\vec{\pi}_p \sim$ Dirichlet $(\ \vec{\alpha}\ )$.
- For each pair of nodes $(p, q) \in \mathcal{N} \times \mathcal{N}$:
  - Draw membership indicator for the initiator, $\vec{z}_{p \to q} \sim \text{Multinomial}\ (\ \vec{\pi}_p\ )$.
  - Draw membership indicator for the receiver, $\vec{z}_{q \to p} \sim \text{Multinomial}\ (\ \vec{\pi}_q\ )$.
  - Sample the value of their interaction, $R(p, q) \sim \text{Bernoulli}\ (\ \vec{z}^{\top}_{p \to q} B\ \vec{z}_{p \leftarrow q}\ )$.

$$
\begin{aligned}
&p(R, \vec{\pi}_{1:N}, Z_{\rightarrow}, Z_{\leftarrow} | \vec{\alpha}, B) \\
&\quad = \prod_{p,q} P(R(p,q) | \vec{z}_{p \to q}, \vec{z}_{p \leftarrow q}, B) P(\vec{z}_{p \to q} | \vec{\pi}_p) P(\vec{z}_{p \leftarrow q} | \vec{\pi}_q) \prod_{p} P(\vec{\pi}_p | \vec{\alpha}).
\end{aligned}
$$

Inference on mixed membership

- Define: observations $Y = R$, latent variables $X = (\Pi, Z)$, and underlying constants $\Theta = (\alpha, B)$

$$
\begin{aligned}
p(Y|\Theta) &= \log \int_{\mathcal{X}} p(Y, X|\Theta)\, dX \\
&= \log \int_{\mathcal{X}} q(X)\, \frac{p(Y, X|\Theta)}{q(X)}\, dX \qquad \text{(for any } q\text{)} \\
&\geq \int_{\mathcal{X}} q(X)\, \log \frac{p(Y, X|\Theta)}{q(X)}\, dX \qquad \text{(Jensen's)} \\
&= \mathbb{E}_q \left[\, \log p(Y, X|\Theta) - \log q(X)\, \right] \quad =: \ \mathcal{L}(q, \Theta)
\end{aligned}
$$

Variational approximation

- The idea is to maximize lower bound over (X, Θ)
- Alas, not possible to compute

$$q^{(t)} = p(X|Y, \Theta^{(t-1)}),$$

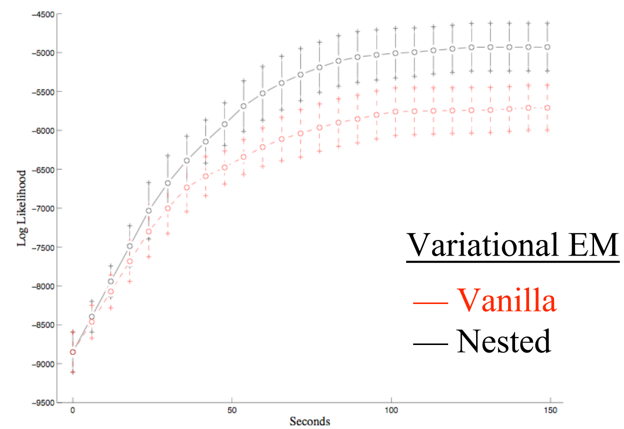
- Posit parametric approximation for q using free parameters Δ

$$q^{(t)} \approx q^{(t)}_{\Delta^*(Y)}(X) = p(X|Y).$$

Large scale computation

- Masses of data
  - 750K observations in a small problem (N=871)
  - 2.5M observations in a medium problem (N=1567)
  - Introduce parameter ρ to deal with sparsity
- Variational inference *[Jordan et al., 2001]*
  - Naïve implementation does not work
  - Develop a novel "nested" variational EM algorithm

Large scale computation



1. Stochastic blockmodel, *B*

- Captures salient structure, at the block level
  (collapse nodes into groups, or blocks)
- Node-specific connectivity patterns are instances of (multiple) block-to-block connectivity patterns
- Connectivity among nodes within the same block is only specified on average

2. Mixed membership, *Π*

- Extends the idea of a mixture
  - Mixture: variability of data *top-down*; global weights
  - MM: variability of data *bottom-up*, unit-specific weights
- Unit-specific descriptions useful for prediction
- Sparsity: to induce parsimony in the mixed membership map between nodes and patterns
  - Enforced via prior distribution, or other means

3. Allocation paradigms, *Pr(Π)*

- Alternative specifications of mixed membership lead to different interpretations
- The simplex
  - Intuition: finite resources, more constrained
- The unit hyper-cube
  - Intuition: relevance, less constrained

Agenda

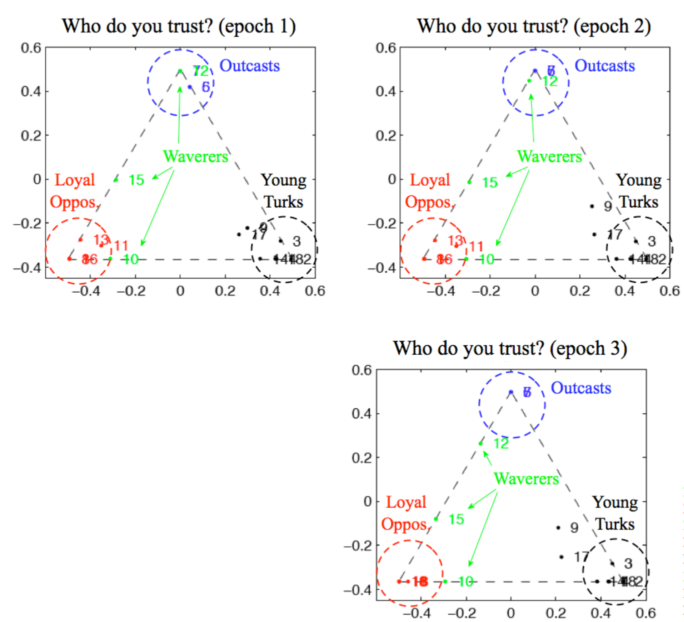
- Static network analysis
- Methodological themes
- Dynamics of social failure
- The exchangeable edge model
- Concluding remarks

Modeling social dynamics

- Mixed membership analysis reduces pair-wise measurements to node-specific attributes
- Introduce smooth temporal evolution

$$X_t(n,m) \quad \text{s.t.} \;\; n,m = 1,\ldots,N = 18 \;\text{ and }\; t = 1,2,3.$$

$$
\begin{aligned}
P\left(\, \vec{\pi}_0(n) \mid \Theta \,\right) &\sim \mathbf{f} \circ Gaussian\left(\, \vec{0}, A \,\right),\\
P\left(\, \vec{\pi}_t(n) \mid \vec{\pi}_{t-1}(n), \Theta \,\right) &\sim \mathbf{f} \circ \left[\, Gaussian\left(\, \vec{0}, A \,\right) + \mathbf{f}^{-1} \circ \vec{\pi}_{t-1}(n) \,\right],\\
P\left(\, X_t(n,m) \mid \Pi_t, \Theta \,\right) &\sim \text{Bernoulli}\left(\, \vec{\pi}_t(n)' B \, \vec{\pi}_t(m) \,\right),
\end{aligned}
$$



1	Ambrose	(9)
2	Boniface	(15)
3	Mark	(7)
4	Winfrid	(12)
5	Elias	(17)
6	Basil	(3)
7	Simplicius	(18)
8	Berthold	(6)
9	John Bosco	(1)
10	Victor	(8)
11	Bonaventure	(5)
12	Amand	(13)
13	Louis	(11)
14	Albert	(16)
15	Ramuald	(10)
16	Peter	(4)
17	Gregory	(2)
18	Hugh	(14)

Social failure in isolated communities

- Analysis suggests elements of a dynamic theory of social failure in isolated communities:

1. Fragmented social structure
2. Progressive polarization
3. Interstitial members as traitors

Agenda

- Static network analysis
- Methodological themes
- Dynamics of social failure
- The exchangeable edge model
- Concluding remarks

An abstraction exercise

- Goal: new model of randomness for graphs
- What are the essential features of our models?
  1. Node attributes
  2. Scarcity (sparsity)
  3. Latent variables

The exchangeable graph model

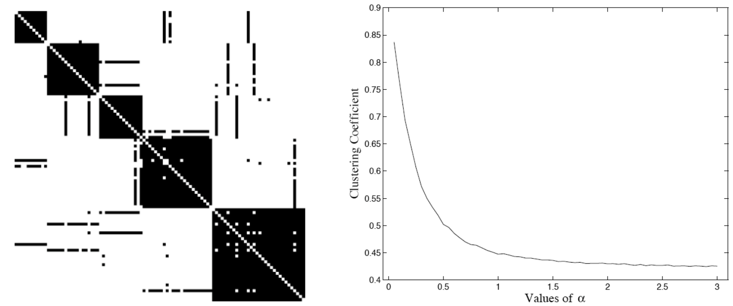
- Random graphs (Erdos-Renyi-Gilbert)
  - For all (n,m) do
    $Y(n,m) \sim \text{Bernoulli}(p)$
- Exchangeable graphs
  - For all n do
    $X_k(n) \sim \text{Bernoulli}(p), \; k = 1 \ldots K$
  - For all (n,m) do
    $Y(n,m) = f(X(n), X(m))$

Some results

- Emergence of the giant component
- Emergence of community structure
  - No phase transition
- Lognormal graphs
  - True limit connectivity
  - Scale-free graphs as approximation
- Study connectivity induced by imputing edges

Emergence of community structure

- As negative correlation* among node-specific bit strings increases, communities emerge



Agenda

- Static network analysis
- Methodological themes
- Dynamics of social failure
- The exchangeable edge model
- Concluding remarks

Related work in biology

1. Inferring protein function from interacts (patches of connectivity correspond to stable complexes)
2. Statistical discovery of signaling pathways from an ensemble of weakly informative data sources

   Data: interactions (e.g. Y2H), node attributes (e.g. microarrays, domains), path constraints (e.g. RNAi)

   Idea: signaling pathways as latent graphs

Take home points

- Mixed membership analysis as a quantitative tool for exploring static/dynamic social networks
- The exchangeable graph model as a new paradigm for theoretical explorations of graph connectivity

Manuscripts on arXiv:

1. Stochastic blockmodel: stat.ME 0705.4485
2. Exchangeable graph model: email me