Topic models for corpora and for graphs

Announcement – [*c.* Jan 27]

- Paper presentations: 3/3 and 3/5
- Projects:
  - see "project info" on wiki
  - 1-2 page writeup of your idea: 2/17
  - Response to my feedback: 3/5
  - Option for 605 students to collaborate:
    - Proposals will be posted; proposers can advertise slots for collaborators, who can be 605 students (1-2 per project max)
    - "Pay": 1 less assignment, no exam

- http://curtis.ml.cmu.edu/w/courses/index.php/Machine_Learning_with_Large_Datasets_10-605_in_Spring_2015#Grading_Policies **updated**

Announcement

- Quiz: https://qna-app.appspot.com/view.html?aglzfnFuYS1hcHByGQsSDFF1ZXN0aW9uTGlzdBiAgICAyvPFCgw

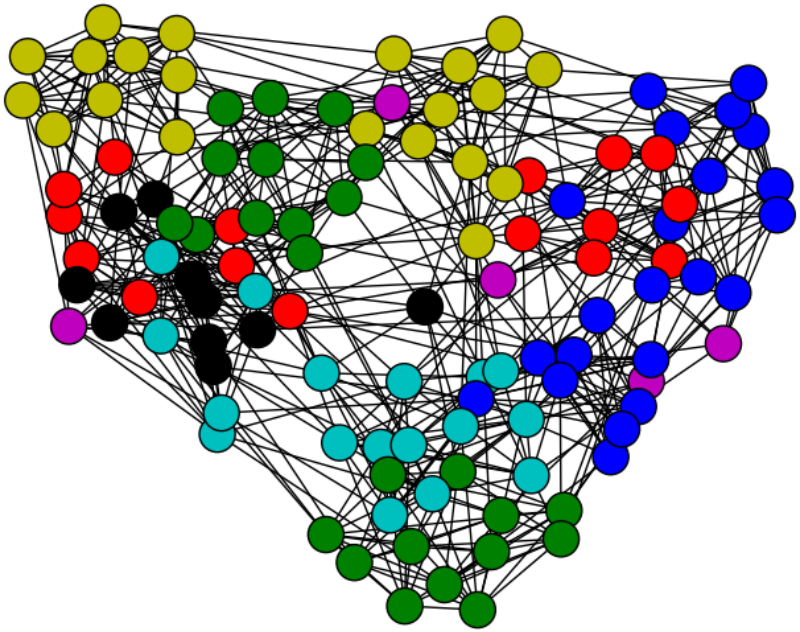
Motivation

- Social graphs seem to have
  - some aspects of randomness
    - small diameter, giant connected components,..
  - some structure
    - homophily, scale-free degree dist?

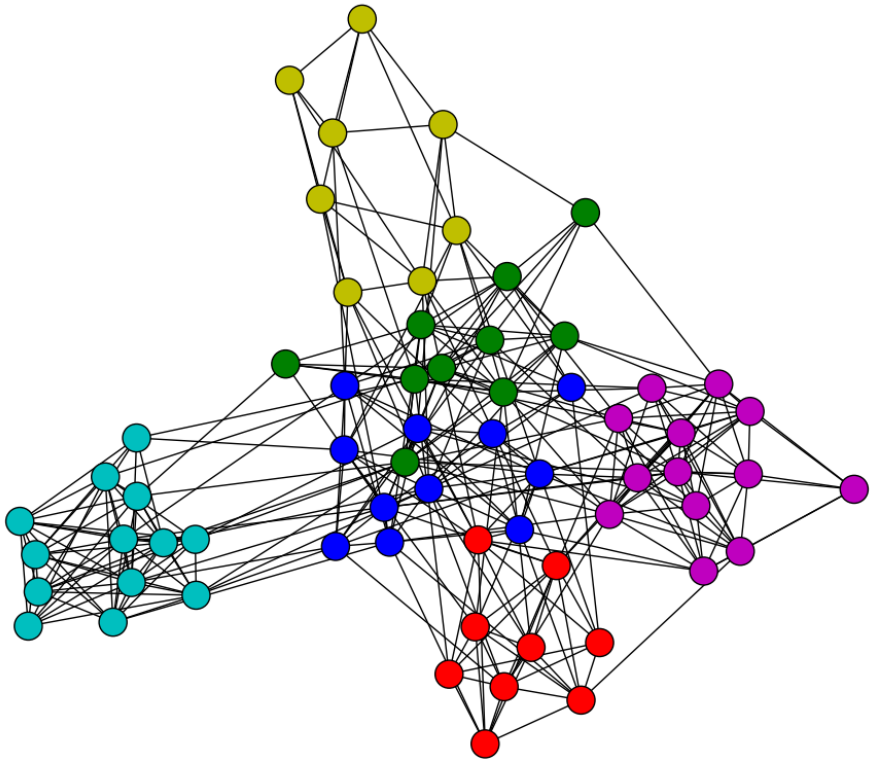
More terms

- "Stochastic block model", aka "Block-stochastic matrix":
  - Draw n_i nodes in block i
  - With probability p_{ij}, connect pairs (u,v) where u is in block i, v is in block j
  - Special, simple case: $p_{ii}=q_i$, and $pij=s$ for all $i\neq j$

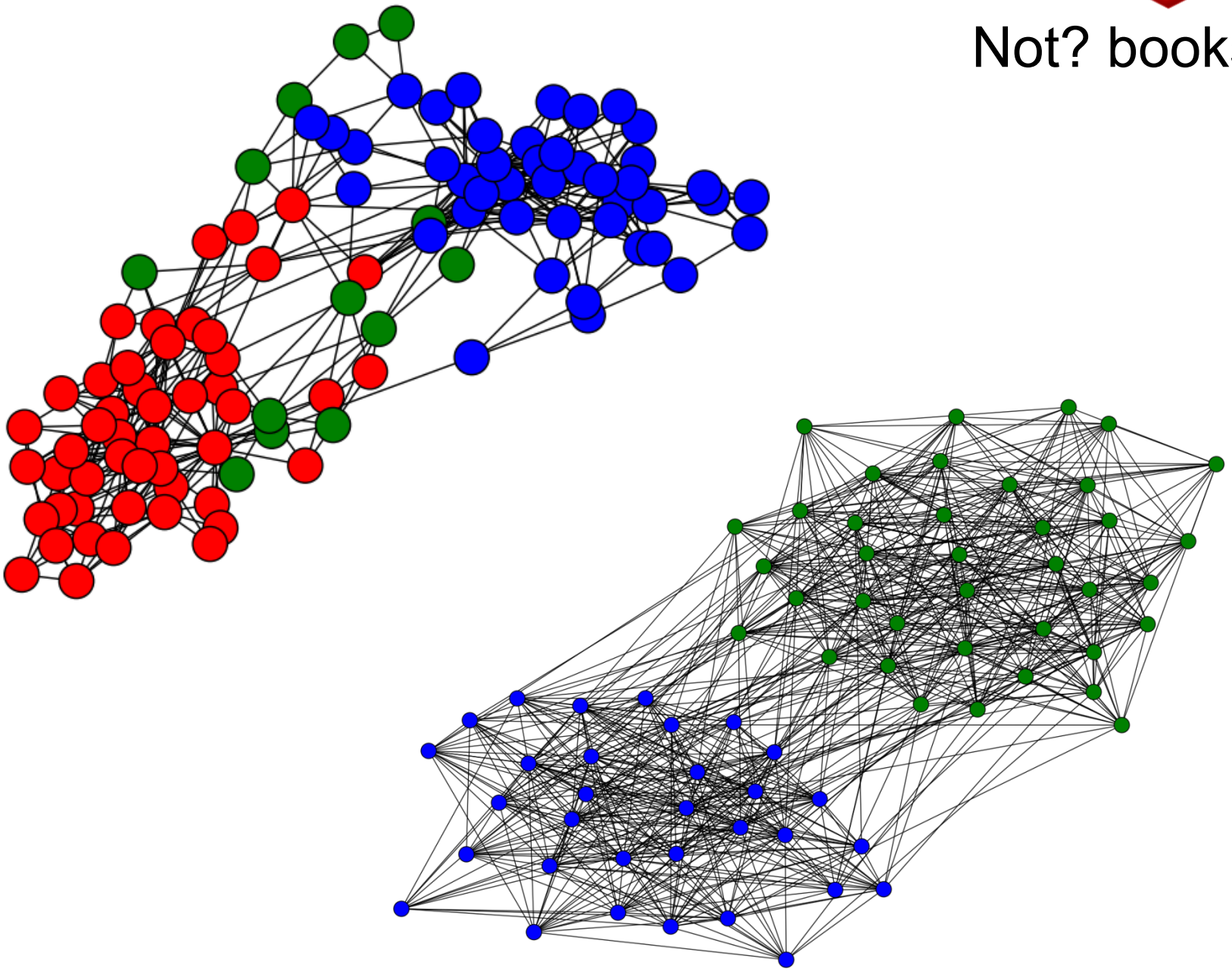
- Question: can you fit this model to a graph?
  - find each p_{ij} and latent node→block mapping



Not? football



Not? books

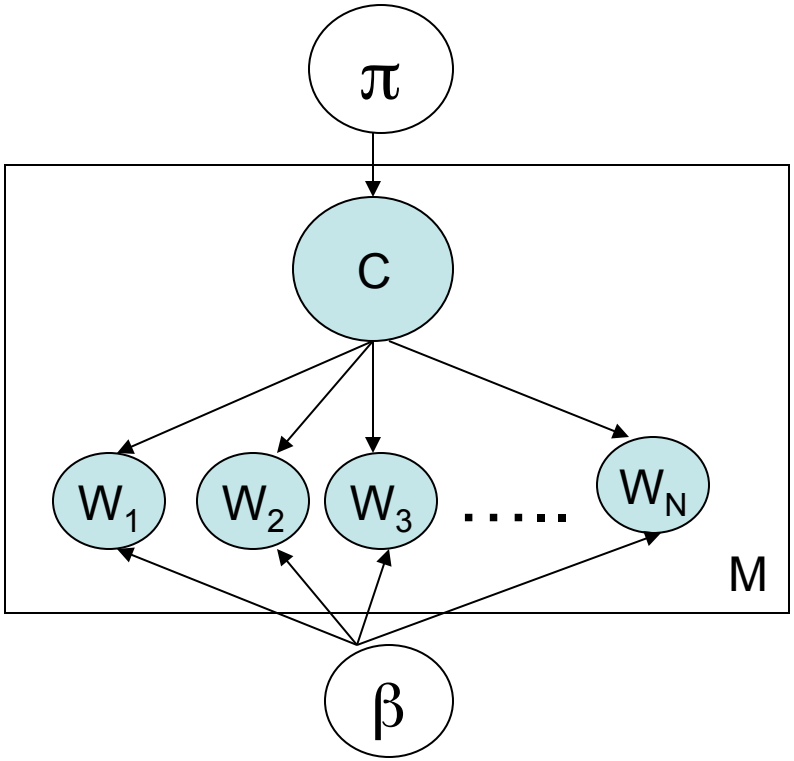


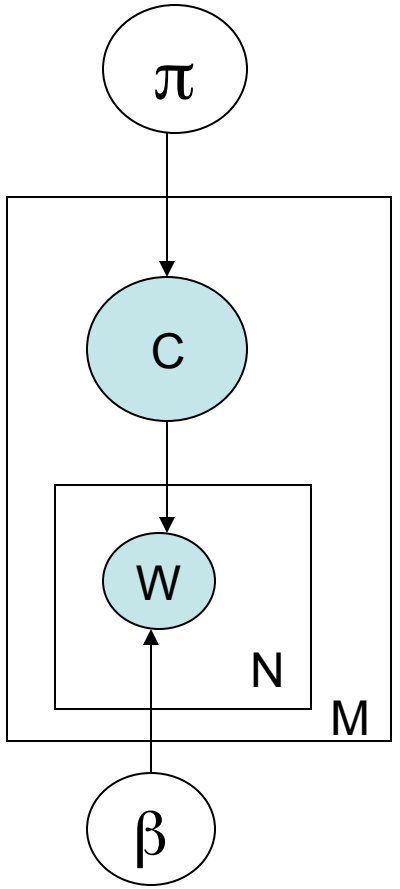
Outline

- 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

Review – supervised Naïve Bayes

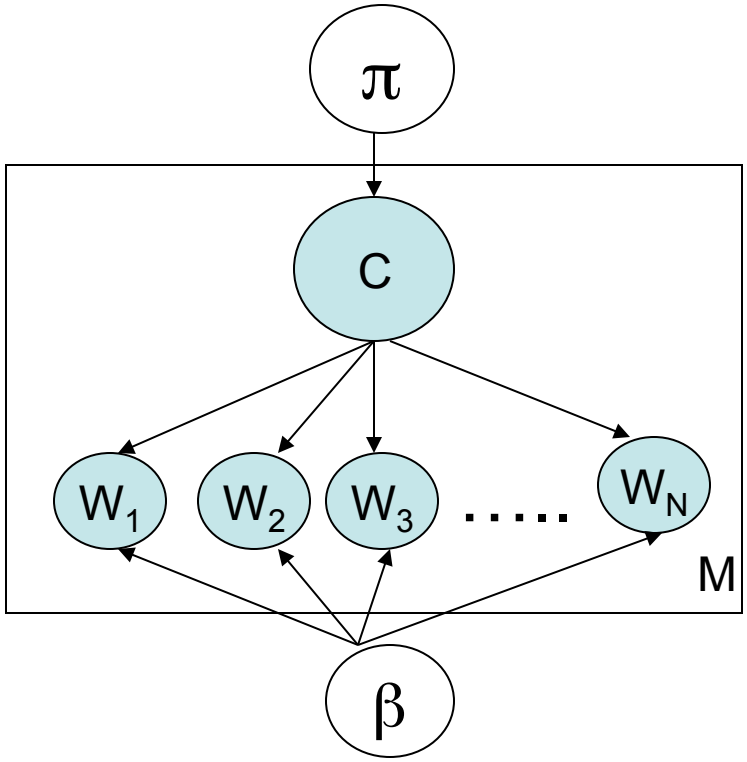
- Naïve Bayes Model: Compact representation





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)

Review – supervised Naïve Bayes

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

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

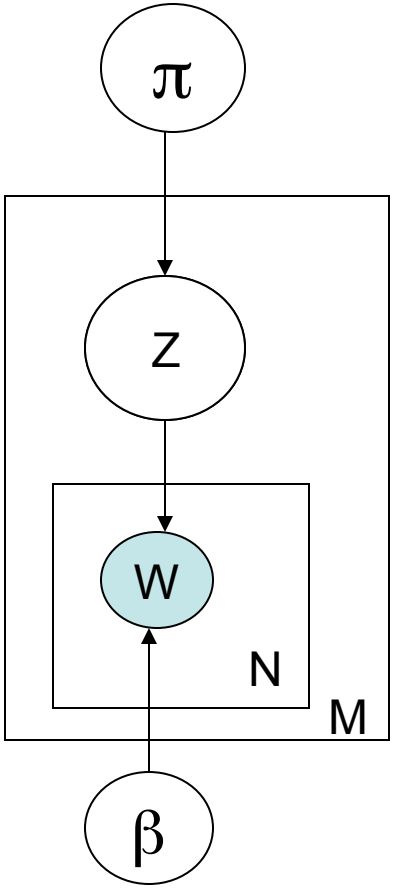
- Convex function: global optimum
- Solution:

$$\pi_C = \frac{\sum_{d=1}^{N} \delta_C(C_d)}{M}$$

$$\beta_{C,w} = \frac{\sum_{d:C_d=C} n(d,w)}{\sum_{d:C_d=C} \sum_w n(d,w)}$$

Review – unsupervised 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:

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

LDA

Latent Dirichlet Allocation

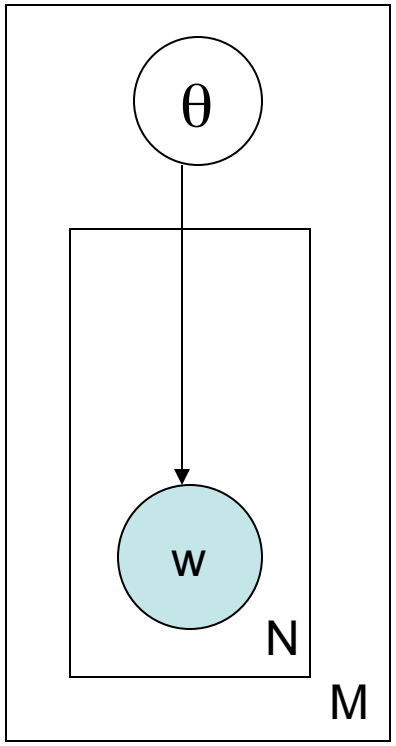
David M. Blei
Computer Science Division
University of California
Berkeley, CA 94720, USA

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

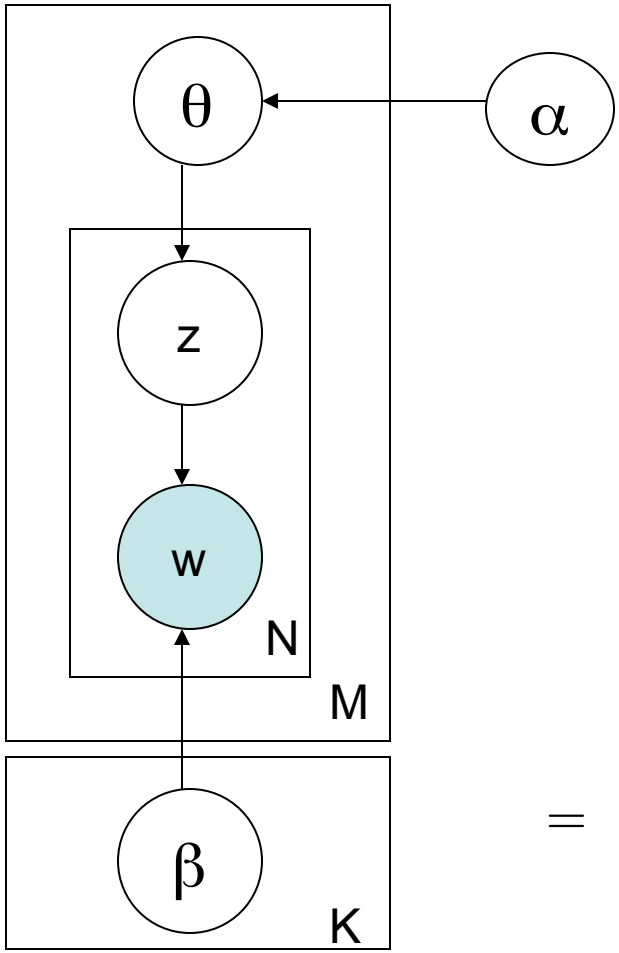
- Generate $\theta_d \sim D_1(\ldots)$
- For each word n = 1,⋯, N_d
  - •generate $w_n \sim D_2(\ \ | \vartheta_{d_n})$

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

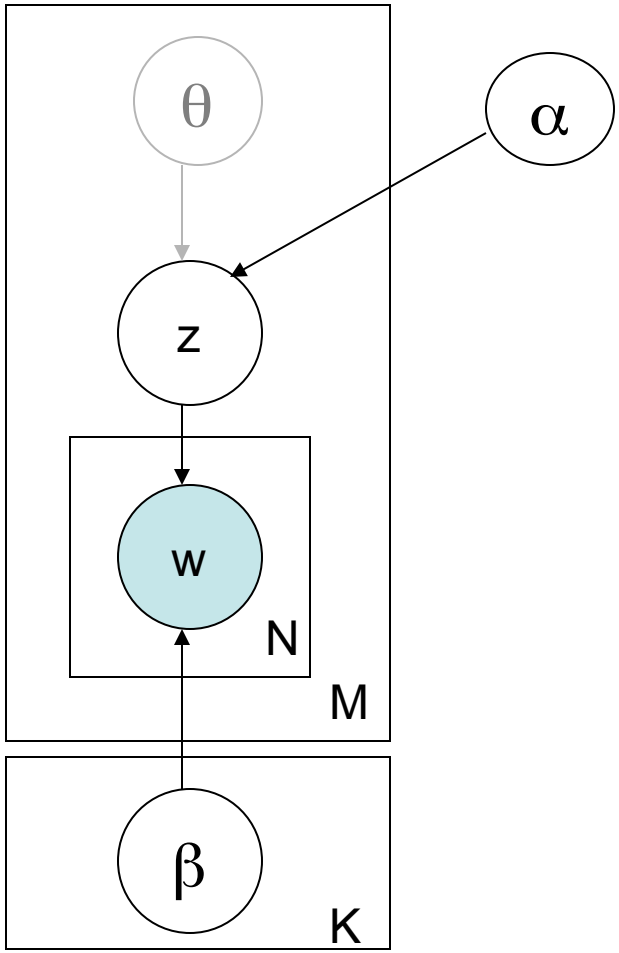


- For each document d = 1,⋯,M
  - Generate θ_d ~ Dir(| α)
  - For each position n = 1,⋯, N_d
    - generate z_n ~ Mult(| θ_d)
    - generate w_n ~ Mult(| β_{z_n})

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

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

- vs Naïve Bayes…



- 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: $n_m^{(k)} - 1; n_m - 1; n_k^{(t)} - 1; n_k - 1$
      sample topic index $\tilde{k} \sim p(z_i|\vec{z}_{\neg i}, \vec{w})$ ep):
      ⊡ 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

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

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

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

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

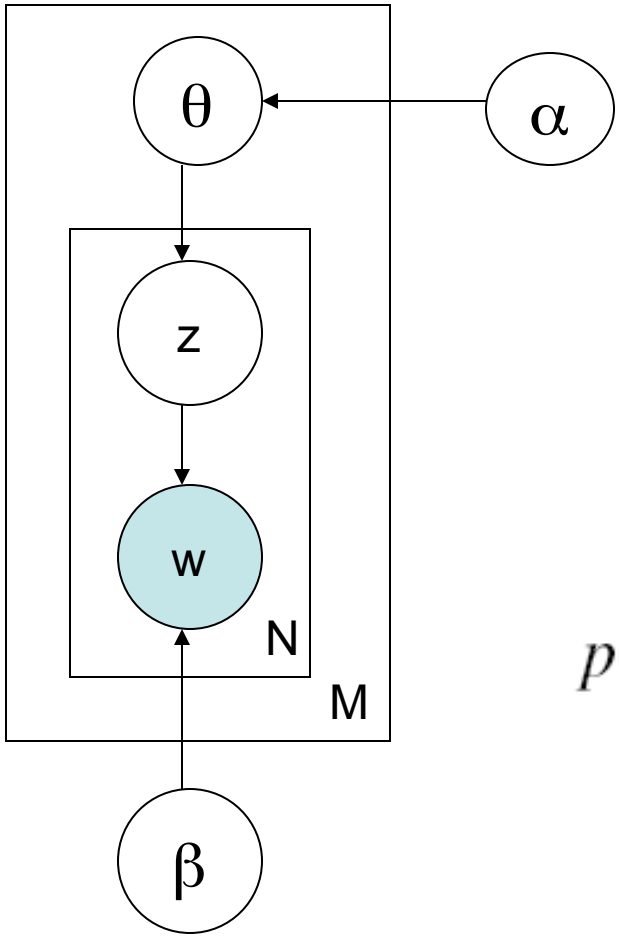
Called “collapsed Gibbs sampling” since you’ve marginalized away some variables

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.

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

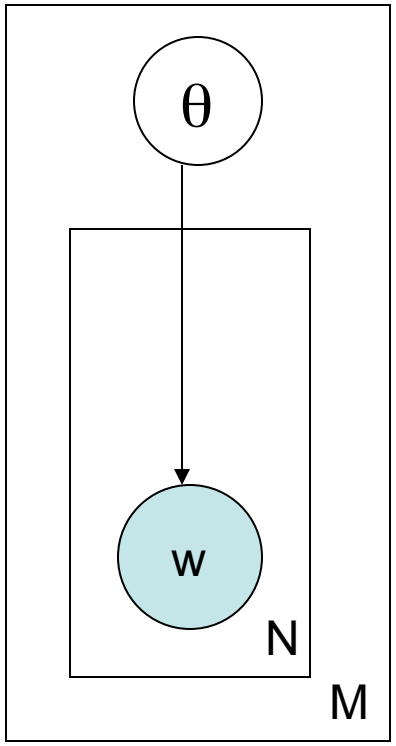
$$\propto \frac{n^{(t)}_{k,\neg i}+\beta_t}{\sum_{t=1}^{V} n^{(t)}_{k,\neg i}+\beta_t} \cdot \frac{n^{(t)}_{m,\neg i}+\alpha_k}{[\sum_{k=1}^{K} n^{(k)}_{m}+\alpha_k]-1}$$

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

Review - LDA

- Motivation



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

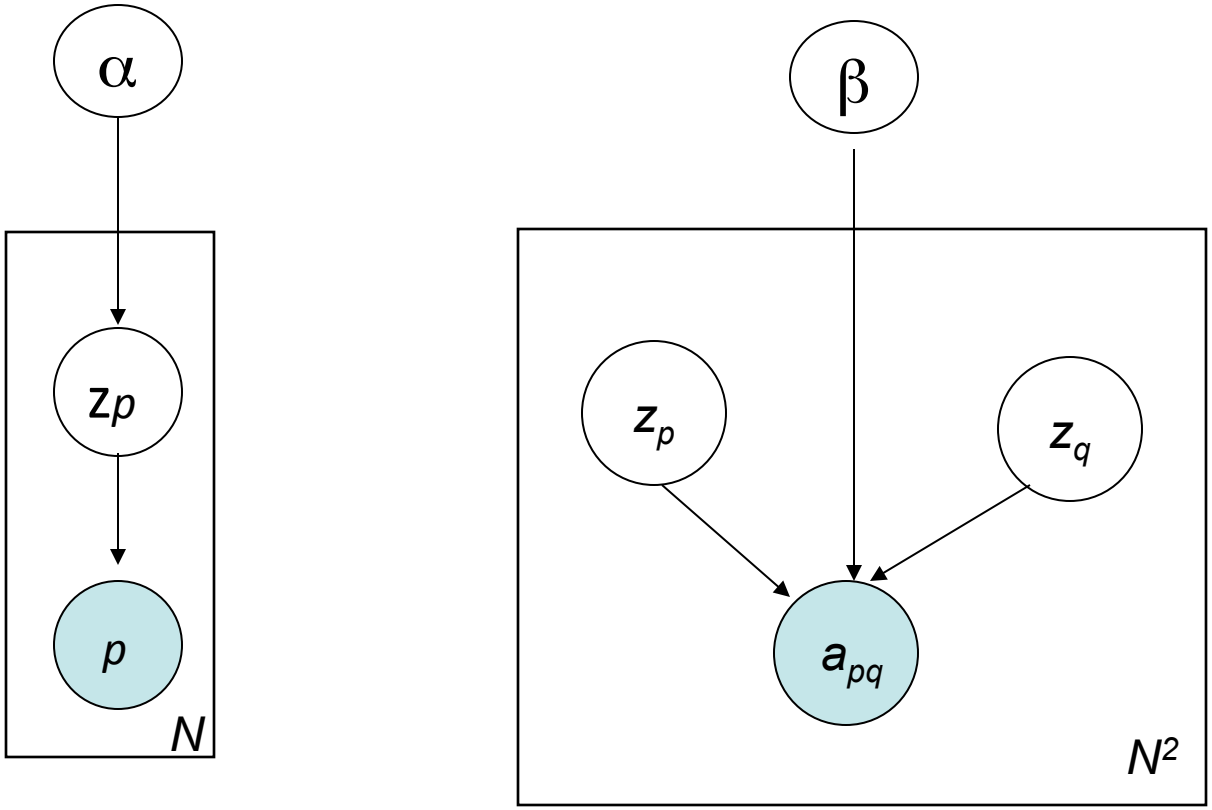
•For each document d = 1,⋯,M

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

Docs and words are *exchangeable*.

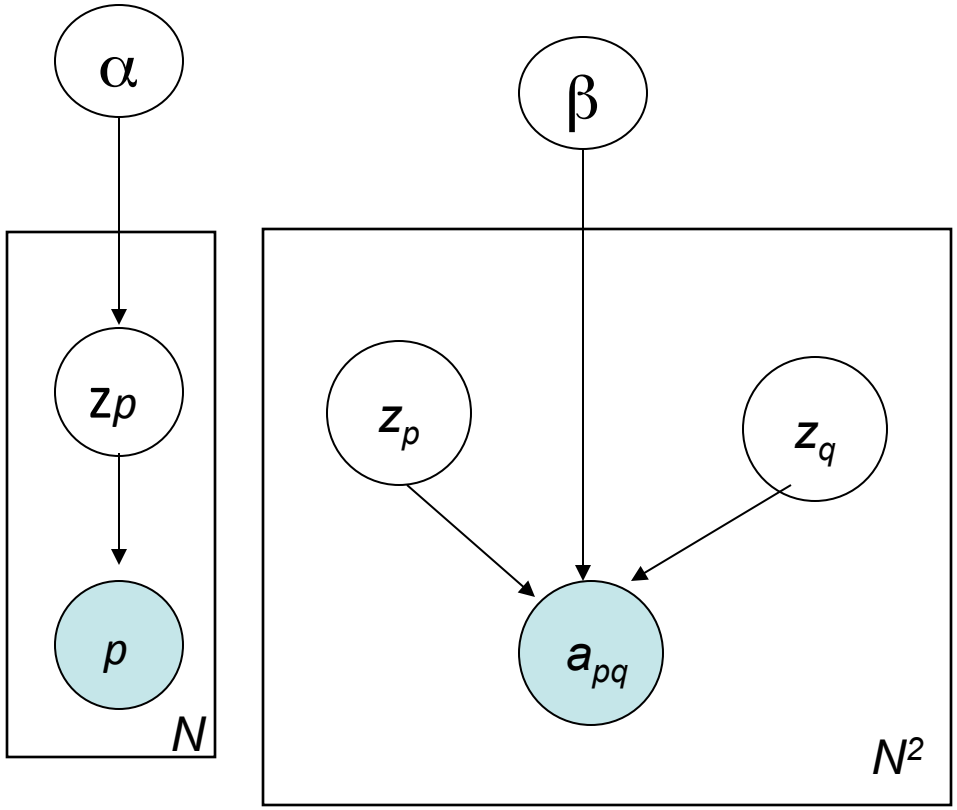
Stochastic Block models:

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*

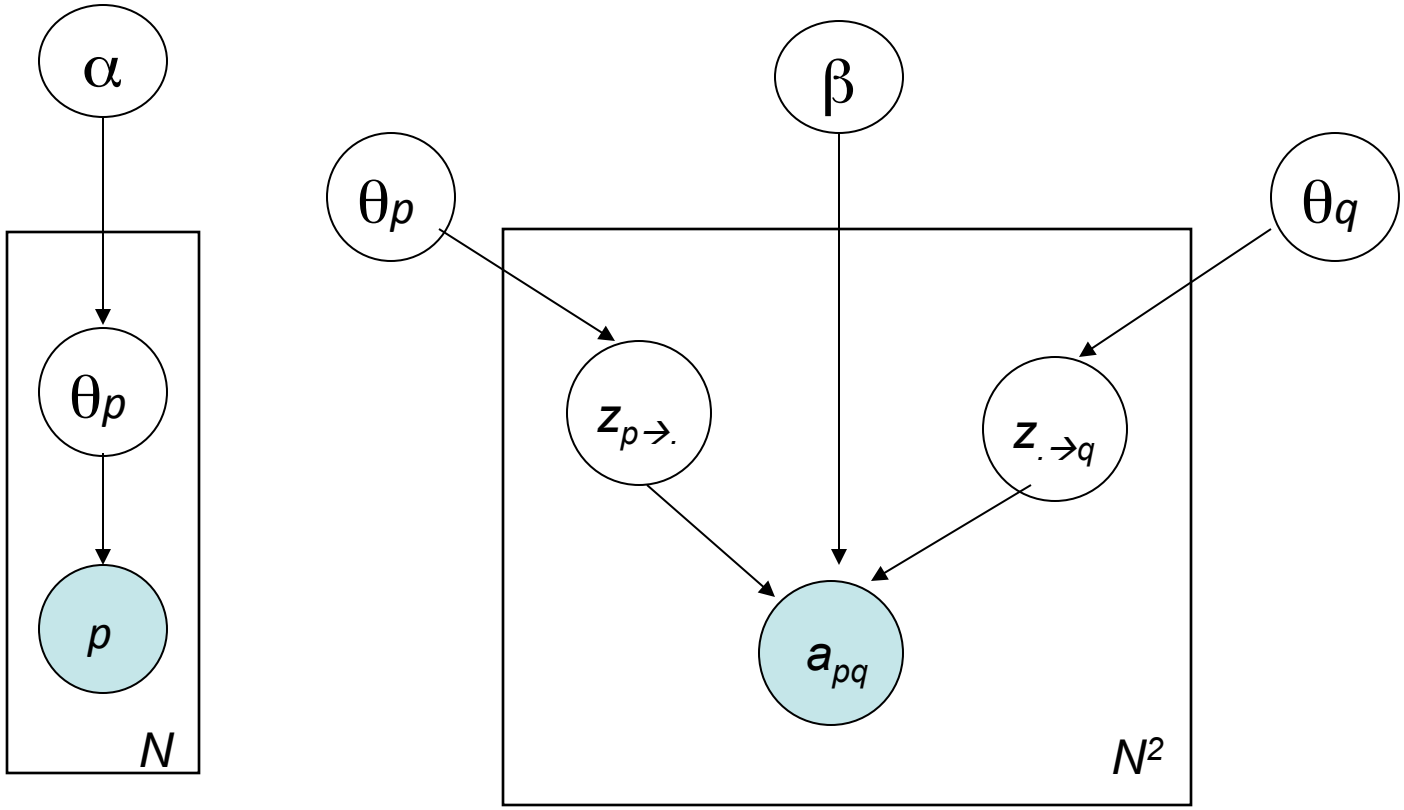


Gibbs sampling:

- 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

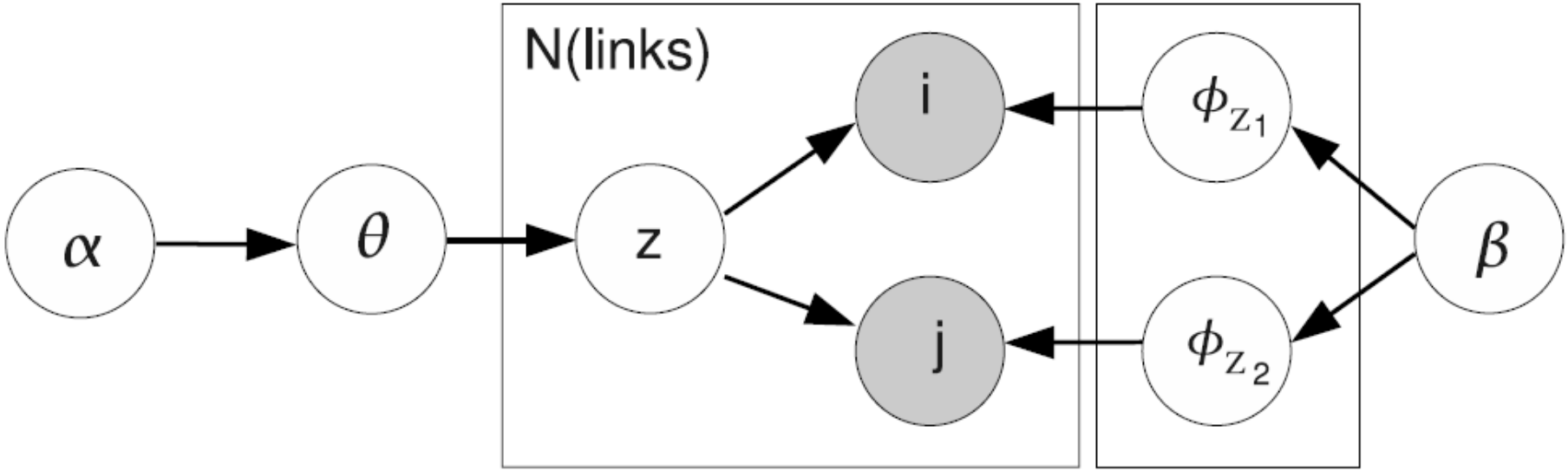
Mixed Membership Stochastic Block models



Airoldi et al, JMLR 2008

Parkkinen et al paper

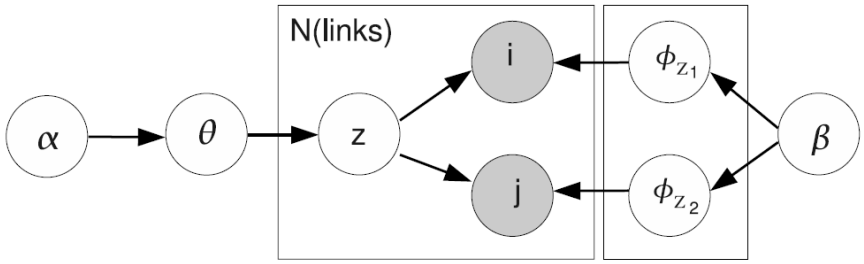
Another mixed membership block model



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

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

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

$q_{z1,.}$ = #outlinks in block z1

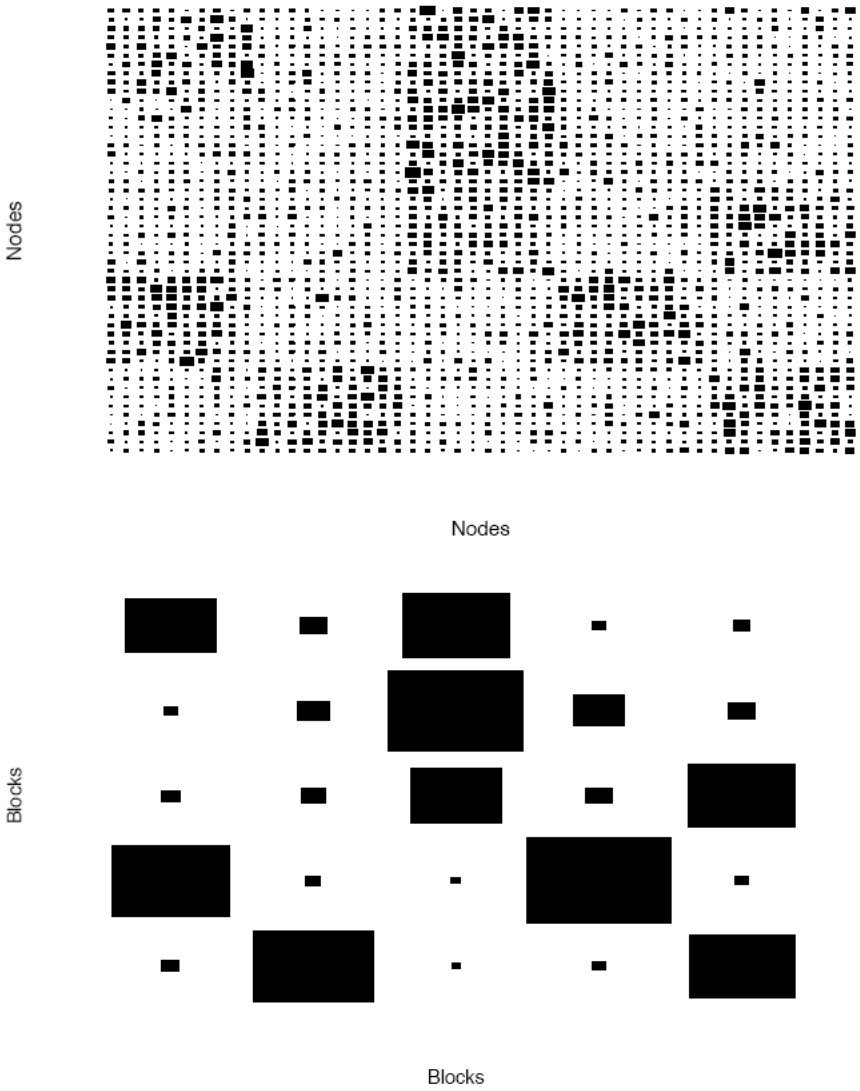
δ = indicator for diagonal

M = #nodes

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

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

Another mixed membership block model



Experiments

Table 1: Dataset Statistics (N/E/C indicates Nodes / Edges / Clusters)

(a) Social network

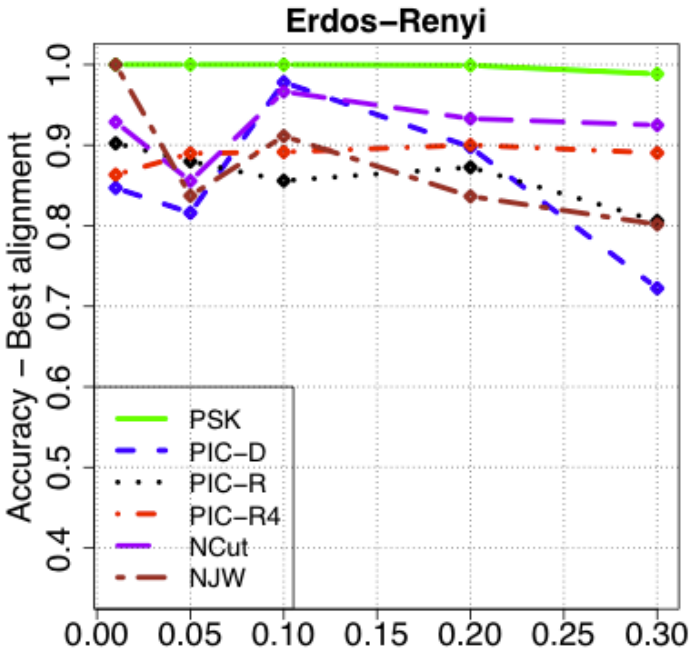
Dataset	N/E/C	Dataset	N/E/C
karate	34 / 156 / 2	umbc	404 / 4764 / 2
polbooks	105 / 882 / 3	mgemail	280 / 1344 / 55
dolphin	62 / 318 / 2	citeseer	2114 / 7396 / 6
football	115 / 1226 / 10	cora	2485 / 10138 / 7
msp	4324 / 37254 / 2		
ag	1222 / 33428 / 2		
senate	98 / 9506 / 2		

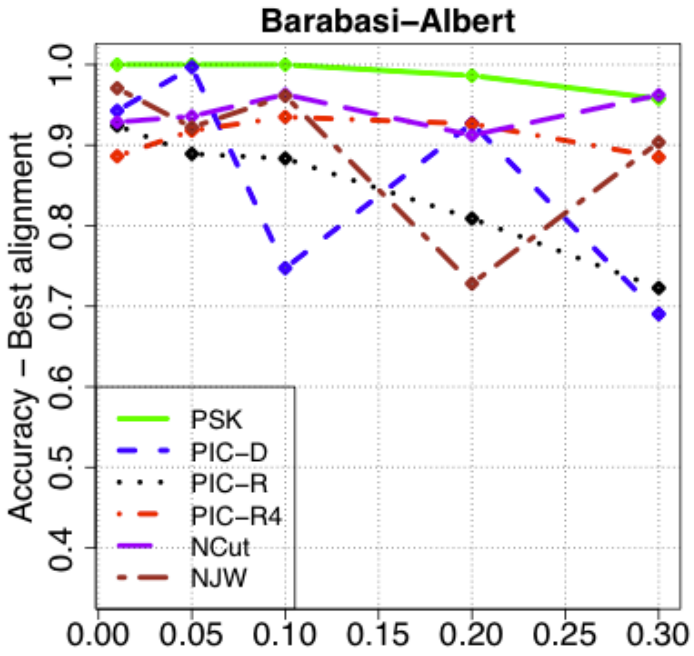
(b) Author disambiguation

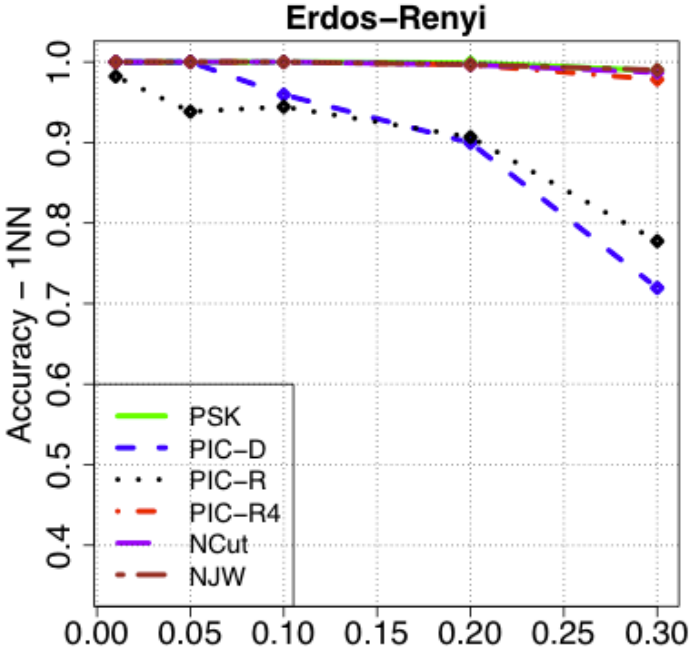
Dataset	N/E/C	Dataset	N/E/C
jsmith	4120 / 21452 / 30	jrobinson	686 / 2846 / 12
akumar	801 / 2476 / 14	ktanaka	827 / 2758 / 10
cchen	424 / 1558 / 16	mbrown	579 / 2112 / 13
djohnson	1381 / 5344 / 15	mmiller	2106 / 9918 / 12
jmartin	424 / 1558 / 16	jlee	5820 / 23110 / 100
agupta	2485 /10208 / 26	ychen	5472 / 25584 / 71
mjones	961 / 3450 / 13	slee	5963 / 23086 / 86

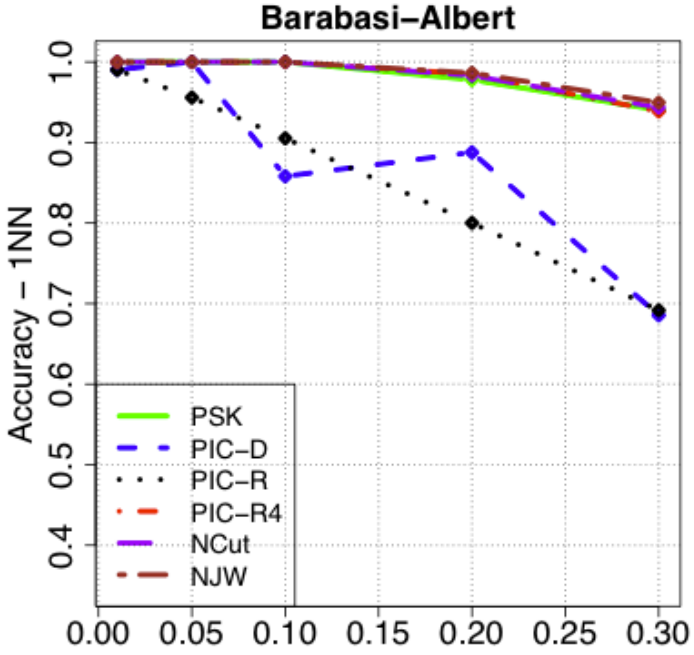
+ lots of synthetic data

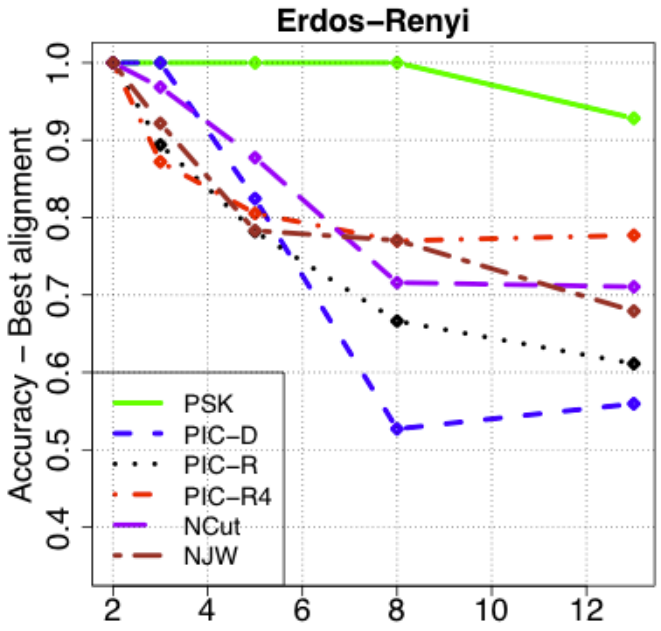
Balasubramanyan, Lin, Cohen, NIPS w/s 2010

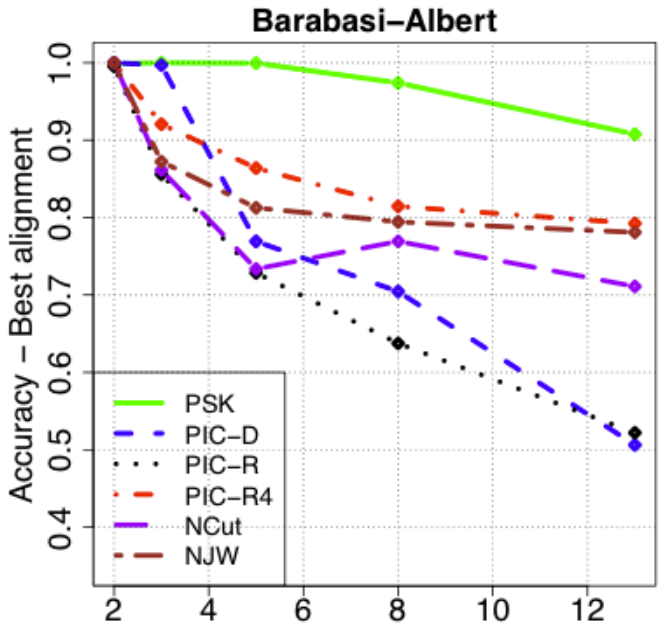

Erdos–Renyi
Accuracy – Best alignment
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.00
0.05
0.10
0.15
0.20
0.25
0.30
PSK
PIC–D
PIC–R
PIC–R4
NCut
NJW

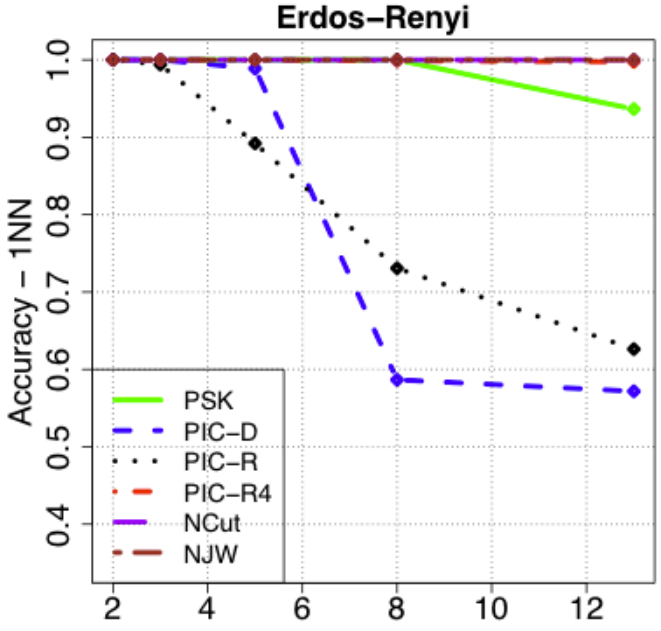

Barabasi–Albert
Accuracy – Best alignment
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.00
0.05
0.10
0.15
0.20
0.25
0.30
PSK
PIC–D
PIC–R
PIC–R4
NCut
NJW

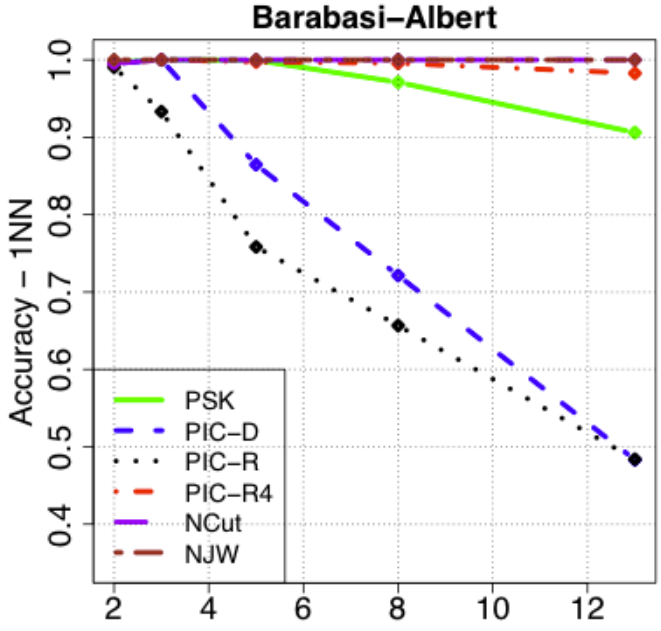

Erdos–Renyi
Accuracy – 1NN
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.00
0.05
0.10
0.15
0.20
0.25
0.30
PSK
PIC–D
PIC–R
PIC–R4
NCut
NJW


Barabasi–Albert
Accuracy – 1NN
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.00
0.05
0.10
0.15
0.20
0.25
0.30
PSK
PIC–D
PIC–R
PIC–R4
NCut
NJW


Erdos–Renyi
Accuracy – Best alignment
1.0
0.9
0.8
0.7
0.6
0.5
0.4
2
4
6
8
10
12
PSK
PIC–D
PIC–R
PIC–R4
NCut
NJW


Barabasi–Albert
Accuracy – Best alignment
1.0
0.9
0.8
0.7
0.6
0.5
0.4
2
4
6
8
10
12
PSK
PIC–D
PIC–R
PIC–R4
NCut
NJW


Erdos–Renyi
Accuracy – 1NN
1.0
0.9
0.8
0.7
0.6
0.5
0.4
2
4
6
8
10
12
PSK
PIC–D
PIC–R
PIC–R4
NCut
NJW


Barabasi–Albert
Accuracy – 1NN
1.0
0.9
0.8
0.7
0.6
0.5
0.4
2
4
6
8
10
12
PSK
PIC–D
PIC–R
PIC–R4
NCut
NJW

Experiments

(c) Best alignment: Social networks

Dataset	PSK	PIC_D	PIC_R	PIC_{R4}	NCut	NJW
Karate	**1.00**	0.91	0.93	0.95	0.95	0.95
Dolphin	0.90	**0.98**	0.98	**0.98**	**0.98**	**0.98**
UMBC	0.95	0.93	0.95	0.95	0.95	**0.96**
AG	**0.95**	0.91	0.94	0.94	0.52	0.51
MSP	**0.88**	0.63	0.63	0.63	0.63	0.64
Senate	0.98	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
PolBook	0.78	0.80	0.81	**0.83**	0.82	0.80
Football	**0.76**	0.47	0.51	0.66	0.72	0.67
MGEmail	0.28	0.39	0.40	**0.64**	0.59	0.56
CiteSeer	0.33	0.51	0.48	**0.55**	0.48	0.52
Cora	**0.47**	0.46	0.40	0.45	0.29	0.42
Average	0.75	0.73	0.73	**0.78**	0.72	0.73

(d) Best alignment: Author disambiguation

Dataset	PSK	PIC_D	PIC_R	PIC_{R4}	NCut	NJW
AGupta	0.13	0.26	0.24	**0.37**	0.26	0.34
AKumar	0.20	0.29	0.31	0.37	0.35	**0.40**
CChen	0.30	0.43	0.44	**0.53**	0.24	0.50
DJohnson	0.15	0.24	0.33	0.46	**0.47**	0.35
JLee	0.11	0.20	0.23	**0.41**	0.17	0.39
JMartin	0.28	0.42	0.43	**0.53**	0.25	0.49
JRobinson	0.26	0.37	0.42	**0.49**	0.26	0.48
JSmith	0.11	0.22	0.21	0.41	0.31	**0.42**
KTanaka	0.19	0.36	0.41	0.45	**0.45**	0.43
MBrown	0.21	0.35	0.41	**0.52**	0.47	0.50
MJones	0.19	0.29	0.34	0.38	**0.38**	0.35
MMiller	0.14	0.30	0.41	0.52	0.52	**0.53**
SLee	0.08	0.19	0.23	**0.41**	0.23	0.39
YChen	0.10	0.23	0.28	**0.47**	0.23	0.46
Average	0.18	0.30	0.34	**0.45**	0.33	0.43

Balasubramanyan, Lin, Cohen, NIPS w/s 2010

Experiments

(e) 1-NN: Social networks

Dataset	PSK	PIC_D	PIC_R	PIC_{R4}	NCut	NJW
Karate	**1.00**	**1.00**	0.99	0.99	1.00	0.97
Dolphin	0.89	0.95	0.95	0.95	0.95	**0.98**
UMBC	0.92	0.93	0.93	0.93	0.92	**0.94**
AG	0.92	**0.94**	0.93	0.93	0.88	0.89
MSP	0.84	0.76	0.73	**0.86**	0.64	0.59
Senate	0.97	1.00	**1.00**	**1.00**	**1.00**	**1.00**
PolBook	0.79	0.68	0.76	0.80	**0.84**	0.78
Football	0.89	0.43	0.45	0.85	0.94	**0.95**
MGEmail	0.22	0.27	0.26	0.72	0.80	**0.81**
CiteSeer	0.34	0.55	0.54	**0.71**	0.69	0.66
Cora	0.45	0.56	0.51	**0.80**	0.47	0.75
Average	0.75	0.73	0.73	**0.87**	0.83	0.85

(f) 1-NN: Author disambiguation

Dataset	PSK	PIC_D	PIC_R	PIC_{R4}	NCut	NJW
AGupta	0.68	0.74	0.72	**0.95**	0.79	0.91
AKumar	0.82	0.69	0.74	**0.85**	0.79	0.81
CChen	0.77	0.73	0.74	**0.89**	0.75	0.85
DJohnson	0.81	0.81	0.83	**0.95**	0.85	0.92
JLee	0.55	0.61	0.68	**0.92**	0.79	0.91
JMartin	0.77	0.73	0.73	**0.88**	0.75	0.85
JRobinson	0.86	0.75	0.80	**0.92**	0.83	0.85
JSmith	0.65	0.75	0.67	**0.93**	0.85	0.91
KTanaka	0.81	0.84	0.86	**0.95**	0.90	0.90
MBrown	0.83	0.78	0.82	**0.93**	0.86	0.89
MJones	0.79	0.69	0.71	**0.91**	0.90	0.89
MMiller	0.81	0.83	0.81	**0.99**	0.97	0.98
SLee	0.59	0.69	0.77	**0.92**	0.85	0.92
YChen	0.57	0.73	0.79	**0.95**	0.84	0.94
Average	0.74	0.74	0.76	**0.92**	0.84	0.90

Balasubramanyan, Lin, Cohen, NIPS w/s 2010