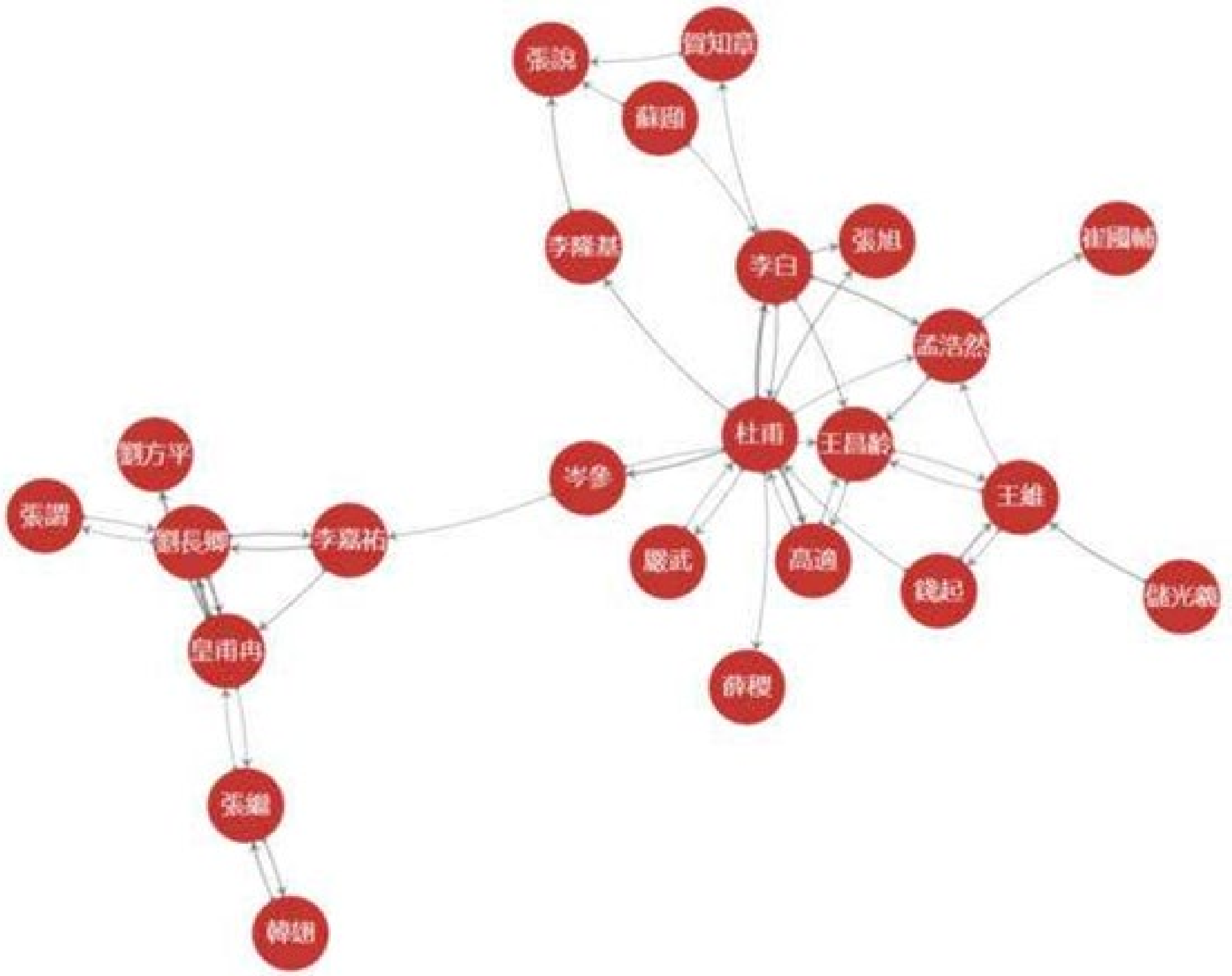
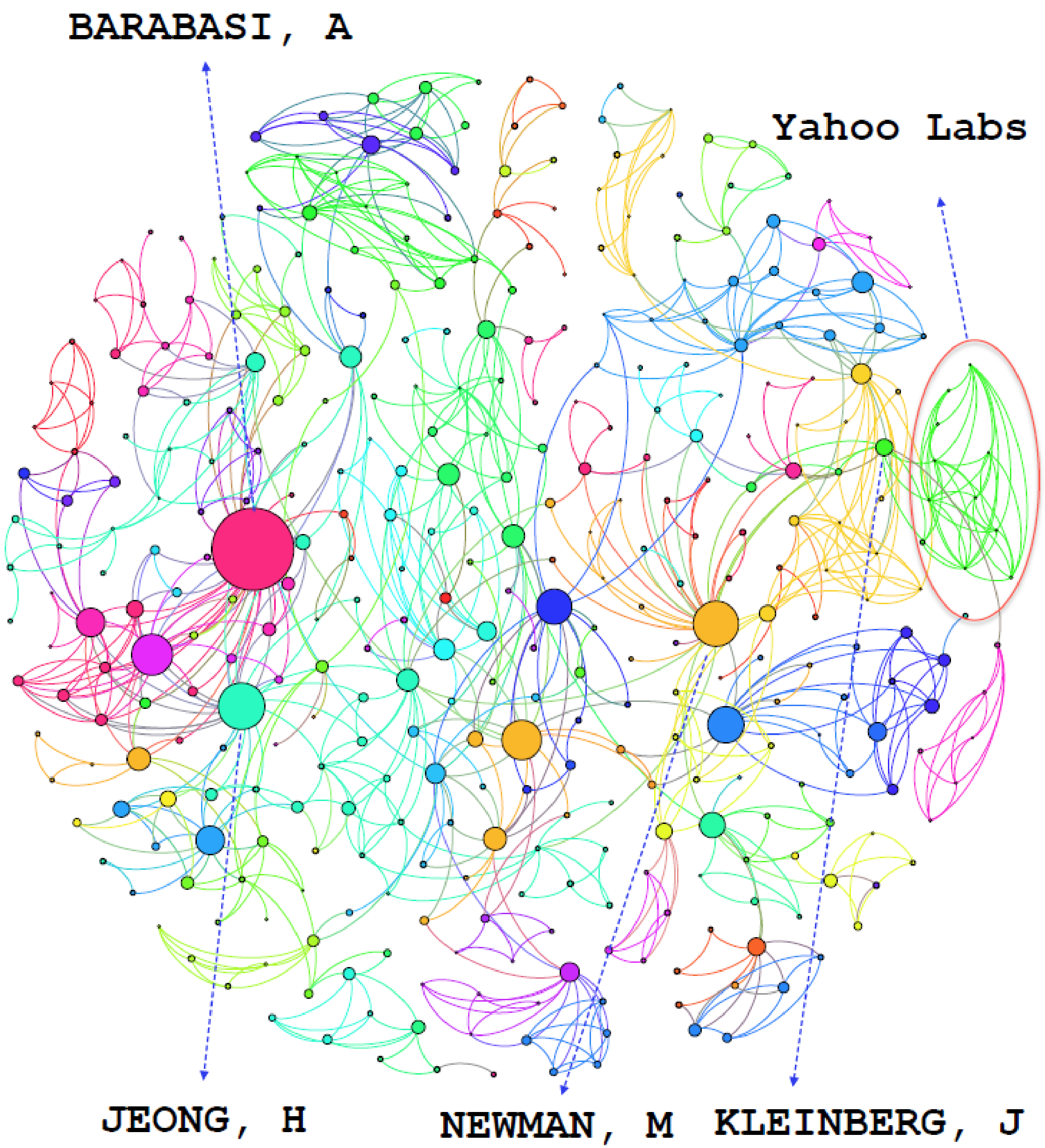
Mixed Membership Stochastic Blockmodels

JMLR2008

2019.10.9


張說
賀知章
蘇頲
李隆基
李白
張旭
崔國輔
孟浩然
杜甫
王昌齡
王維
岑參
嚴武
高適
錢起
儲光羲
劉方平
張謂
劉長卿
李嘉祐
皇甫冉
張繼
韓翃


BARABASI, A
Yahoo Labs
JEONG, H
NEWMAN, M
KLEINBERG, J

The Mixed Membership Stochastic Blockmodel

$G = (\mathcal{N}, Y)$ represent observed relational data as a graph

$Y(p, q) \in \{0, 1\}$ maps pairs of nodes to values, that is, edge weights.

Measure a collection of sociometric relations among a group of monks by repeatedly asking questions such as “whom do you like?” and “whom do you dislike?” to determine asymmetric social relationships within the group.

In analyzing this data, the goal is to determine the social structure within the monastery.

The data can be thought of as a directed graph.

We assume K factions, that is, latent groups, exist in the monastery, the observed network is generated according to distributions of group-membership for each monk and a matrix of group-group interaction strength.

Notations

Each monk is associated with a randomly drawn vector $\vec{\pi}_i$ for monk i.

$\pi_{i,g}$ denotes the probability of monk i belonging to group g

That is, each monk can simultaneously belong to multiple groups with different degrees of affiliation strength.

The probabilities of interactions between different groups are defined by a matrix of Bernoulli rates $B_{(K \times K)}$

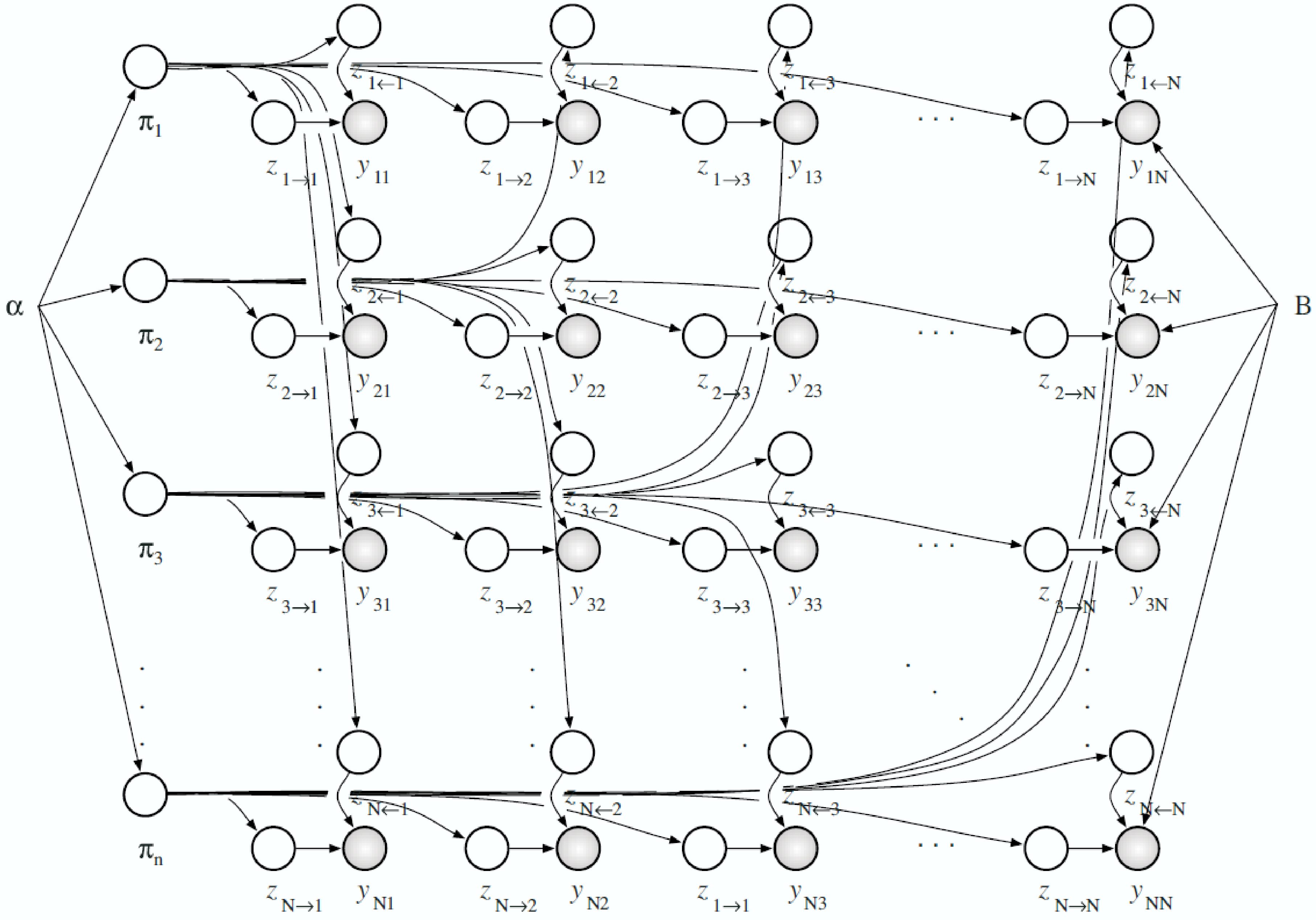
$B(g,h)$ represents the probability of having a link between a monk from group g and a monk from group h.

$\vec{z}_{p \to q}$ denotes the group membership of monk p when he responds to survey questions about monk q

N denotes the number of monks in the monastery

The mixed membership stochastic blockmodel (MMB) posits that a graph $G = (\mathcal{N}, Y)$ is drawn from the following procedure.

- 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, $\vec{z}_{p \rightarrow q} \sim$ Multinomial ($\vec{\pi}_p$).
  - Draw membership indicator for the receiver, $\vec{z}_{q \rightarrow p} \sim$ Multinomial ($\vec{\pi}_q$).
  - Sample the value of their interaction, $Y(p,q) \sim$ Bernoulli ($\vec{z}_{p \rightarrow q}^{\top} B \, \vec{z}_{p \leftarrow q}$).



α
π_1
π_2
π_3
π_n
B
$z_{1\leftarrow 1}$
$z_{1\leftarrow 2}$
$z_{1\leftarrow 3}$
$z_{1\leftarrow N}$
$z_{1\rightarrow 1}$
y_{11}
$z_{1\rightarrow 2}$
y_{12}
$z_{1\rightarrow 3}$
y_{13}
$z_{1\rightarrow N}$
y_{1N}
$z_{2\leftarrow 1}$
$z_{2\leftarrow 2}$
$z_{2\leftarrow 3}$
$z_{2\leftarrow N}$
$z_{2\rightarrow 1}$
y_{21}
$z_{2\rightarrow 2}$
y_{22}
$z_{2\rightarrow 3}$
y_{23}
$z_{2\rightarrow N}$
y_{2N}
$z_{3\leftarrow 1}$
$z_{3\leftarrow 2}$
$z_{3\leftarrow 3}$
$z_{3\leftarrow N}$
$z_{3\rightarrow 1}$
y_{31}
$z_{3\rightarrow 2}$
y_{32}
$z_{3\rightarrow 3}$
y_{33}
$z_{3\rightarrow N}$
y_{3N}
$z_{N\leftarrow 1}$
$z_{N\leftarrow 2}$
$z_{N\leftarrow 3}$
$z_{N\leftarrow N}$
$z_{N\rightarrow 1}$
y_{N1}
$z_{N\rightarrow 2}$
y_{N2}
$z_{1\rightarrow 1}$
y_{N3}
$z_{N\rightarrow N}$
y_{NN}

The two sets of latent group indicators are denoted by $\{\vec{z}_{p\to q} : p,q \in \mathcal{N}\} =: Z_{\to}$ $\{\vec{z}_{p\leftarrow q} : p,q \in \mathcal{N}\} =: Z_{\leftarrow}$.

Under the MMB, the joint probability of the data Y and the latent variables $\{\vec{\pi}_{1:N}, Z_{\to}, Z_{\leftarrow}\}$ can be written

in the following factored form

$$p(Y, \vec{\pi}_{1:N}, Z_{\to}, Z_{\leftarrow} | \vec{\alpha}, B)$$
$$= \prod_{p,q} P(Y(p,q) | \vec{z}_{p\to q}, \vec{z}_{p\leftarrow q}, B) P(\vec{z}_{p\to q} | \vec{\pi}_p) P(\vec{z}_{p\leftarrow q} | \vec{\pi}_q) \prod_p P(\vec{\pi}_p | \vec{\alpha})$$

This model generalizes to two important cases

First, multiple networks among the same actors can be generated by the same latent vectors. This may be useful, for instance, to analyze multivariate sociometric relations.

Second, in the MMB the data generating distribution is a Bernoulli, but B can be a matrix that parameterizes any kind of distribution.

Modeling Sparsity

Adjacency matrices encoding binary pairwise measurements are often sparse, that is, they contain many zeros or non-interactions.

It is useful to distinguish two sources of non-interaction:

- they maybe the result of the rarity of interactions in general.
- they may be an indication that the pair of relevant blocks rarely interact.

A good estimate of the portion of zeros that should not be explained by the blockmodel B reduces the bias of the estimates of its elements.

Thus, we introduce a sparsity parameter $\rho \in [0, 1]$ in the MMB to characterize the source of non-interaction.

we down-weight the probability of successful interaction to $(1 - \rho) \cdot \vec{z}_{p\rightarrow q}^{\top} B \, \vec{z}_{p\leftarrow q}$

The sparsity parameter ρ can be estimated. Its maximum likelihood estimate provides the best data-driven guess about the proportion of zeros that the blockmodel can explain.

Summarizing and De-Noising Pairwise Measurements

- First, MMB can be used to summarize the data, Y, in terms of the global blockmodel B and the node-specific mixed memberships Πs.
- Second, MMB can be used to de-noise the data, Y, in terms of the global blockmodel B and interaction-specific single memberships Zs.

In both cases the model depends on a small set of unknown constants to be estimated: α, and B.

The likelihood is the same in both cases, although, the rationale for including the set of latent variables Zs differs.

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

When summarizing data, we could integrate out the Zs analytically; We choose to keep the Zs

inference.

When de-noising, the Zs are instrumental in estimating posterior expectations of each interactions individually—a network analog to the Kalman Filter.

The posterior expectations of an interaction is computed as follows, in the two cases,

$$\mathbb{E}\left[\,Y(p,q)=1\,\right] \approx \widehat{\vec{\pi}}_p{}'\,\widehat{B}\,\widehat{\vec{\pi}}_q \qquad\qquad \mathbb{E}\left[\,Y(p,q)=1\,\right] \approx \widehat{\vec{\phi}}_{p\to q}{}'\,\widehat{B}\,\widehat{\vec{\phi}}_{p\leftarrow q}$$

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

An Illustration: Crisis in a Cloister

Sampson spent several months in a monastery in New England

- the loyal opposition (whose members joined the monastery first)
- the young turks (who joined later on)
- the outcasts (who were not accepted in the two main factions)
- the waverers (who did not take sides).

Here, we use the following approximation to BIC to choose the number of groups in the MMB:

$$BIC = 2 \cdot \log p(Y) \approx 2 \cdot \log p(Y|\widehat{\vec{\pi}}, \widehat{Z}, \widehat{\vec{\alpha}}, \widehat{B}) - |\vec{\alpha}, B| \cdot \log |Y|$$

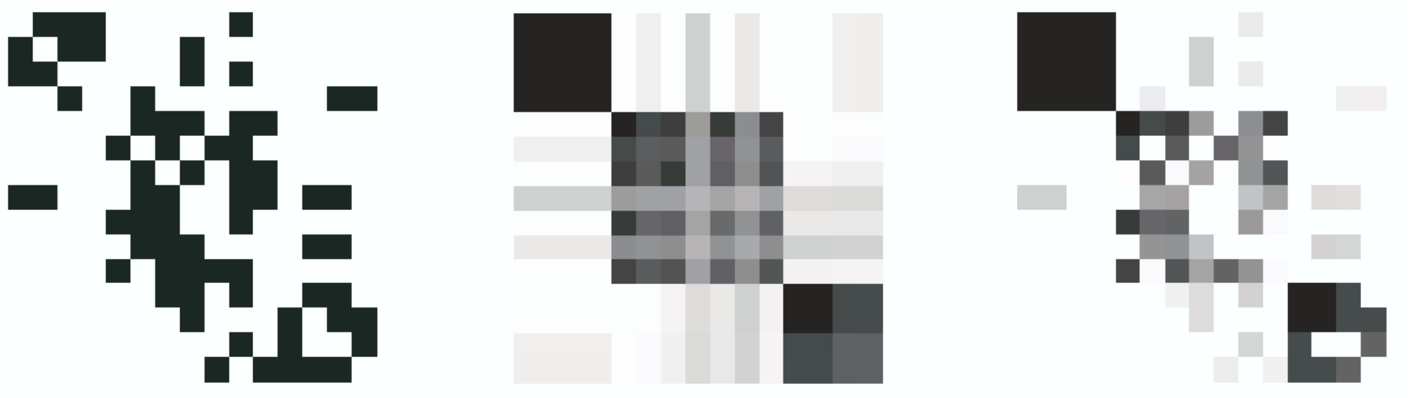
$|\vec{\alpha}, B|$ is the number of hyper-parameters in the model

$|Y|$ is the number of positive relations observed.

The three panels contrast the different resolution of the original adjacency matrix of whom-do-like sociometric relations (left panel) obtained in different uses of MMB.

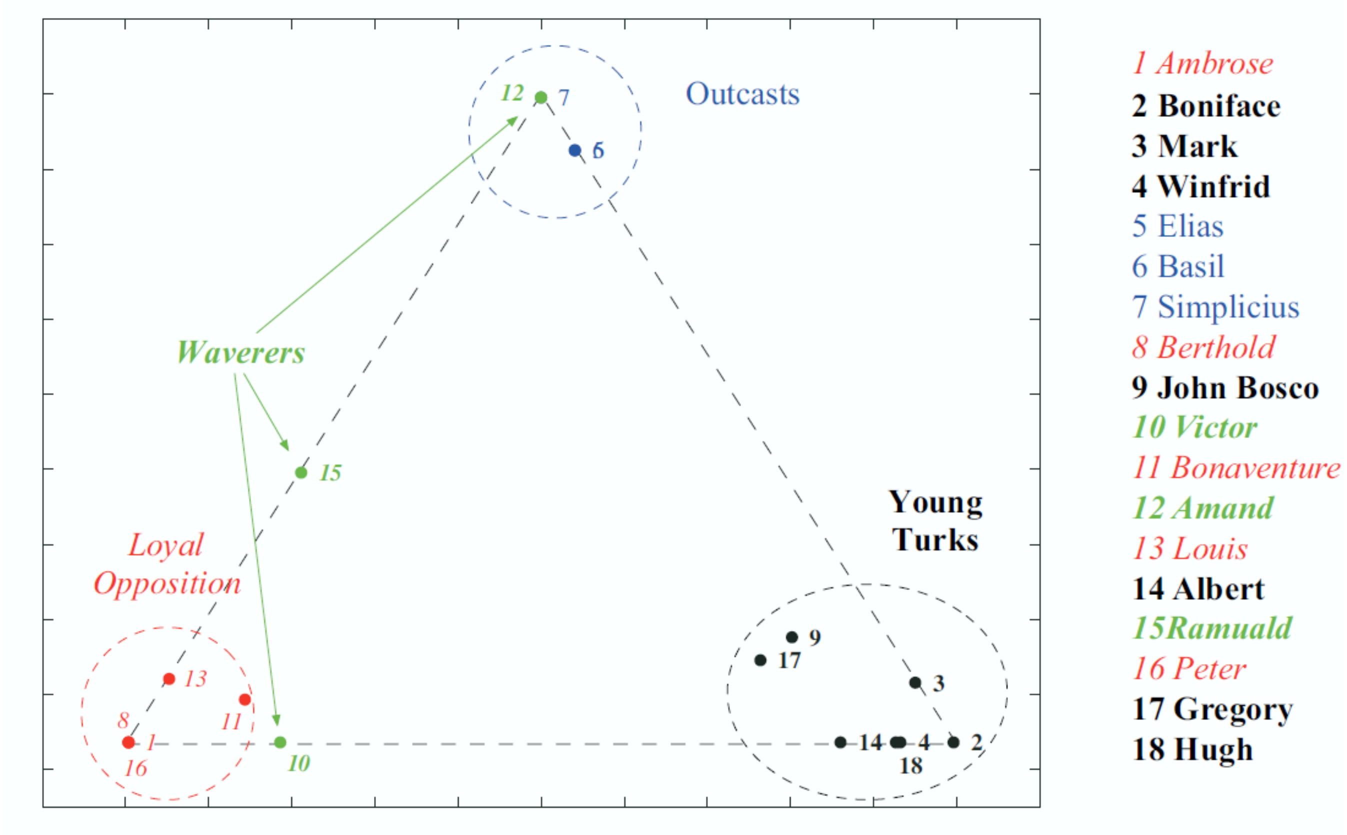
If the goal of the analysis if to find a parsimonious summary of the data, the amount of relational information that is captured by in $\hat{\alpha}, \hat{B}$, and $\mathbb{E}[\vec{\pi}|Y]$ leads to a coarse reconstruction of the original sociomatrix (central panel).

If the goal of the analysis if to de-noising a collection of pairwise measurements, the amount of relational information that is revealed by $\hat{\alpha}, \hat{B}, \mathbb{E}[\vec{\pi}|Y]$ and $\mathbb{E}[Z_{\rightarrow}, Z_{\leftarrow}|Y]$ leads to a finer reconstruction of the original sociomatrix, relations in Y Are re-weighted according to how much they make sense to the model (right panel).



Original adjacency matrix of whom-do-like sociometric relations (left), relations predicted using approximate MLEs for $\vec{\pi}_{1:N}$ and B (center), and relations de-noised using the model including Zs indicators (right).

We can see the central role played by John Bosco and Gregory, who exhibit relations in all three groups, as well as the uncertain affiliations of Ramuald and Victor. (Amand's uncertain affiliation, however, is not captured.)



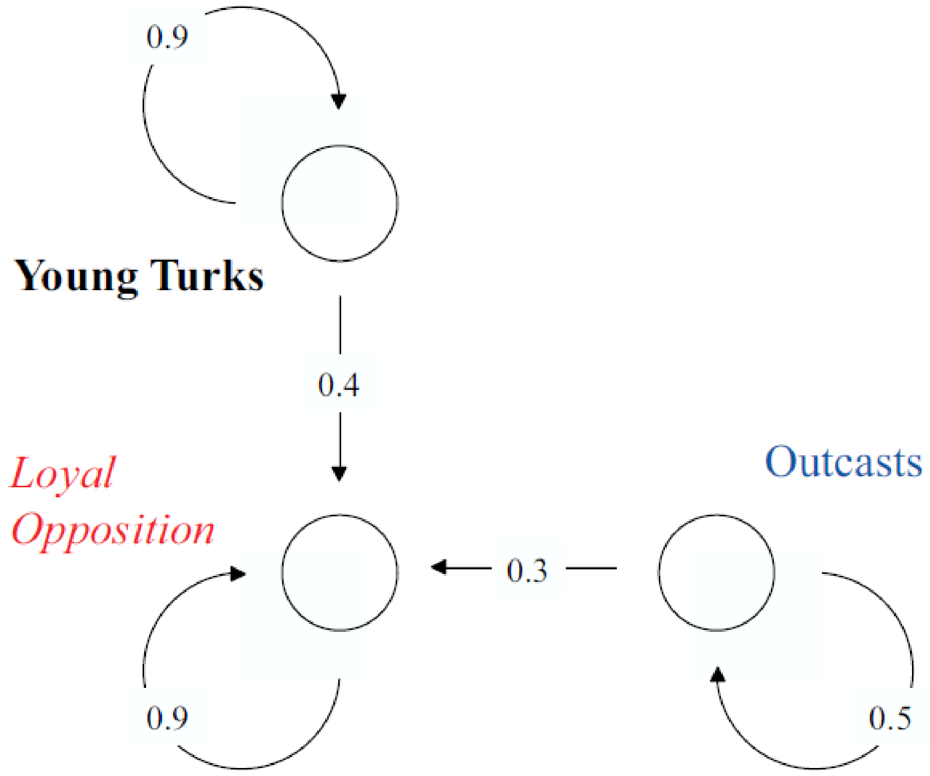


Figure 3: Posterior mixed membership vectors, $\vec{\pi}_{1:18}$, projected in the simplex. Numbered points can be mapped to monks' names using the legend on the right. The colors identify the four factions defined by Sampson's anthropological observations.

Parameter Estimation and Posterior Inference

Two computational problems are central to the MMB:

- posterior inference of the per-node mixed membership vectors and per-pair roles
- parameter estimation of the Dirichlet parameters and Bernoulli rate matrix

We derive empirical Bayes estimates of the parameters $(\vec{\alpha}, B)$ and employ a mean-field approximation scheme for posterior inference.

Posterior Inference

The posterior inference problem is to compute the posterior distribution of the latent variables given a collection of observations.

The normalizing constant of the posterior distribution is the marginal probability of the data, which requires an integral over the simplicial vectors $\vec{\pi}_p$

$$p(Y|\vec{\alpha},B) = \int_{\Pi} \sum_{Zs} \left(\prod_{p,q} P(Y(p,q)|\vec{z}_{p\to q},\vec{z}_{p\leftarrow q},B)P(\vec{z}_{p\to q}|\vec{\pi}_p)P(\vec{z}_{p\leftarrow q}|\vec{\pi}_q) \prod_{p} P(\vec{\pi}_p|\vec{\alpha}) \right) d\vec{\pi}.$$

which is not solvable in closed form.

- mean-field variational methods
- expectation propagation
- Monte Carlo Markov chain sampling (MCMC)

We appeal to **variational methods**. The main idea behind variational methods is to first posit a **distribution** of the latent variables with free parameters, and then fit those parameters such that the distribution is close in Kullback-Leibler divergence to the true posterior.

The variational distribution is simpler than the true posterior so that the optimization problem can be approximately solved.

In the MMB, we begin by bounding the log of the marginal probability of the data with Jensen's inequality

$$\log p(Y \,|\, \alpha, B) \geq \mathbb{E}_q \left[\log p(Y, \vec{\pi}_{1:N}, Z_{\rightarrow}, Z_{\leftarrow} | \alpha, B) \right] - \mathbb{E}_q \left[\log q(\vec{\pi}_{1:N}, Z_{\rightarrow}, Z_{\leftarrow}) \right]$$

We have introduced a distribution of the latent variables q We specify q as the mean-field fully-factorized family

$$q(\vec{\pi}_{1:N}, Z_{\rightarrow}, Z_{\leftarrow} | \vec{\gamma}_{1:N}, \Phi_{\rightarrow}, \Phi_{\leftarrow}) = \prod_p q_1(\vec{\pi}_p | \vec{\gamma}_p) \prod_{p,q} \Big(q_2(\vec{z}_{p\rightarrow q} | \vec{\phi}_{p\rightarrow q}) \; q_2(\vec{z}_{p\leftarrow q} | \vec{\phi}_{p\leftarrow q}) \Big)$$

q_1 is a Dirichlet q_2 is a multinomial, $\{\vec{\gamma}_{1:N}, \Phi_{\rightarrow}, \Phi_{\leftarrow}\}$ are the set of free variational parameters that are optimized to tighten the bound.

$$\hat{\phi}_{p\rightarrow q,g} \;\propto\; e^{\,\mathbb{E}_q\left[\log \pi_{p,g}\right]} \cdot \prod_h \left(B(g,h)^{Y(p,q)} \cdot \left(\, 1 - B(g,h)\,\right)^{1-Y(p,q)} \right)^{\phi_{p\leftarrow q,h}}$$

$$\hat{\phi}_{p\leftarrow q,h} \;\propto\; e^{\,\mathbb{E}_q\left[\log \pi_{q,h}\right]} \cdot \prod_g \left(B(g,h)^{Y(p,q)} \cdot \left(\, 1 - B(g,h)\,\right)^{1-Y(p,q)} \right)^{\phi_{p\rightarrow q,g}}$$

$$\hat{\gamma}_{p,k} = \alpha_k + \sum_q \phi_{p\rightarrow q,k} + \sum_q \phi_{p\leftarrow q,k}$$

1. initialize $\vec{\gamma}^0_{pk} = \frac{2N}{K}$ for all p,k
2. **repeat**
3. **for** $p = 1$ to N
4. **for** $q = 1$ to N
5. get **variational** $\vec{\phi}^{t+1}_{p\rightarrow q}$ and $\vec{\phi}^{t+1}_{p\leftarrow q} = f\left(\, Y(p,q), \vec{\gamma}^t_p, \vec{\gamma}^t_q, B^t \,\right)$
6. partially update $\vec{\gamma}^{t+1}_p$, $\vec{\gamma}^{t+1}_q$ and B^{t+1}
7. **until** convergence

5.1. initialize $\phi^0_{p\rightarrow q,g} = \phi^0_{p\leftarrow q,h} = \frac{1}{K}$ for all g,h
5.2. **repeat**
5.3. **for** $g = 1$ to K
5.4. update $\phi^{s+1}_{p\rightarrow q} \propto f_1\left(\, \vec{\phi}^s_{p\leftarrow q}, \vec{\gamma}_p, B \,\right)$
5.5. normalize $\vec{\phi}^{s+1}_{p\rightarrow q}$ to sum to 1
5.6. **for** $h = 1$ to K
5.7. update $\phi^{s+1}_{p\leftarrow q} \propto f_2\left(\, \vec{\phi}^s_{p\rightarrow q}, \vec{\gamma}_q, B \,\right)$
5.8. normalize $\vec{\phi}^{s+1}_{p\leftarrow q}$ to sum to 1
5.9. **until** convergence

Parameter Estimation

We compute the empirical Bayes estimates of the model hyper-parameters $\{\vec{\alpha}, B\}$ with a variational expectation-maximization (EM) algorithm.

Empirical Bayes, guides the posterior inference towards a region of the hyper-parameter space that is supported by the data.

参数估计：最大似然估计（MLE），最大后验估计(MAP)，贝叶斯估计，经验贝叶斯(Empirical Bayes)与全贝叶斯(Full Bayes)

Variational EM uses the lower bound in Equation 5 as a surrogate for the likelihood. To find a local optimum of the bound, we iterate between fitting the variational distribution.

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

In EM, the lower bound $\mathcal{L}(q,\Theta)$ is then iteratively maximized with respect to Θ, in the M step, and q in the E step (Dempster et al., 1977). In particular, at the t-th iteration of the E step we set

$$q^{(t)} = p(X|Y,\Theta^{(t-1)}), \tag{5}$$

A closed form solution for the approximate maximum likelihood estimate of $\vec{\alpha}$ does not exist

We use a linear-time Newton-Raphson method, where the gradient and Hessian are

$$\frac{\partial \mathcal{L}_{\vec{\alpha}}}{\partial \alpha_k} = N \left(\psi\left(\sum_k \alpha_k\right) - \psi(\alpha_k) \right) + \sum_p \left(\psi(\gamma_{p,k}) - \psi\left(\sum_k \gamma_{p,k}\right) \right)$$

$$\frac{\partial \mathcal{L}_{\vec{\alpha}}}{\partial \alpha_{k_1} \alpha_{k_2}} = N \left(\mathbb{I}_{(k_1=k_2)} \cdot \psi'(\alpha_{k_1}) - \psi'\left(\sum_k \alpha_k\right) \right)$$

The approximate MLE of B

$$\hat{B}(g,h) = \frac{\sum_{p,q} Y(p,q) \cdot \phi_{p\rightarrow qg} \phi_{p\leftarrow qh}}{(1-\rho) \cdot \sum_{p,q} \phi_{p\rightarrow qg} \phi_{p\leftarrow qh}}$$

Finally, the approximate MLE of the sparsity parameter ρ is

$$\hat{\rho} = \frac{\sum_{p,q} \left(1 - Y(p,q) \right) \cdot \left(\sum_{g,h} \phi_{p\rightarrow qg} \phi_{p\leftarrow qh} \right)}{\sum_{p,q} \sum_{g,h} \phi_{p\rightarrow qg} \phi_{p\leftarrow qh}}$$