Analysis of Social Media

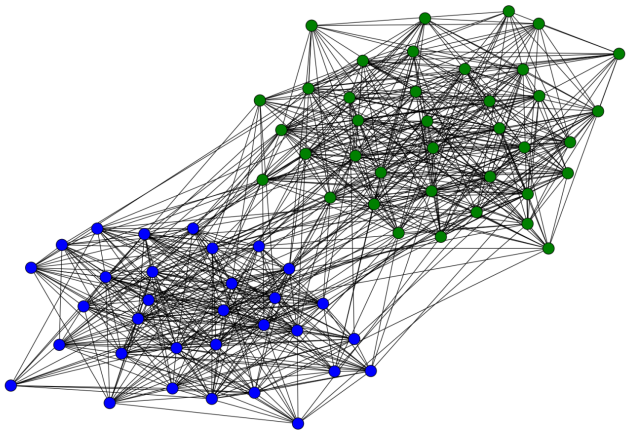
MLD 10-802, LTI 11-772

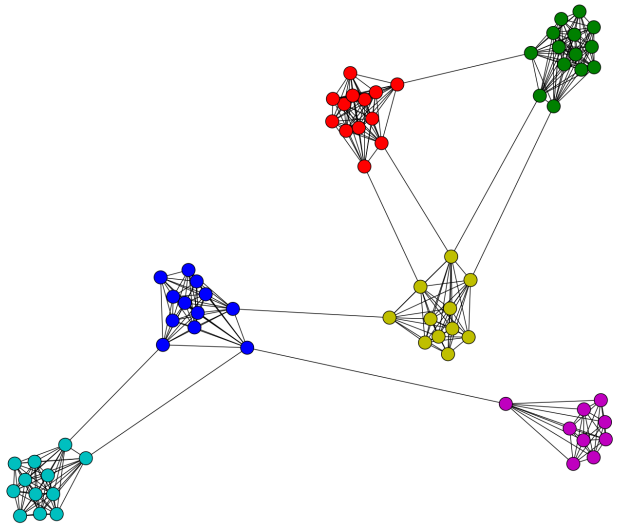
William Cohen

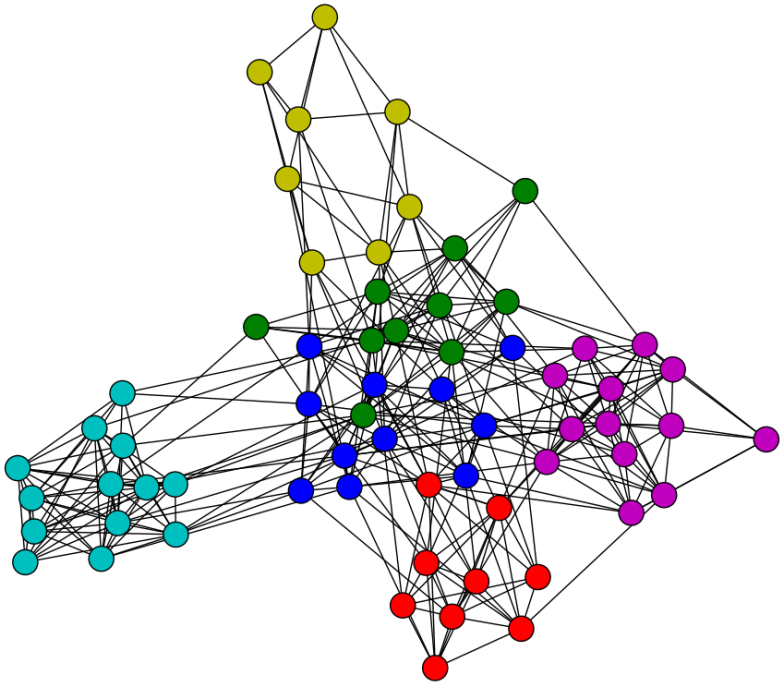
10-09-010

Stochastic blockmodel graphs

- Last week: spectral clustering
- *Theory* suggests it will work for graphs produced by a *particular generative model*
- Question: can you *directly maximize* Pr(structure,parameters|data) for that model?





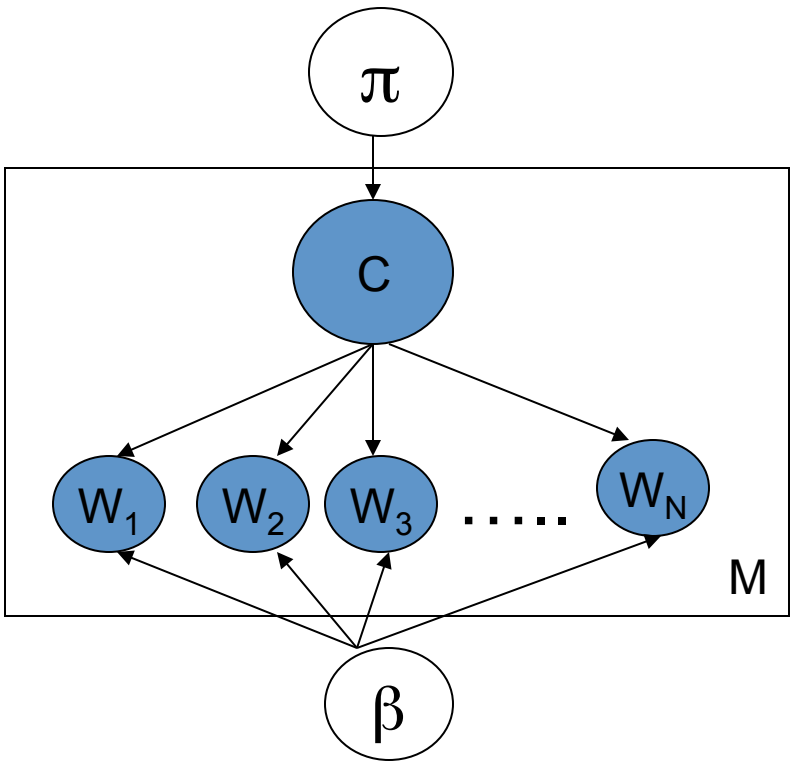


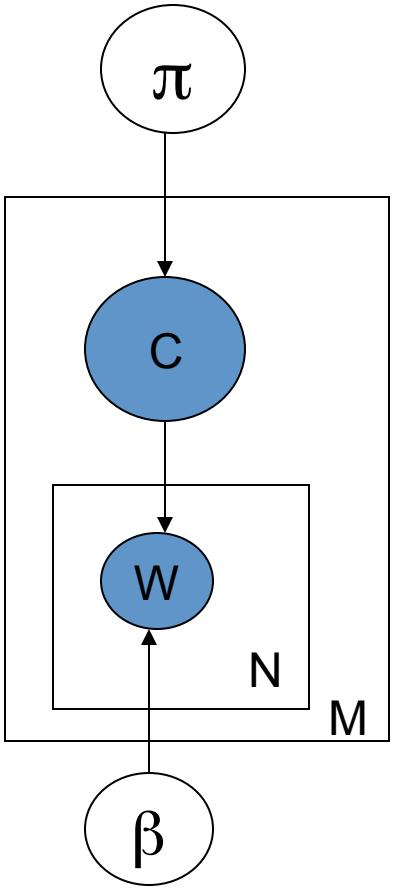
Outline

- Stochastic block models & inference question
- Review of text models
  - Mixture of multinomials & EM
  - LDA and Gibbs (or variational EM)
- Block models and inference
- Mixed-membership block models
- Multinomial block models and inference w/ Gibbs
- Beastiary of other probabilistic graph models
  - Latent-space models, exchangeable graphs, p1, ERGM

Review – supervised Naïve Bayes

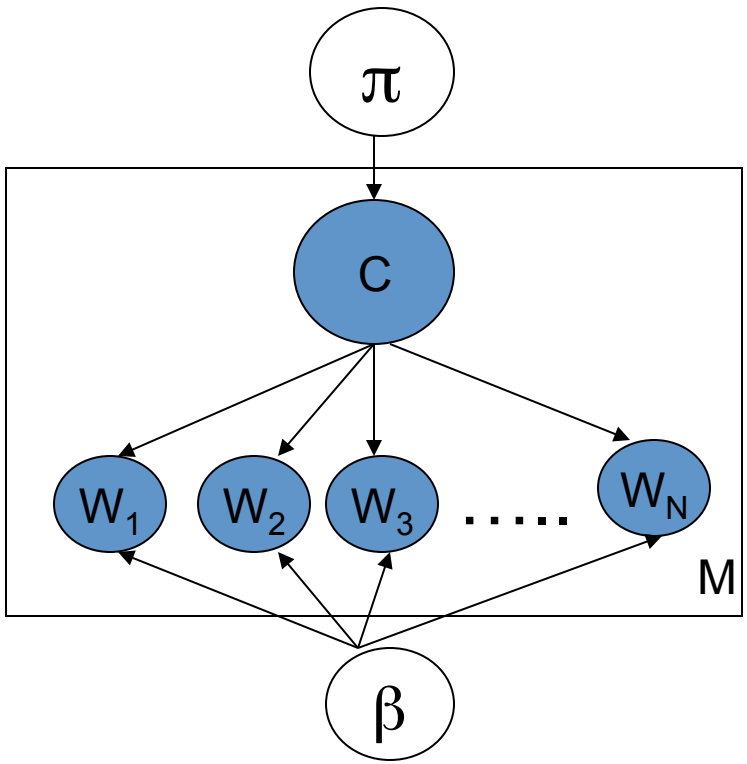
- Naïve Bayes Model: Compact representation





Review – supervised Naïve Bayes

- Multinomial Naïve Bayes



- For each document d = 1,⋯, M
  - Generate C_d ~ Mult(¢ | π)
  - For each position n = 1,⋯, N_d
    - Generate w_n ~ Mult(¢|β,C_d)

$$\prod_{d=1}^{M} P(w_1, \cdots, w_{N_d}, C_d | \beta, \pi)$$

$$= \prod_{d=1}^{M} \left\{ P(C_d|\pi) \prod_{n=1}^{N_d} P(w_n|\beta, C_d) \right\} = \prod_{d=1}^{M} \left\{ \pi_{C_d} \prod_{n=1}^{N_d} \beta_{C_d, w_n} \right\}$$

Review – supervised Naïve Bayes

- Multinomial naïve Bayes: Learning
  - Maximize the log-likelihood of observed variables w.r.t. the parameters:

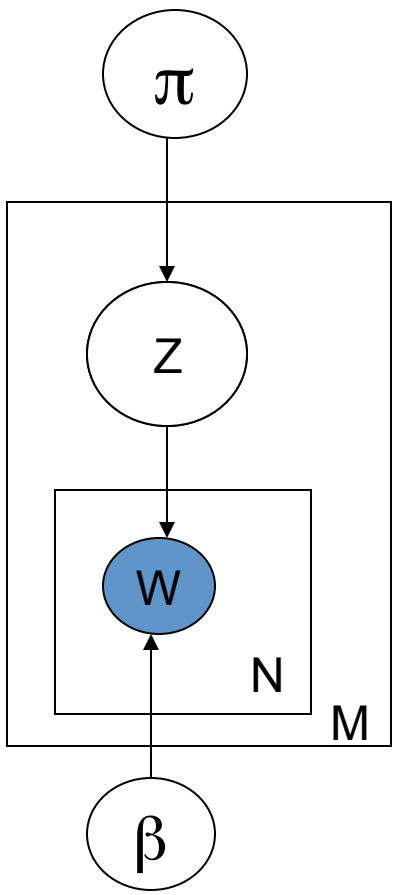
$$\sum_{d=1}^{M} \log P(w_1, \cdots, w_{N_d}, C_d | \beta, \pi) = \sum_{d=1}^{M} \left\{ \log(\pi_{C_d}) + \sum_{n=1}^{N_d} \log(\beta_{C_d, w_n}) \right\}$$

- Convex function: global optimum
- Solution:

$$\begin{aligned} \pi_C &= \frac{\sum_{d=1}^{N} \delta_C(C_d)}{M} \\ \beta_{C,w} &= \frac{\sum_{d:C_d=C} n(d,w)}{\sum_{d:C_d=C} \sum_w n(d,w)} \end{aligned}$$

Review – **un**supervised Naïve Bayes

- Mixture model: unsupervised naïve Bayes model



- Joint probability of words and classes:

$$\prod_{d=1}^{M} P(w_1, \cdots, w_{N_d}, z_d | \beta, \pi) = \prod_{d=1}^{M} \left\{ \pi_{z_d} \prod_{n=1}^{N_d} \beta_{z_d, w_n} \right\}$$

- But classes are not visible:

$$\prod_{d=1}^{M} P(w_1, \cdots, w_{N_d} | \pi, \beta) \quad = \quad \prod_{d=1}^{N_d} \left\{ \sum_{k=1}^{K} \left(\pi_k \prod_{n=1}^{N_d} \beta_{k, w_n} \right) \right\}$$

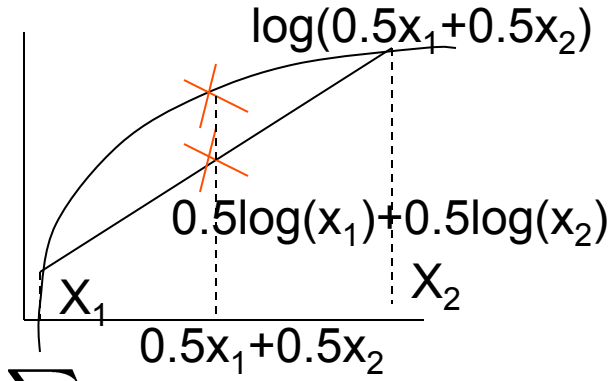
Review – **un**supervised Naïve Bayes

- Mixture model: learning

$$\sum_{d=1}^{M} \log P(w_1, \cdots, w_{N_d} | \pi, \beta) \quad = \quad \sum_{d=1}^{N_d} \left\{ \log \left(\sum_{k=1}^{K} \left[\pi_k \prod_{n=1}^{N_d} \beta_{k,w_n} \right] \right) \right\}$$

  - Not a convex function
    - No global optimum solution
  - Solution: Expectation Maximization
    - Iterative algorithm
    - Finds local optimum
    - Guaranteed to maximize a lower-bound on the log-likelihood of the observed data

Review – **un**supervised Naïve Bayes



- Quick summary of EM:
  - Log is a concave function

$$\log(\sum_i \gamma_i x_i) \quad \geq \quad \sum_i \gamma_i \log(x_i) \text{ where } \gamma_i \geq 0 \ \& \ \sum_i \gamma_i = 1$$

H(γ)

$$\log(\sum_i x_i) \quad = \quad \log(\sum_i \gamma_i \frac{x_i}{\gamma_i}) \geq \sum_i (\gamma_i \log(x_i) - \gamma_i \log(\gamma_i))$$

$$\log\left(\sum_{k=1}^{K}\left[\pi_k \prod_{n=1}^{N_d} \beta_{k,w_n}\right]\right) \quad \geq \quad \sum_{k=1}^{K}\left\{\gamma_k \log\left(\pi_k \prod_{n=1}^{N_d} \beta_{k,w_n}\right)\right\} + H(\gamma)$$

  - Lower-bound is convex!
  - Optimize this lower-bound w.r.t. each variable instead

Review – **un**supervised Naïve Bayes

$$\log\left(\sum_{k=1}^{K}\left[\pi_k \prod_{n=1}^{N_d} \beta_{k,w_n}\right]\right) \geq \sum_{k=1}^{K}\left\{\gamma_k \log\left(\pi_k \prod_{n=1}^{N_d} \beta_{k,w_n}\right)\right\} + H(\gamma)$$

- Mixture model: EM solution

E-step:

$$\gamma_{dk}^{(t+1)} = \frac{\pi_k^{(t)} \prod_{n=1}^{N_d} \beta_{k,w_n}^{(t)}}{\sum_k \pi_k^{(t)} \prod_{n=1}^{N_d} \beta_{k,w_n}^{(t)}}$$

M-step:

$$\begin{cases} \pi_k^{(t+1)} = \dfrac{\sum_{d=1}^{M} \gamma_{dk}^{(t)}}{M} \\[2ex] \beta_{k,w}^{(t+1)} = \dfrac{\sum_{d=1}^{M} \gamma_{dk}^{(t)} n(d,w)}{\sum_{d=1}^{M} \gamma_{nk}^{(t)} \sum_w n(d,w)} \end{cases}$$

Key capability: estimate distribution of **latent variables** given **observed variables**

Review - LDA

Latent Dirichlet Allocation

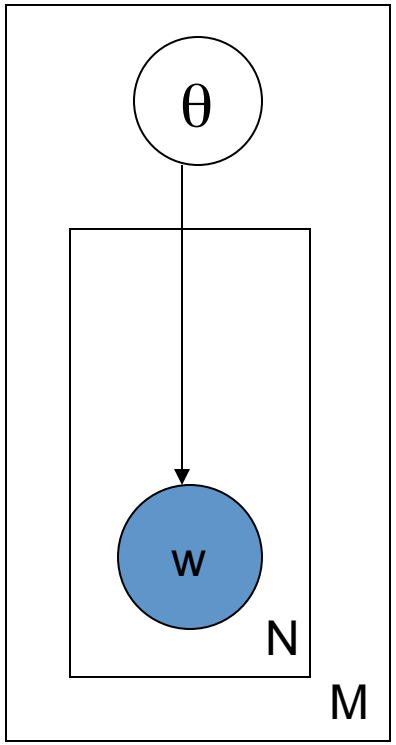
David M. Blei
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

Review - LDA

- Motivation



Assumptions: 1) documents are i.i.d 2) *within* a document, words are i.i.d. (bag of words)

•For each document d = 1,⋯,M

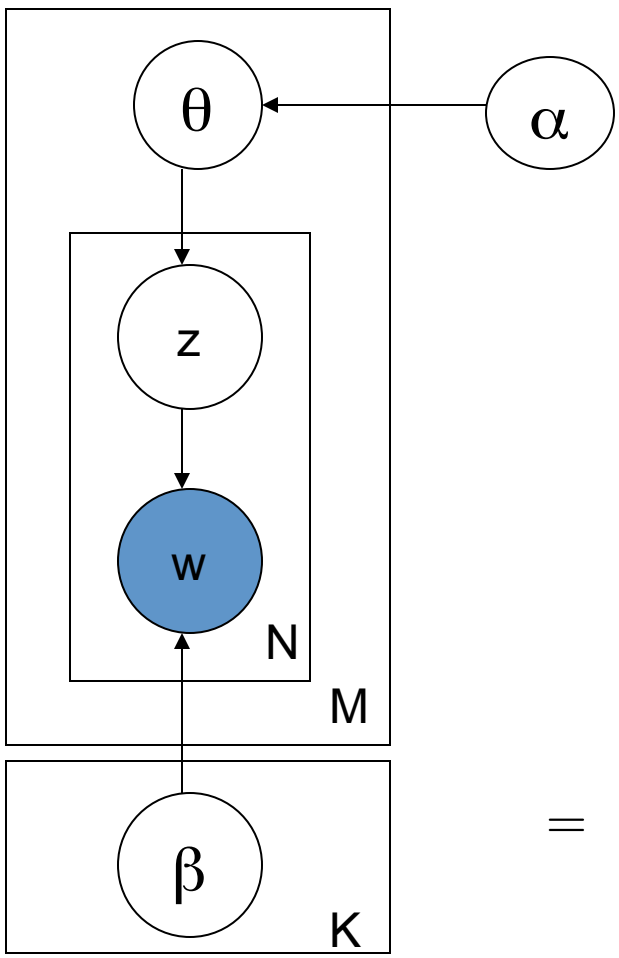
- Generate $\theta_d \sim D_1(\ldots)$
- For each word n = 1,⋯, N_d
  - •generate $w_n \sim D_2(\phi \mid \vartheta_{d_n})$

Now pick your favorite distributions for D_1, D_2

$$\Pr(z = j \mid n_1, n_2, \ldots, n_k, \alpha) = \frac{n_j + \alpha_j}{n_1 + \alpha_1 + \ldots n_k + \alpha_k}$$

"Mixed membership"

- Latent Dirichlet Allocation



- For each document d = 1,⋯,M
  - Generate θ_d ~ Dir(¢ | α)
  - For each position n = 1,⋯, N_d
    - generate z_n ~ Mult(¢ | θ_d)
    - generate w_n ~ Mult(¢ | β_{z_n})

$$\prod_{d=1}^{N_d} P(w_1, \cdots, w_{N_d} | \beta, \alpha)$$

$$= \prod_{d=1}^{N_d} \int_{\theta_d} P(\theta_d | \alpha) \left\{ \prod_{n=1}^{N_d} \left(\sum_k \theta_{dk} \beta_{k w_n} \right) \right\} d\theta_d$$

- LDA's view of a document

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

"Arts" "Budgets" "Children" "Education"

- LDA topics

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Review - LDA

- Latent Dirichlet Allocation
  - Parameter learning:
    - Variational EM
      - Numerical approximation using lower-bounds
      - Results in biased solutions
      - Convergence has numerical guarantees
    - Gibbs Sampling
      - Stochastic simulation
      - unbiased solutions
      - Stochastic convergence

Review - LDA

- Gibbs sampling
  - Applicable when joint distribution is hard to evaluate but conditional distribution is known
  - Sequence of samples comprises a Markov Chain
  - Stationary distribution of the chain is the joint distribution

1. Initialise $x_{0,1:n}$.
2. For $i = 0$ to $N - 1$
   - Sample $x_1^{(i+1)} \sim p(x_1 | x_2^{(i)}, x_3^{(i)}, \ldots, x_n^{(i)})$.
   - Sample $x_2^{(i+1)} \sim p(x_2 | x_1^{(i+1)}, x_3^{(i)}, \ldots, x_n^{(i)})$.

$$\vdots$$

   - Sample $x_j^{(i+1)} \sim p(x_j | x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)})$.

$$\vdots$$

   - Sample $x_n^{(i+1)} \sim p(x_n | x_1^{(i+1)}, x_2^{(i+1)}, \ldots x_{n-1}^{(i+1)})$.

Key capability: estimate distribution of **one** latent variables given **the other latent variables** and observed variables.

Why does Gibbs sampling work?

- What's the fixed point?
  - Stationary distribution of the chain is the joint distribution
- When will it converge (in the limit)?
  - Graph defined by the chain is connected
- How long will it take to converge?
  - Depends on second eigenvector of that graph

⊡ initialisation

zero all count variables, $n_m^{(k)}, n_m, n_k^{(t)}, n_k$

for all documents $m \in [1, M]$ **do**

  for all words $n \in [1, N_m]$ in document m **do**

    sample topic index $z_{m,n}=k \sim \text{Mult}(1/K)$

    increment document–topic count: $n_m^{(k)} + 1$

    increment document–topic sum: $n_m + 1$

    increment topic–term count: $n_k^{(t)} + 1$

    increment topic–term sum: $n_k + 1$

  end for

end for

⊡ Gibbs sampling over burn-in period and sampling period

while not finished **do**

  for all documents $m \in [1, M]$ **do**

    for all words $n \in [1, N_m]$ in document m **do**

      ⊡ for the current assignment of k to a term t for word $w_{m,n}$:

      decrement counts and sums: $n_m^{(k)} - 1; n_m - 1; n_k^{(t)} - 1; n_k - 1$

      ep):

      sample topic index $\tilde{k} \sim p(z_i | \vec{z}_{\neg i}, \vec{w})$

      ⊡ use the new assignment of $z_{m,n}$ to the term t for word $w_{m,n}$ to:

      increment counts and sums: $n_m^{(\tilde{k})} + 1; n_m + 1; n_{\tilde{k}}^{(t)} + 1; n_{\tilde{k}} + 1$

    end for

  end for

  ⊡ check convergence and read out parameters

$$p(z_i{=}k|\vec{z}_{\neg i}, \vec{w}) = \frac{p(\vec{w}, \vec{z})}{p(\vec{w}, \vec{z}_{\neg i})} = \frac{p(\vec{w}|\vec{z})}{p(\vec{w}_{\neg i}|\vec{z}_{\neg i})p(w_i)} \cdot \frac{p(\vec{z})}{p(\vec{z}_{\neg i})}$$

$$\propto \frac{\Delta(\vec{n}_z + \vec{\beta})}{\Delta(\vec{n}_{z,\neg i} + \vec{\beta})} \cdot \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{n}_{m,\neg i} + \vec{\alpha})}$$

$$\propto \frac{\Gamma(n_k^{(t)} + \beta_t)\,\Gamma(\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t)}{\Gamma(n_{k,\neg i}^{(t)} + \beta_t)\,\Gamma(\sum_{t=1}^{V} n_k^{(t)} + \beta_t)} \cdot \frac{\Gamma(n_m^{(k)} + \alpha_k)\,\Gamma(\sum_{k=1}^{K} n_{m,\neg i}^{(k)} + \alpha_k)}{\Gamma(n_{m,\neg i}^{(k)} + \alpha_k)\,\Gamma(\sum_{k=1}^{K} n_m^{(k)} + \alpha_k)}$$

$$\propto \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t} \cdot \frac{n_{m,\neg i}^{(t)} + \alpha_k}{[\sum_{k=1}^{K} n_m^{(k)} + \alpha_k] - 1}$$

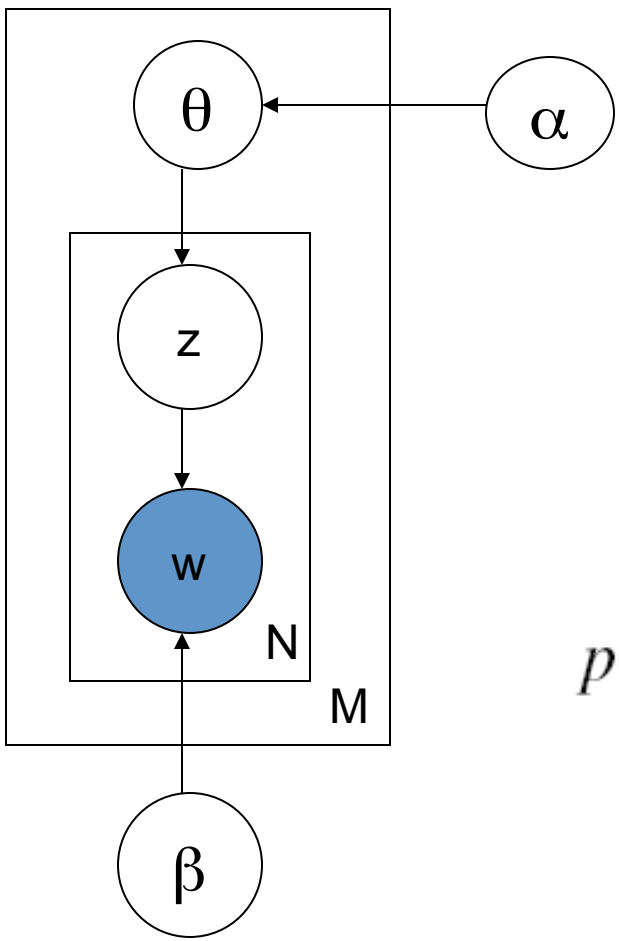
Called “collapsed Gibbs sampling” since you’ve marginalized away some variables

Fr*: Parameter estimation for text analysis* - Gregor Heinrich

Review - LDA

"Mixed membership"

- Latent Dirichlet Allocation



- Randomly initialize each $z_{m,n}$
- Repeat for t=1,….
  - For each doc *m,* word *n*
    - Find Pr(z_{mn}=*k*|other z's)
    - Sample z_{mn} according to that distr.

$$p(z_i{=}k|\vec{z}_{\neg i}, \vec{w}) =$$

$$\propto \frac{n^{(t)}_{k,\neg i} + \beta_t}{\sum_{t=1}^{V} n^{(t)}_{k,\neg i} + \beta_t} \cdot \frac{n^{(t)}_{m,\neg i} + \alpha_k}{[\sum_{k=1}^{K} n^{(k)}_{m} + \alpha_k] - 1}$$

Outline

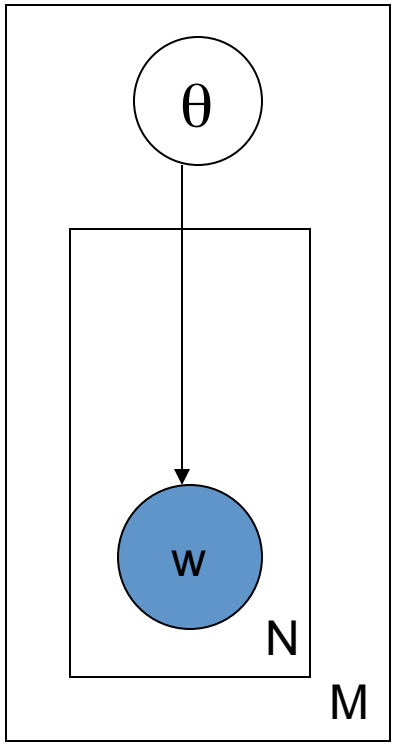
- Stochastic block models & inference question
- Review of text models
  - Mixture of multinomials & EM
  - LDA and Gibbs (or variational EM)
- **Block models and inference**
- Mixed-membership block models
- Multinomial block models and inference w/ Gibbs
- Beastiary of other probabilistic graph models
  - Latent-space models, exchangeable graphs, p1, ERGM

Statistical Models of Networks

- Want a generative probabilistic model that's amenable to analysis….
- … but more expressive than Erdos-Renyi
- One approach: *exchangeable graph model*
  - *Exchangeable*: X1,X2 are exchangable if Pr(X1,X2,W)=Pr(X2,X1,W).
  - The generalizes of i.i.d.-ness
  - It's a Bayesian thing

Review - LDA

- Motivation



Assumptions: 1) documents are i.i.d 2) *within* a document, words are i.i.d. (bag of words)

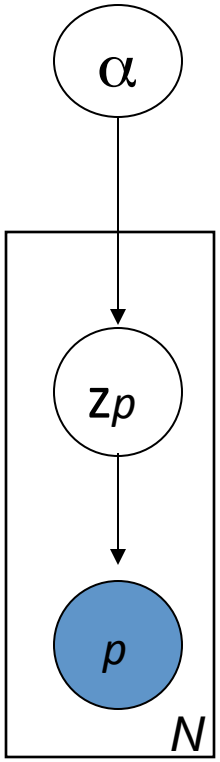
•For each document d = 1,⋯,M

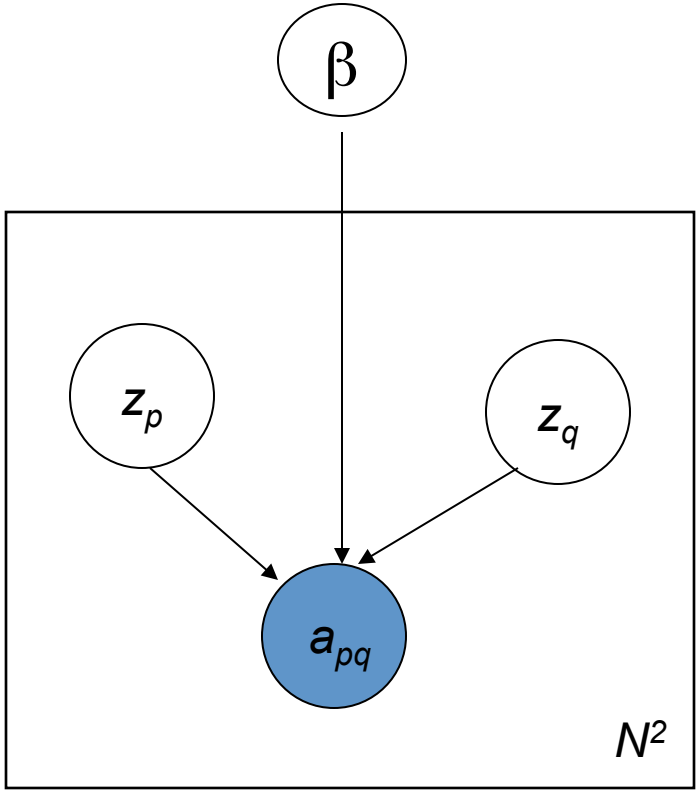
- Generate $\theta_d \sim D_1(\ldots)$
- For each word n = 1,⋯, N_d
  - •generate $w_n \sim D_2(\phi \mid \vartheta_{d_n})$

Docs and words are *exchangeable*.

Stochastic Block models:

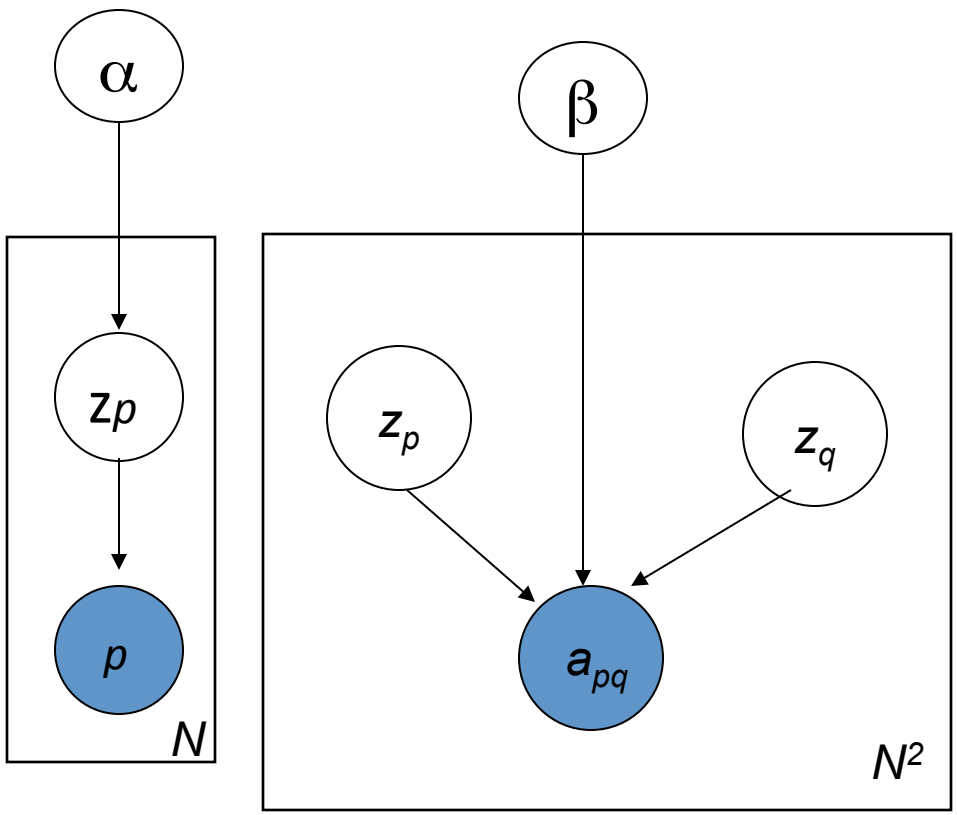
assume 1) nodes w/in a block *z* and
2) edges between blocks z_p, z_q are *exchangeable*





Stochastic Block models:

assume 1) nodes w/in a block z and
2) edges between blocks z_p, z_q are *exchangeable*

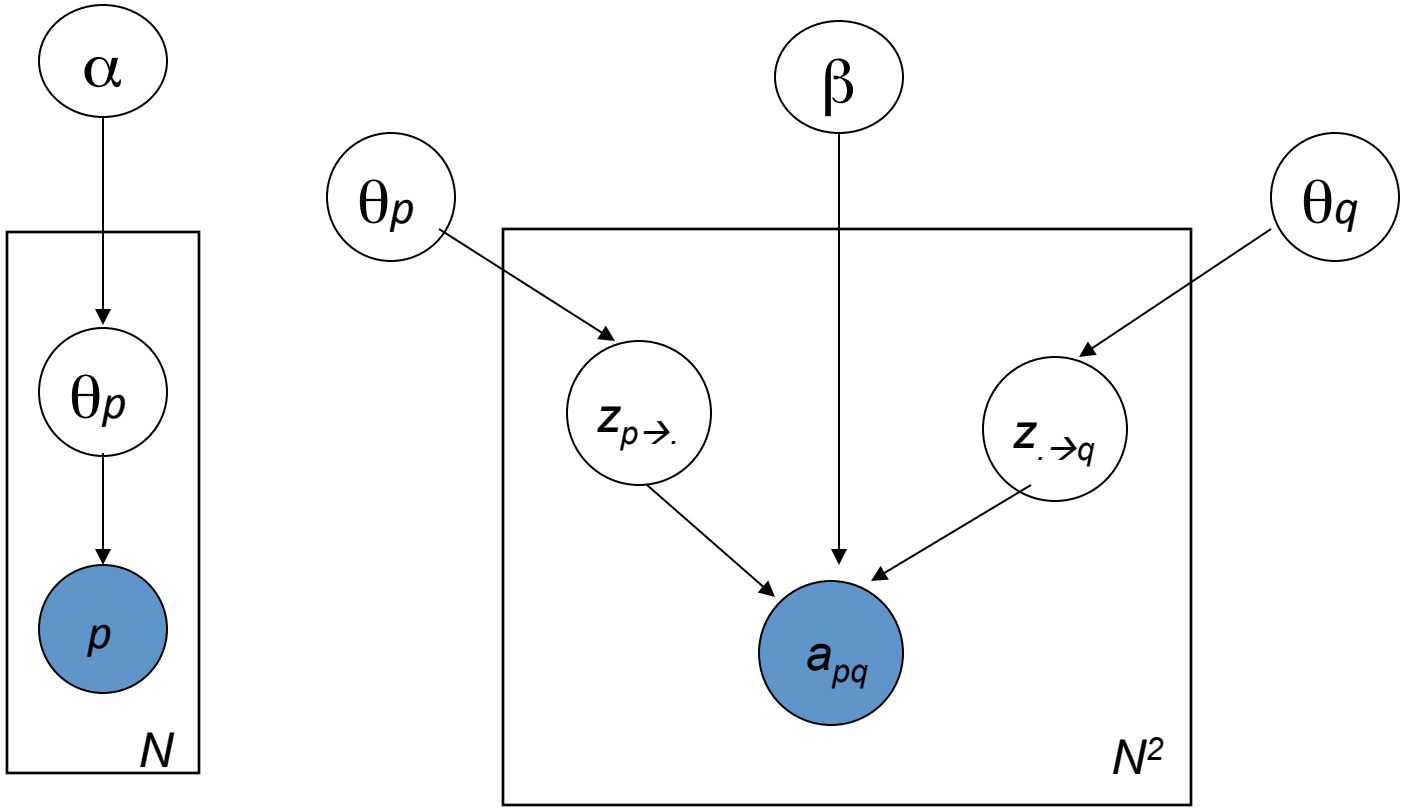


Gibbs sampling:

- Randomly initialize z_p for each node p.
- For t = 1…
  - For each node p
    - Compute z_p given other z's
    - Sample z_p

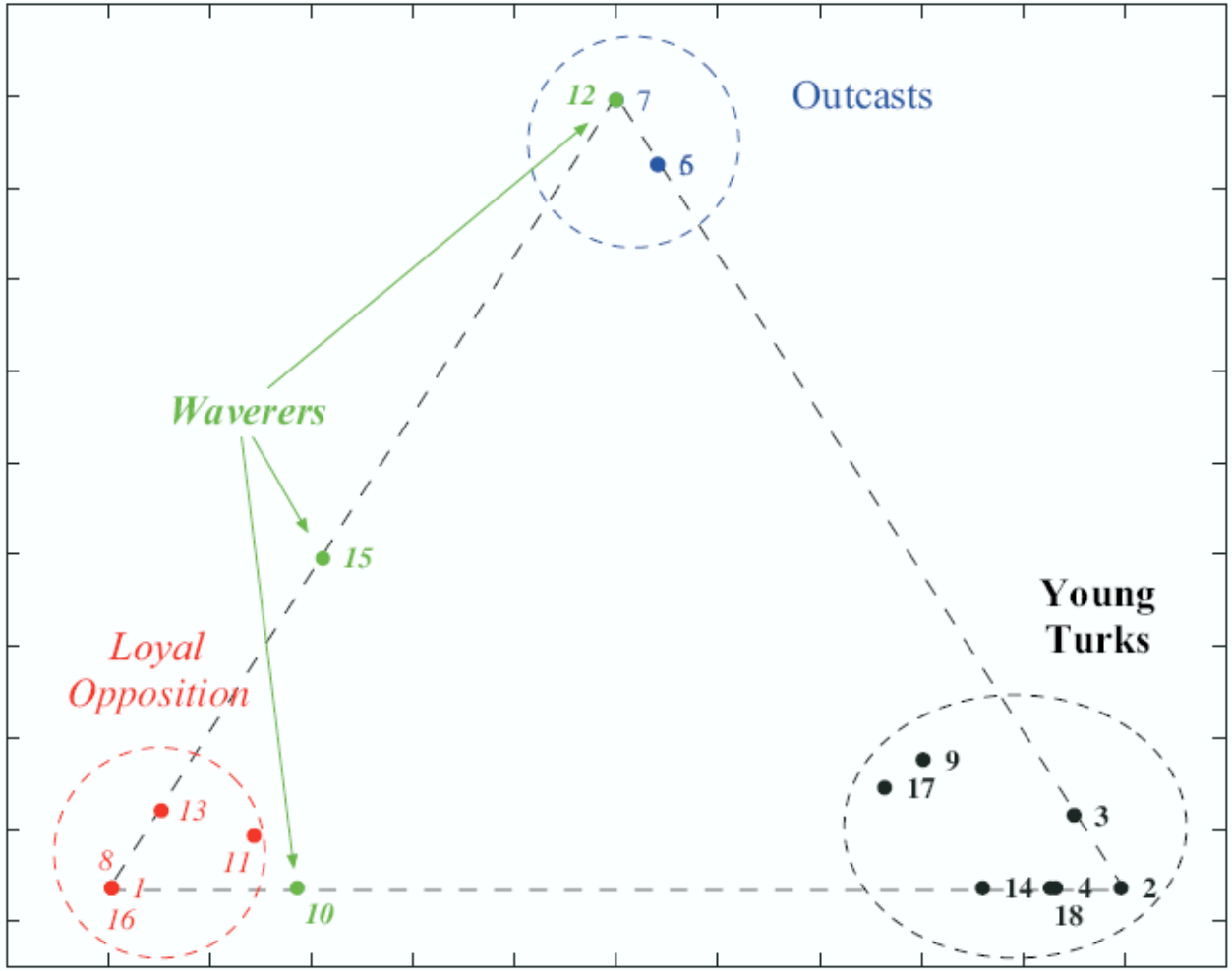
See: Snijders & Nowicki, 1997, Estimation and Prediction for Stochastic Blockmodels for Groups with Latent Graph Structure

Mixed Membership Stochastic Block models



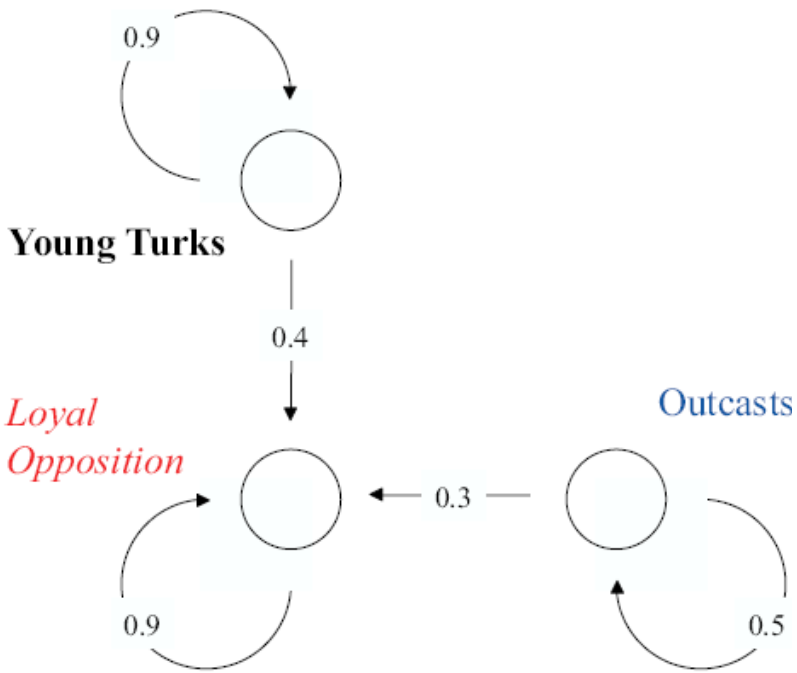
Airoldi et al, JMLR 2008

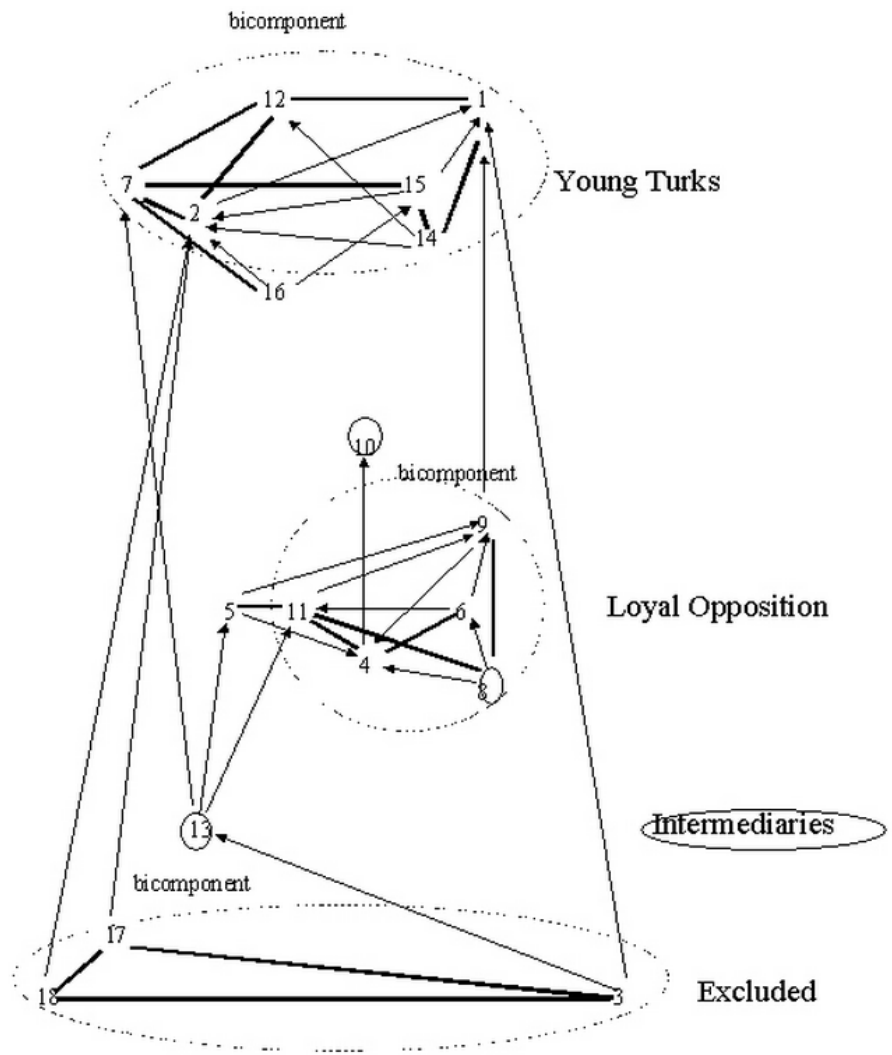
Mixed Membership Stochastic Block models



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

Mixed Membership Stochastic Block models

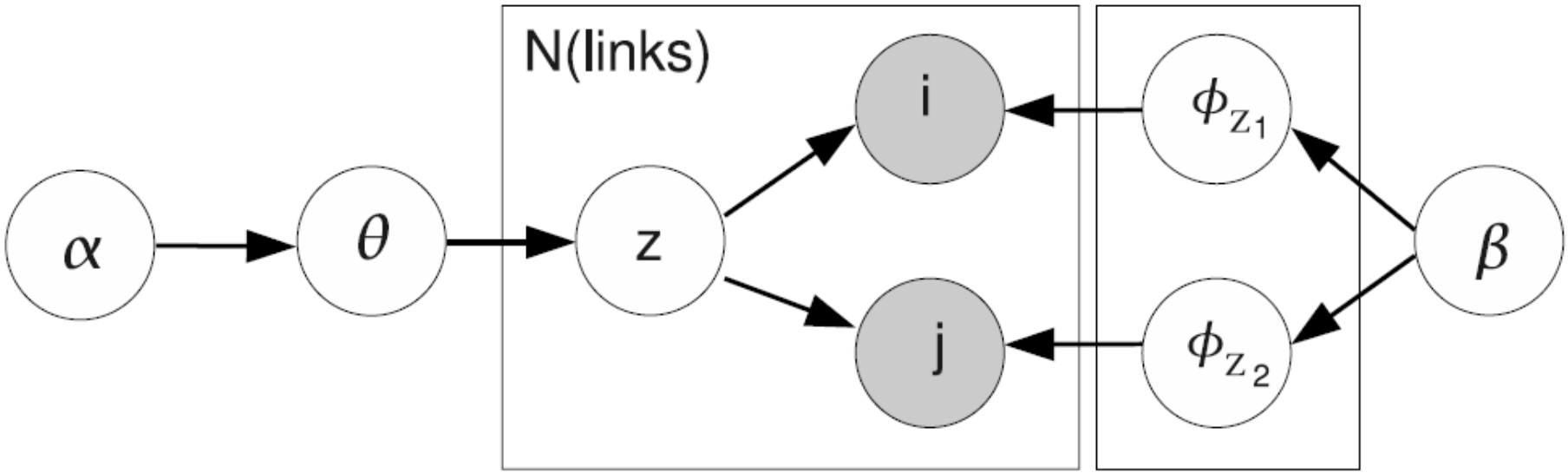




Stopped 10/9

Parkkinen et al paper

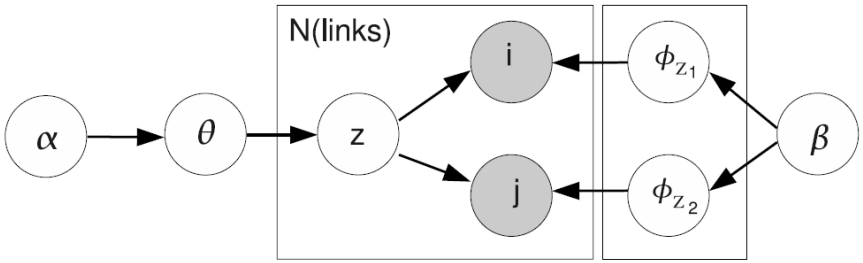
Another mixed membership block model



$$p(z_l|\{z\}^{\neg l}, \{(i,j)\}^{\neg l}, \alpha, \beta) \propto$$

$$(n_z^{\neg l} + \alpha) \cdot \frac{(q_{z_1 i}^{\neg l} + \beta)(q_{z_2 j}^{\neg l} + \beta)}{(q_{z_1 \cdot}^{\neg l} + M\beta)(q_{z_2 \cdot}^{\neg l} + M\beta + \delta_z)},$$

Another mixed membership block model



z=(zi,zj) is a pair of block ids

n_z = #pairs *z*

$q_{z1,i}$ = #links to *i* from block z1

$q_{z1,\cdot}$ = #outlinks in block z1

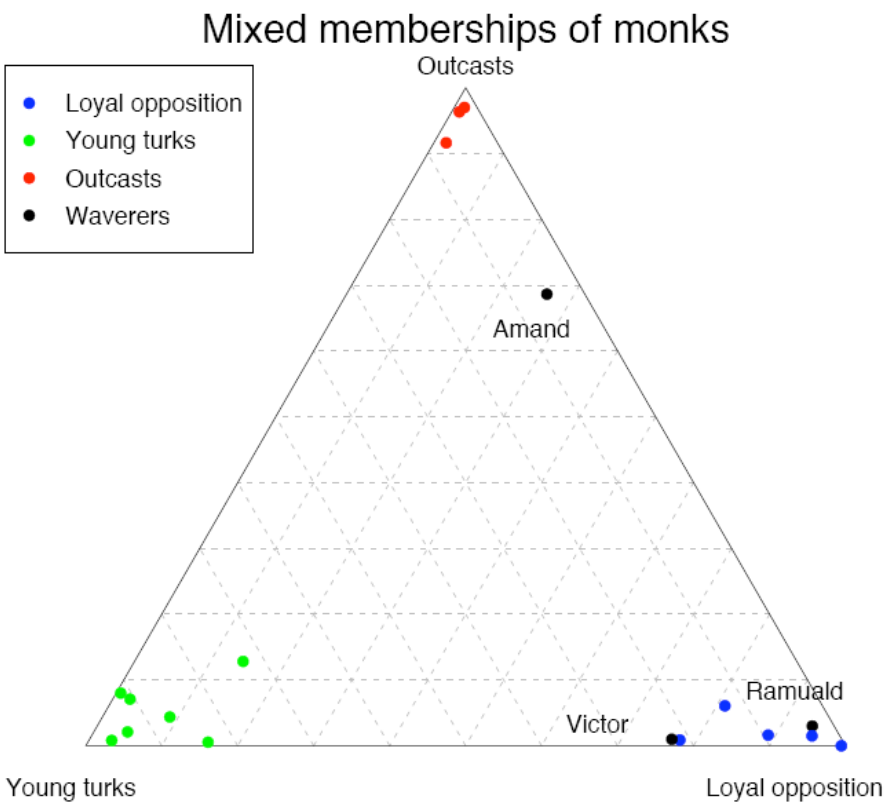
δ = indicator for diagonal

M = #nodes

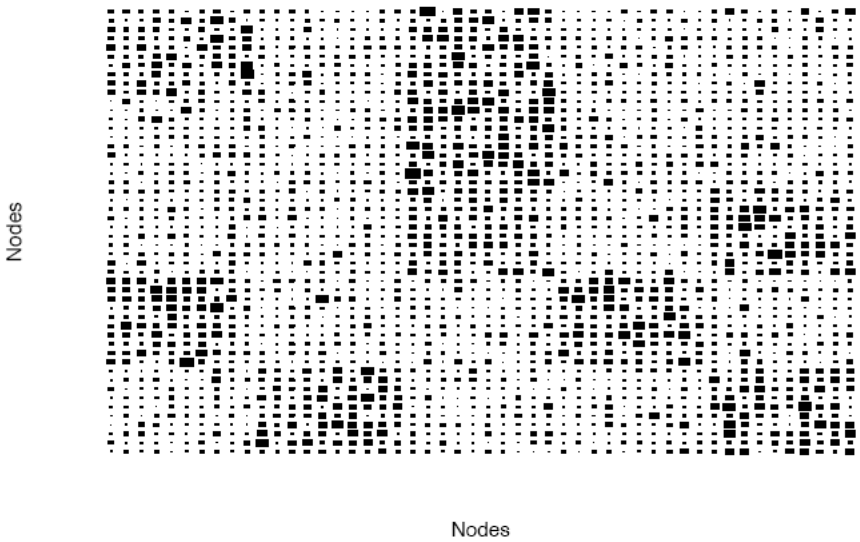
$$p(z_l|\{z\}^{\neg l}, \{(i,j)\}^{\neg l}, \alpha, \beta) \propto$$

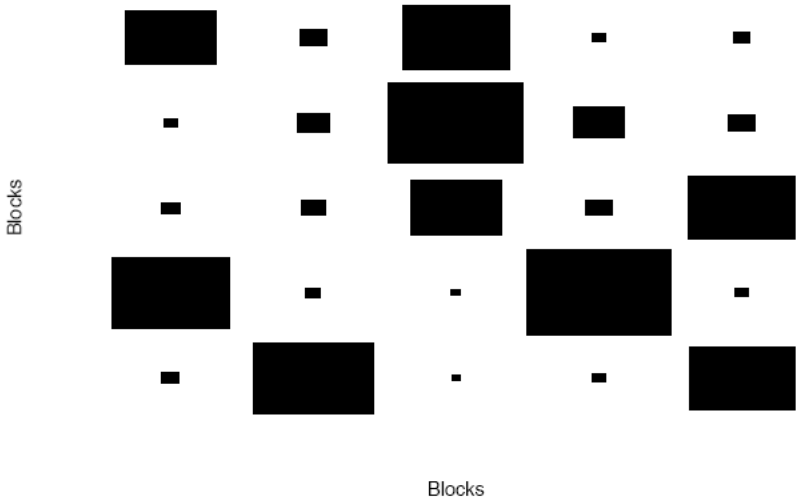
$$(n_z^{\neg l} + \alpha) \cdot \frac{(q_{z_1 i}^{\neg l} + \beta)(q_{z_2 j}^{\neg l} + \beta)}{(q_{z_1 \cdot}^{\neg l} + M\beta)(q_{z_2 \cdot}^{\neg l} + M\beta + \delta_z)},$$

Another mixed membership block model



Another mixed membership block model





Outline

- Stochastic block models & inference question
- Review of text models
  - Mixture of multinomials & EM
  - LDA and Gibbs (or variational EM)
- Block models and inference
- Mixed-membership block models
- Multinomial block models and inference w/ Gibbs
- **Beastiary of other probabilistic graph models**
  - Latent-space models, exchangeable graphs, p1, ERGM

Exchangeable Graph Model

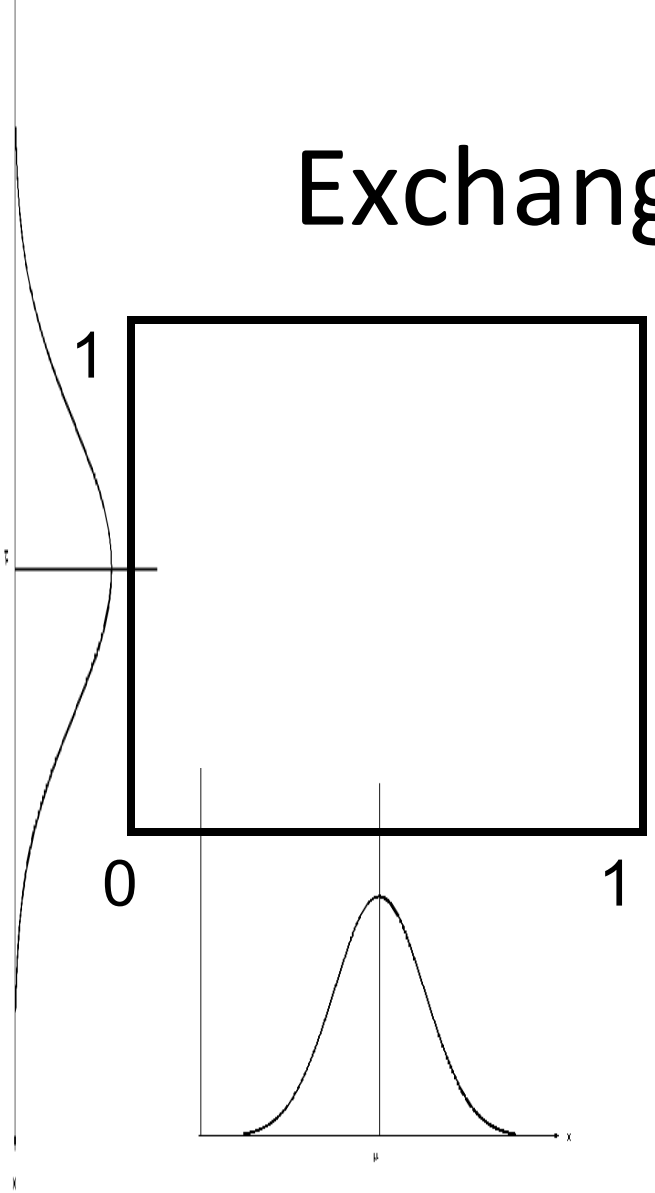
- Defined by a 2^k x 2^k table $q(b_1,b_2)$
- Draw a length-k bit string $b(n)$ like 01101 for each node n from a ~~*uniform*~~ distribution.

complicated

- For each pair o
  - Flip a coin wit
  - If it's heads co

- Pick k-dimensional vector u from a multivariate normal w/ variance α and covariance β – so u_i's are correlated.
- Pass each u_i thru a sigmoid so it's in [0,1] – call that p_i
- Pick b_i using p_i

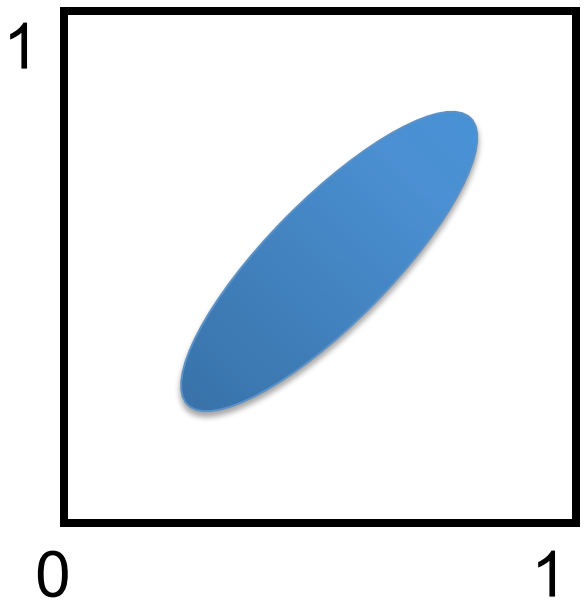
Exchangeable Graph Model



If α is big then *ux,uy* are really big (or small) so *px,py* will end up in a corner.

- Pick *k*-dimensional vector *u* from a multivariate normal w/ variance α and covariance β – so u_i's are correlated.
- Pass each u_i thru a sigmoid so it's in [0,1] – call that p_i
- Pick b_i using p_i

Exchangeable Graph Model



If α is big then *ux,uy* are really big (or small) so *px,py* will end up in a corner.

- Pick *k*-dimensional vector *u* from a multivariate normal w/ variance α and covariance β – so u_i's are **correlated**.
- Pass each u_i thru a sigmoid so it's in [0,1] – call that p_i
- Pick b_i using p_i

The p_1 model for a directed graph

- Parameters, per node *i:*
  - Θ: background edge probability
  - α_i: "expansiveness" – how extroverted is *i?*
  - β_i: "popularity" – how much do others want to be with *i?*
  - ρ_i: "reciprocation" – how likely is *i* to respond to an incomping link with an outgoing one?

$$\log \Pr(i \ldots j) = \lambda_{ij}$$

$$\log \Pr(i \rightarrow j) = \lambda_{ij} + \boxed{\alpha_i + \beta_j + \theta}$$

$$\log \Pr(i \leftarrow j) = \lambda_{ij} + \boxed{\alpha_j + \beta_i + \theta}$$

$$\log \Pr(i \leftrightarrow j) = \lambda_{ij} + \boxed{} + \boxed{}$$

Logistic-regression like procedure can be used to fit this to data from a graph

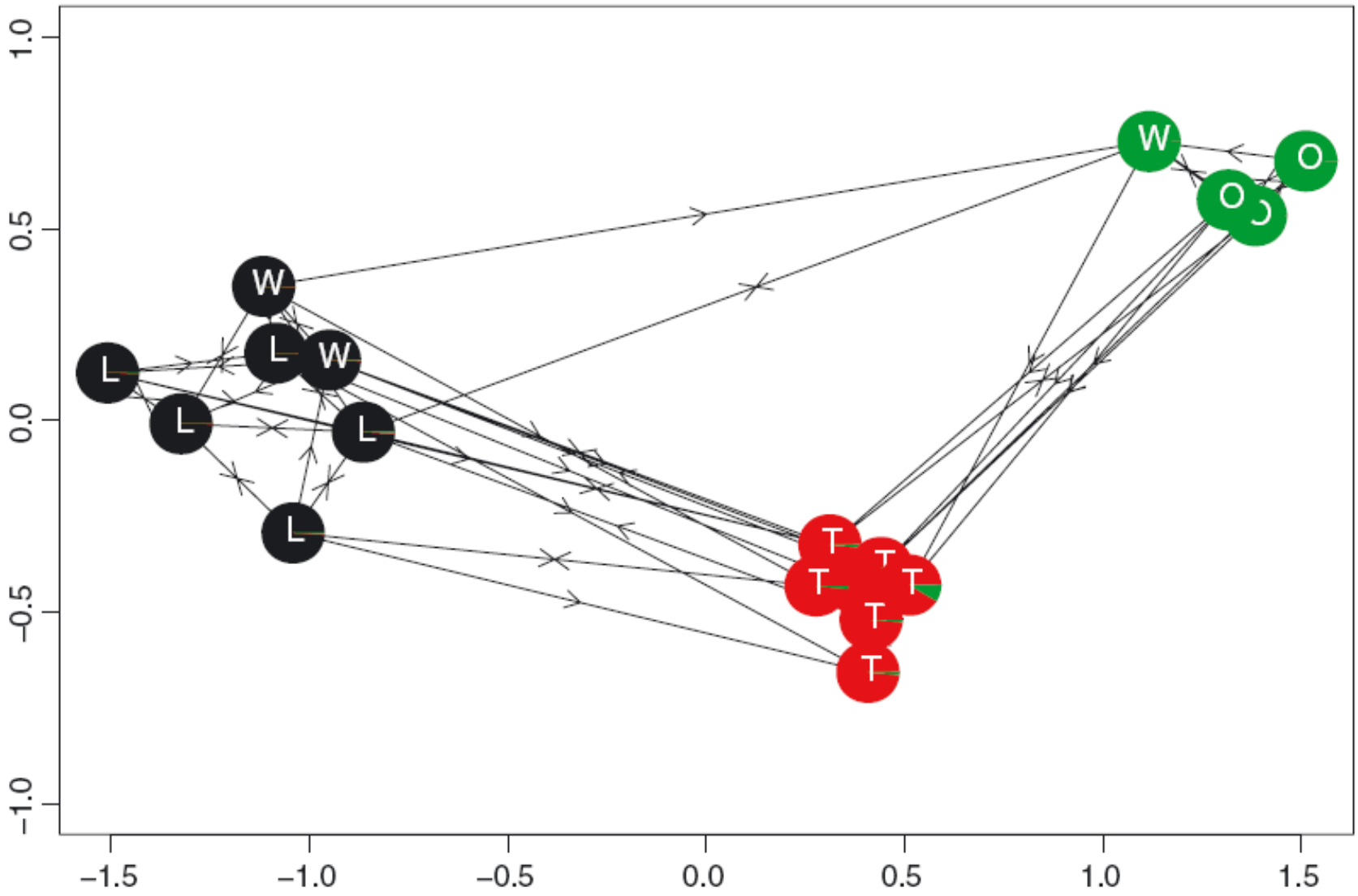
$$\log Pr_{p_1}(y) \propto y_{++}\theta + \sum_i y_{i+}\alpha_i + \sum_j y_{+j}\beta_j + \sum_{ij} y_{ij}y_{ji}\rho,$$

Exponential Random Graph Model

- Basic idea:
  - Define some features of the graph (e.g., number of edges, number of triangles, …)
  - Build a MaxEnt-style model based on these features

Latent Space Model

- Each node *i* has a latent position in Euclidean space, *z(i)*
- *z(i)'s* drawn from a mixture of Gaussians
- Probability of interaction between *i* and *j* depend on the *distance* between *z(i)* and *z(j)*
- Inference is a little more complicated… [Handcock & Raftery, 2007]


1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
W
O
O
O
W
L
L
W
L
L
L
T
T
T
T
T
T

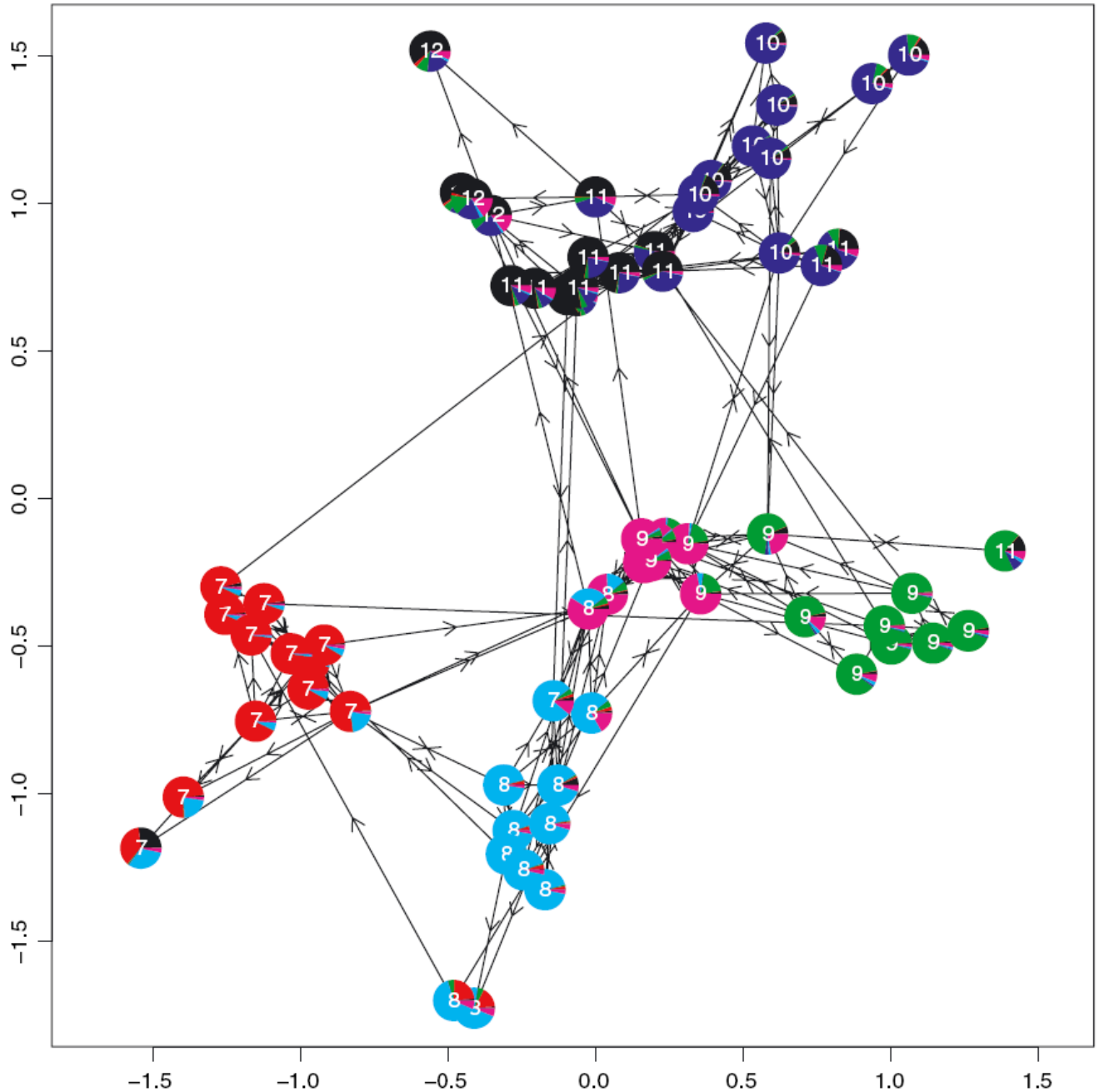


Fig. 8. Pie charts for posterior probabilities of cluster assignment for each actor, at the Bayesian estimates of posterior latent positions for the friendship network in the adolescent health school: the students' grades are shown as numbers

Outline

- Stochastic block models & inference question
- Review of text models
  - Mixture of multinomials & EM
  - LDA and Gibbs (or variational EM)
- Block models and inference
- Mixed-membership block models
- Multinomial block models and inference w/ Gibbs
- Beastiary of other probabilistic graph models
  - Latent-space models, exchangeable graphs, p1, ERGM