Introduction to Graphical Models for Data Mining

Arindam Banerjee
banerjee@cs.umn.edu

Dept of Computer Science & Engineering
University of Minnesota, Twin Cities

16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining

July 25, 2010

Introduction

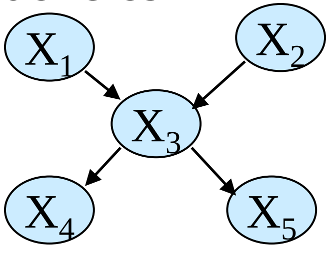
- Graphical Models
  - Brief Overview
- Part I: Tree Structured Graphical Models
  - Exact Inference
- Part II: Mixed Membership Models
  - Latent Dirichlet Allocation
  - Generalizations, Applications
- Part III: Graphical Models for Matrix Analysis
  - Probabilistic Matrix Factorization
  - Probabilistic Co-clustering
  - Stochastic Block Models

Graphical Models: What and Why

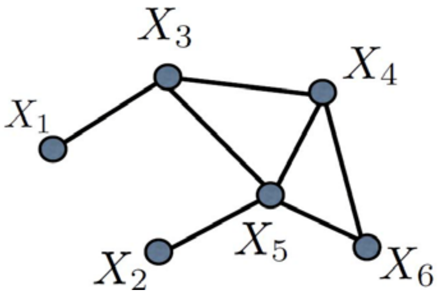
- Statistical Data Analaysis
  - Build diagnostic/predictive models from data
  - Uncertainty quantification based on (minimal) assumptions
- The I.I.D. assumption
  - Data is independently and identically distributed
  - Example: Words in a doc drawn i.i.d. from the dictionary
- Graphical models
  - Assume (graphical) dependencies between (random) variables
  - Closer to reality, domain knowledge can be captured
  - Learning/inference is much more difficult

Flavors of Graphical Models

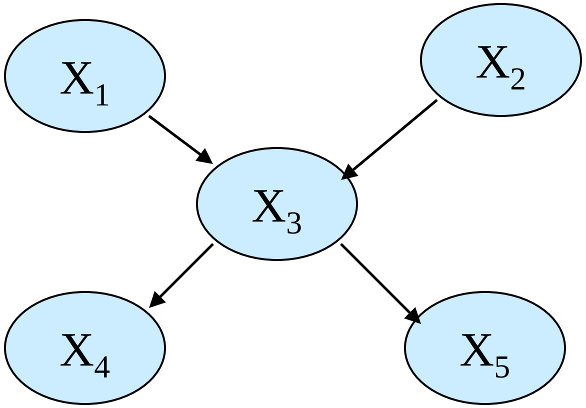
- Basic nomenclature
  - Node = random variable, maybe observed/hidden
  - Edge = statistical dependency
- Two popular flavors: 'Directed' and 'Undirected'
- Directed Graphs
  - A *directed* graph between random variables, causal dependencies
  - Example: Bayesian networks, Hidden Markov Models
  - Joint distribution is a product of P(child|parents)



- Undirected Graphs
  - An *undirected* graph between random variables
  - Example: Markov/Conditional random fields
  - Joint distribution in terms of potential functions



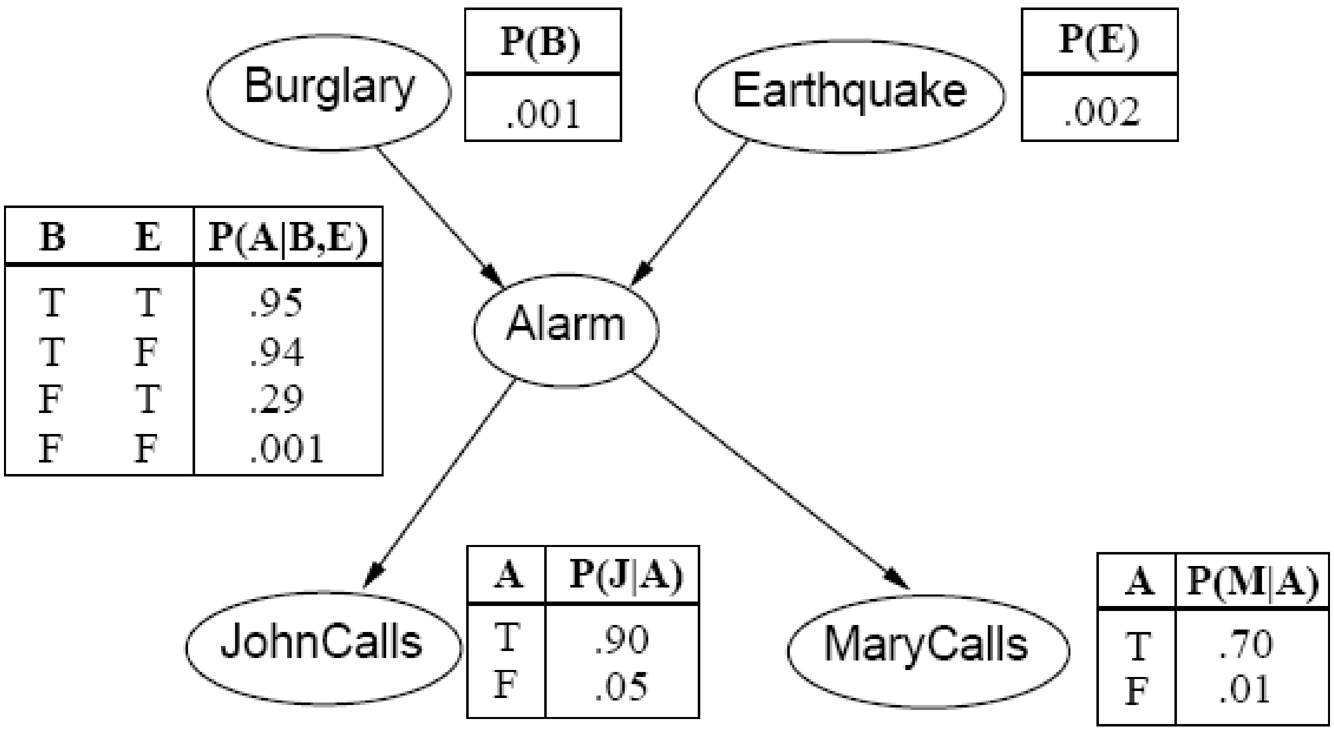
Bayesian Networks



- Joint distribution in terms of *P(X|Parents(X))*

$$
\begin{aligned}
P(X_1, \ldots, X_n) &= \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1}) \\
&= \prod_{i=1}^{n} P(X_i | Parents(X_i))
\end{aligned}
$$

Example I: Burglary Network



P(B)
.001

P(E)
.002

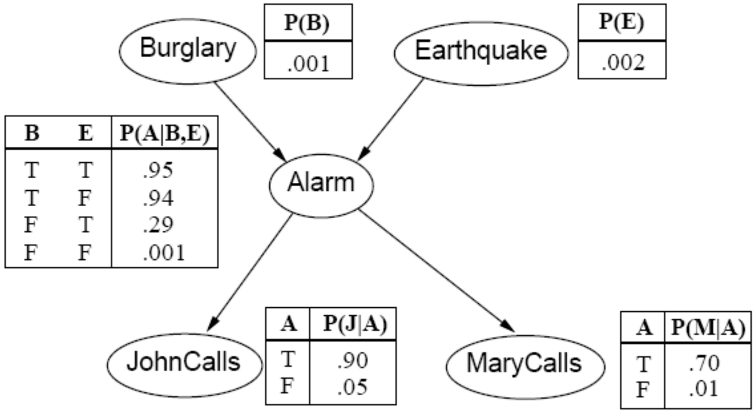
B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

This and several other examples are from the Russell-Norvig AI book

Computing Probabilities of Events



- Probability of any event can be computed:

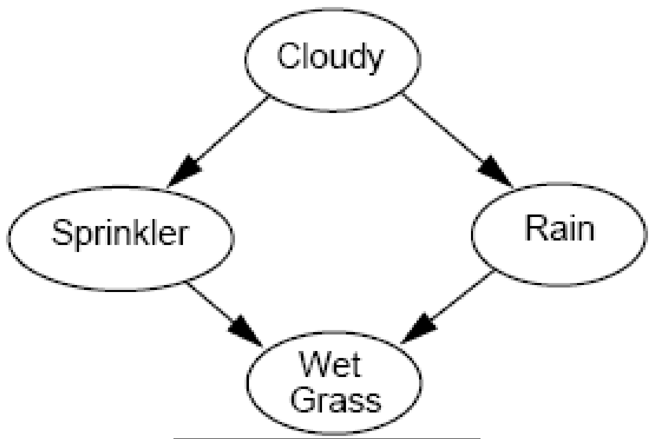
$$P(B,E,A,J,M) = P(B)\ P(E|B)\ P(A|B,E)\ P(J|B,E,A)\ P(M|B,E,A,J)$$
$$= P(B)\ P(E)\ P(A|B,E)\ P(J|A)\ P(M|A)$$

- Example:

$$P(b,\neg e,a,\neg j,m) = P(b)\ P(\neg e) P(a|b,\neg e)\ P(\neg j|a)\ P(m|a)$$

Example II: Rain Network

P(C)
.50

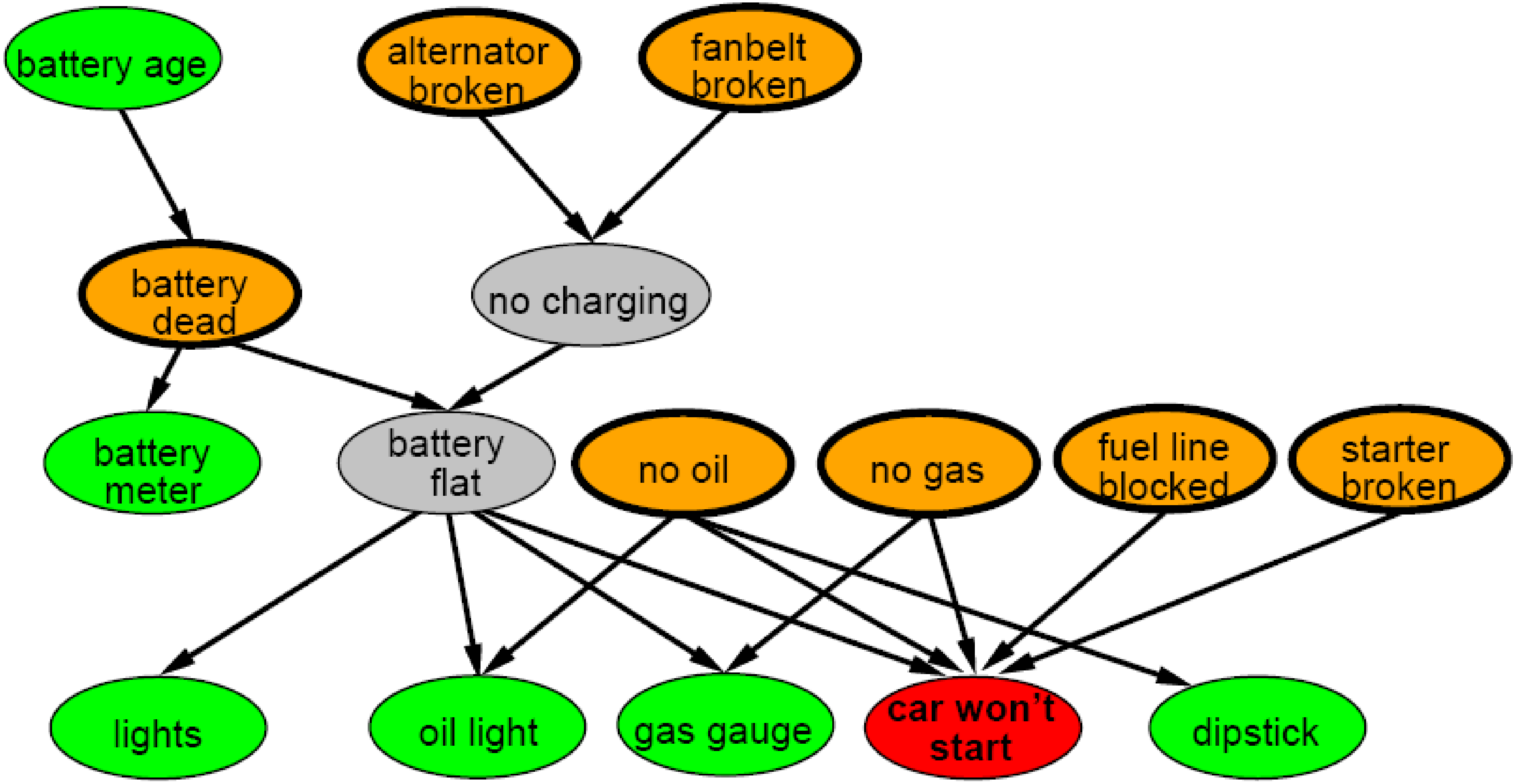


C	P(S|C)
T	.10
F	.50

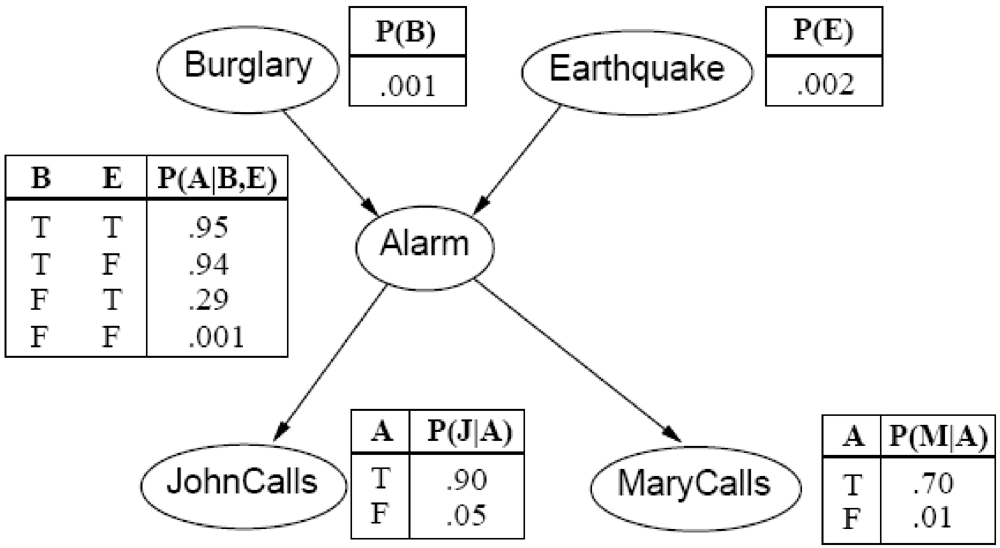
C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Example III: “Car Won’t Start” Diagnosis



Inference



- Some variables in the Bayes net are observed
  - the evidence/data, e.g., John has not called, Mary has called
- Inference
  - How to compute value/probability of other variables
  - Example: What is the probability of Burglary, i.e., $P(b|\neg j,m)$

Inference Algorithms

- Graphs without loops: Tree-structured Graphs
  - Efficient exact inference algorithms are possible
  - Sum-product algorithm, and its special cases
    - Belief propagation in Bayes nets
    - Forward-Backward algorithm in Hidden Markov Models (HMMs)
- Graphs with loops
  - Junction tree algorithms
    - Convert into a graph without loops
    - May lead to exponentially large graph
  - Sum-product/message passing algorithm, ‘disregarding loops’
    - Active research topic, correct convergence ‘not guaranteed’
    - Works well in practice
  - Approximate inference

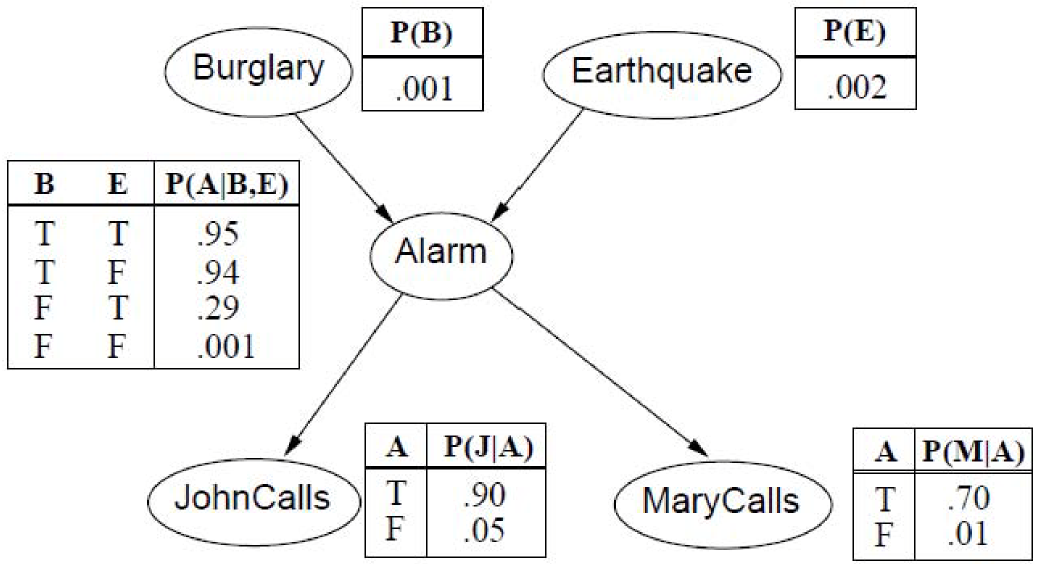
Approximate Inference

- Variational Inference
  - Deterministic approximation
  - Approximate complex true distribution over latent variables
  - Replace with family of simple/tractable distributions
    - Use the best approximation in the family
  - Examples: Mean-field, Bethe, Kikuchi, Expectation Propagation
- Stochastic Inference
  - Simple sampling approaches
  - Markov Chain Monte Carlo methods (MCMC)
    - Powerful family of methods
  - Gibbs sampling
    - Useful special case of MCMC methods

Part I: Tree Structured Graphical Models

- The Inference Problem
- Factor Graphs and the Sum-Product Algorithm
- Example: Hidden Markov Models
- Generalizations

The Inference Problem



How can we compute $P(b|j, m)$?

Complexity of Naïve Inference

- Simple query can be answered using Bayes rule
  - From Bayes Rule

$$P(b|j,m) = \frac{P(b,j,m)}{P(j,m)}$$

  - Each marginal can be obtained from the joint distribution

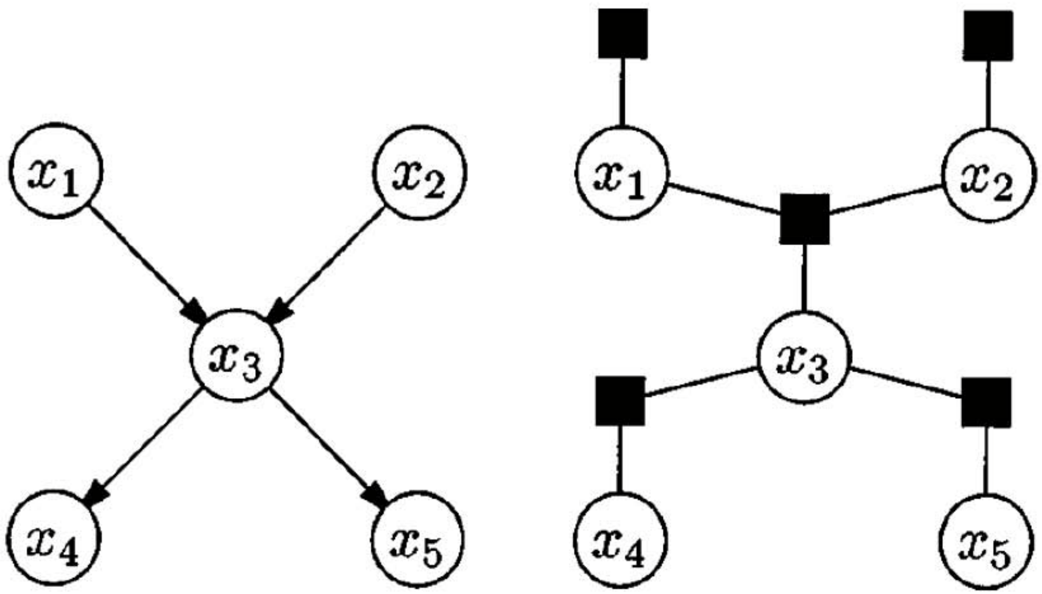
$$\begin{aligned} P(b,j,m) &= \sum_E \sum_A P(b,E,A,j,m) \\ P(j,m) &= \sum_B \sum_E \sum_A P(B,E,A,j,m) \end{aligned}$$

  - Each term can be written as product of conditionals

$$P(b,E,A,j,m) = P(b)P(E)P(A|b,E)P(j|A)P(m|A)$$

- The complexity of the simple approach is $O(n2^n)$

Bayes Nets to Factor Graphs

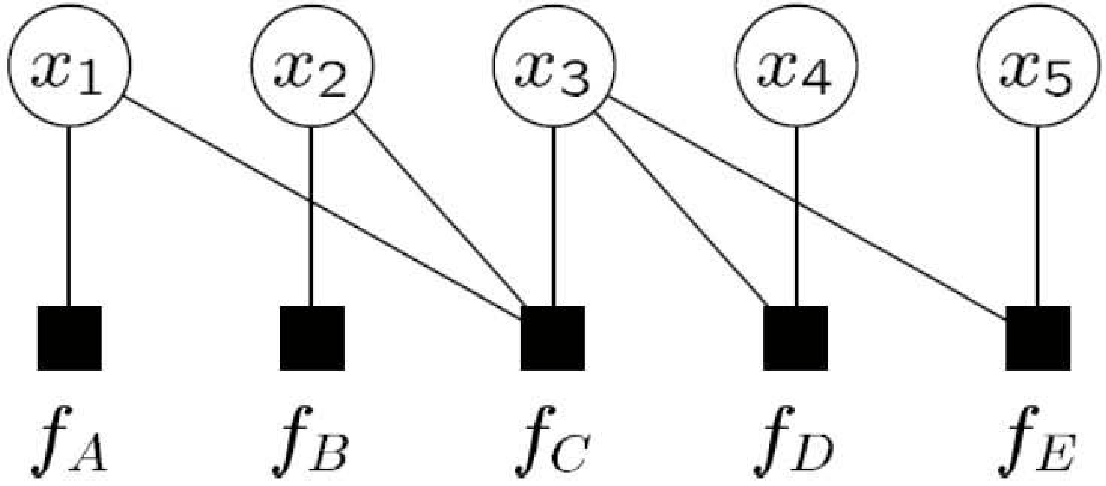


$f_A(x_1) = p(x_1)$ $\quad$ $f_B(x_2) = p(x_2)$ $\quad$ $f_C(x_1, x_2, x_3) = p(x_3|x_1, x_2)$

$f_D(x_3, x_4) = p(x_4|x_3)$ $\quad$ $f_E(x_3, x_5) = p(x_5|x_3)$

Factor Graphs: Product of Local Functions

$$g(x_1, x_2, x_3, x_4, x_5) = f_A(x_1) f_B(x_2) f_C(x_1, x_2, x_3) f_D(x_3, x_4) f_E(x_3, x_5)$$



Marginalize Product of Functions (MPF)

- Marginalize product of functions

$$g(x_1, x_2, x_3, x_4, x_5) = f_A(x_1) f_B(x_2) f_C(x_1, x_2, x_3) f_D(x_3, x_4) f_E(x_3, x_5)$$

- Computing marginal functions

$$g_i(x_i) = \sum_{\sim x_i} g(x_1, x_2, x_3, x_4, x_5)$$

- The "not-sum" notation

$$\sum_{\sim x_2} f(x_1, x_2, x_3) = \sum_{x_1, x_3} f(x_1, x_2, x_3)$$

MPF using Distributive Law

- We focus on two examples: $g_1(x_1)$ and $g_3(x_3)$
- Main Idea: Distributive law

$$ab + ac = a(b+c)$$

- For $g_1(x_1)$, we have

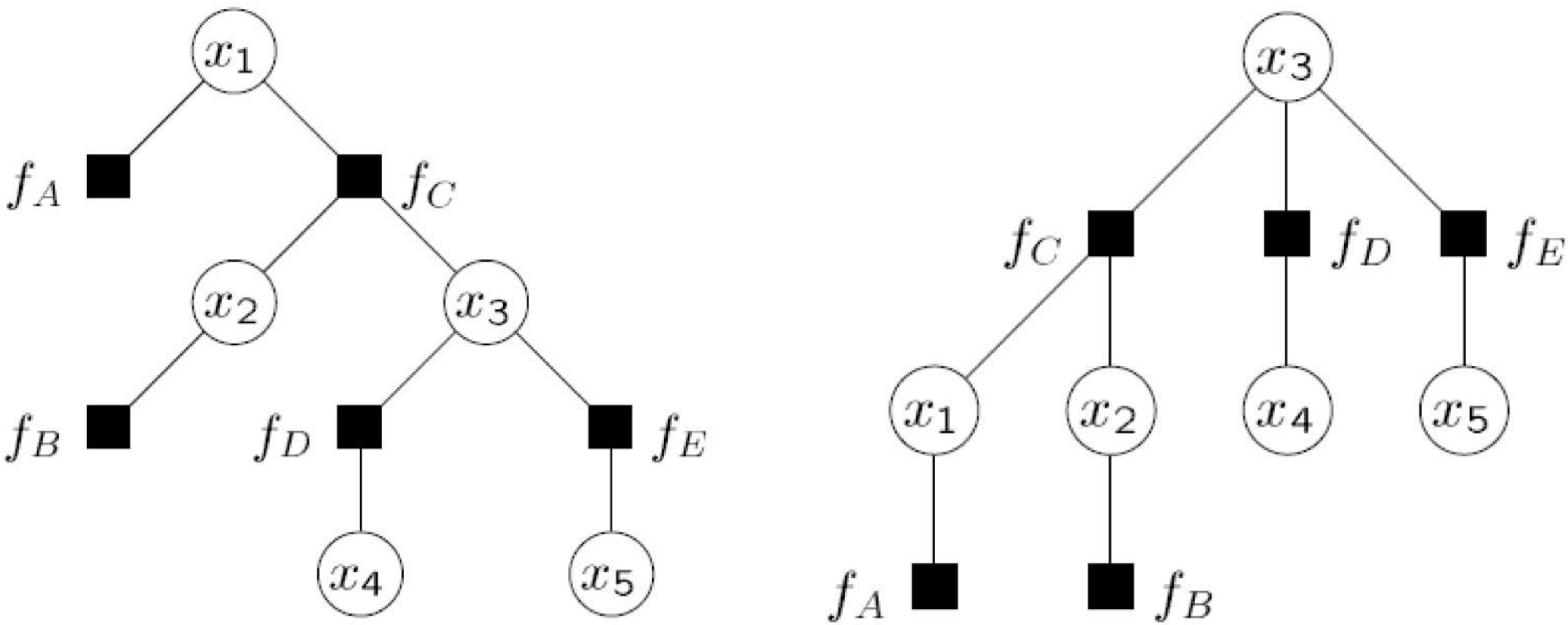
$$g_1(x_1) = f_A(x_1) \sum_{\sim x_1} \left(f_B(x_2) f_C(x_1, x_2, x_3) \left(\sum_{\sim x_3} f_D(x_3, x_4) \right) \left(\sum_{\sim x_3} f_E(x_3, x_5) \right) \right)$$

- For $g_3(x_3)$, we have

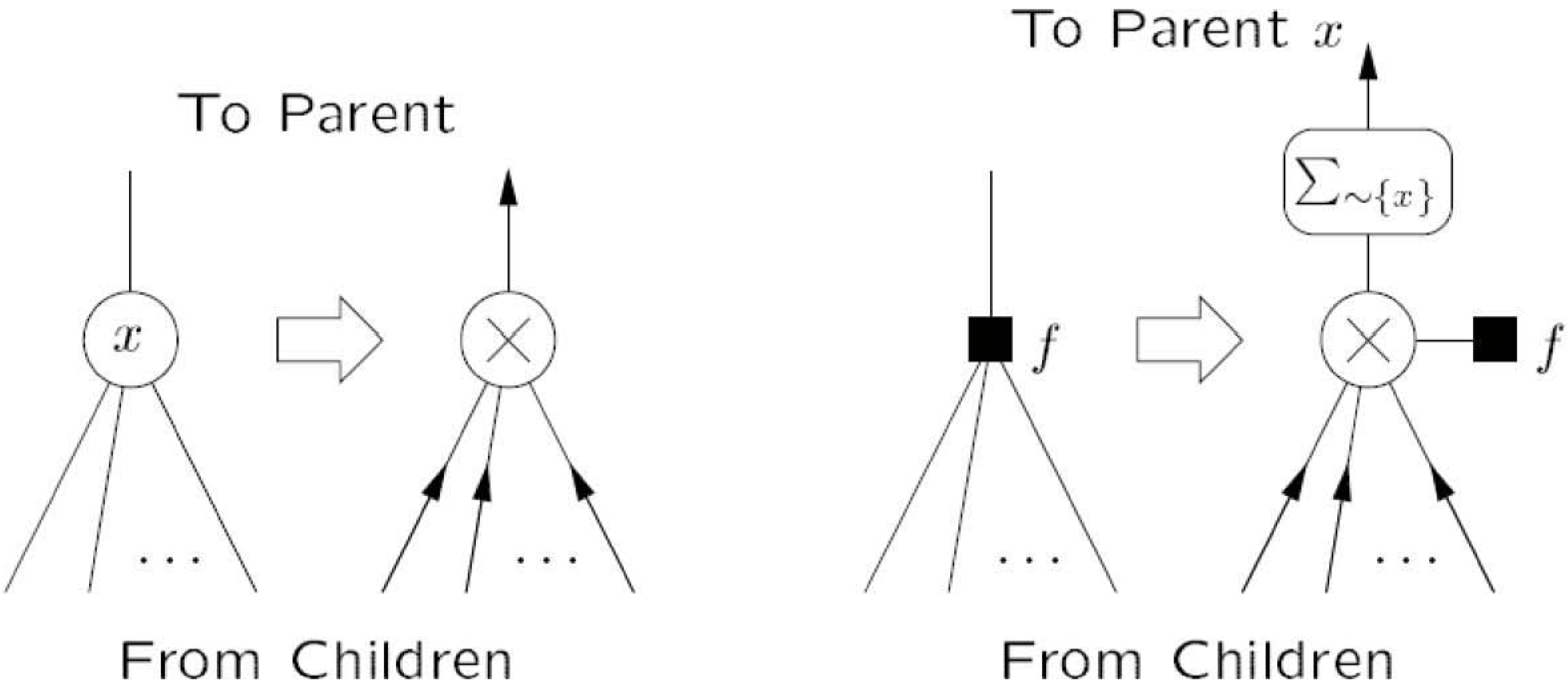
$$g_3(x_3) = \left(\sum_{\sim x_3} f_A(x_1) f_B(x_2) f_C(x_1, x_2, x_3) \right) \left(\sum_{\sim x_3} f_D(x_3, x_4) \right) \left(\sum_{\sim x_3} f_E(x_3, x_5) \right)$$

Computing Single Marginals

- Main Idea:
  - Target node becomes the root
  - Pass messages from leaves up to the root



Message Passing

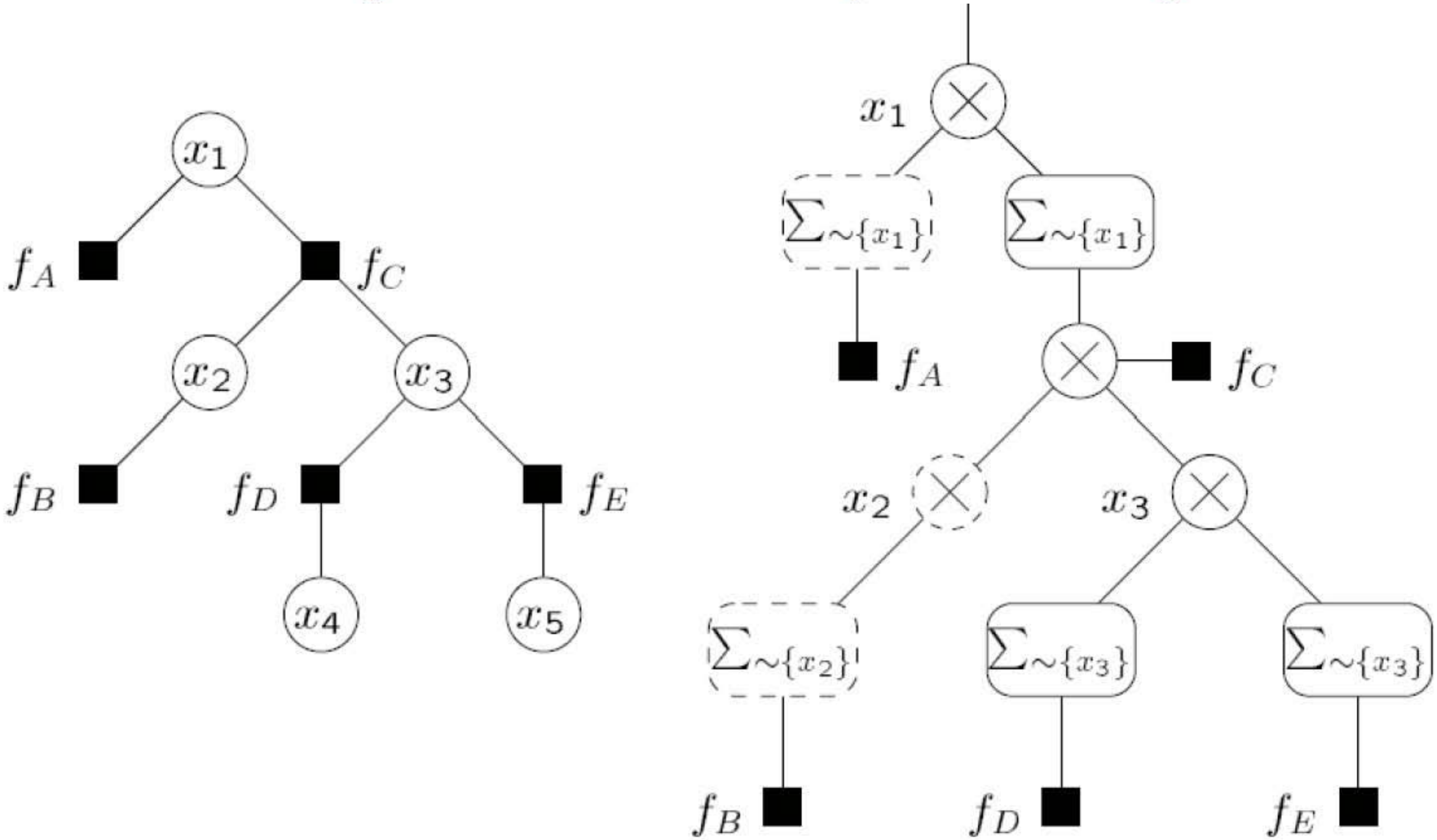


Compute product of descendants

Compute product of descendants with f
Then do not-sum over part

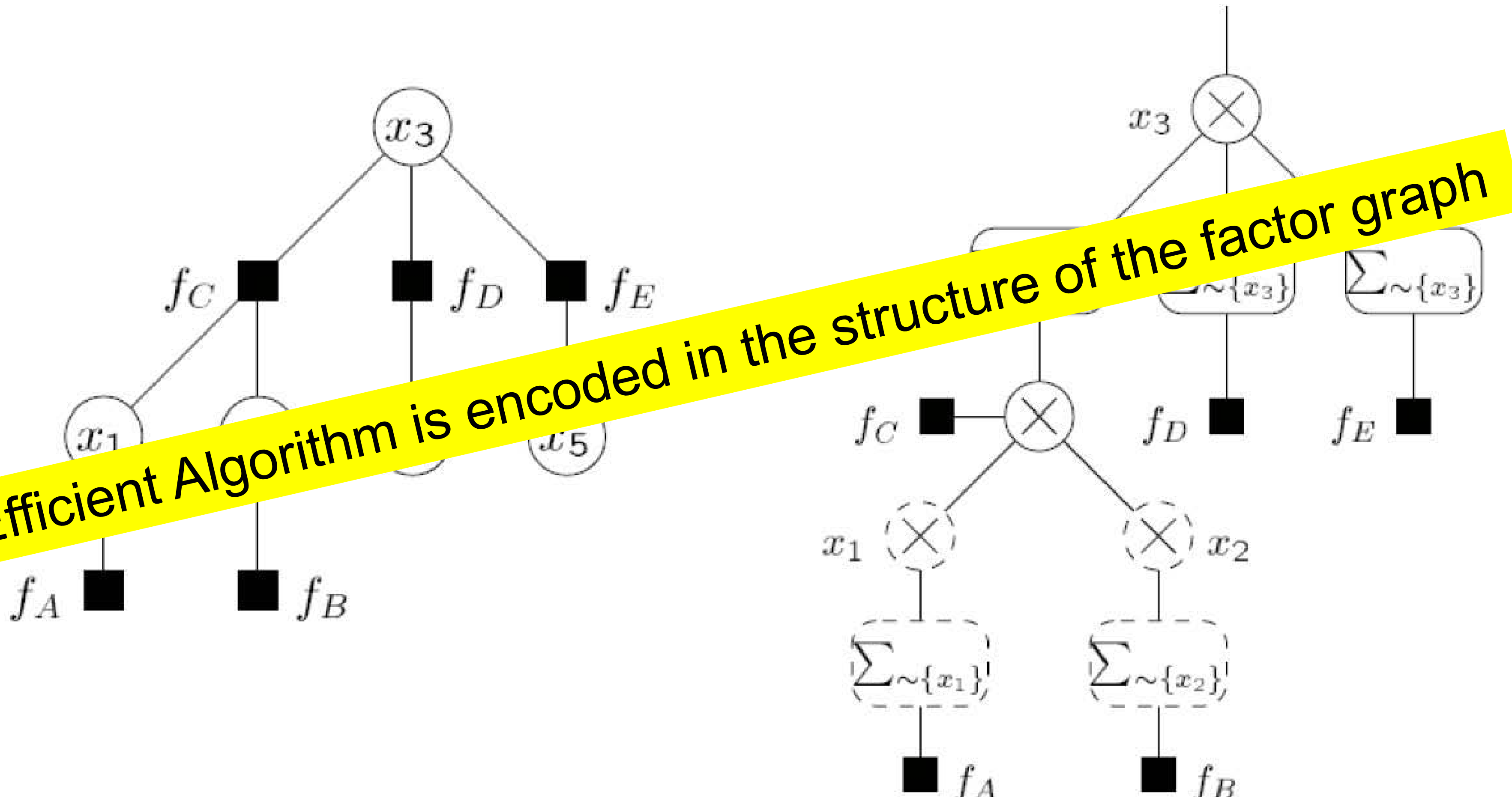
Example: Computing $g_1(x_1)$

$$g_1(x_1) = f_A(x_1) \sum_{\sim x_1} \left(f_B(x_2) f_C(x_1, x_2, x_3) \left(\sum_{\sim x_3} f_D(x_3, x_4) \right) \left(\sum_{\sim x_3} f_E(x_3, x_5) \right) \right)$$



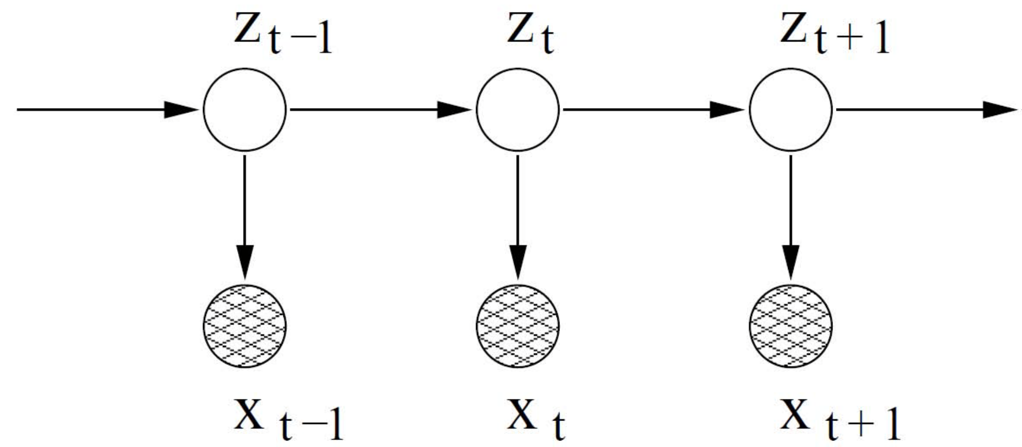
Example: Computing $g_3(x_3)$

$$g_3(x_3) = \left(\sum_{\sim x_3} f_A(x_1) f_B(x_2) f_C(x_1, x_2, x_3)\right) \left(\sum_{\sim x_3} f_D(x_3, x_4)\right) \left(\sum_{\sim x_3} f_E(x_3, x_5)\right)$$



Efficient Algorithm is encoded in the structure of the factor graph

Hidden Markov Models (HMMs)

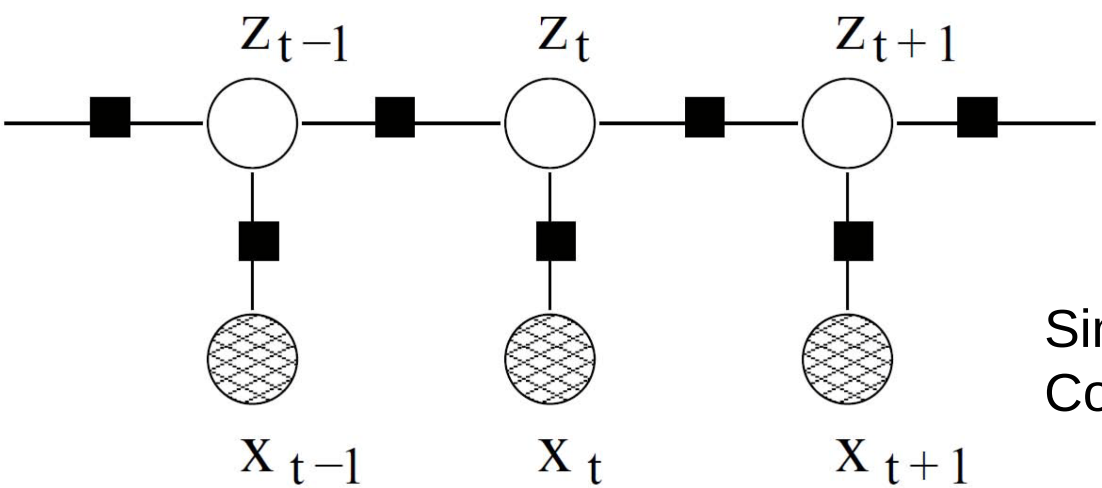


Latent variables:

$z_0, z_1, \ldots, z_{t-1}, z_t, z_{t+1}, \ldots, z_T$

Observed variables:

$x_1, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_T$



Inference Problems:

1. Compute $p(x_{1:T})$
2. Compute $p(z_t | x_{1:T})$
3. Find $\max_{z_{1:T}} p(z_{1:T} | x_{1:T})$

Similar problem for chain-structured Conditional Random Fields (CRFs)

The Sum-Product Algorithm

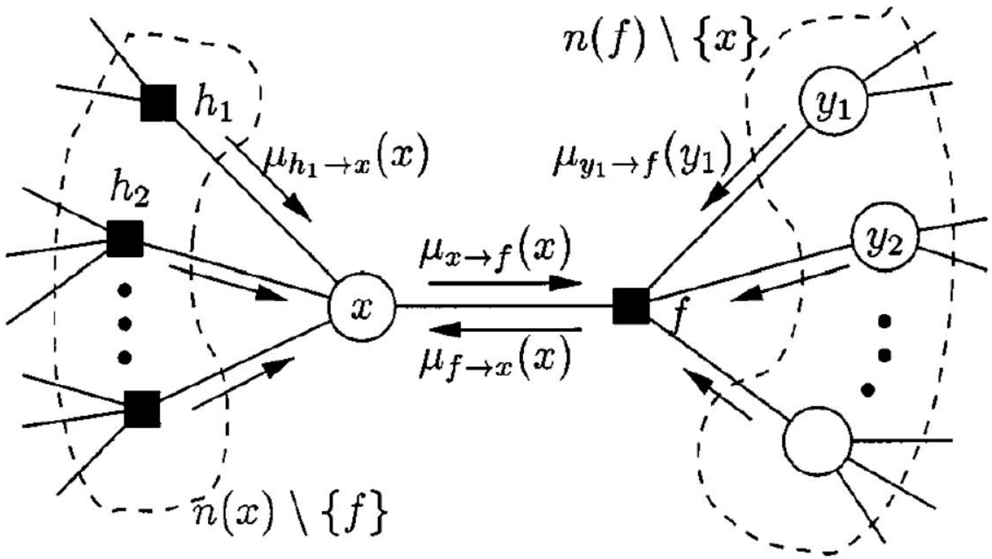
- To compute $g_i(x_i)$, form a tree rooted at x_i
- Starting from the leaves, apply the following two rules
  - Product Rule:

    At a variable node, take the product of descendants
  - Sum-product Rule:

    At a factor node, take the product of f with descendants;

    then perform not-sum over the parent node
- To compute all marginals
  - Can be done one at a time; repeated computations, not efficient
  - Simultaneous message passing following the sum-product algorithm
  - Examples: Belief Propagation, Forward-Backward algorithm, etc.

Sum-Product Updates



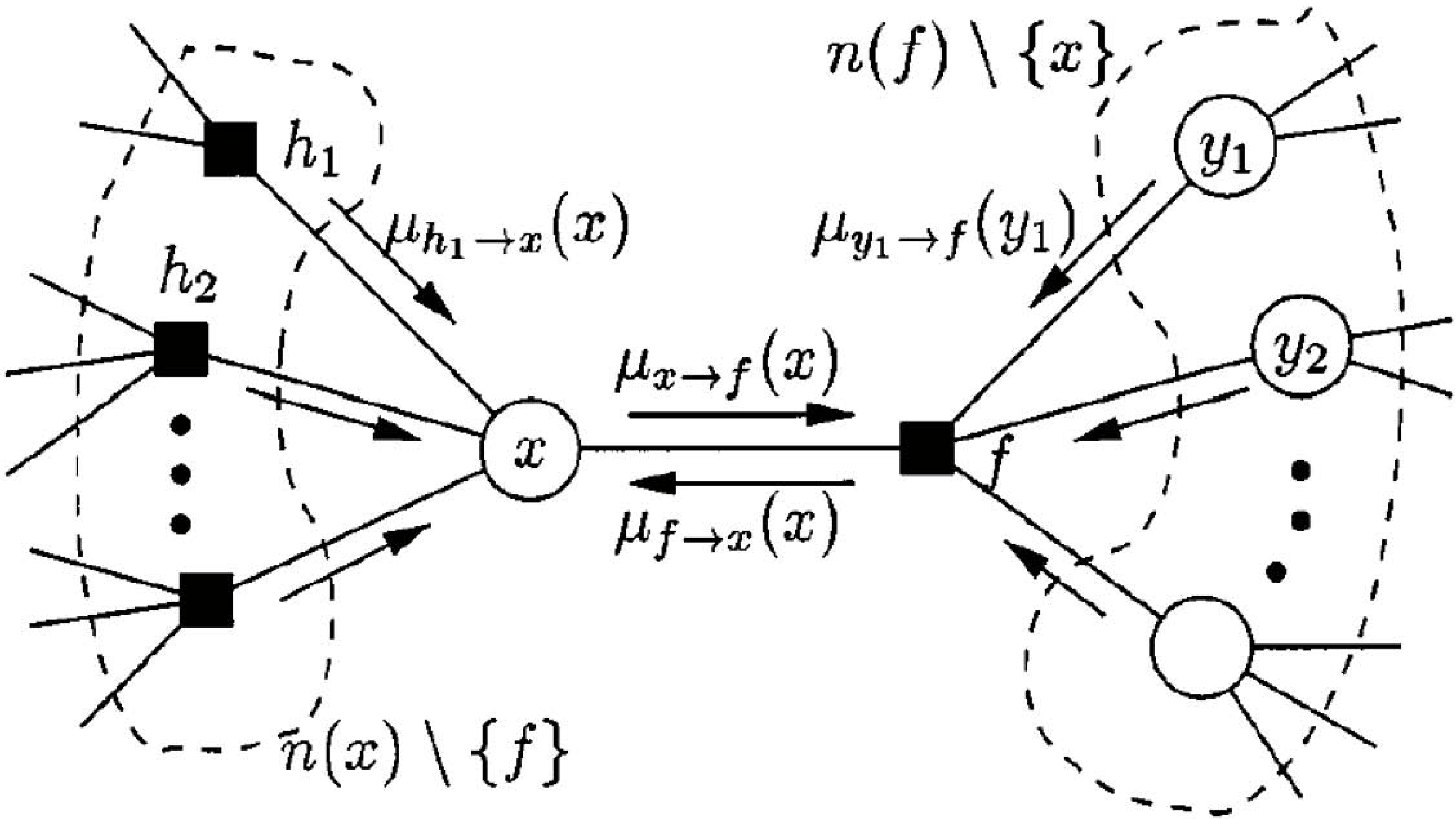
- Variable to local function:

$$\mu_{x \to f}(x) = \prod_{h \in n(x) \setminus f} \mu_{h \to x}$$

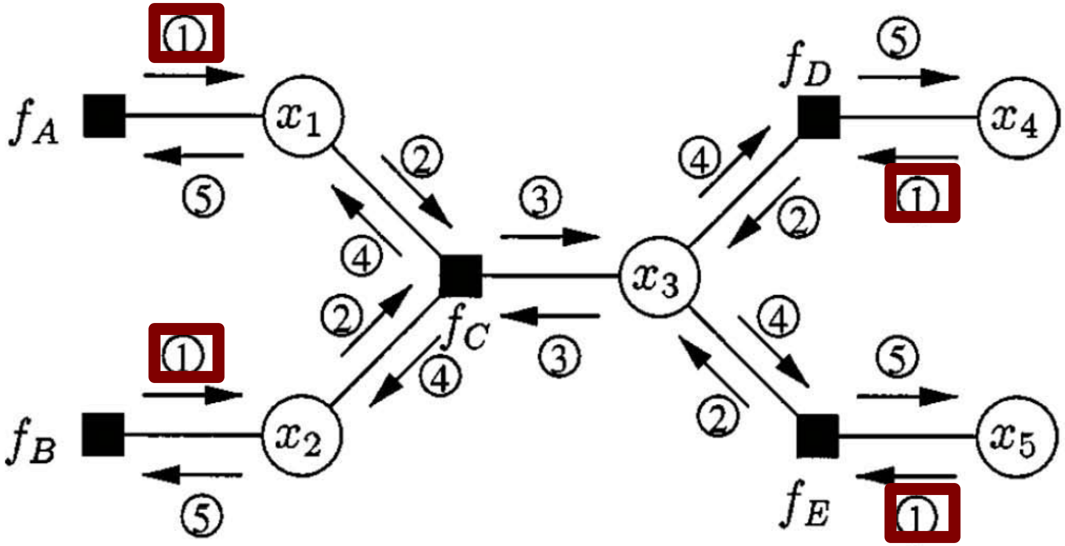
- Local function to variable:

$$\mu_{f \to x}(x) = \sum_{\sim x} \left(f(x) \prod_{y \in n(f) \setminus \{x\}} \mu_{y \to f}(y) \right)$$

Sum-Product Updates

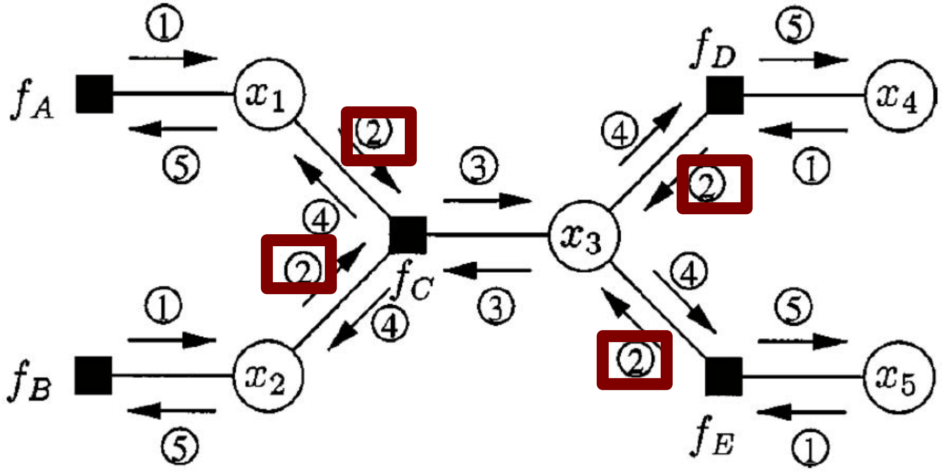


Example: Step 1



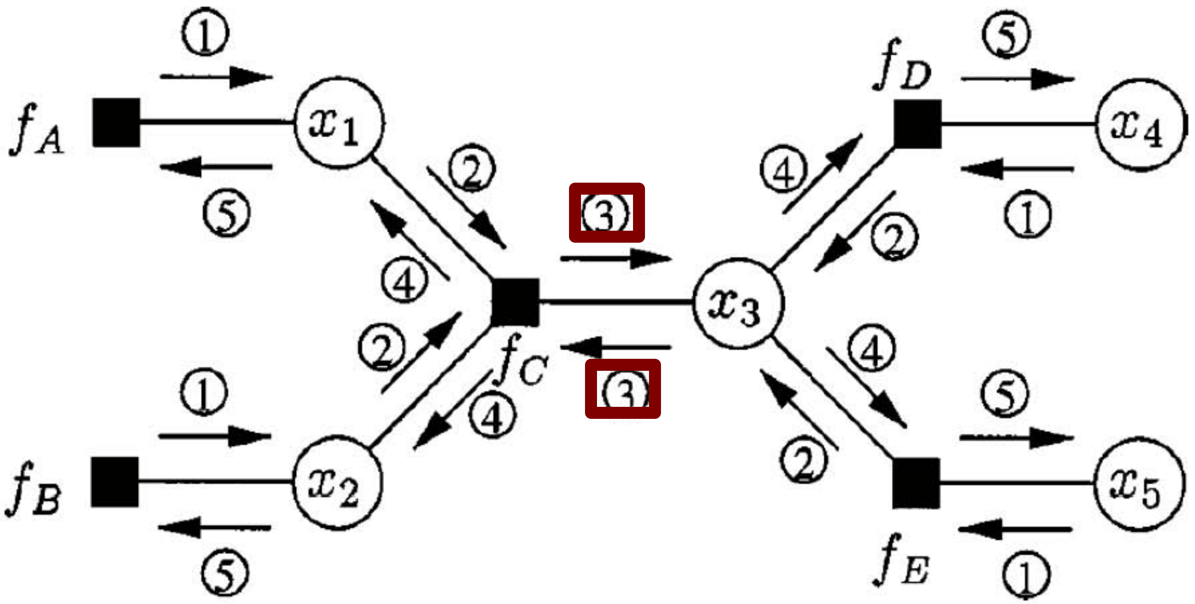
$$
\begin{aligned}
\mu_{f_A \to x_1}(x_1) &= f_A(x_1) \\
\mu_{f_B \to x_2}(x_2) &= f_B(x_2) \\
\mu_{x_4 \to f_D}(x_4) &= 1 \\
\mu_{x_5 \to f_E}(x_5) &= 1
\end{aligned}
$$

Example: Step 2



$$\mu_{x_1 \to f_C}(x_1) = \mu_{f_A \to x_1}(x_1)$$
$$\mu_{x_2 \to f_C}(x_2) = \mu_{f_B \to x_2}(x_2)$$
$$\mu_{f_D \to x_3}(x_3) = \sum_{\sim x_3} f_D(x_3, x_4)\mu_{x_4 \to f_D}(x_4)$$
$$\mu_{f_E \to x_3}(x_3) = \sum_{\sim x_3} f_D(x_3, x_5)\mu_{x_5 \to f_E}(x_5)$$

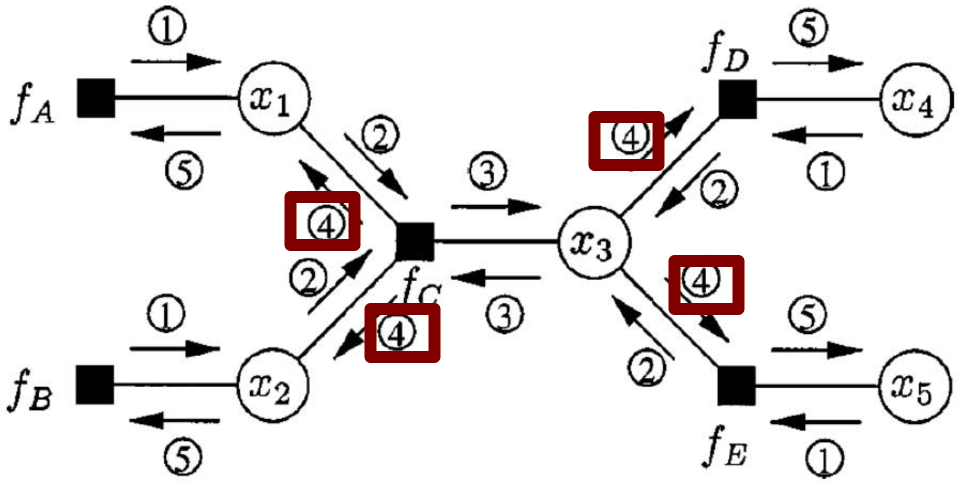
Example: Step 3



$$\mu_{f_C \to x_3}(x_3) = \sum_{\sim x_3} f_C(x_1, x_2, x_3)\mu_{x_1 \to f_C}(x_1)\mu_{x_2 \to f_C}(x_2)$$

$$\mu_{x_3 \to f_C}(x_3) = \mu_{f_D \to x_3}(x_3)\mu_{f_E \to x_3}(x_3)$$

Example: Step 4



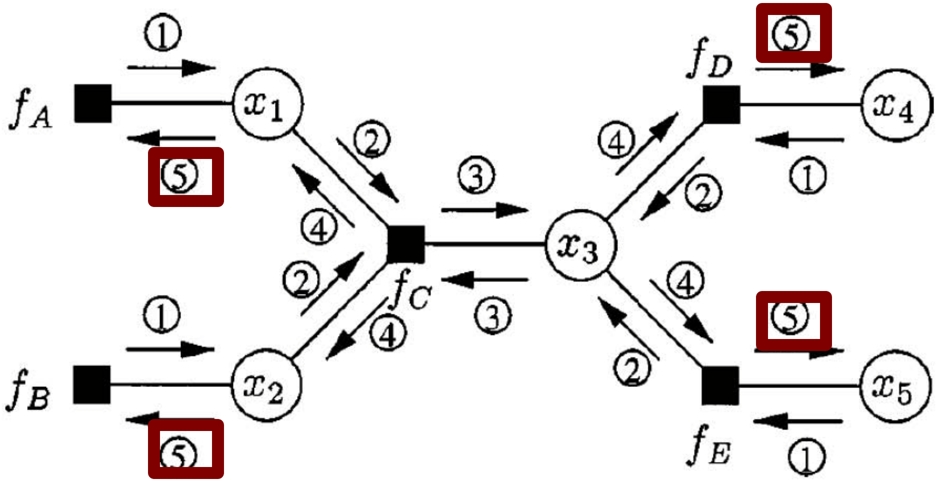
$$\mu_{f_C \to x_1}(x_1) = \sum_{\sim x_1} f_C(x_1, x_2, x_3)\mu_{x_2 \to f_C}(x_2)\mu_{x_3 \to f_C}(x_3)$$

$$\mu_{f_C \to x_2}(x_2) = \sum_{\sim x_2} f_C(x_1, x_2, x_3)\mu_{x_1 \to f_C}(x_1)\mu_{x_3 \to f_C}(x_3)$$

$$\mu_{x_3 \to f_D}(x_3) = \mu_{f_C \to x_3}(x_3)\mu_{f_E \to x_3}(x_3)$$

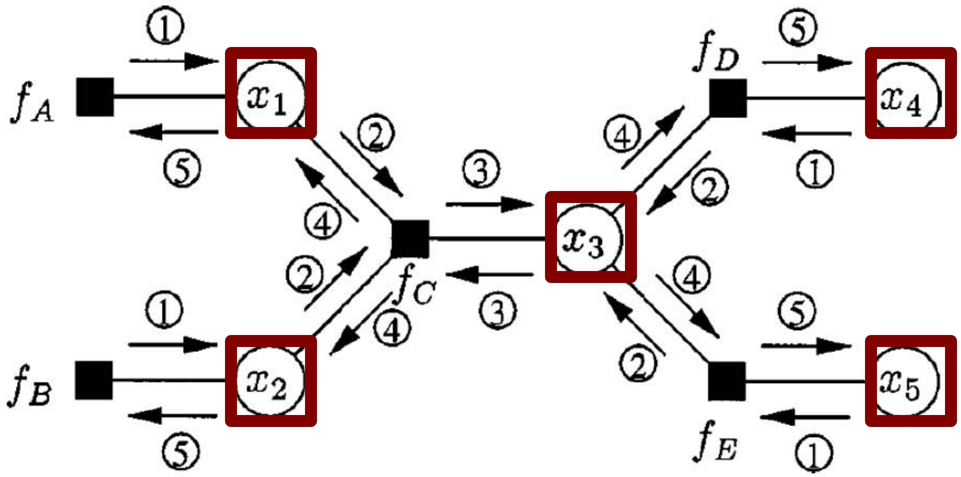
$$\mu_{x_3 \to f_E}(x_3) = \mu_{f_C \to x_3}(x_3)\mu_{f_D \to x_3}(x_3)$$

Example: Step 5



$$
\begin{aligned}
\mu_{x_1 \to f_A}(x_1) &= \mu_{f_C \to x_1}(x_1) \\
\mu_{x_2 \to f_B}(x_2) &= \mu_{f_C \to x_2}(x_2) \\
\mu_{f_D \to x_4}(x_4) &= \sum_{\sim x_4} f_D(x_3, x_4)\mu_{x_3 \to f_D}(x_4) \\
\mu_{f_E \to x_5}(x_5) &= \sum_{\sim x_5} f_D(x_3, x_5)\mu_{x_3 \to f_E}(x_5)
\end{aligned}
$$

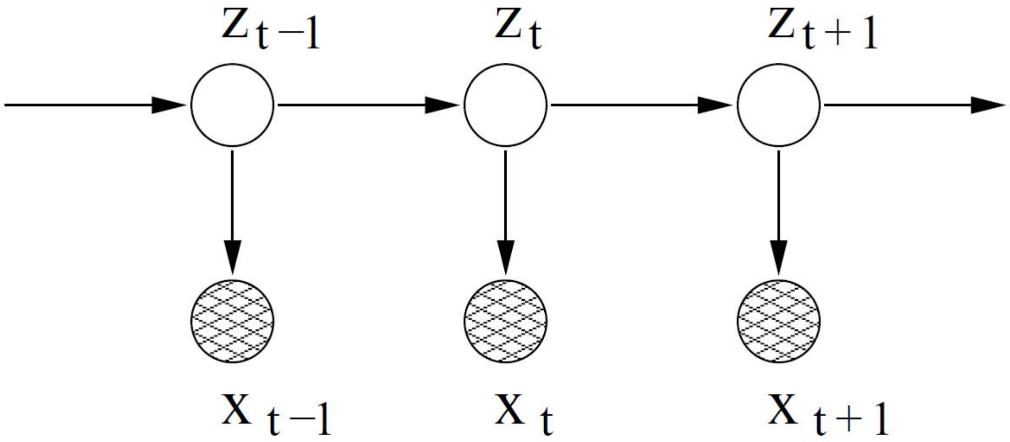
Example: Termination



Marginal function is the product of all incoming messages

$$
\begin{aligned}
g_1(x_1) &= \mu_{f_A \to x_1}(x_1)\mu_{f_C \to x_1}(x_1) \\
g_2(x_2) &= \mu_{f_B \to x_2}(x_2)\mu_{f_C \to x_2}(x_2) \\
g_3(x_3) &= \mu_{f_C \to x_3}(x_3)\mu_{f_D \to x_3}(x_3)\mu_{f_E \to x_3}(x_3) \\
g_2(x_2) &= \mu_{f_D \to x_4}(x_4) \\
g_5(x_5) &= \mu_{f_E \to x_5}(x_5)
\end{aligned}
$$

HMMs Revisited

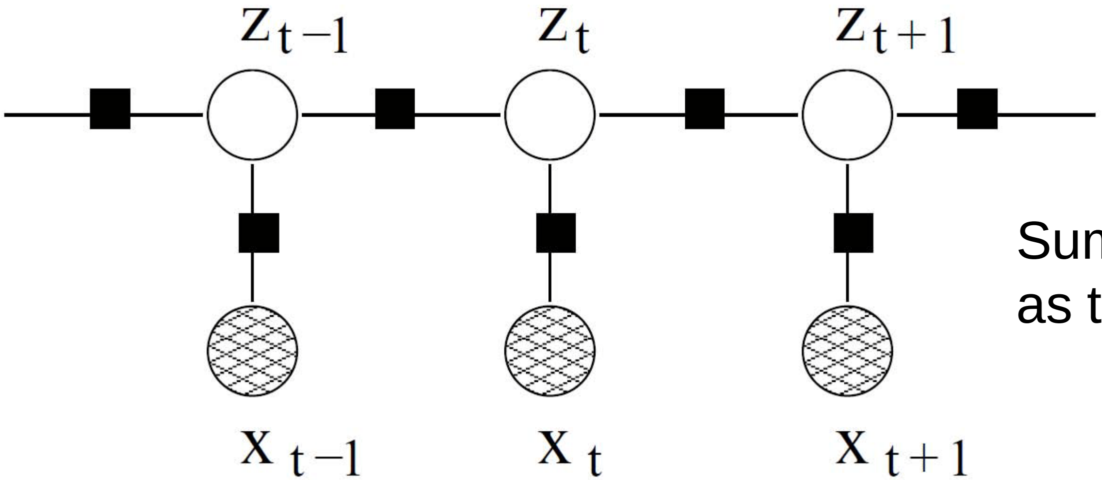


Latent variables:

$z_0, z_1, \ldots, z_{t-1}, z_t, z_{t+1}, \ldots, z_T$

Observed variables:

$x_1, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_T$



Inference Problem:

1. Compute $p(x_{1:T})$
2. Compute $p(z_t | x_{1:T})$

Sum-product algorithm is known as the `forward-backward' algorithm

Smoothing in Kalman Filtering

Distributive Law on Semi-Rings

- Idea can be applied to any commutative semi-ring
- Semi-ring 101
  - Two operations (+,×): Associative, Commutative, Identity
  - Distributive law: a×b + a×c = a×(b+c)

	K	"$(+,0)$"	"$(\cdot,1)$"	short name
1.	A	$(+,0)$	$(\cdot,1)$	
2.	$A[x]$	$(+,0)$	$(\cdot,1)$	
3.	$A[x,y,\ldots]$	$(+,0)$	$(\cdot,1)$	
4.	$[0,\infty)$	$(+,0)$	$(\cdot,1)$	sum-product
5.	$(0,\infty]$	$(\min,\infty)$	$(\cdot,1)$	min-product
6.	$[0,\infty)$	$(\max,0)$	$(\cdot,1)$	max-product
7.	$(-\infty,\infty]$	$(\min,\infty)$	$(+,0)$	min-sum
8.	$[-\infty,\infty)$	$(\max,-\infty)$	$(+,0)$	max-sum
9.	$\{0,1\}$	$(\mathtt{OR},0)$	$(\mathtt{AND},1)$	Boolean
10.	2^S	$(\cup,\emptyset)$	$(\cap,S)$	
11.	Λ	$(\vee,0)$	$(\wedge,1)$	
12.	Λ	$(\wedge,1)$	$(\vee,0)$.	

- Belief Propagation in Bayes nets
- MAP inference in HMMs
- Max-product algorithm
- Alternative to Viterbi Decoding
- Kalman Filtering
- Error Correcting Codes
- Turbo Codes
- …

Message Passing in General Graphs

- Tree structured graphs
  - Message passing is guaranteed to give correct solutions
  - Examples: HMMs, Kalman Filters
- General Graphs
  - Active research topic
    - Progress has been made in the past 10 years
  - Message passing
    - May not converge
    - May converge to a ‘local minima’ of ‘Bethe variational free energy’
  - New approaches to convergent and correct message passing
- Applications
  - True Skill: Ranking System for Xbox Live
  - Turbo Codes: 3G, 4G phones, satellite comm, Wimax, Mars orbiter

Part II: Mixed Membership Models

- Mixture Models vs Mixed Membership Models

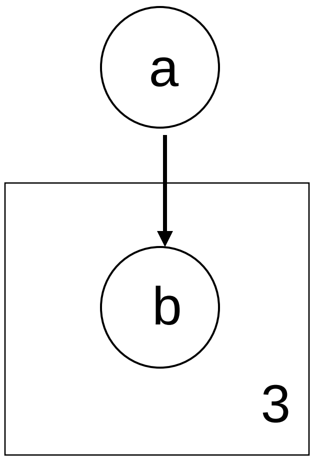
- Latent Dirichlet Allocation

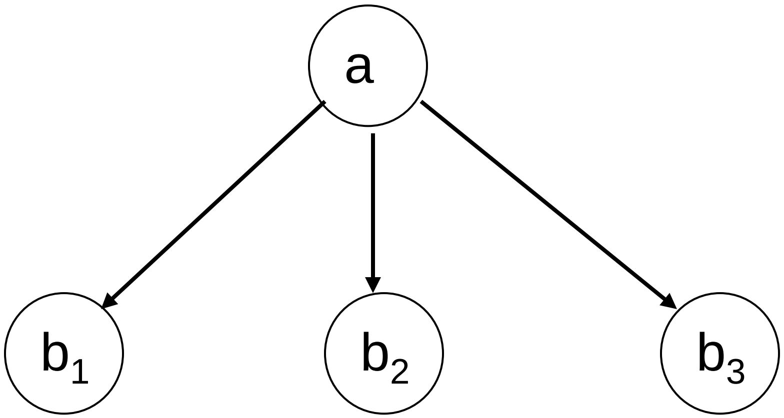
- Inference
  - Mean-Field and Collapsed Variational Inference
  - MCMC/Gibbs Sampling

- Applications

- Generalizations

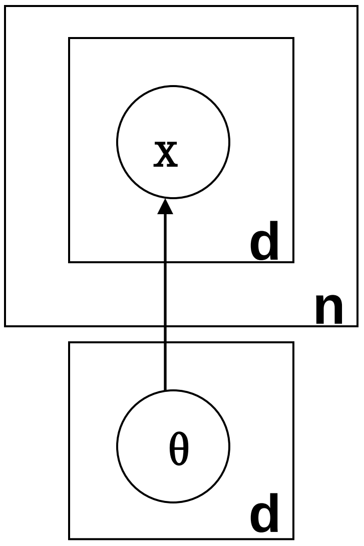
Background: Plate Diagrams

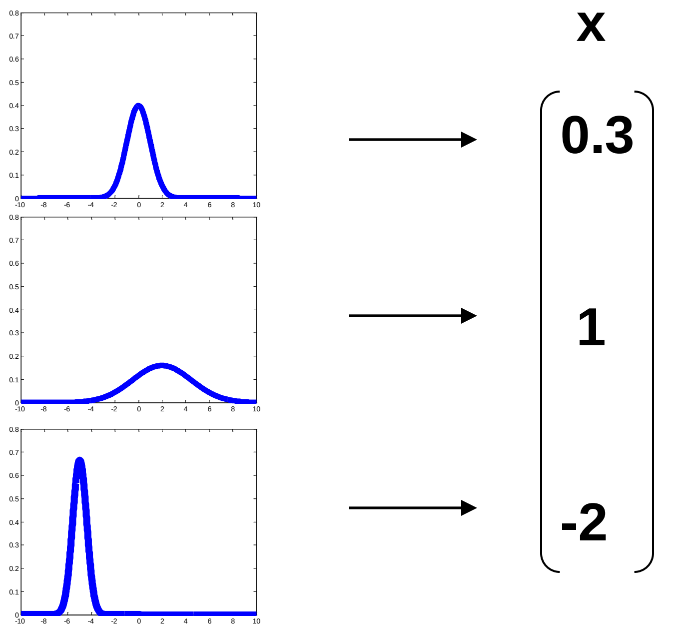




Compact representation of large Bayesian networks

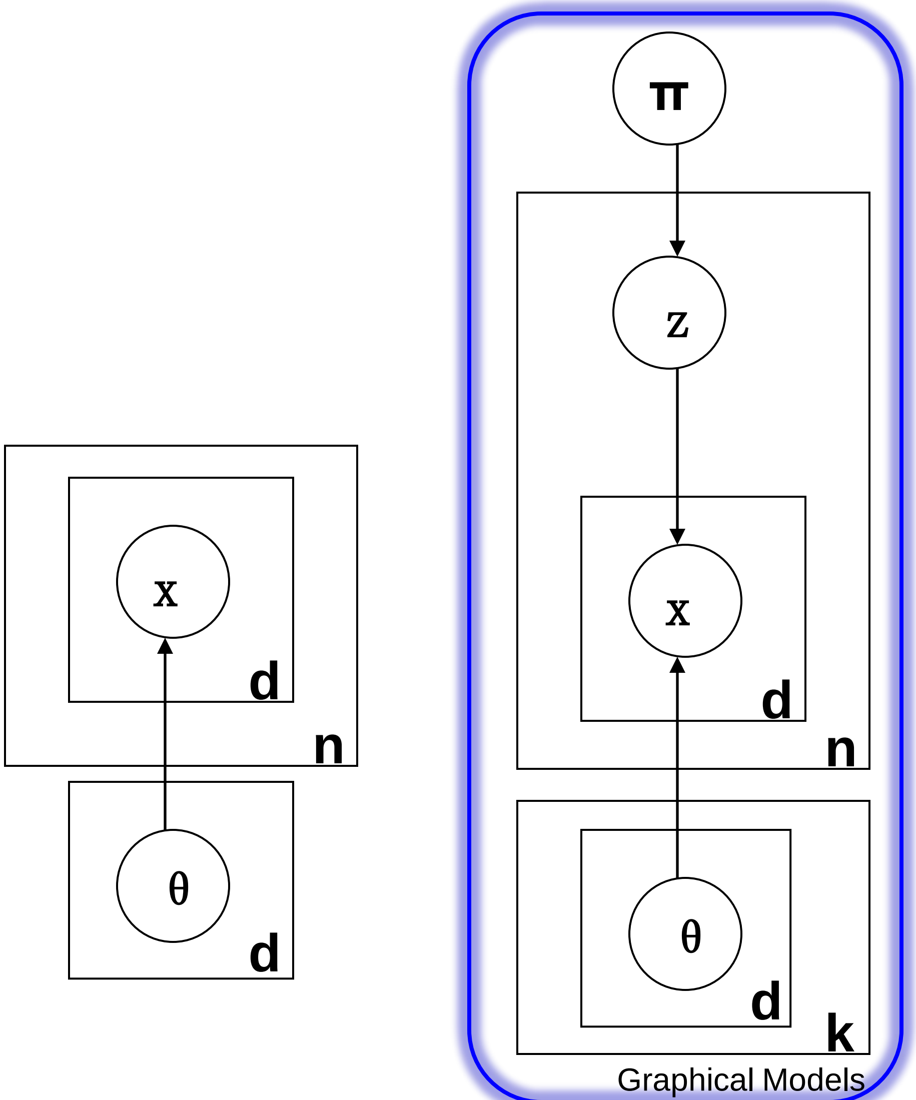
Model 1: Independent Features



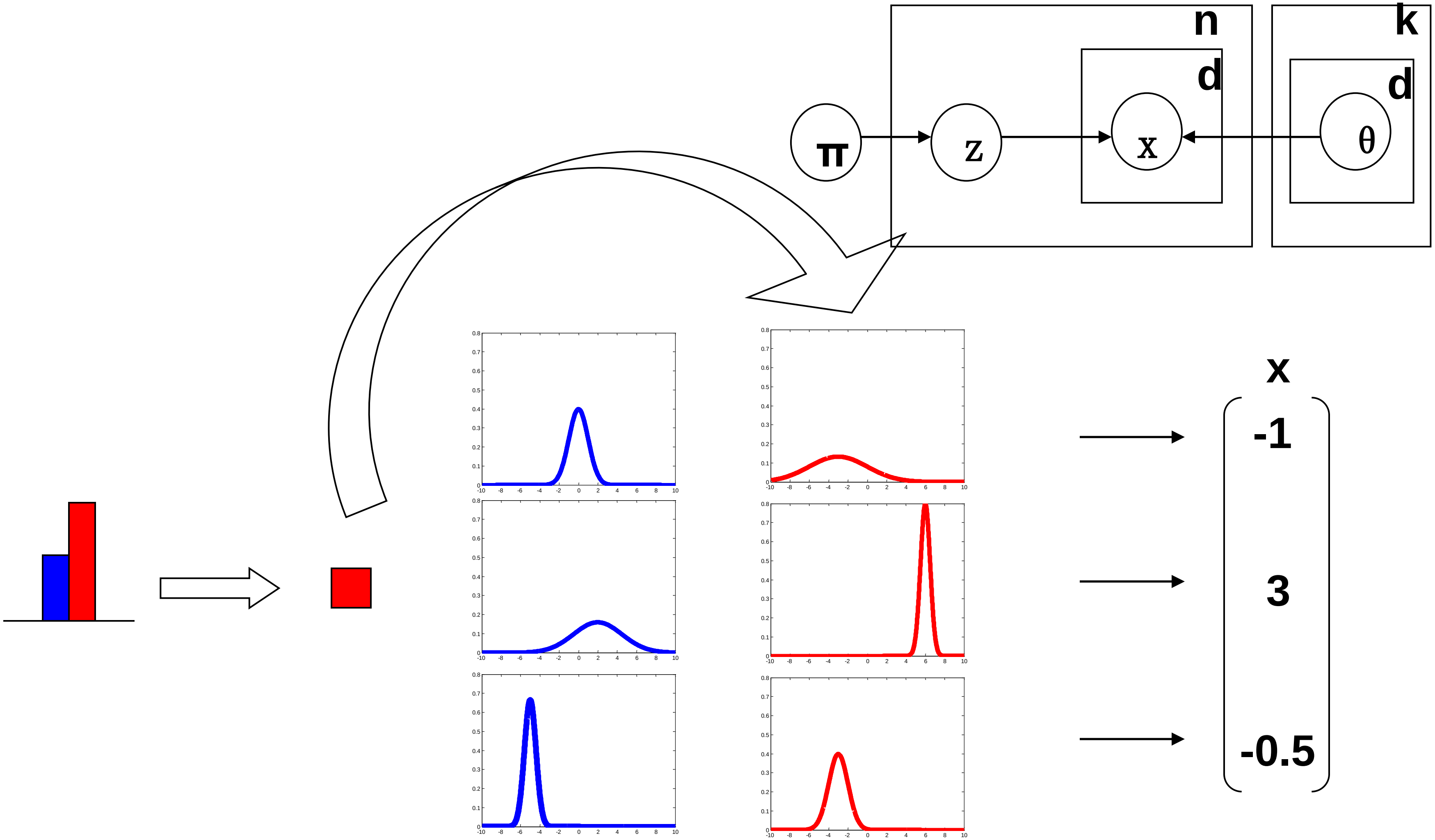


d=3, n=1

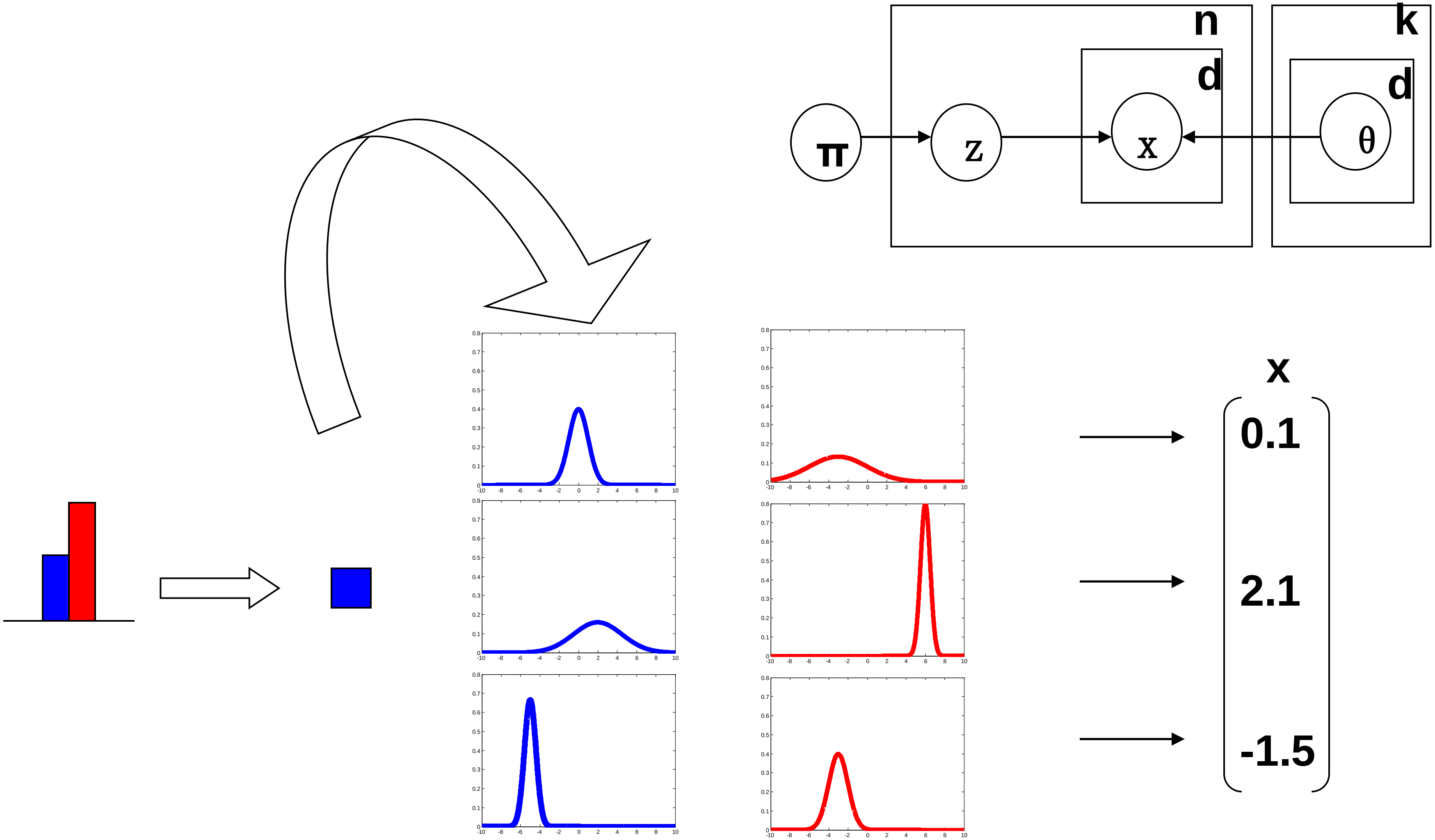
Model 2: Naïve Bayes (Mixture Models)



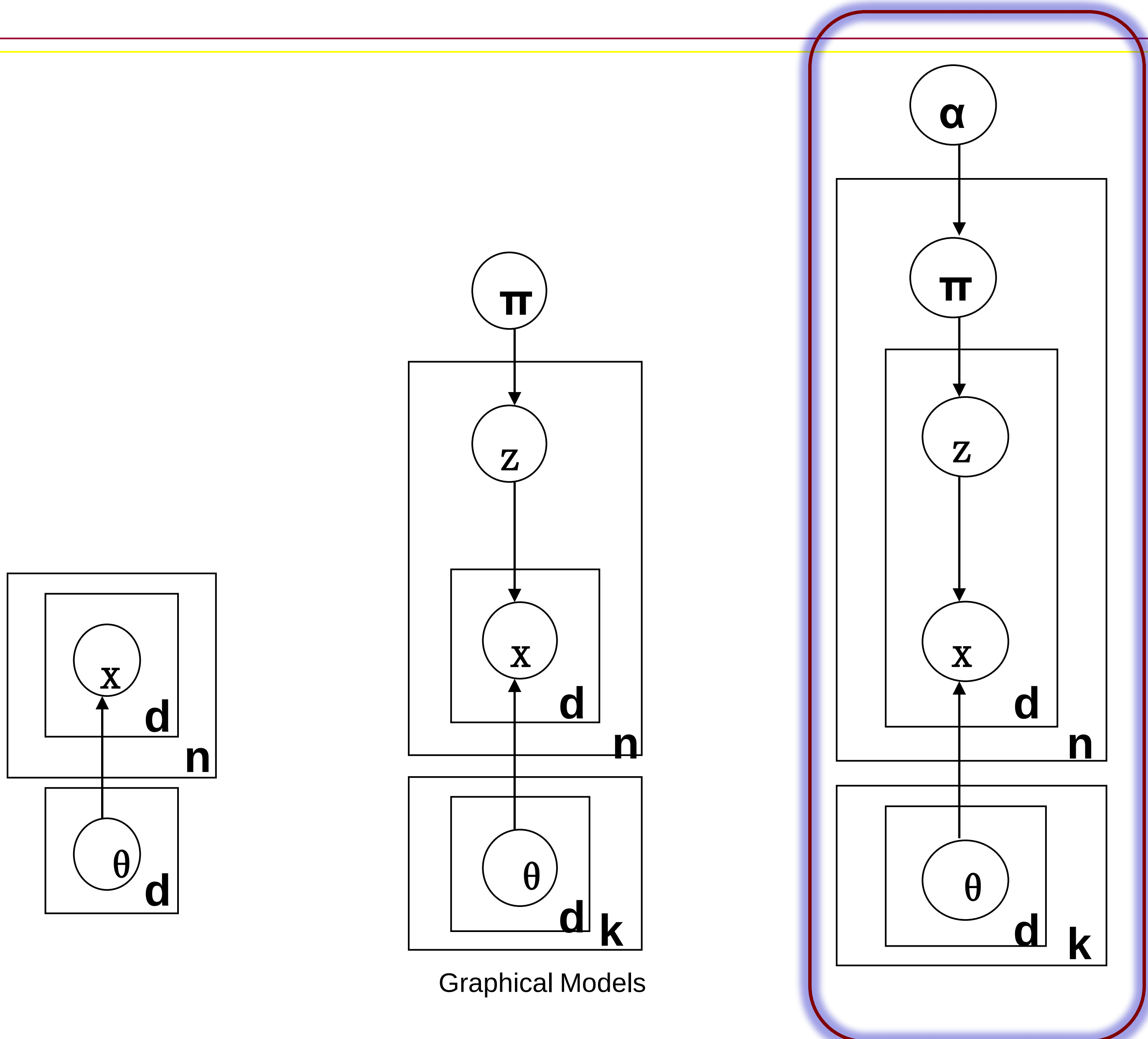
Naïve Bayes Model



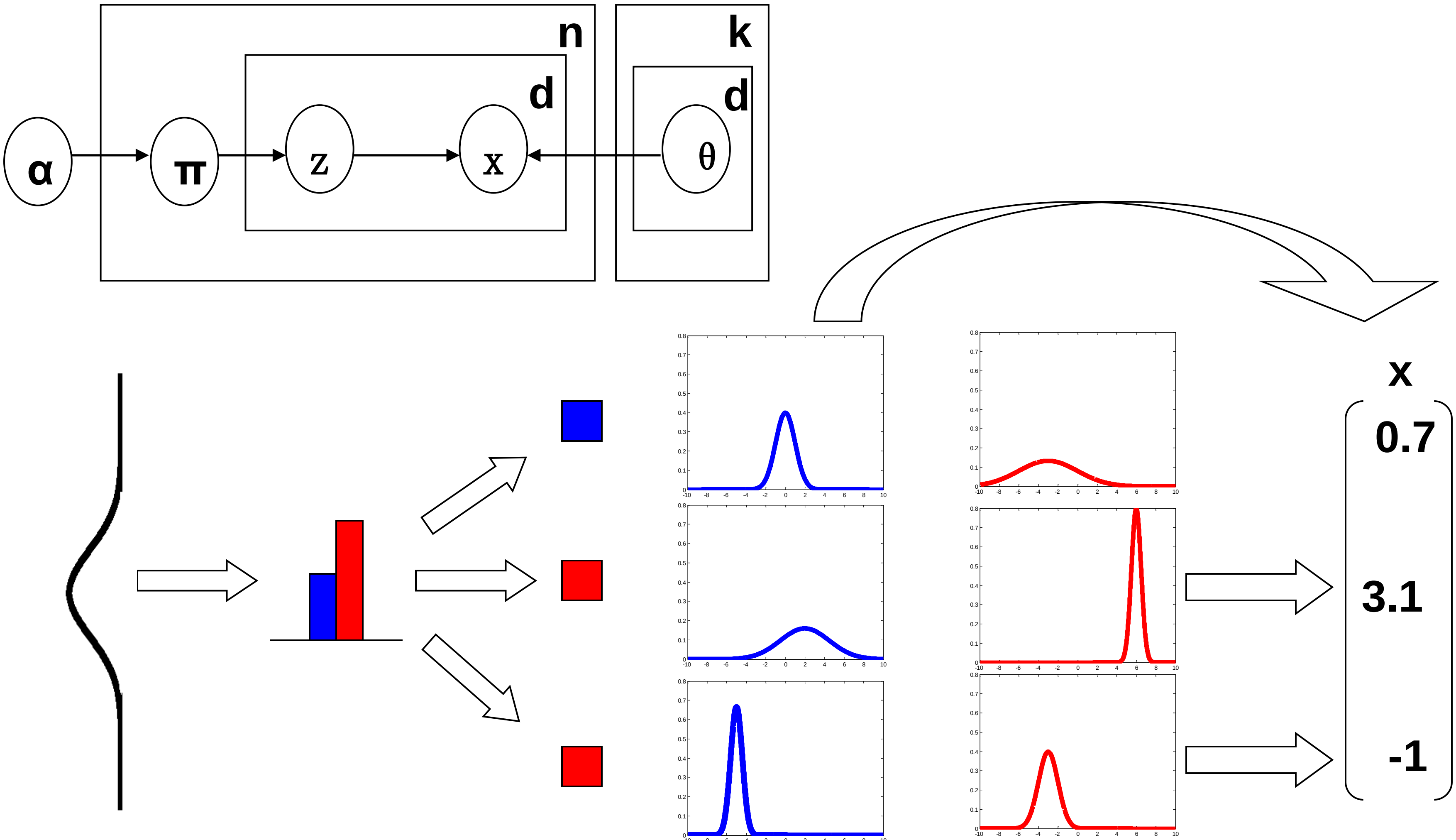
Naïve Bayes Model



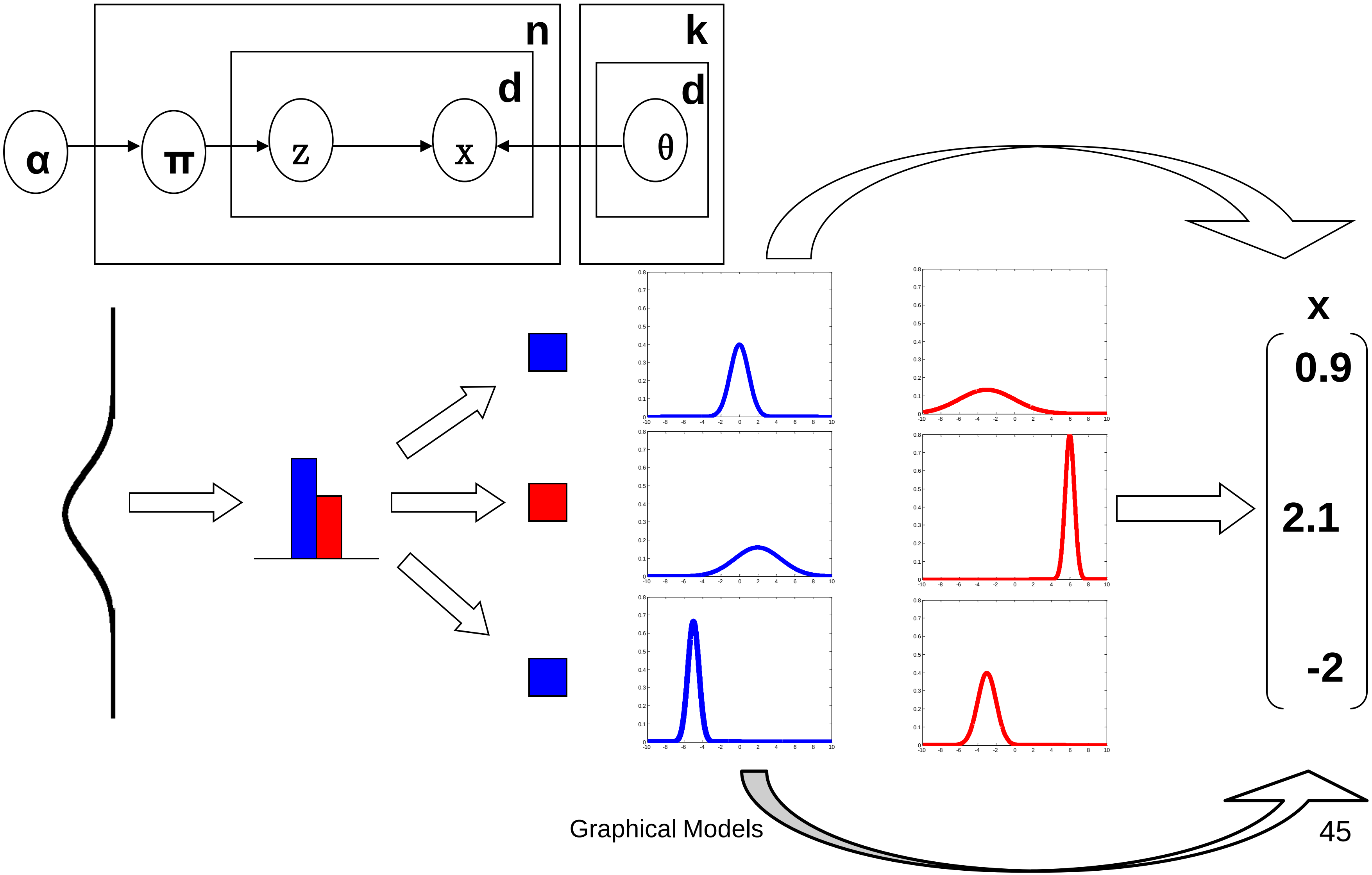
Model 3: Mixed Membership Model



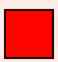
Mixed Membership Models



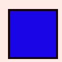
Mixed Membership Models



Mixture Model vs Mixed Membership Model

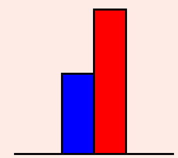


$$\begin{array}{c} x \\ \begin{bmatrix} -1 \\ 3 \\ -0.5 \end{bmatrix} \end{array}$$

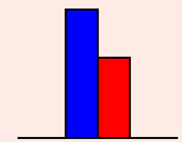


$$\begin{array}{c} x \\ \begin{bmatrix} 0.1 \\ 2.1 \\ -1.5 \end{bmatrix} \end{array}$$

Single component membership



$$\begin{array}{c} x \\ \begin{bmatrix} 0.7 \\ 3.1 \\ -1 \end{bmatrix} \end{array}$$



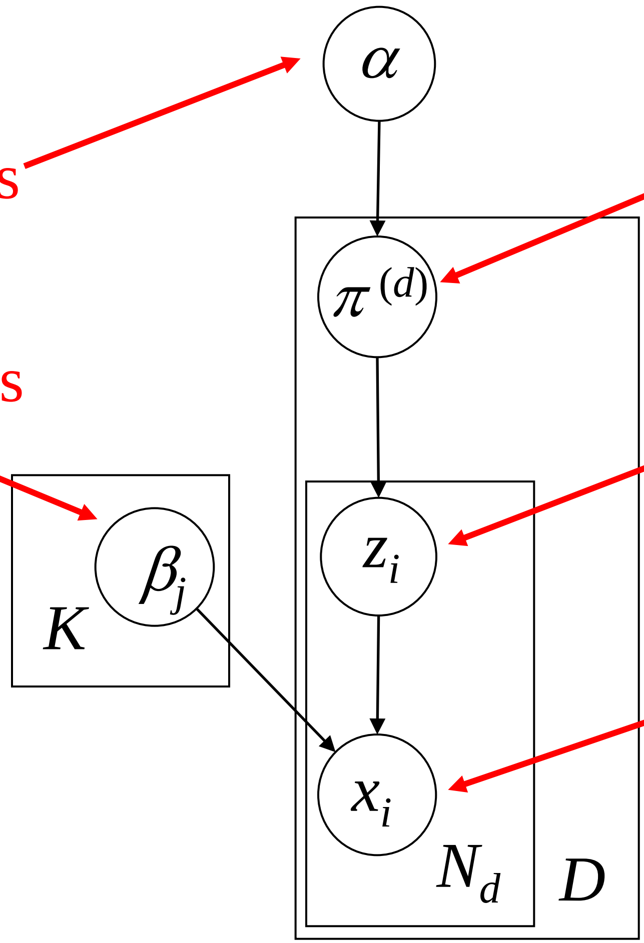
$$\begin{array}{c} x \\ \begin{bmatrix} 0.9 \\ 2.1 \\ -2 \end{bmatrix} \end{array}$$

Multi-component mixed membership

Latent Dirichlet Allocation (LDA)

Dirichlet priors

distribution over words
for each topic



distribution over topics
for each document

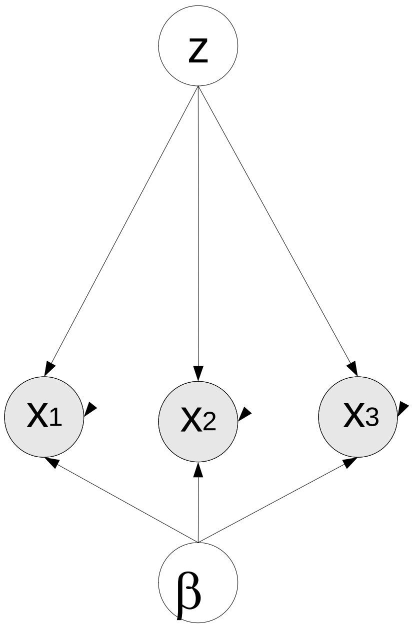
$p^{(d)} \sim \text{Dirichlet}(\alpha)$

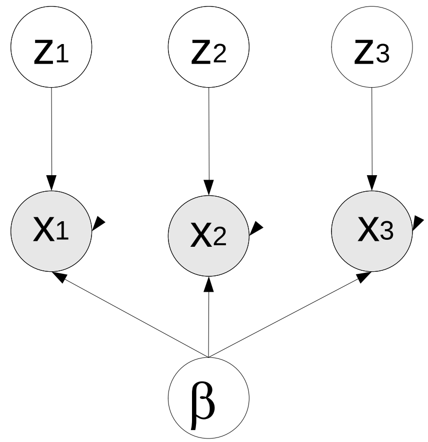
topic assignment
for each word

$z_i \sim \text{Discrete}(\pi^{(d)})$

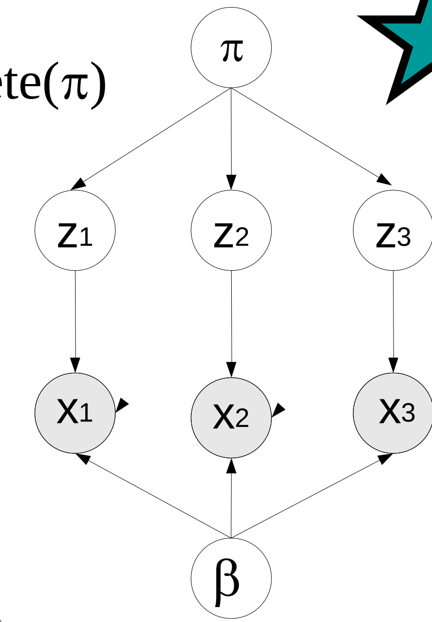
word generated from
assigned topic

$x_i \sim \text{Discrete}(\beta^{(z_i)})$

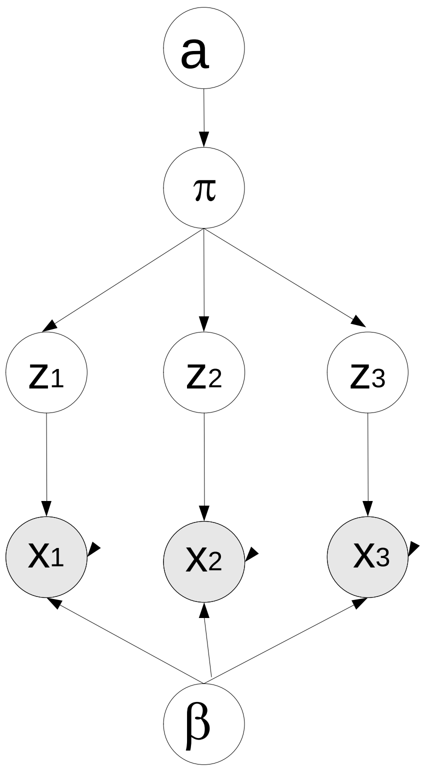


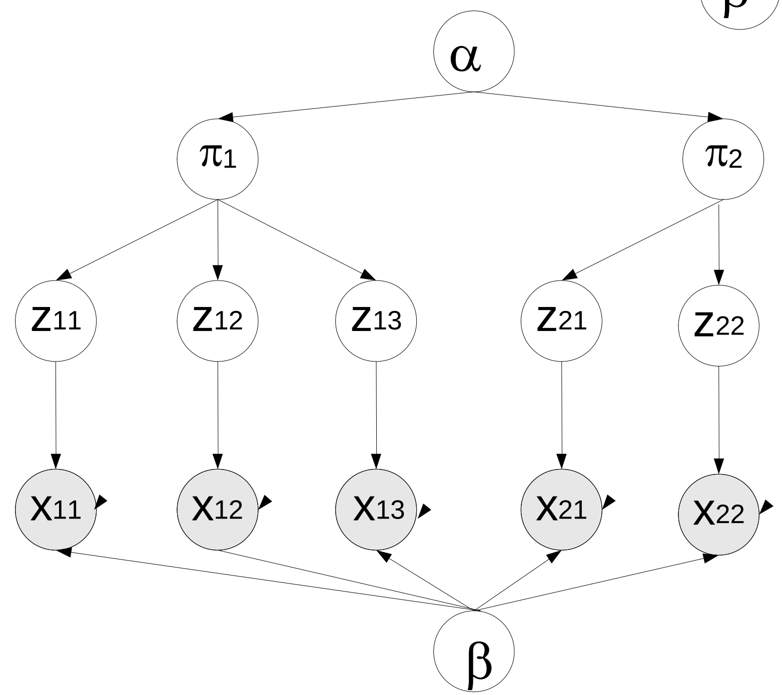


z~Discrete(π)

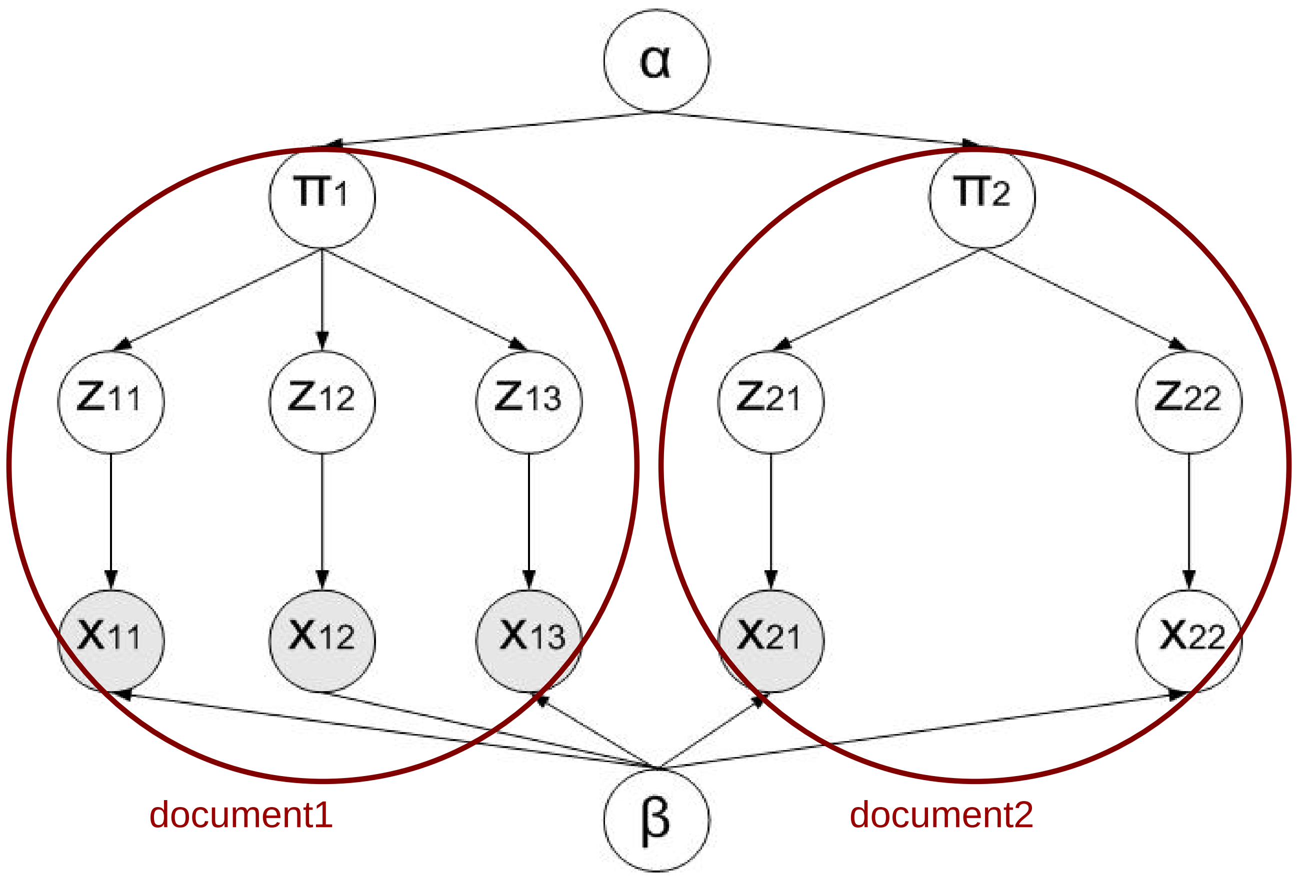


π~Drichlet(α)

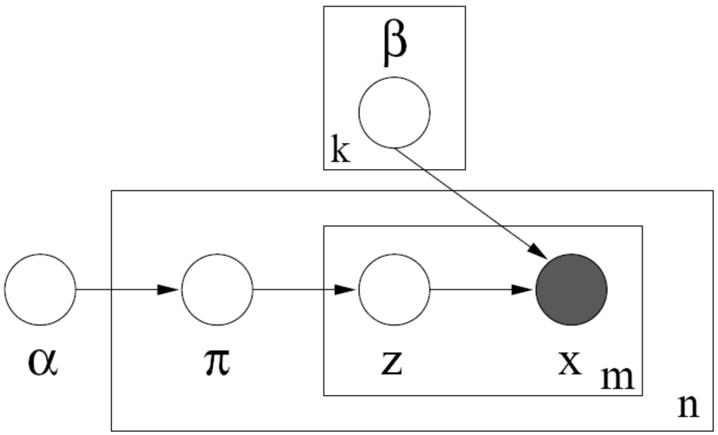


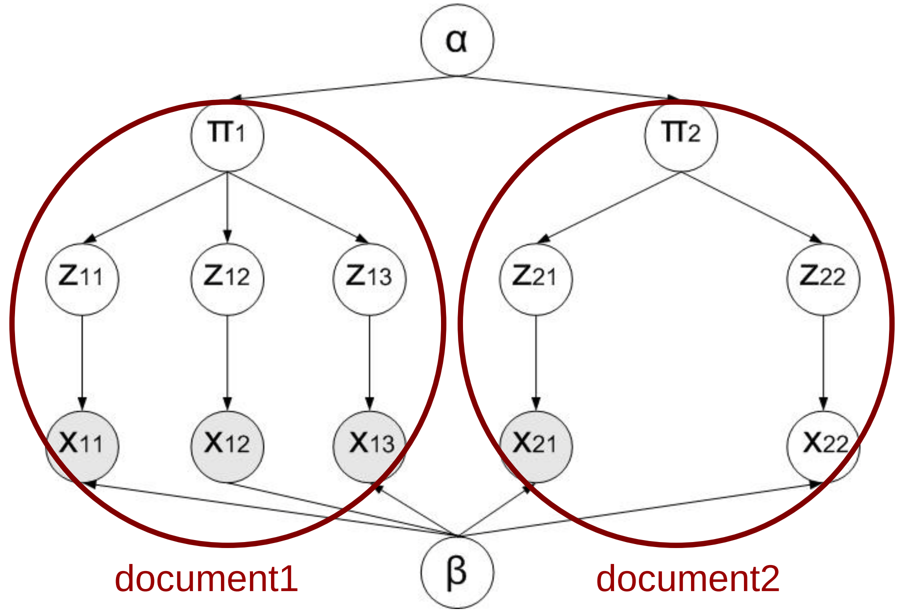


LDA Generative Model



LDA Generative Model





1. Choose $\pi \sim \mathrm{Dir}(\alpha)$
2. For each of d tokens $(x_j, [j]_1^m)$ in $\mathbf{x}$:
   (a) Choose a component $z_j \sim \mathrm{Discrete}(\pi)$.
   (b) Choose x_j from $p(x_j|\beta_{z_j})$, a Discrete distribution conditioned on the topic z_j.

Learning: Inference and Estimation

- Learning
  - Estimate model parameters (α, β) to maximize log-likelihood
  - Infer 'mixed-memberships' of documents
- Expectation Maximization
  - E-step: Calculate posterior probability $p(\pi, \mathbf{z}|\mathbf{x}, \alpha, \beta)$ to obtain

$$L(\alpha, \beta) = \log p(\mathbf{x}|\alpha, \beta) = \log \int_\pi \sum_{\mathbf{z}} p(\mathbf{x}, \pi, \mathbf{z}|\alpha, \beta) d\pi$$

$$= \log \int_\pi \sum_{\mathbf{z}} p(\mathbf{x}|\alpha, \beta) p(\pi, \mathbf{z}|\mathbf{x}, \alpha, \beta) d\pi$$

  - M-step: Maximize $L(\alpha, \beta)$ w.r.t. (α, β)
- Issues: Posterior probability cannot be obtained in closed form

Variational Inference

- Introduce a variational distribution $q(\pi, z|\gamma, \phi)$ to approximate $p(\pi, z|\mathbf{x}, \alpha, \beta)$
- Use Jensen's inequality to get a tractable lower bound

$$\log p(\mathbf{x}|\alpha, \beta) \geq E_q[\log p(\mathbf{x}, \pi, \mathbf{z}|\alpha, \beta)] + H(q(\pi, \mathbf{z}))$$

- Obtain a family of lower bounds
  - A lower bound for each (γ, ϕ)
  - Maximize the lower bounds w.r.t. (γ, ϕ)
  - Equivalent to minimizing $KL(q(\pi, z|\gamma, \phi) \| p(\pi, z|\mathbf{x}, \alpha, \beta))$
- Maximize the *best lower bound* w.r.t. (α, β)

Variational EM for LDA

$$L(\gamma, \phi; \alpha, \beta) = \text{lower bound to log-likelihood } L(\alpha, \beta)$$

- E-step: Given model parameters $(\alpha^{(t)}, \beta^{(t)})$, find variational parameters:

$$(\gamma^{(t+1)}, \phi^{(t+1)}) = \underset{(\gamma,\phi)}{\operatorname{argmax}}\ L(\gamma, \phi; \alpha^{(t)}, \beta^{(t)})$$

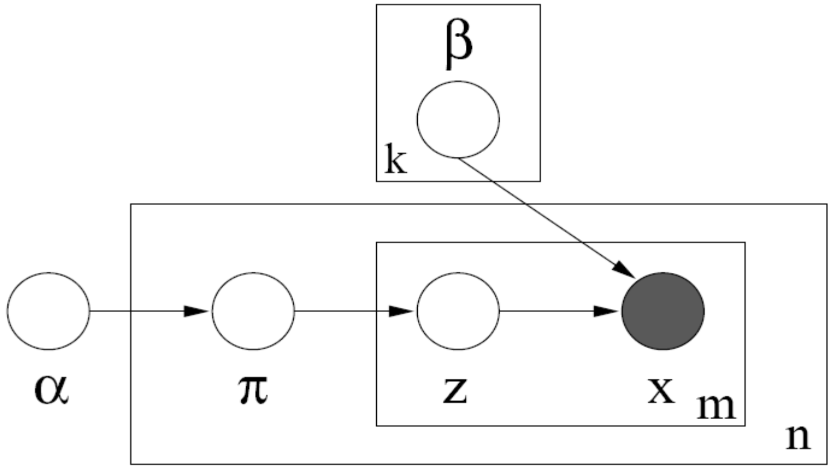
- Now $L(\gamma^{(t+1)}, \phi^{(t+1)}; \alpha, \beta)$ serves as a lower bound for $L(\alpha, \beta)$
- M-step: Obtain an improved estimate of the model parameters:

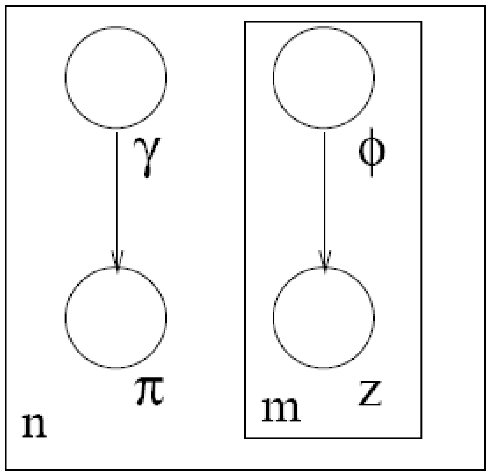
$$(\alpha^{(t+1)}, \beta^{(t+1)}) = \underset{(\alpha,\beta)}{\operatorname{argmax}}\ L(\gamma^{(t+1)}, \phi^{(t+1)}; \alpha, \beta)$$

E-step: Variational Distribution and Updates

- Fully factorized distribution over the latent variables

$$q(\pi, z | \gamma, \phi) = q_{\text{Dirichlet}}(\pi | \gamma) \prod_{j=1}^{m} q_{\text{discrete}}(z_j | \phi_j)$$





M-step: Parameter Estimation

- For fixed (γ_d, ϕ_d), the lower bound is optimized over (α, β)
- Updates for word distributions

$$\beta_h(v) \propto \sum_{d=1}^{D} \sum_{j=1}^{m} \phi_{d,j}(h) \mathbb{1}_{w_{d,j}=v}$$

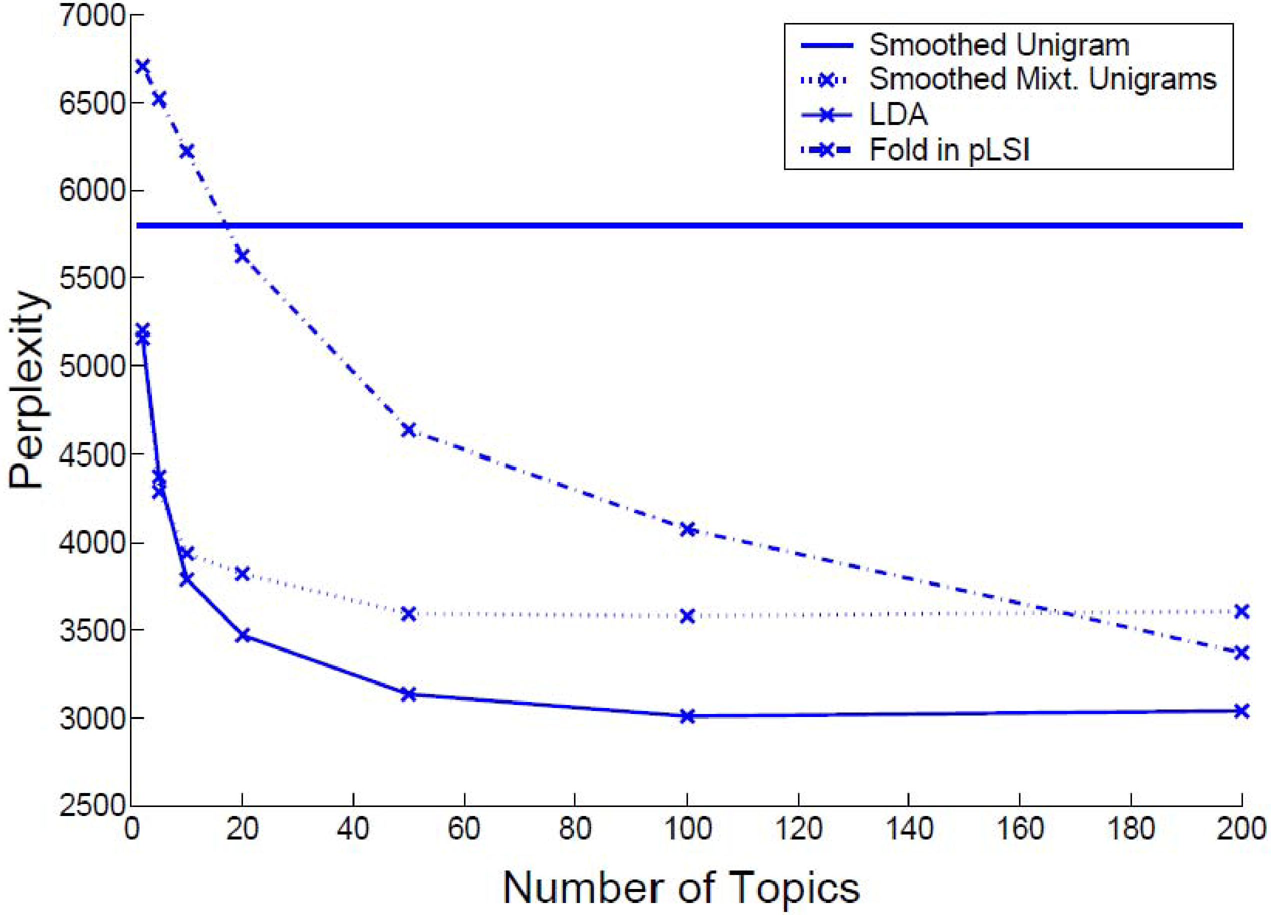
- α can be estimated using an efficient Newton method
- Alternate E- and M-steps till convergence

Results: Topics Inferred

"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

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.

Results: Perplexity Comparison



$$Perplexity(X) = \exp\left\{ -\frac{\sum_{i=1}^{n} \log p(\mathbf{x}_i)}{\sum_{i=1}^{n} m_i} \right\}$$

Aviation Safety Reports (NASA)

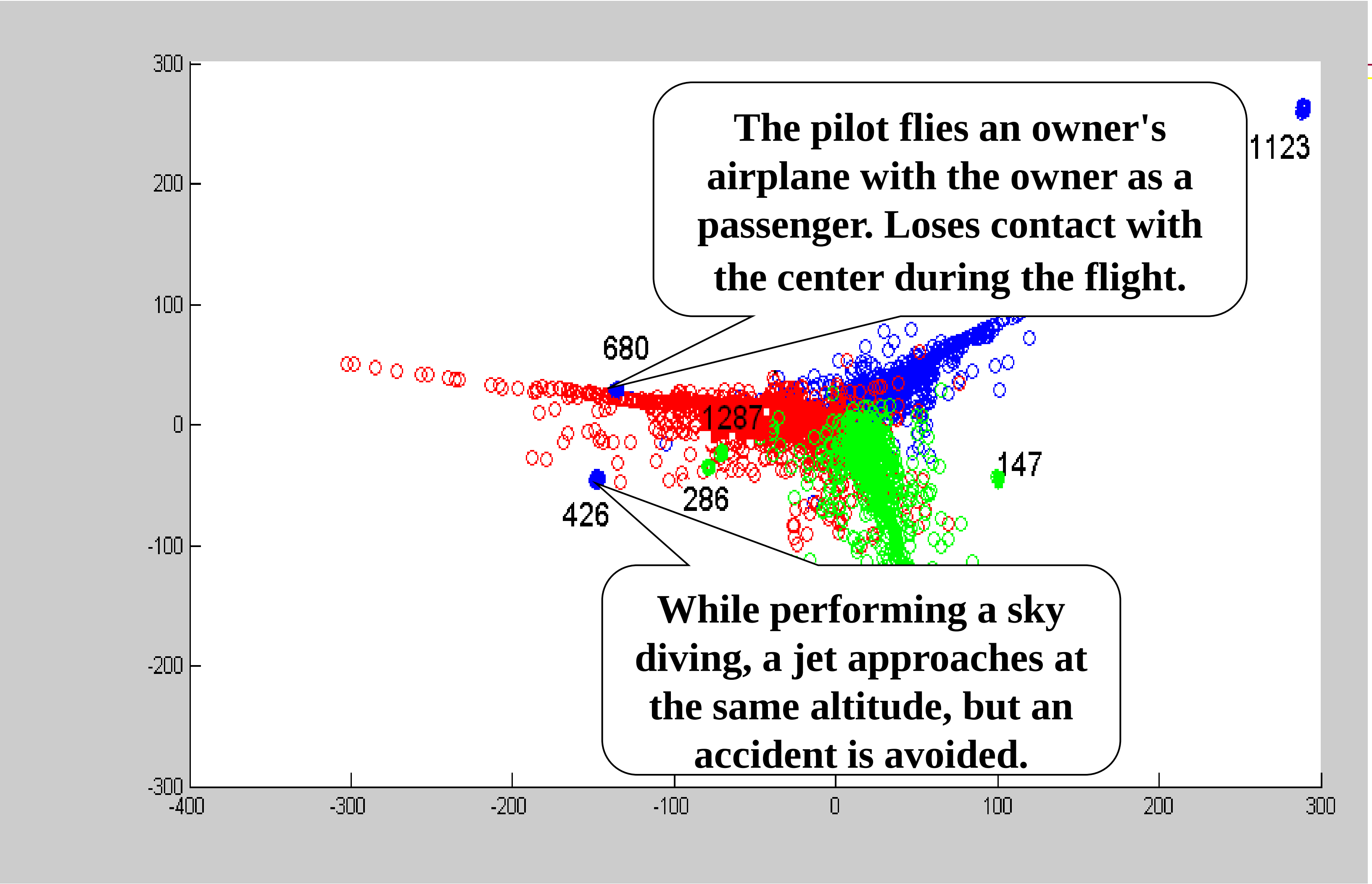
Results: NASA Reports I

Arrival Departure	**Passenger**	**Maintenance**
runway	passenger	maintenance
approach	attendant	engine
departure	flight	mel
altitude	seat	zzz
turn	medical	air craft
tower	captain	installed
air traffic control	attendants	check
heading	lavatory	inspection
taxi way	told	fuel
flight	police	Work

Results: NASA Reports II

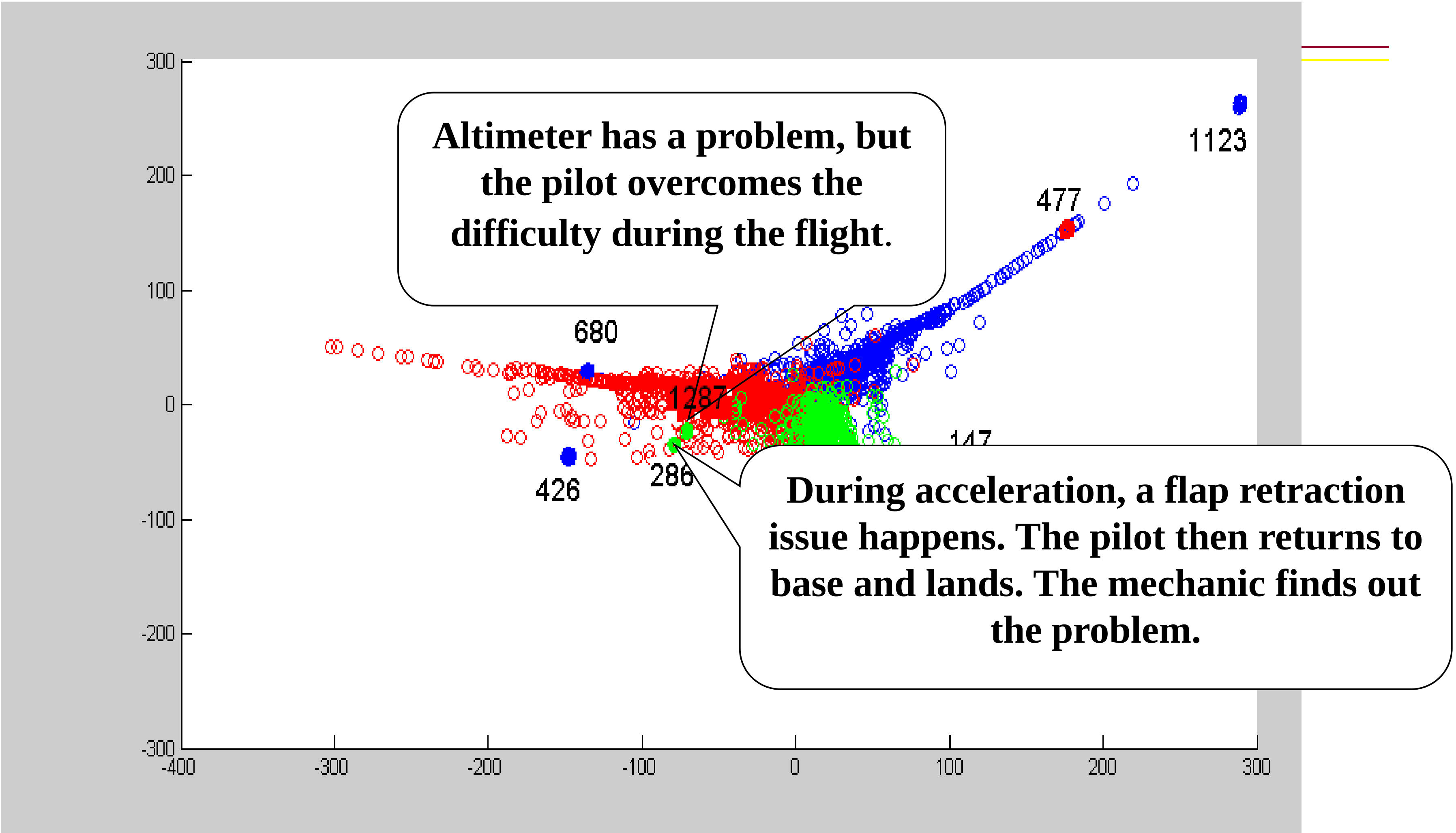
Medical Emergency	Wheel Maintenance	Weather Condition	Departure
medical	tire	knots	departure
passenger	wheel	turbulence	sid
doctor	assembly	aircraft	dme
attendant	nut	degrees	altitude
oxygen	spacer	ice	climbing
emergency	main	winds	mean sea level
paramedics	axle	wind	heading
flight	bolt	speed	procedure
nurse	missing	air speed	turn
aed	tires	conditions	degree

Two-Dimensional Visualization for Reports



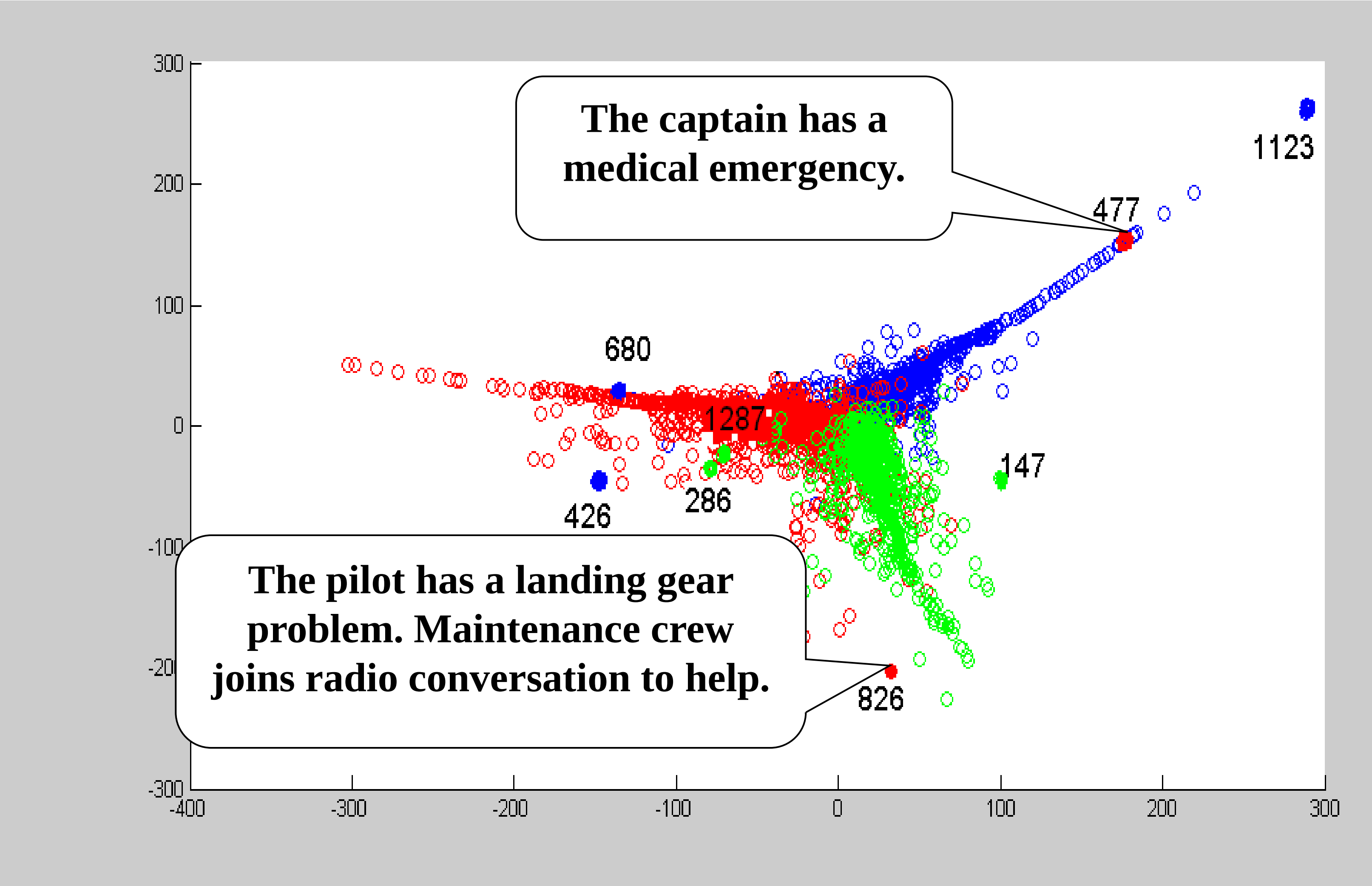
Red: Flight Crew Blue: Passenger Green: Maintenance

Two-Dimensional Visualization for Reports



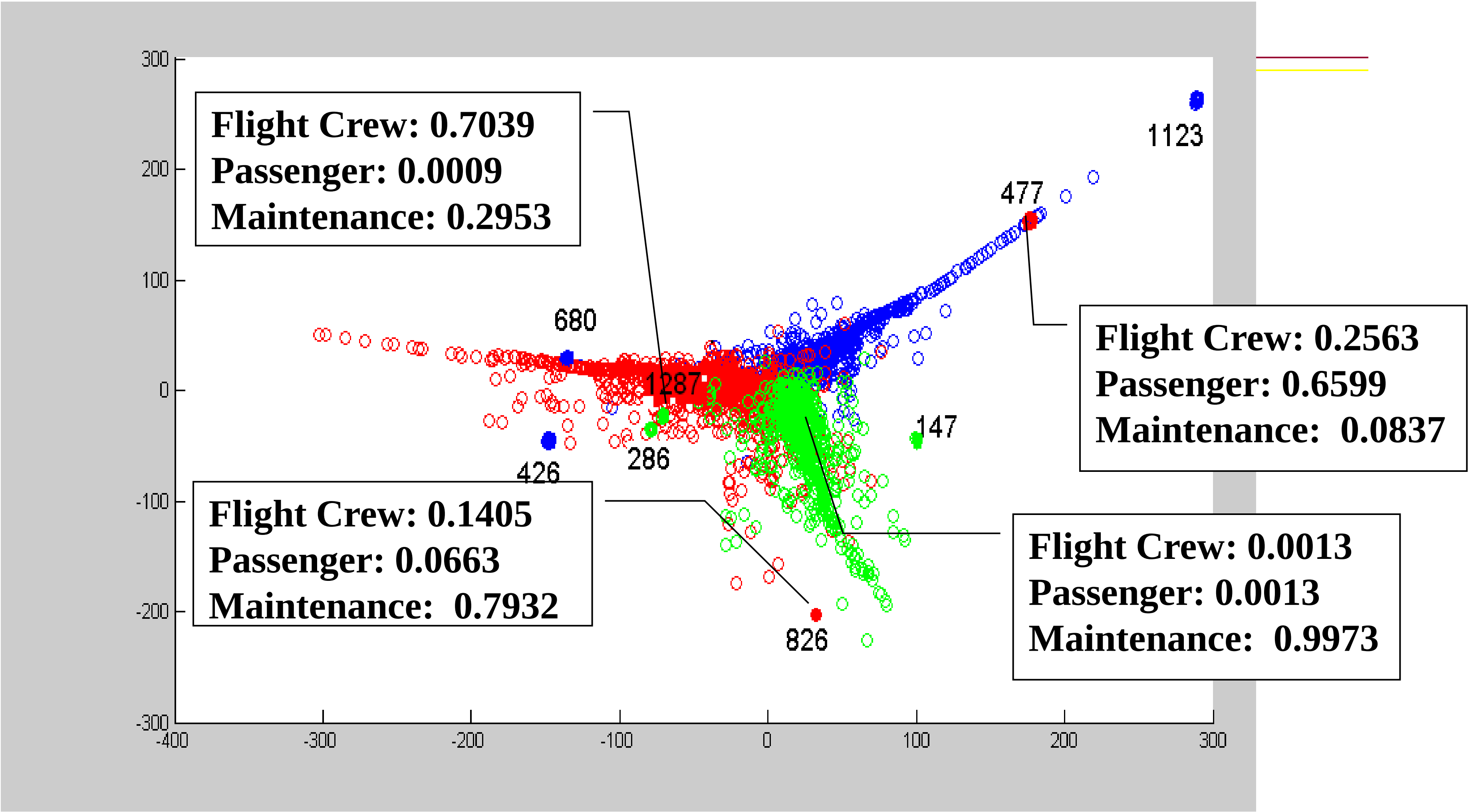
Red: Flight Crew Blue: Passenger Green: Maintenance

Two-Dimensional Visualization for Reports



Red: Flight crew Blue: Passenger Green: Maintenance

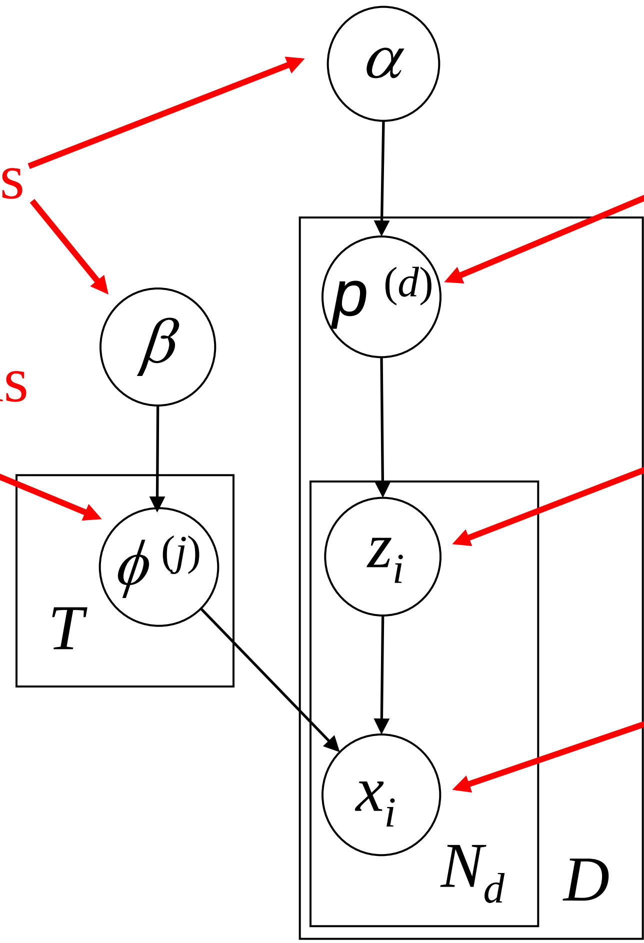
Mixed Membership of Reports



Red: Flight Crew Blue: Passenger Green: Maintenance

Smoothed Latent Dirichlet Allocation

Dirichlet priors



distribution over topics for each document

$p^{(d)} \sim \text{Dirichlet}(\alpha)$

distribution over words for each topic

$\phi^{(j)} \sim \text{Dirichlet}(\beta)$

topic assignment for each word

$z_i \sim \text{Discrete}(p^{(d)})$

word generated from assigned topic

$x_i \sim \text{Discrete}(\phi^{(z_i)})$

Stochastic Inference using Markov Chains

- Powerful family of approximate inference methods
  - Markov Chain Monte Carlo, Gibbs Sampling
- The basic idea
  - Need to marginalize over complex latent variable distribution

$$p(x|\theta) = \int_z p(x,z|\theta) = \int_z p(x|\theta)\, p(z|x,\theta) = E_{z\sim p(z|x,\theta)}[p(x|\theta)]$$

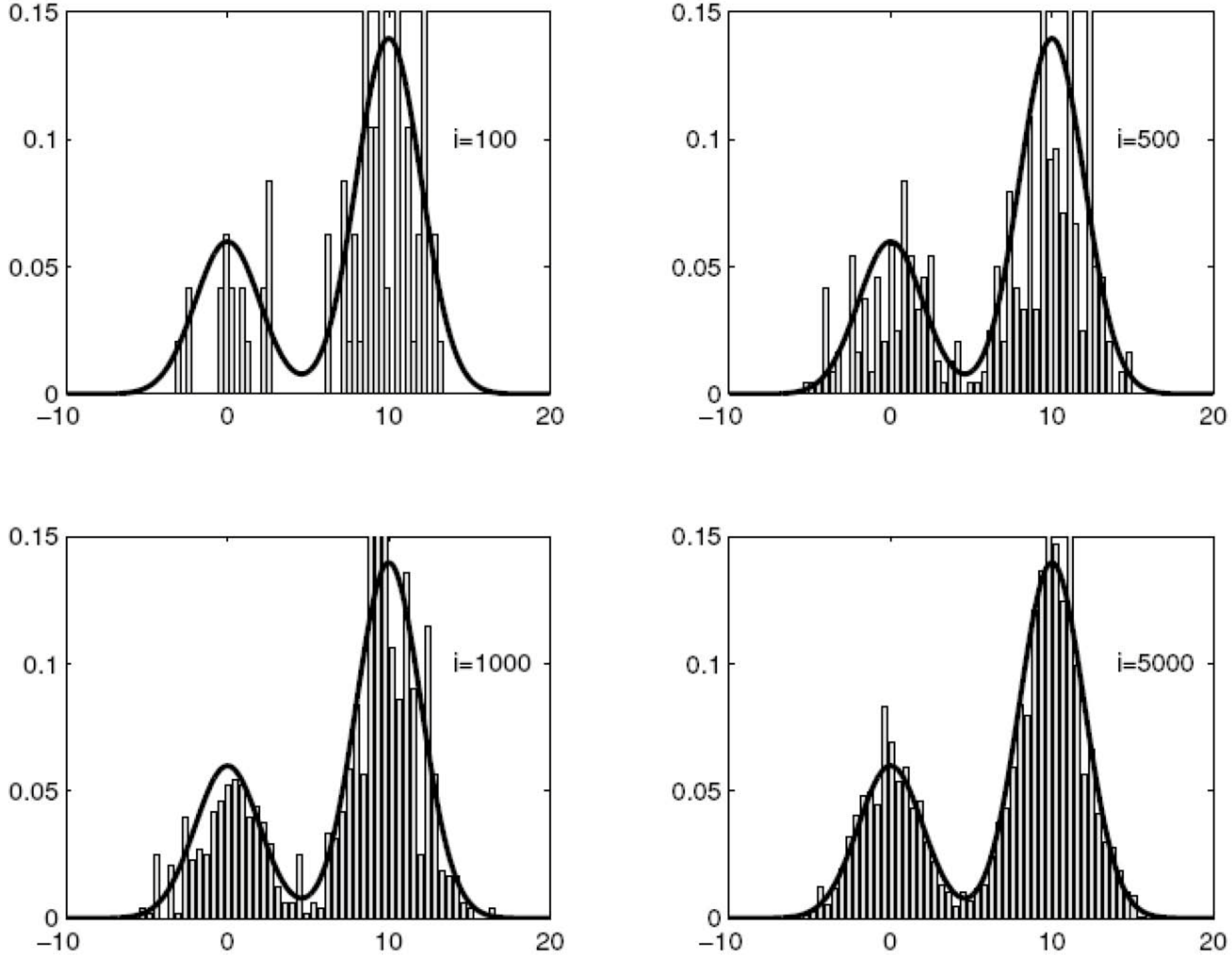
  - Draw 'independent' samples from $p(z|x,\theta)$
  - Compute sample based average instead of the full integral
- Main Issue: How to draw samples?
  - Difficult to directly draw samples from $p(z|x,\theta)$
  - Construct a Markov chain whose stationary distribution is $p(z|x,\theta)$
  - Run chain till 'convergence'
  - Obtain samples from $p(z|x,\theta)$

The Metropolis-Hastings Algorithm

- Most popular MCMC method
- Based on a proposal distribution $q(x^*|x)$
- Algorithm: For $i = 0, \ldots, (n-1)$
  - Sample $u \sim \mathcal{U}(0,1)$
  - Sample $x^* \sim q(x^*|x_i)$
  - Then

$$x_{i+1} = \begin{cases} x^* & \text{if } \; u < A(x_i, x^*) = \min\left\{1, \frac{p(x^*)q(x_i|x^*)}{p(x_i)q(x^*|x_i)}\right\} \\ x_i & \text{otherwise} \end{cases}$$

The Metropolis-Hastings Algorithm (Contd)



The Gibbs Sampler

- For a d-dimensional vector x, assume we know

$$p(x_j|x_{-j}) = p(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, \cdots, x_d)$$

- Gibbs sampler uses the following proposal distribution

$$q(x^*|x^{(i)}) = \begin{cases} p(x_j^*|x_{-j}^{(i)}) & \text{if } x_{-j}^* = x_{-j}^{(i)} \\ 0 & \text{otherwise} \end{cases}$$

- The acceptance probability

$$A(x^{(i)}, x^*) = \min\left\{1, \frac{p(x^*)q(x^{(i)}|x^*)}{p(x^{(i)})q(x^*|x^{(i)})}\right\} = 1$$

- Deterministic scan: All samples are accepted

Collapsed Gibbs Sampling for LDA

- Naive MCMC would sample all latent variables: (z, ϕ, θ)
- Observation: (ϕ, θ) can be marginalized in closed form
- We can obtain $p(x, z|\alpha, \beta)$ but cannot marginalize z
- Conditional distribution can be obtained in closed form:

$$P(z_{ij} = h | x_{ij} = w, z_{-ij}, x_{-ij}, \alpha, \beta) \propto \frac{n_{w,h}^{-ij} + \beta}{n_{\cdot,h}^{-ij} + D\beta}(n_{h,j}^{-ij} + \alpha)$$

  where, not including the current token,
  $n_{w,h}^{-ij}$ = # times word w got assigned to topic h
  $n_{\cdot,h}^{-ij}$ = total number of words assigned to topic h
  $n_{h,j}^{-ij}$ = # words from document j assigned to topic h
- Perform Gibbs sampling using the conditional distributions

Collapsed Variational Inference for LDA

- Recall that $p(x, z|\alpha, \beta)$ can be obtained in closed form
- However, we cannot marginalize over z
- We approximate $p(z|x, \alpha, \beta)$ with $q(z|x, \alpha, \beta)$
- As before, we have a variational lower bound on $\log p(x|\alpha, \beta)$
- The variational distribution is fully factorized

$$q(z|\gamma) = \prod_{d=1}^{D} \prod_{j=1}^{m} p_{\text{discrete}}(z_{dj}|\gamma_{dj})$$

- Exact variational inference can be expensive
- Approximations for efficient inference
  - Approximate sum of large number of Bernoulli variables with Gaussian
  - Second order Taylor approximation

Collapsed Variational Inference for LDA

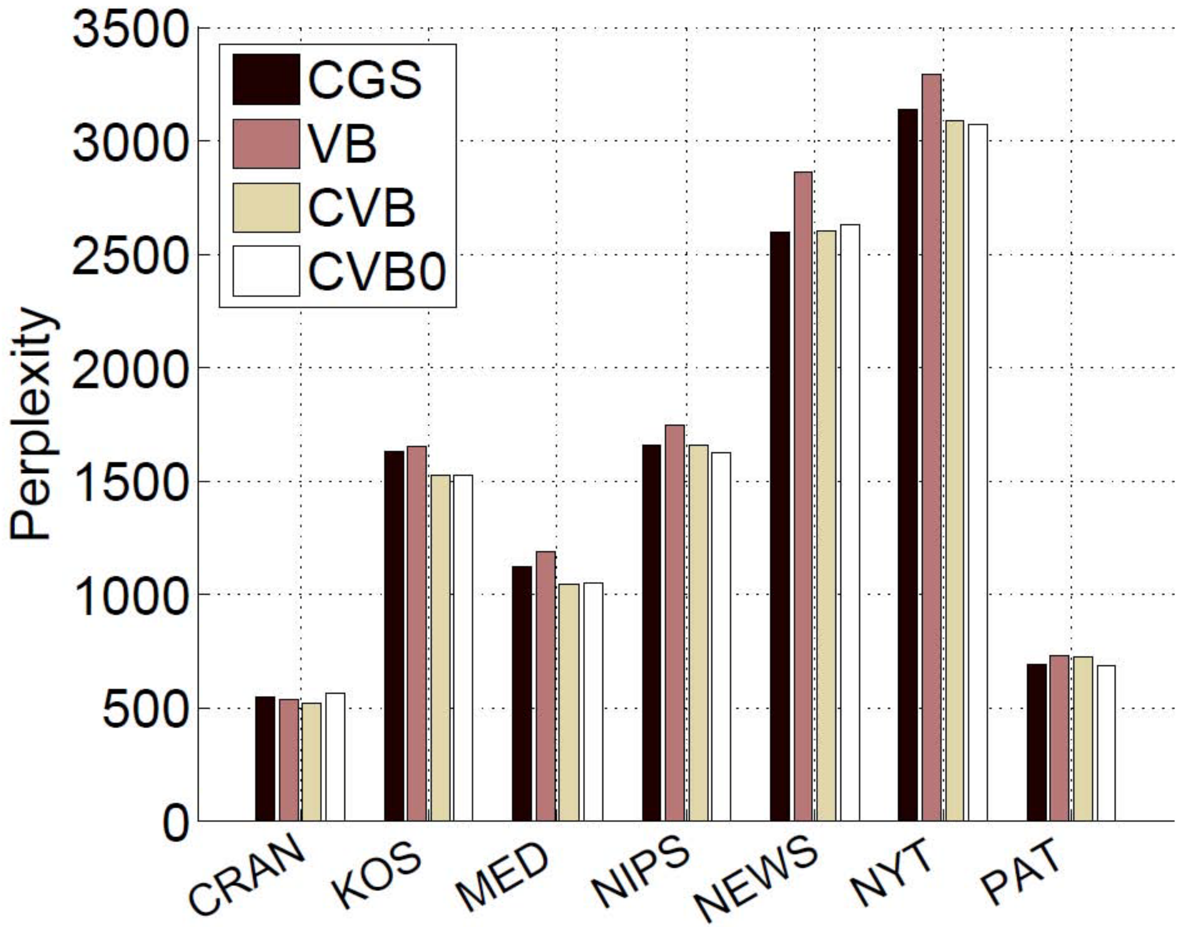
- With these approximations, the variational update is

$$\gamma_{d,j}(h) \propto \frac{n_{w,h}^{-ij} + \beta}{n_{\cdot,h}^{-ij} + D\beta}(n_{h,j}^{-ij} + \alpha) \exp\left(- \frac{v_{h,j}^{-ij}}{2(n_{h,j}^{-ij} + \alpha)^2} - \frac{v_{wh}^{-ij}}{2(n_{wh}^{-ij} + \beta)^2} + \frac{v_{\cdot,h}^{-ij}}{2(n_{\cdot,h}^{-ij} + D\beta)^2} \right)$$

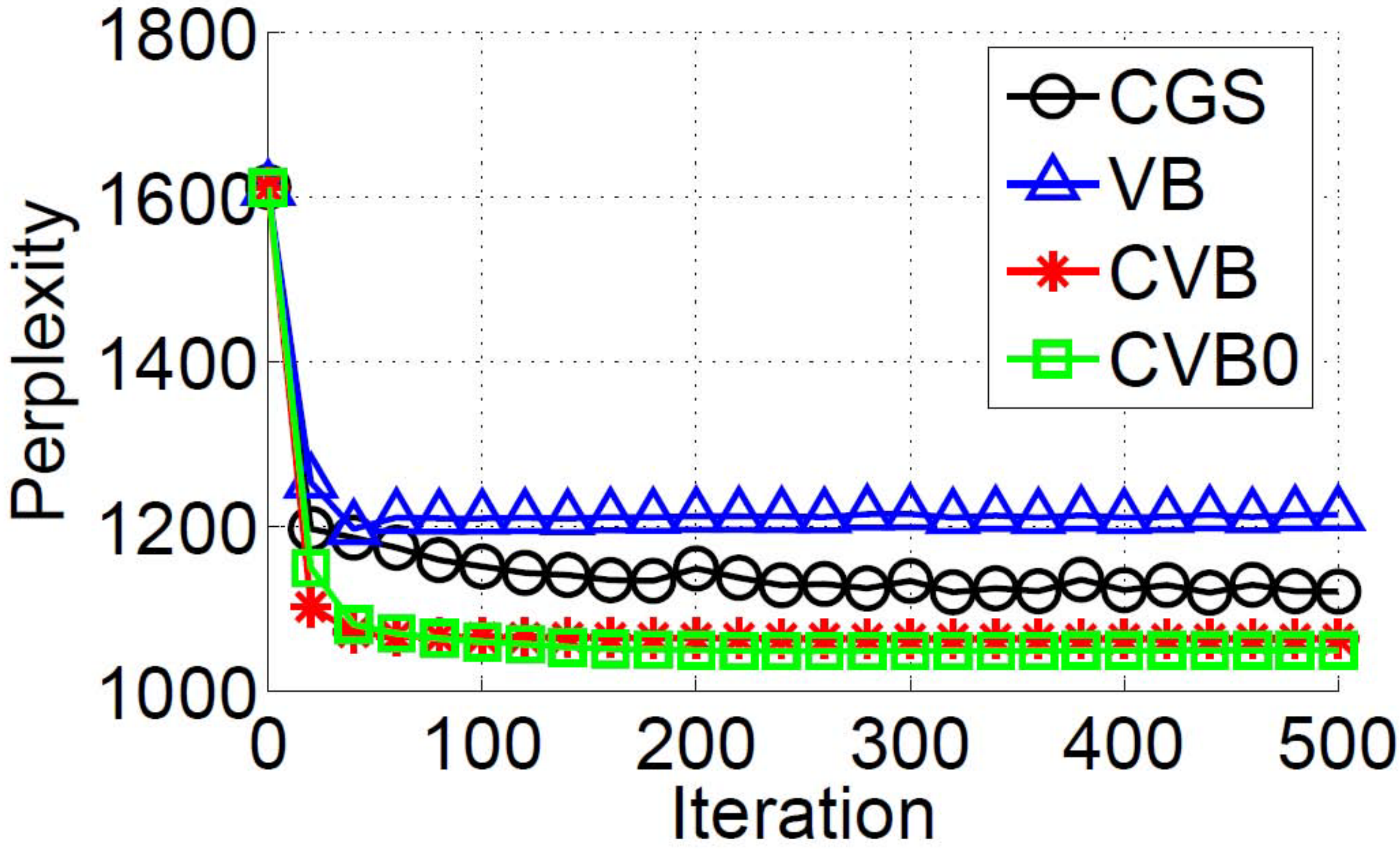
  where, not including the current token,
  - $n_{h,j}^{-ij} = \sum_{i' \neq i} \gamma_{i'jh}$, the expected number of tokens in document j assigned to topic h;
  - $v_{h,j}^{-ij} = \sum_{i' \neq i} \gamma_{i'jh}(1 - \gamma_{i'jh})$, the variance associated with the expected count; and similarly for other terms
- Ignoring the higher order information

$$\gamma_{d,j}(h) \propto \frac{n_{w,h}^{-ij} + \beta}{n_{\cdot,h}^{-ij} + D\beta}(n_{h,j}^{-ij} + \alpha)$$

Results: Comparison of Inference Methods



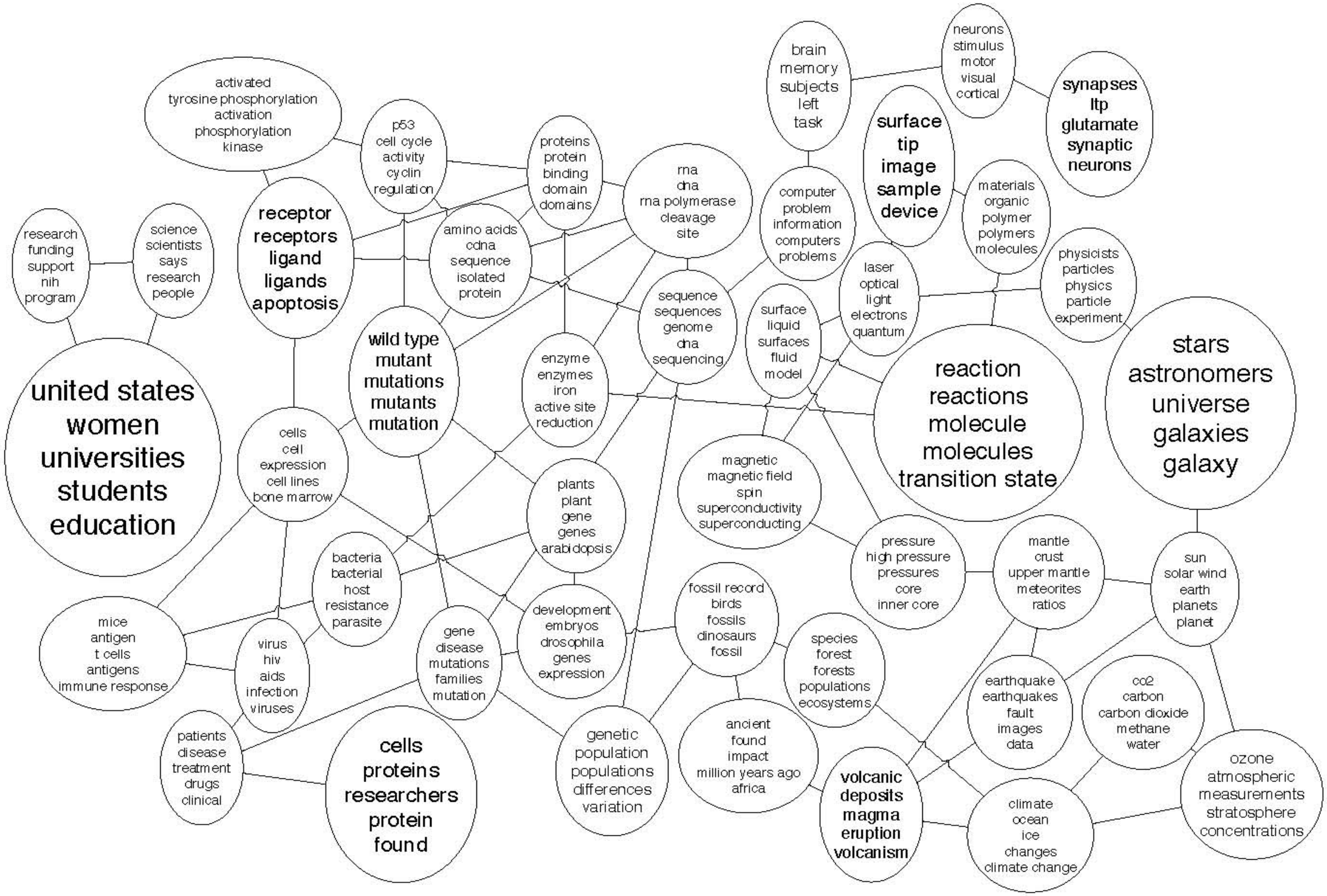
Results: Comparison of Inference Methods



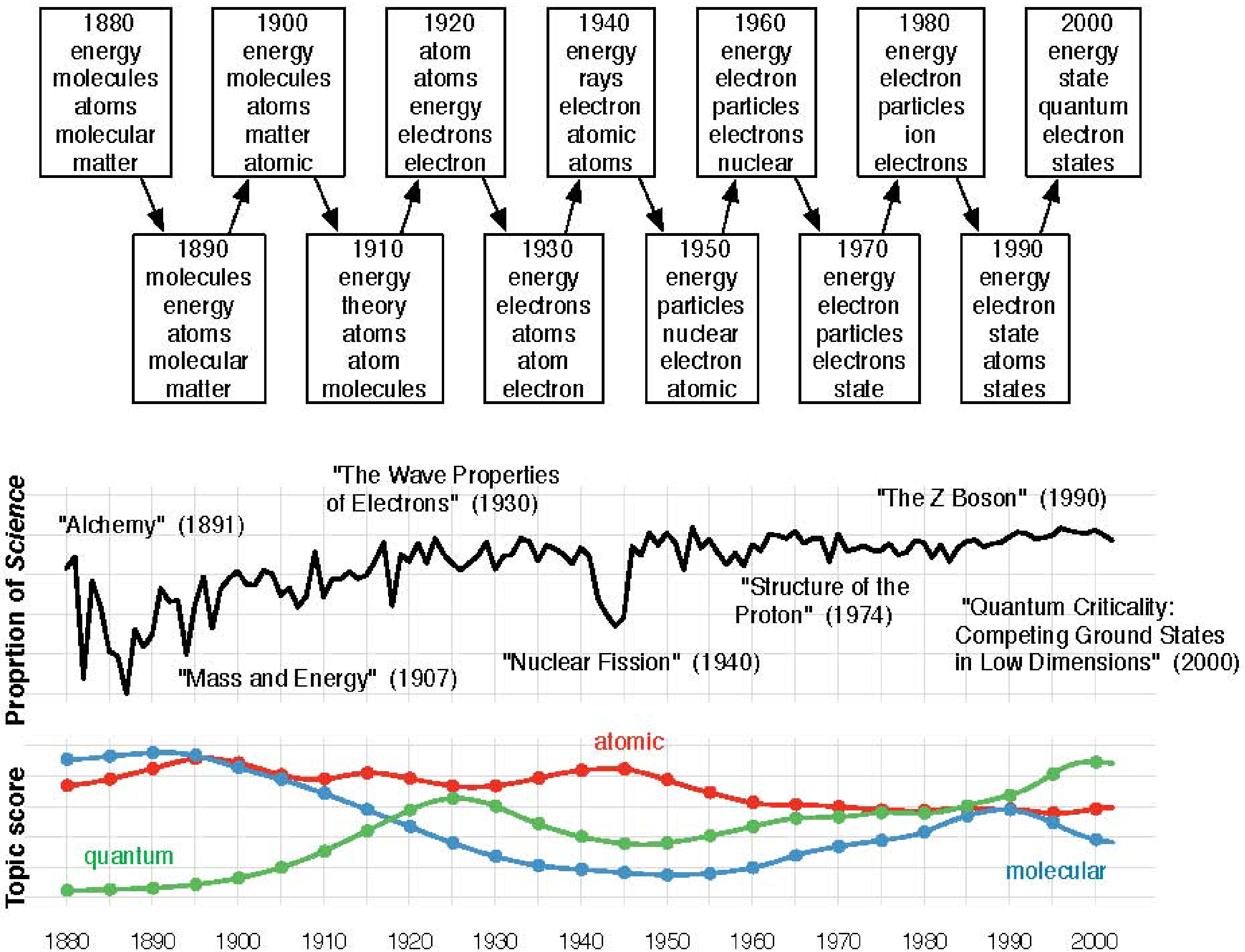
Generalizations

- Generalized Topic Models
  - Correlated Topic Models
  - Dynamic Topic Models, Topics over Time
  - Dynamic Topics with birth/death
- Mixed membership models over non-text data, applications
  - Mixed membership naïve-Bayes
  - Discriminative models for classification
  - Cluster Ensembles
- Nonparametric Priors
  - Dirichlet Process priors: Infer number of topics
  - Hierarchical Dirichlet processes: Infer hierarchical structures
  - Several other priors: Pachinko allocation, Gaussian Processes, IBP, etc.

CTM Results

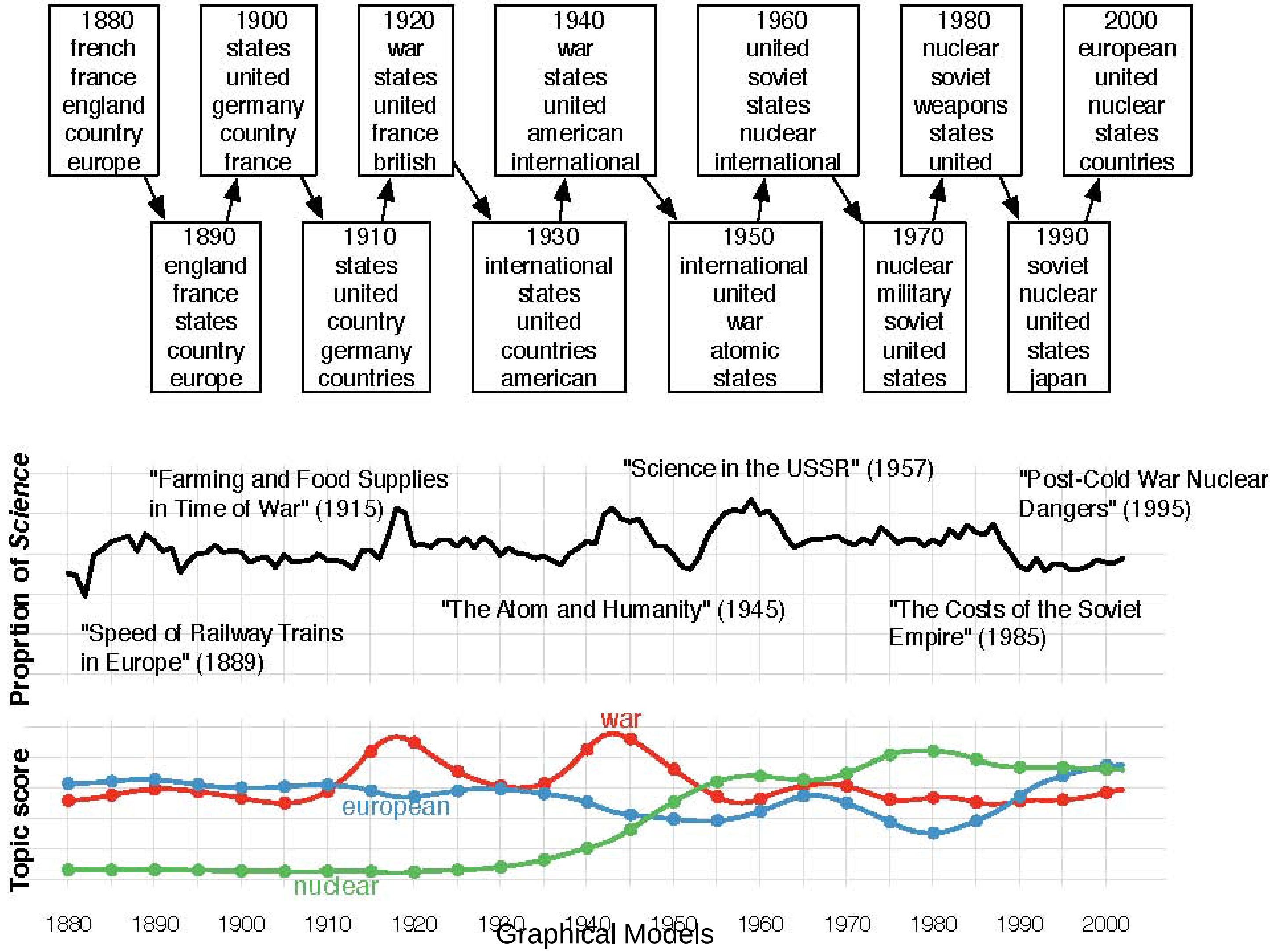

activated tyrosine phosphorylation activation phosphorylation kinase
p53 cell cycle activity cyclin regulation
proteins protein binding domain domains
rna dna rna polymerase cleavage site
brain memory subjects left task
neurons stimulus motor visual cortical
synapses ltp glutamate synaptic neurons
surface tip image sample device
computer problem information computers problems
materials organic polymer polymers molecules
research funding support nih program
science scientists says research people
receptor receptors ligand ligands apoptosis
amino acids cdna sequence isolated protein
physicists particles physics particle experiment
laser optical light electrons quantum
sequence sequences genome dna sequencing
surface liquid surfaces fluid model
wild type mutant mutations mutants mutation
enzyme enzymes iron active site reduction
stars astronomers universe galaxies galaxy
reaction reactions molecule molecules transition state
united states women universities students education
cells cell expression cell lines bone marrow
magnetic magnetic field spin superconductivity superconducting
plants plant gene genes arabidopsis
pressure high pressure pressures core inner core
mantle crust upper mantle meteorites ratios
sun solar wind earth planets planet
bacteria bacterial host resistance parasite
development embryos drosophila genes expression
fossil record birds fossils dinosaurs fossil
species forest forests populations ecosystems
mice antigen t cells antigens immune response
virus hiv aids infection viruses
gene disease mutations families mutation
earthquake earthquakes fault images data
co2 carbon carbon dioxide methane water
patients disease treatment drugs clinical
cells proteins researchers protein found
genetic population populations differences variation
ancient found impact million years ago africa
volcanic deposits magma eruption volcanism
climate ocean ice changes climate change
ozone atmospheric measurements stratosphere concentrations

DTM Results



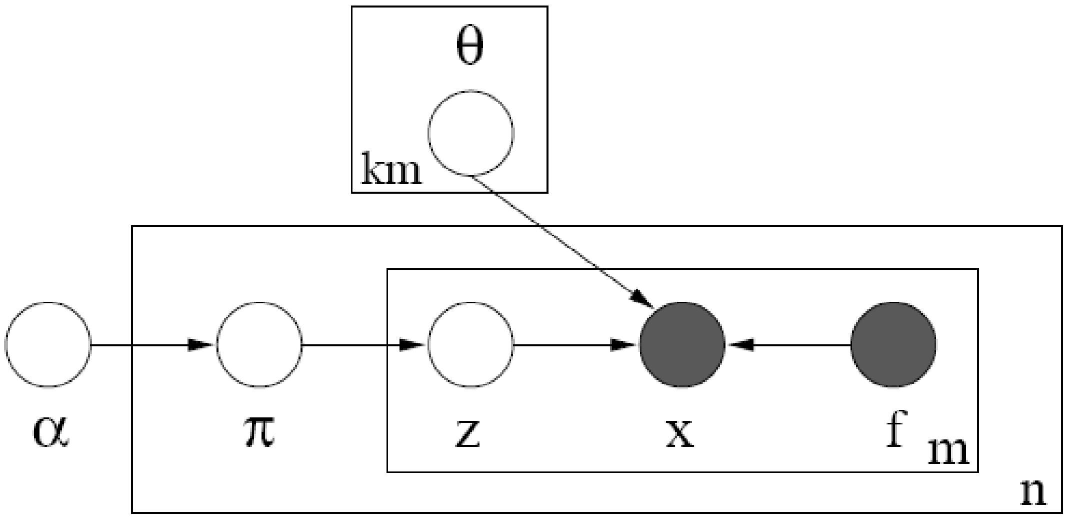
1880 energy molecules atoms molecular matter
1890 molecules energy atoms molecular matter
1900 energy molecules atoms matter atomic
1910 energy theory atoms atom molecules
1920 atom atoms energy electrons electron
1930 energy electrons atoms atom electron
1940 energy rays electron atomic atoms
1950 energy particles nuclear electron atomic
1960 energy electron particles electrons nuclear
1970 energy electron particles electrons state
1980 energy electron particles ion electrons
1990 energy electron state atoms states
2000 energy state quantum electron states
Proprtion of *Science*
"Alchemy" (1891)
"The Wave Properties of Electrons" (1930)
"The Z Boson" (1990)
"Mass and Energy" (1907)
"Nuclear Fission" (1940)
"Structure of the Proton" (1974)
"Quantum Criticality: Competing Ground States in Low Dimensions" (2000)
Topic score
atomic
quantum
molecular
1880 1890 1900 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000

DTM Results II

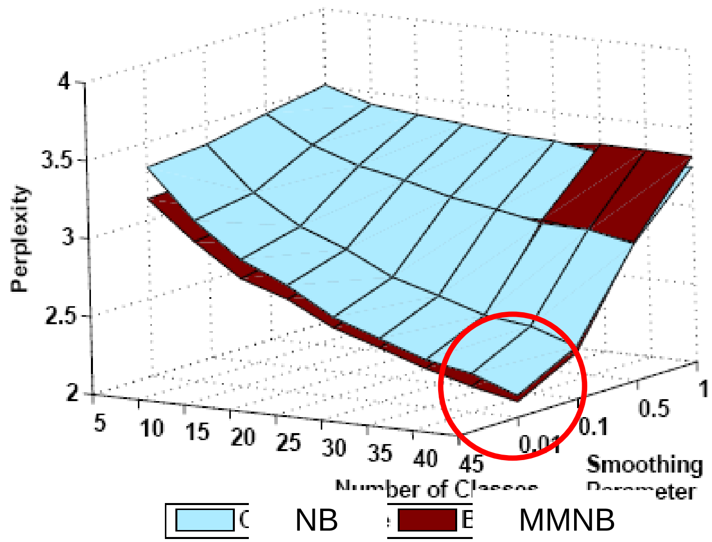


Mixed Membership Naïve Bayes

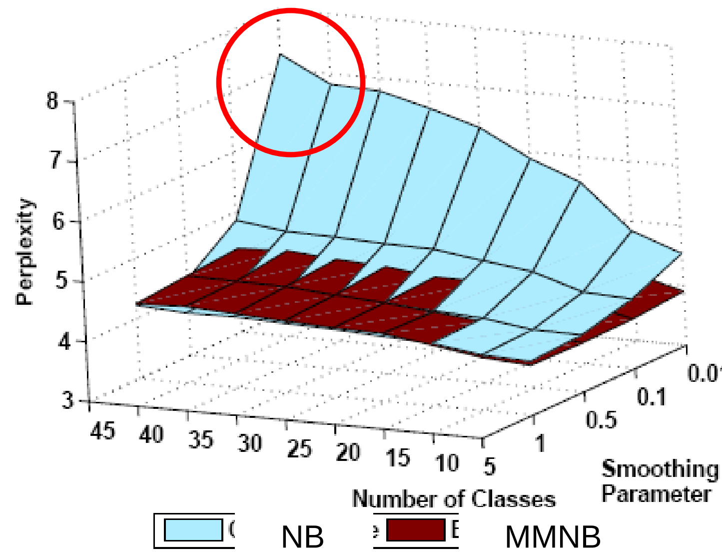
- For each data point,
  - Choose $\pi \sim \text{Dirichlet}(\alpha)$
- For each of observed features f_n:
  - Choose a class $z_n \sim \text{Discrete}(\pi)$
  - Choose a feature value x_n from $p(x_n|z_n,f_n,\Theta)$, which could be Gaussian, Poisson, Bernoulli…



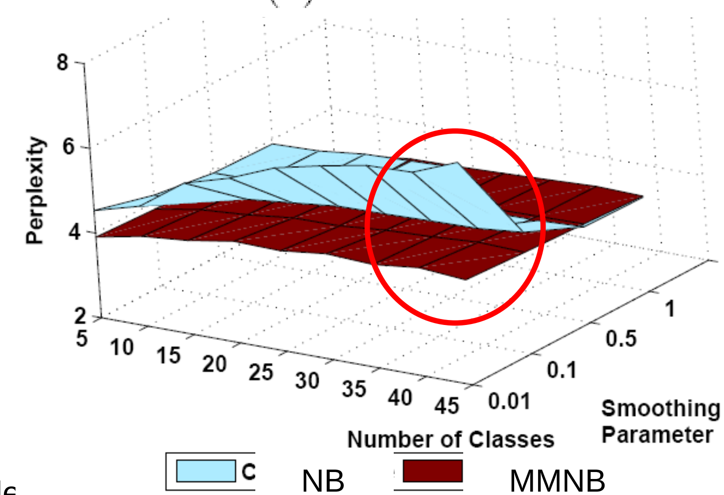
MMNB vs NB: Perplexity Surfaces



(a) Training set



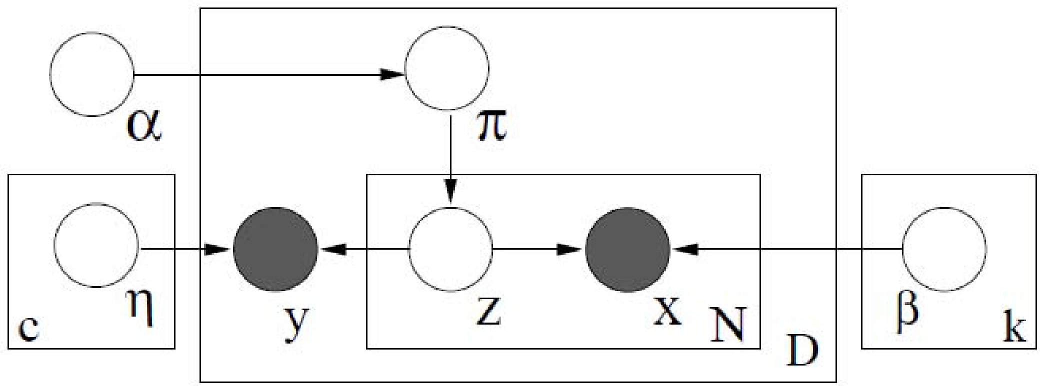
(b) Test set



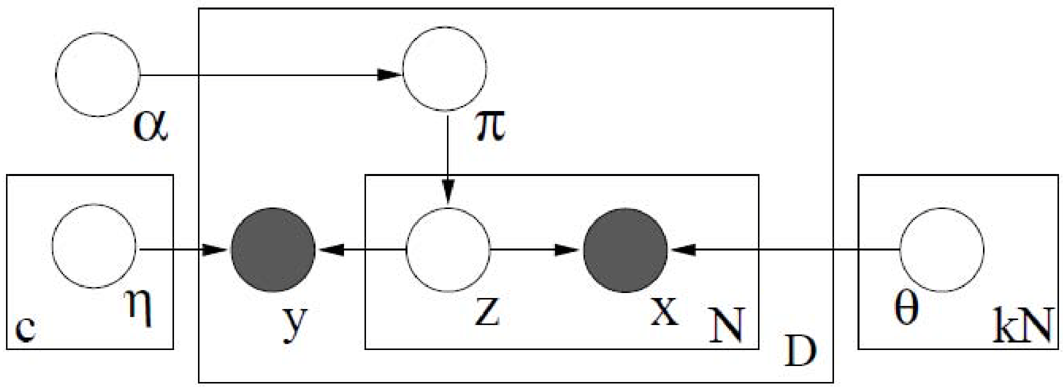
•MMNB typically achieves a lower perplexity than NB

•On test set, NB shows overfitting, but MMNB is stable and robust.

Discriminative Mixed Membership Models



(a) DLDA



(b) DMNB

Results: DLDA for text classification

	Nasa	Classic3	Diff	Sim	Same
Fast DLDA	0.9301±0.0128	**0.6866±0.0245**	**0.9823±0.0083**	**0.8718±0.0182**	**0.8468±0.0190**
vMF	0.9216±0.0113	0.6509±0.0246	0.9530±0.0071	0.7447±0.0214	0.7600±0.0347
NB	**0.9334±0.0094**	0.6766±0.0230	0.9813±0.0069	0.8613±0.0216	0.8410±0.0262
LR	0.9209±0.0157	0.6396±0.0252	0.9553±0.0157	0.6750±0.1330	0.4823±0.1283
SVM	0.9192±0.0146	0.6854±0.0278	0.9563±0.0105	0.8357±0.0156	0.8120±0.2030

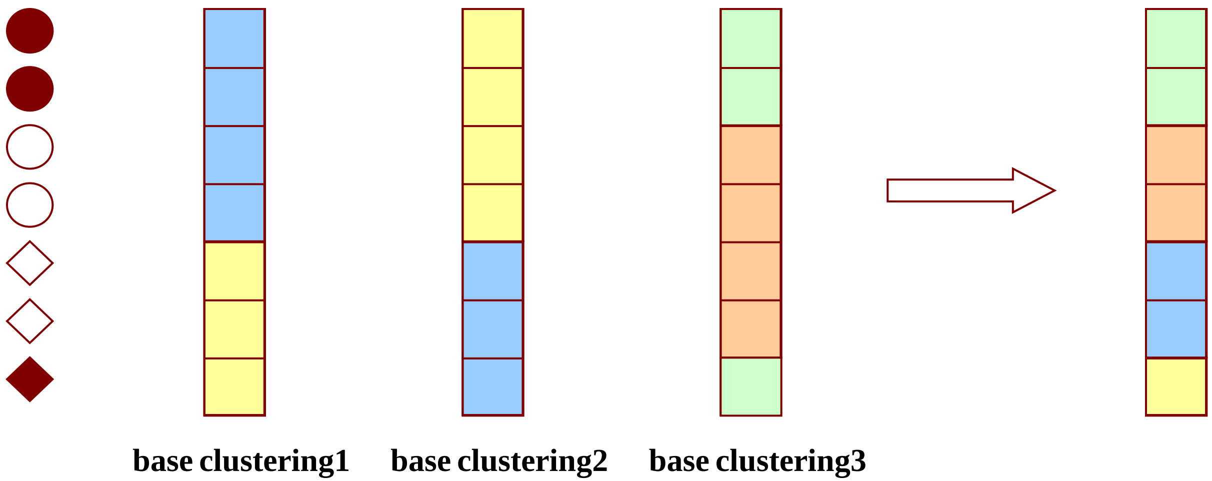
Generally, Fast DLDA has a higher accuracy on most of the datasets

Topics from DLDA

cabin	flight	ice	aircraft	flight
descent	hours	aircraft	gate	smoke
pressurization	time	flight	ramp	cabin
emergency	crew	wing	wing	passenger
flight	day	captain	taxi	aircraft
aircraft	duty	icing	stop	captain
pressure	rest	engine	ground	cockpit
oxygen	trip	anti	parking	attendant
atc	zzz	time	area	smell
masks	minutes	maintenance	line	emergency

Cluster Ensembles

- Combining multiple base clusterings of a dataset



- Robust and stable
- Distributed and scalable
- Knowledge reuse, privacy preserving

Problem Formulation

- Input & Output

Data points

	λ_1	λ_2	$\cdots$	λ_M
x_1	2	1	$\cdots$	2
x_2	1	3	$\cdots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_N	3	2	$\cdots$	6

Base clusterings

λ^*
1
3
$\vdots$
6

Consensus clustering

Results: State-of-the-art vs Bayesian Ensembles

algorithms / dataset	The results of base clusterings K-means		MCLA		CSPA		HGPA		MM		K-means cluster ensemble		G-BCE		V-BCE random initialization	
	Max	average	Max	average	Max	average	Max	average	Max	average	Max	average	Max	average	Max	average
iris	0.8867	0.6267	0.8867	0.8867	0.9533	**0.9167**	0.7333	0.7333	0.9067	0.8867	0.5267	0.5267	0.9533	0.8697	**0.9600**	0.8911
wdbc	0.8541	0.7595	0.8840	0.8840	0.8840	0.8840	0.5518	0.5188	0.8840	0.8840	0.8840	0.8689	**0.8893**	**0.8893**	**0.8893**	0.8840
ionosphere	0.7123	0.6906	0.7123	0.7046	0.6952	0.6952	0.6353	0.6063	0.7179	0.7111	0.7094	0.7094	0.7236	0.7073	**0.7749**	**0.7123**
glass	0.5421	0.5140	0.5187	0.4766	0.4393	0.4393	0.4439	0.4234	0.5748	0.5519	0.5093	0.4363	0.5514	0.4867	**0.6121**	**0.5526**
bupa	0.4841	0.4537	0.5652	0.5652	0.5710	**0.5710**	0.5188	0.5075	0.5710	0.5586	0.5565	0.5164	0.5710	**0.5710**	**0.5942**	0.5664
pima	0.6602	0.5751	0.6602	0.6602	0.5065	0.5065	0.5260	0.5163	0.6654	0.6503	0.6029	0.6029	0.6615	0.6445	**0.7044**	**0.6612**
wine	0.6629	0.5904	0.7247	0.7247	**0.7416**	**0.7416**	0.5562	0.5250	0.7247	0.7129	0.4775	0.4775	0.6966	0.6559	0.7247	0.7247
magic04	0.6491	0.6252	0.6491	0.6491	×	×	0.6491	0.6235	0.6530	0.6231	0.6491	0.6250	0.6491	0.6491	**0.6531**	**0.6497**
balance	0.5936	0.5114	0.5216	0.5188	0.5408	0.5408	0.4256	0.4256	**0.6016**	**0.5514**	0.5824	0.5824	0.5714	0.5150	0.5968	0.5293
segmentation	0.5710	0.5574	0.5657	0.5657	0.5810	0.5810	0.5419	0.4543	0.6233	0.5817	0.5710	0.5142	0.5233	0.5233	**0.6362**	**0.5854**

Part III: Graphical Models for Matrix Analysis

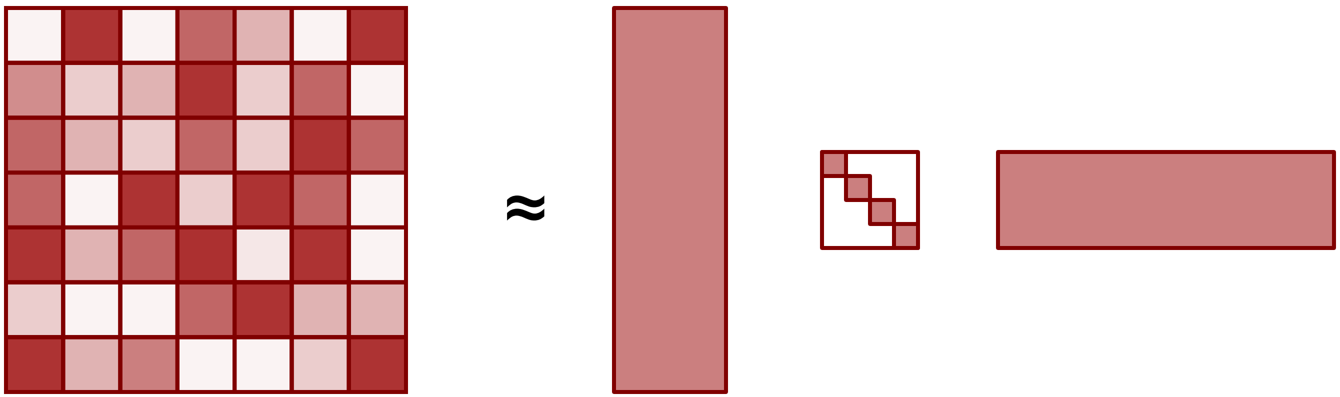
- Probabilistic Matrix Factorizations

- Probabilistic Co-clustering

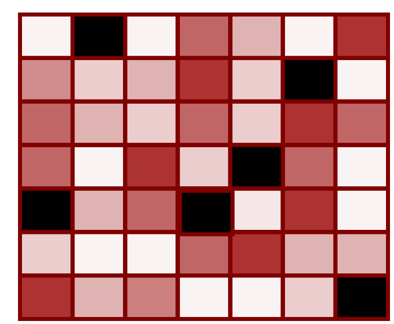
- Stochastic Block Structures

Matrix Factorization

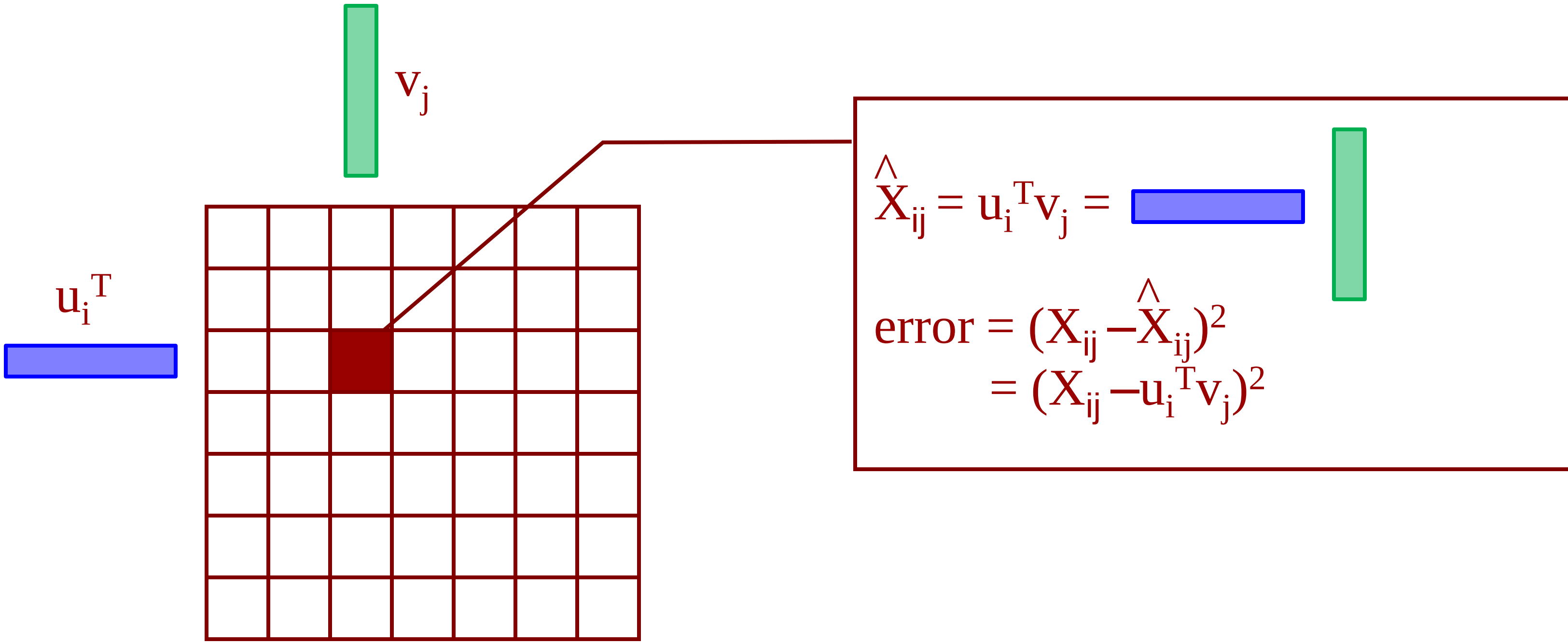
- Singular value decomposition



- Problems
  - Large matrices, with millions of row/colums
    - SVD can be rather slow
  - Sparse matrices, most entries are missing
    - Traditional approaches cannot handle missing entries

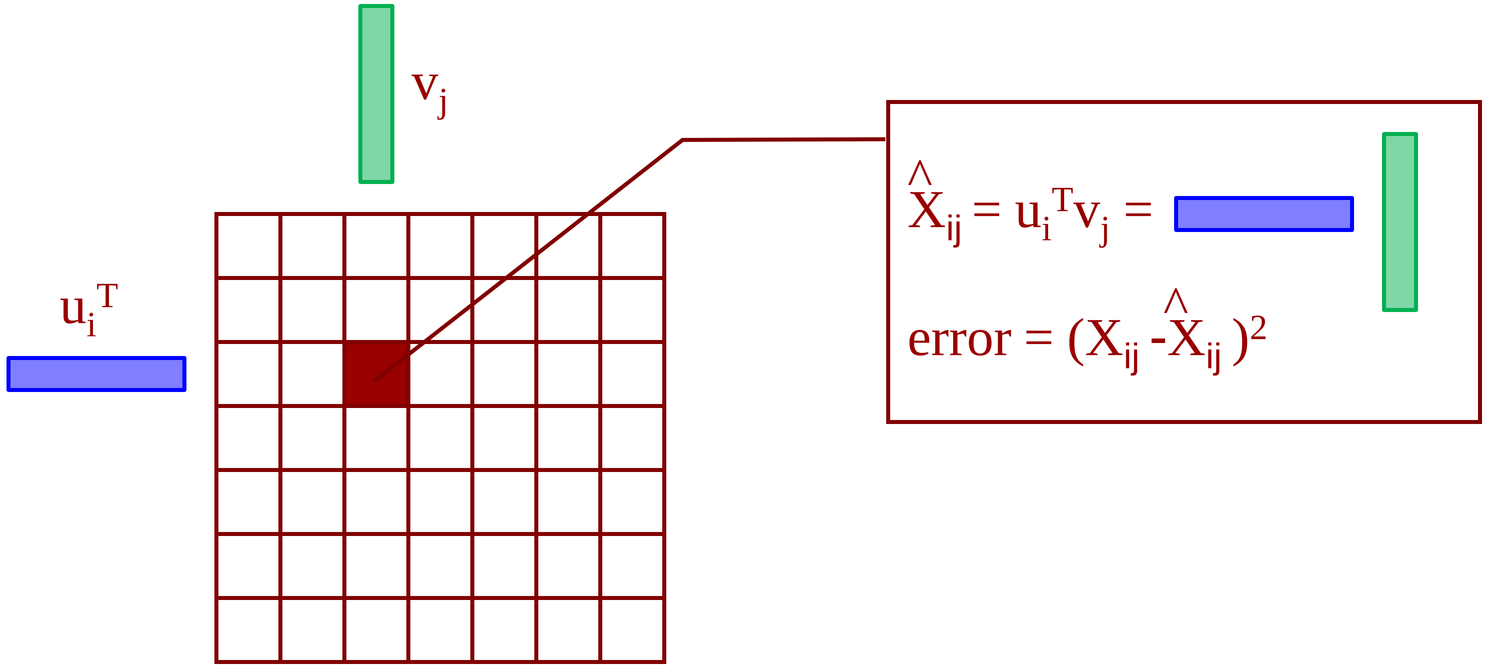


Matrix Factorization: “Funk SVD”



- Model $X \in R^{n \times m}$ as UV^T where
  - U is a $R^{n \times k}$, V is $R^{m \times k}$
  - Alternatively optimize U and V

Matrix Factorization (Contd)



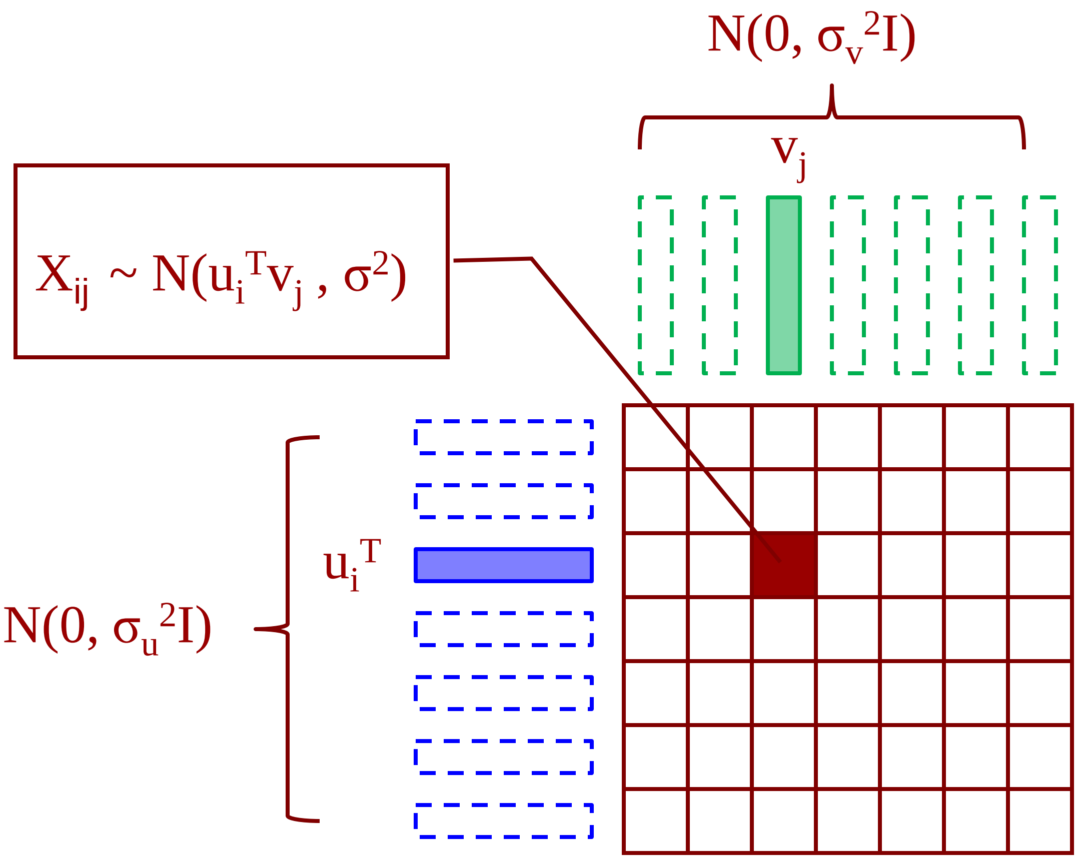
- Gradient descent updates

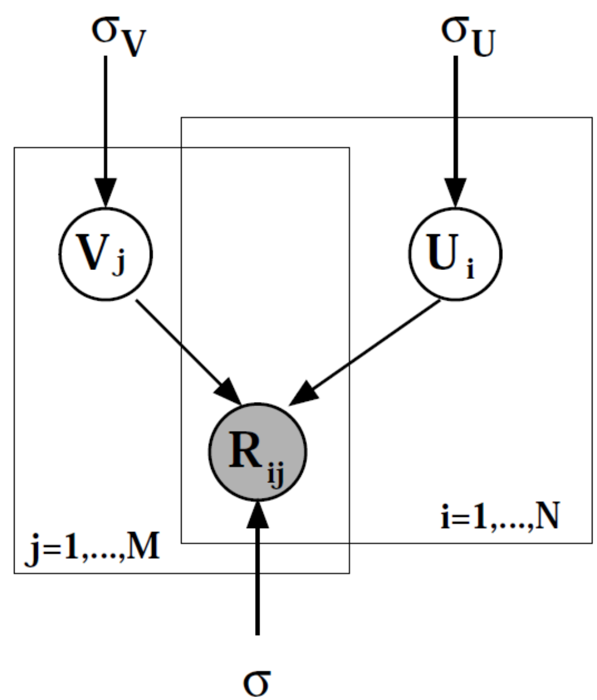
$$u_{ik}^{(t+1)} = u_{ik}^{(t)} + \eta\, (X_{ij} - \hat{X}_{ij})\, v_{jk}^{(t)}$$

$$v_{jk}^{(t+1)} = v_{jk}^{(t)} + \eta\, (X_{ij} - \hat{X}_{ij})\, u_{jk}^{(t)}$$

Probabilistic Matrix Factorization (PMF)

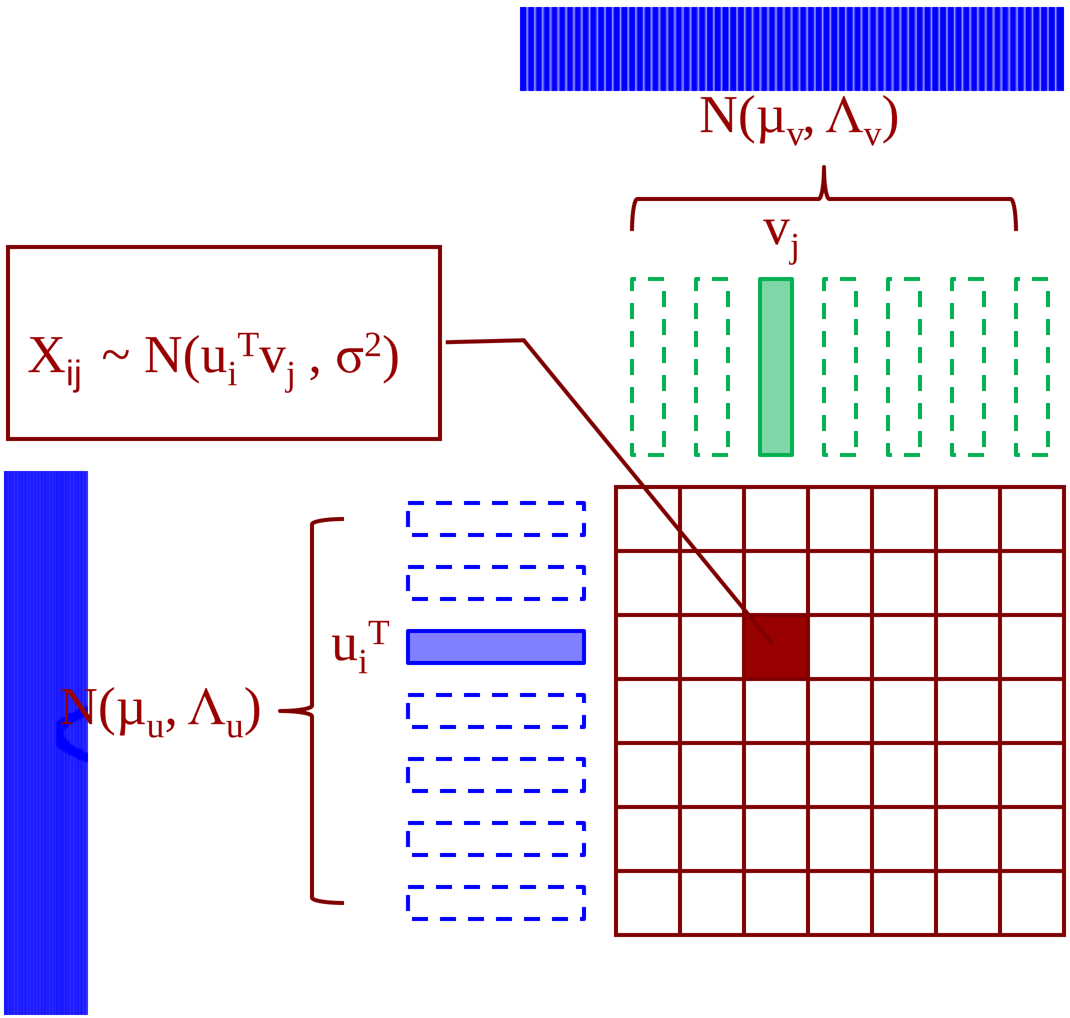
$u_i^T \sim N(0, \sigma_u^2 I)$
$v_j \sim N(0, \sigma_v^2 I)$
$R_{ij} \sim N(u_i^T v_j , \sigma^2)$



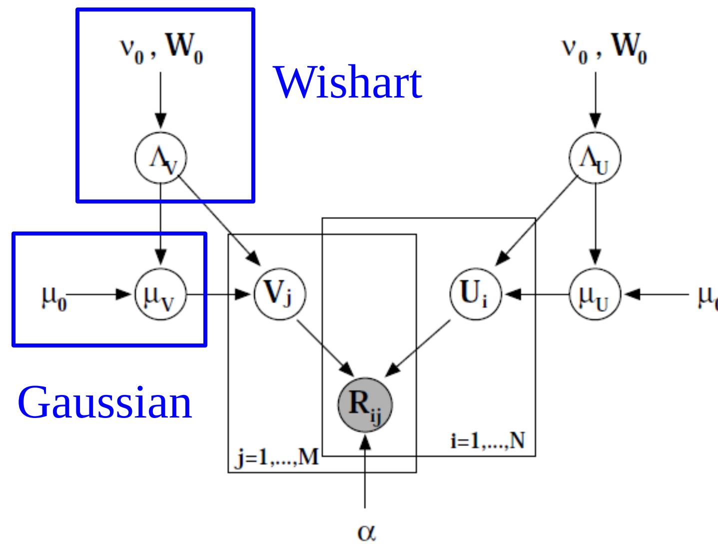


Inference using gradient descent

Bayesian Probabilistic Matrix Factorization

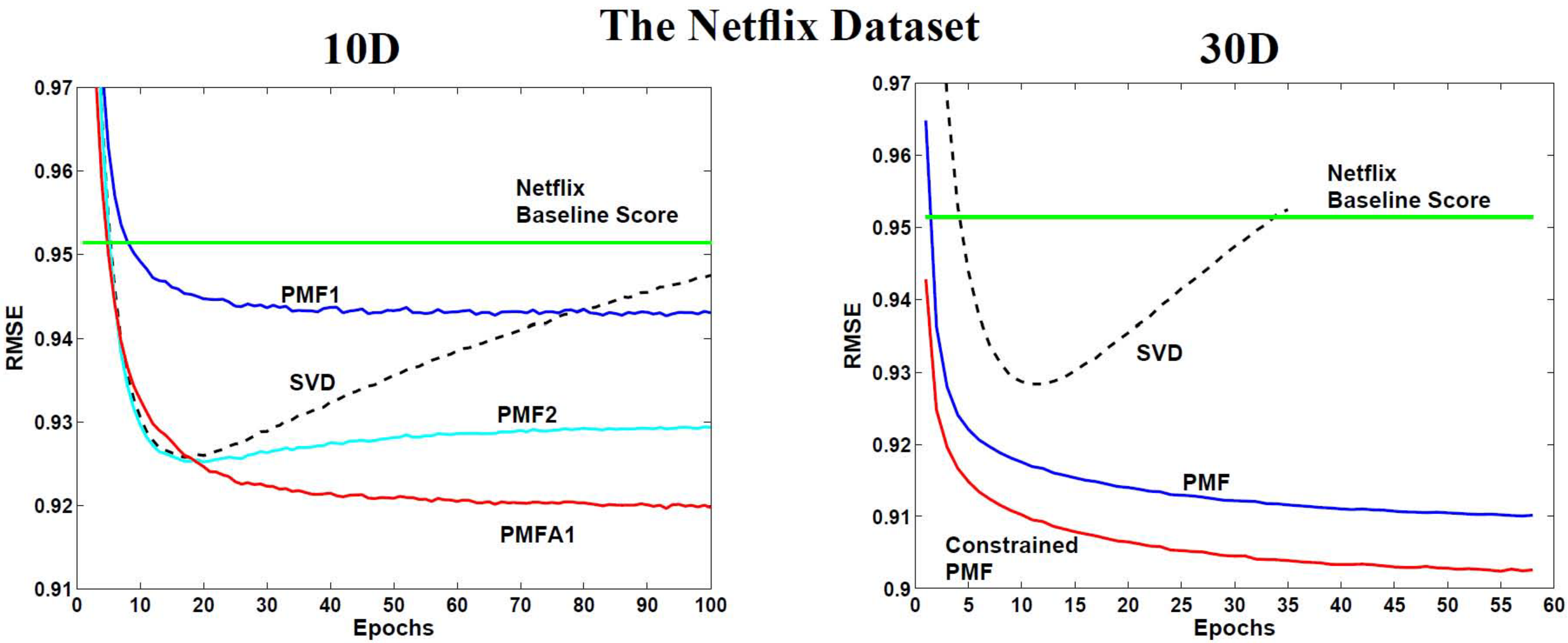


$$\mu_u \sim N(\mu_0, \Lambda_u),\ \Lambda_u \sim W(\nu_0, W_0)$$
$$\mu_v \sim N(\mu_0, \Lambda_v),\ \Lambda_v \sim W(\nu_0, W_0)$$
$$u_i \sim N(\mu_u, \Lambda_u)$$
$$v_j \sim N(\mu_v, \Lambda_v)$$
$$R_{ij} \sim N(u_i^T v_j, \sigma^2)$$

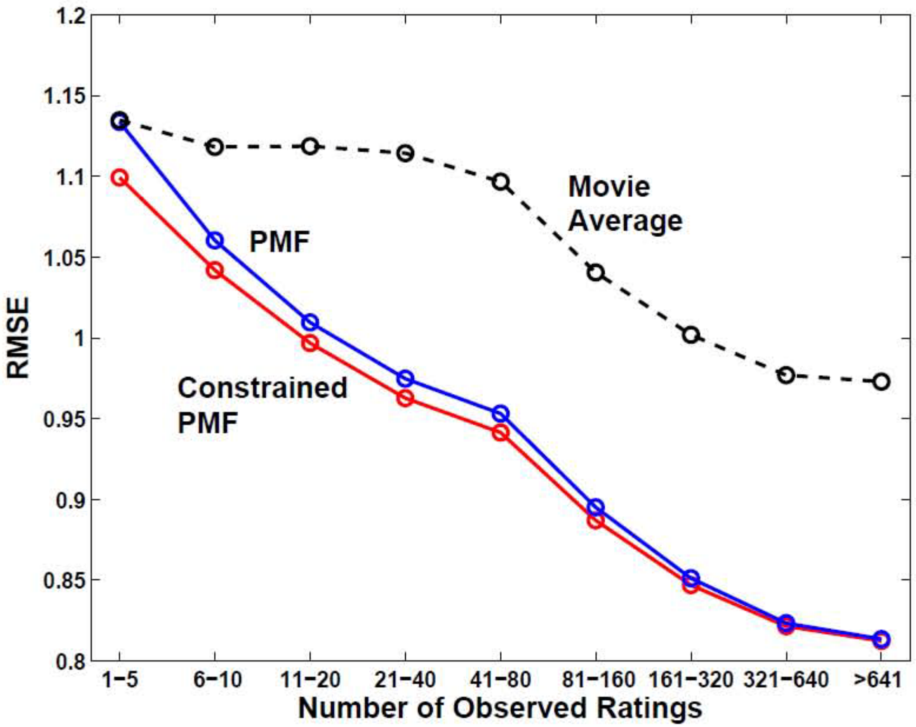


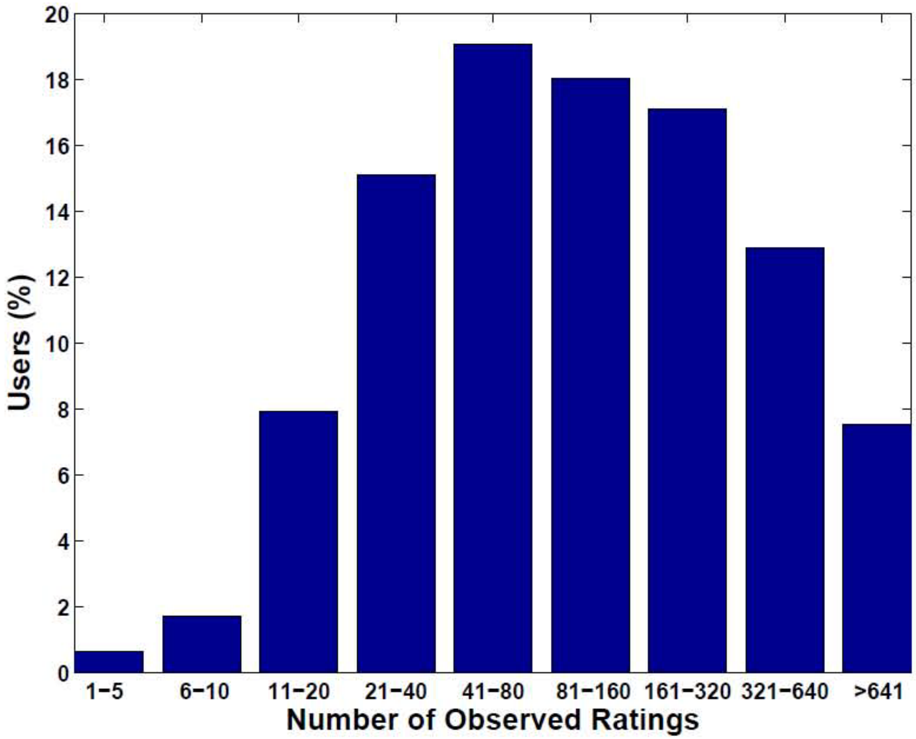
Inference using MCMC

Results: PMF on the Netflix Dataset

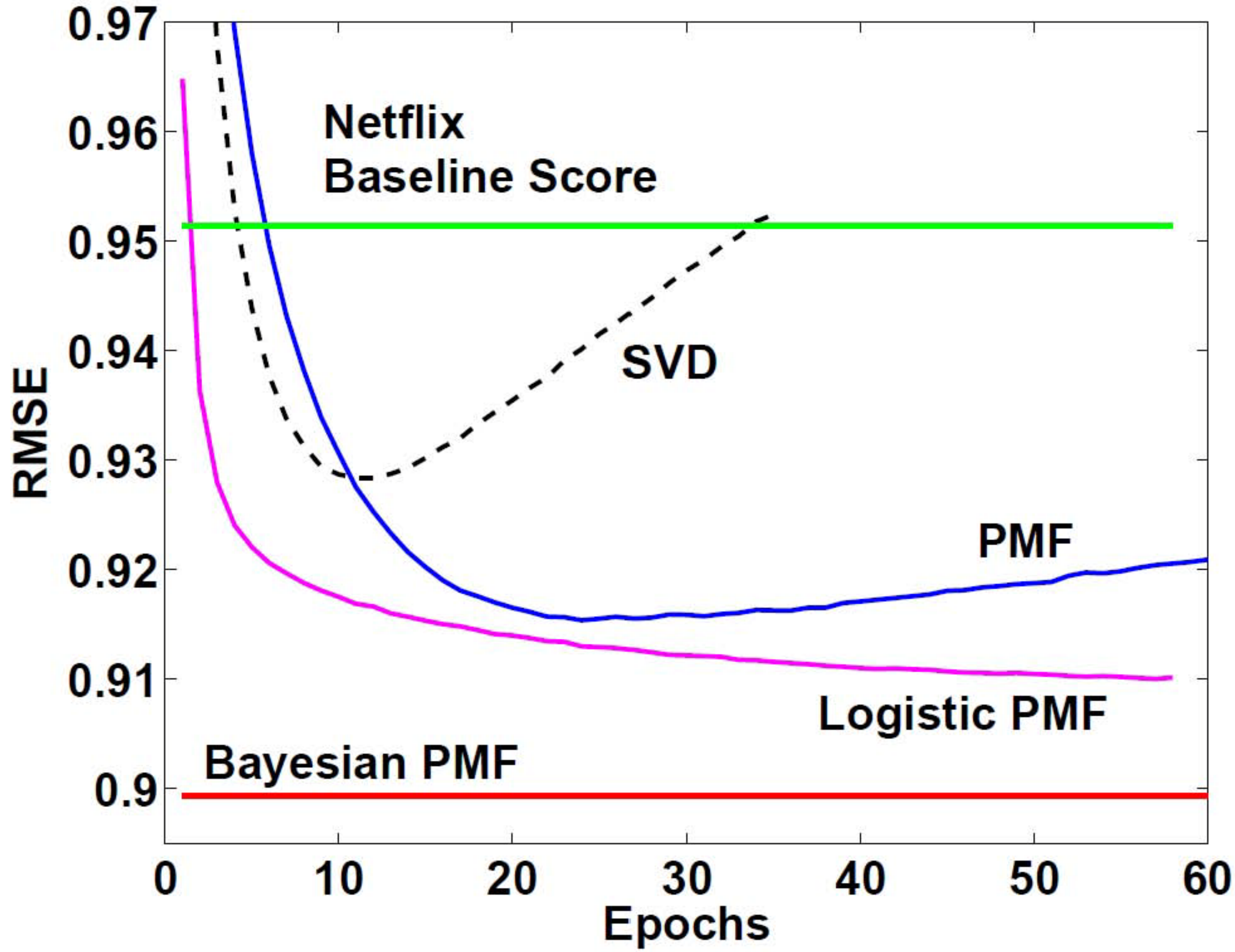


Results: PMF on the Netflix Dataset

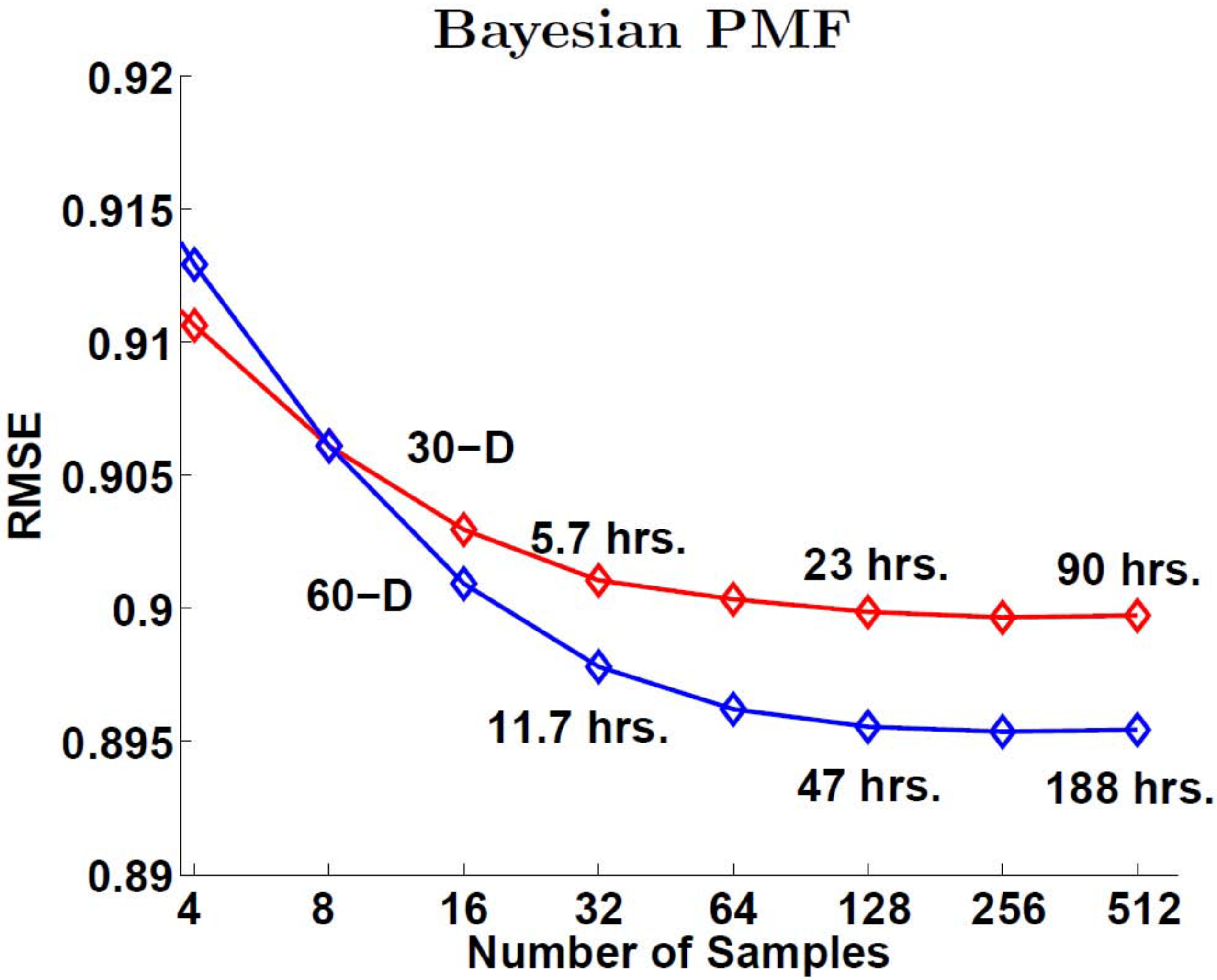




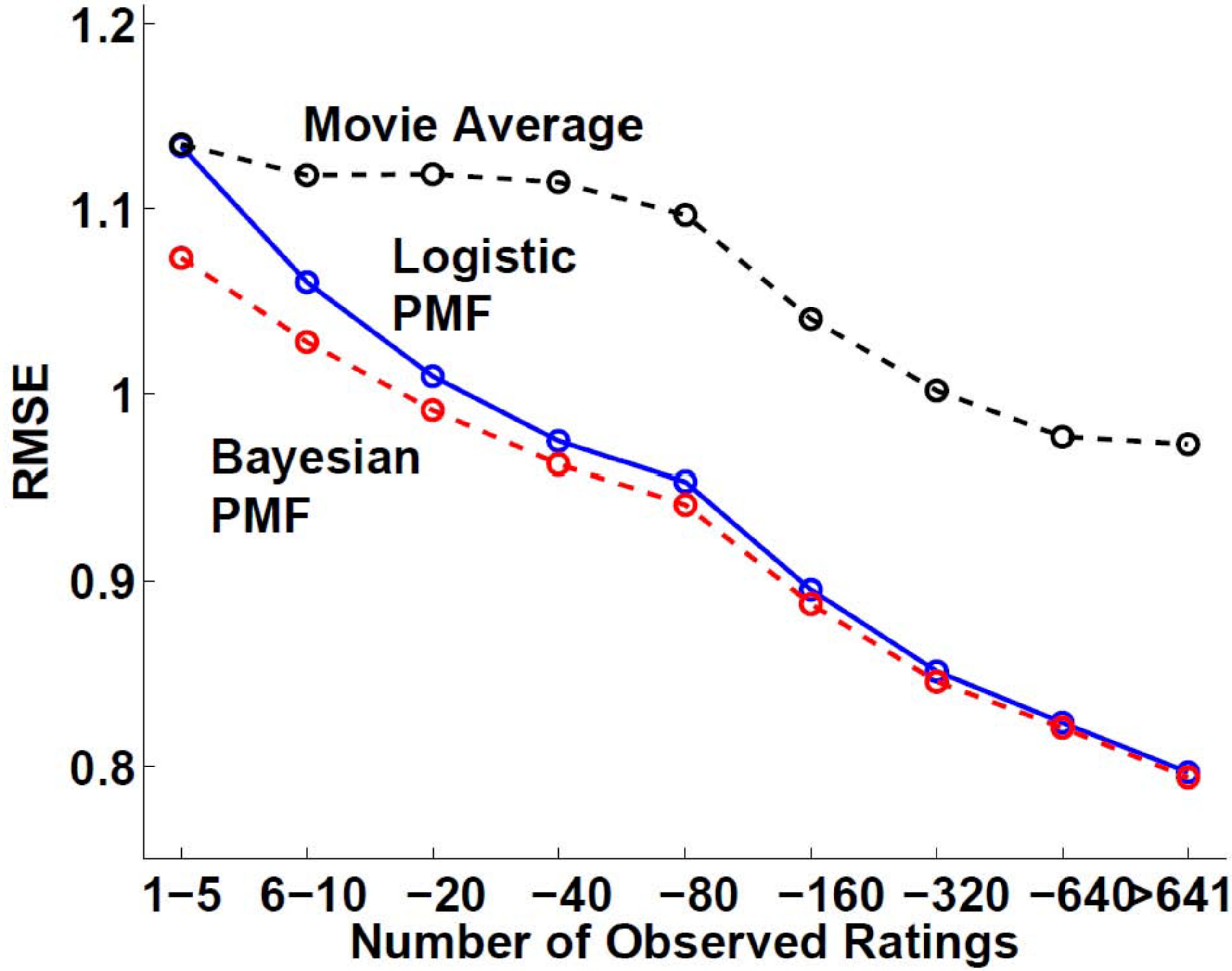
Results: Bayesian PMF on Netflix



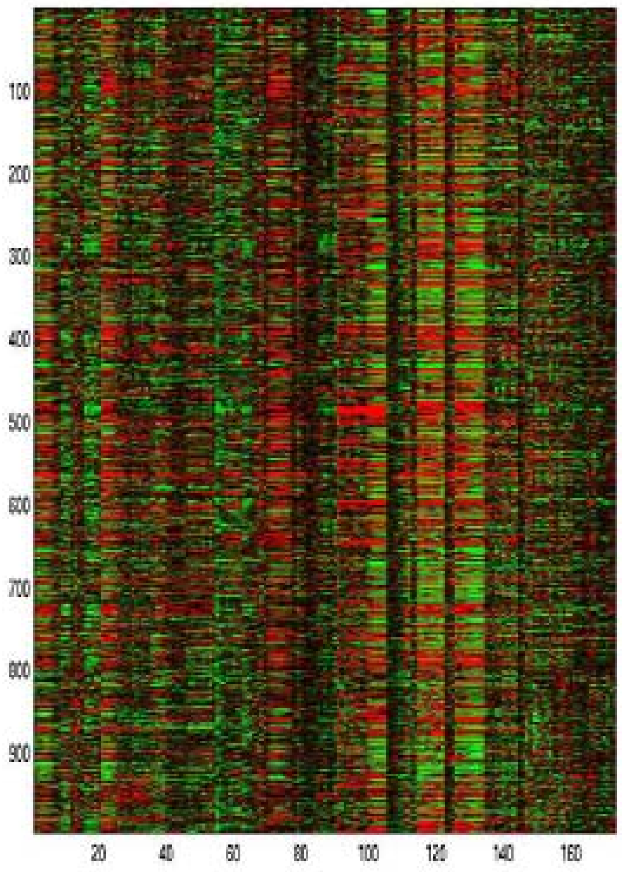
Results: Bayesian PMF on Netflix



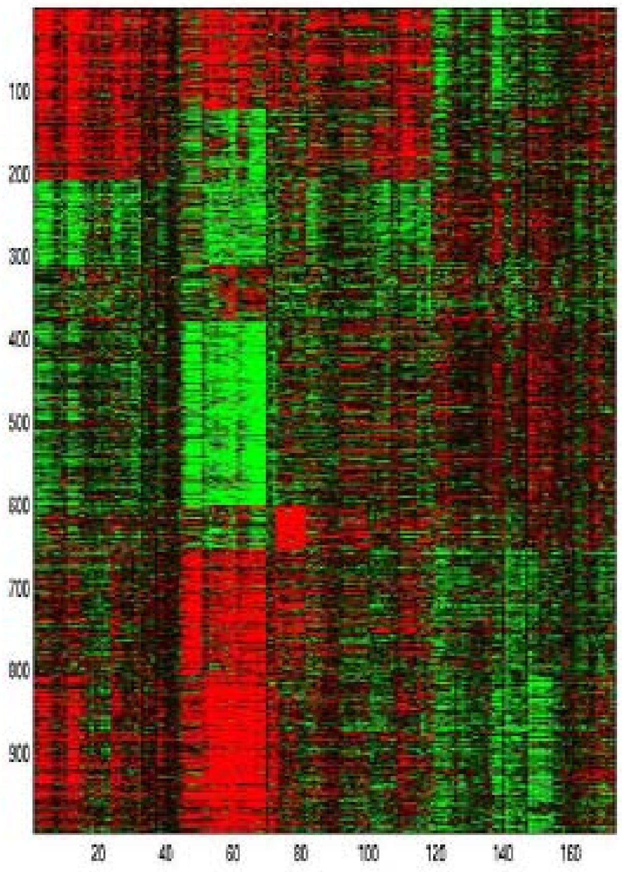
Results: Bayesian PMF on Netflix



Co-clustering: Gene Expression Analysis



Original



Co-clustered

Co-clustering and Matrix Approximation

U, V	1	2	3	4	5	6
1	-66	54	-63	93	51	96
2	35	87	37	-26	84	-22
3	-68	56	-64	92	52	94
4	30	83	32	-24	80	-21
5	-63	55	-60	92	53	95

Original Matrix Z

U, V	1	3	5	2	4	6
4	30	32	80	83	-24	-21
2	35	37	84	87	-26	-22
5	-63	-60	53	55	92	95
1	-66	-63	51	54	93	96
3	-68	-64	52	56	92	94

Reordered Matrix $\tilde{Z}$

$U, \hat{U}$	1	2
1	0	1
2	1	0
3	0	1
4	1	0
5	0	1

Row Clustering

$\times$

$\hat{U}, \hat{V}$	1	2	3
1	33.5	83.5	-23.3
2	-64.0	53.5	93.7

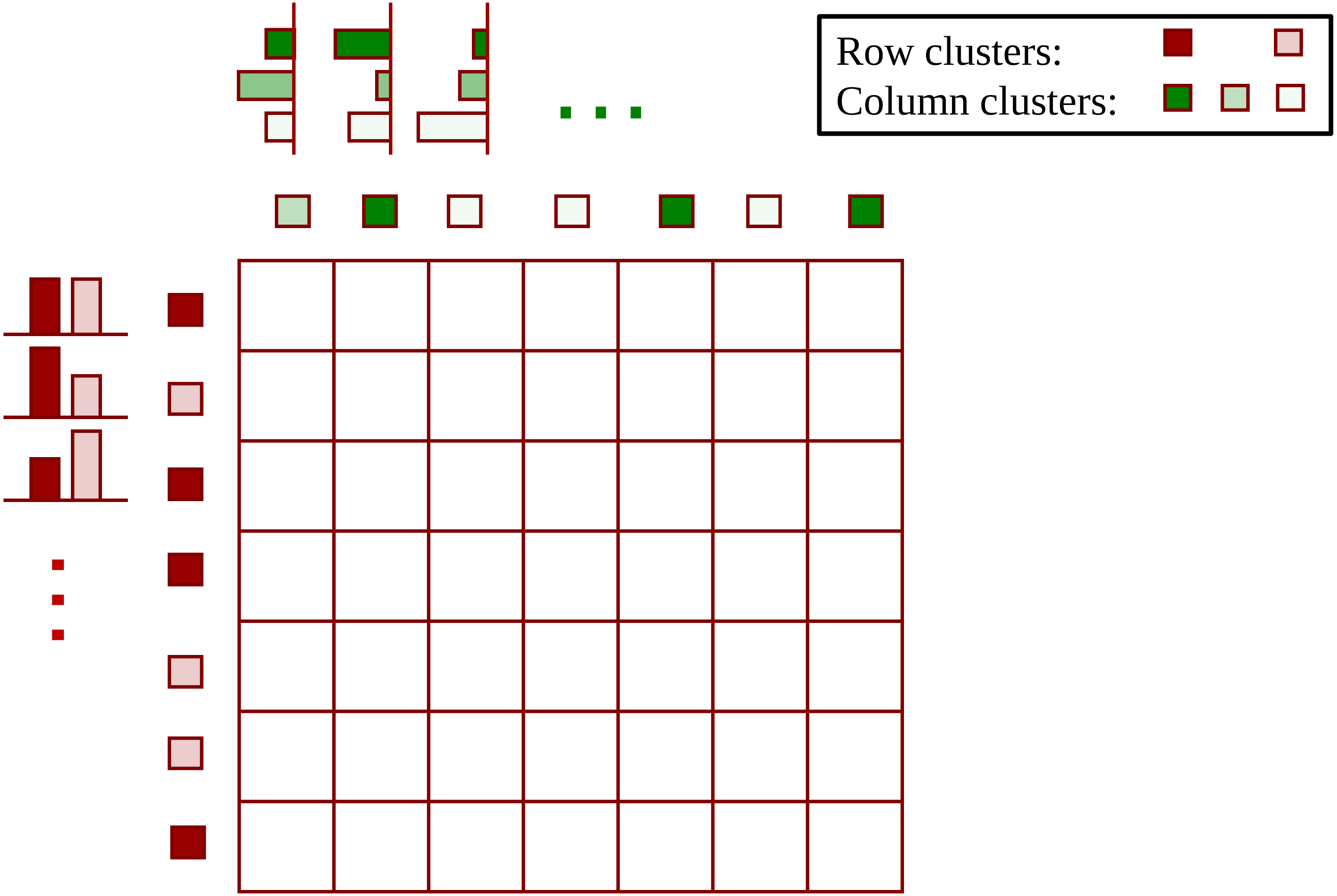
Low Parameter Matrix

$\times$

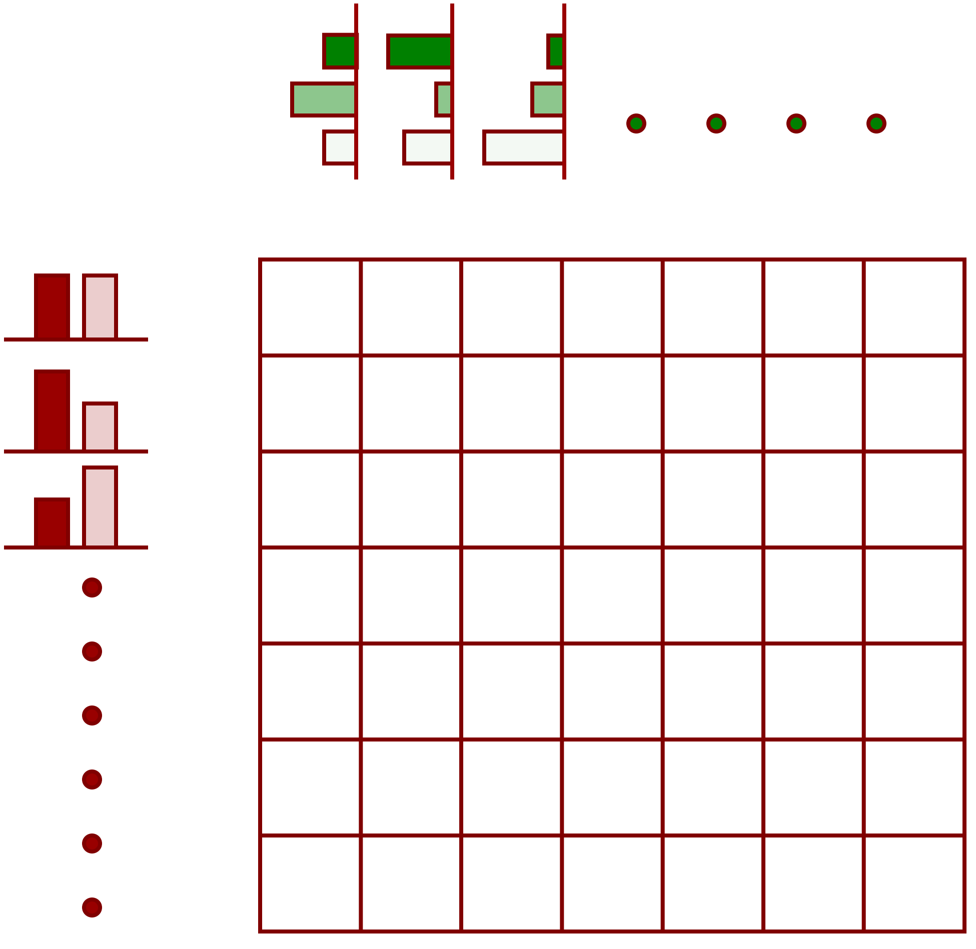
$\hat{V}, V$	1	2	3	4	5	6
1	1	0	1	0	0	0
2	0	1	0	0	1	0
3	0	0	0	1	0	1

Column Clustering

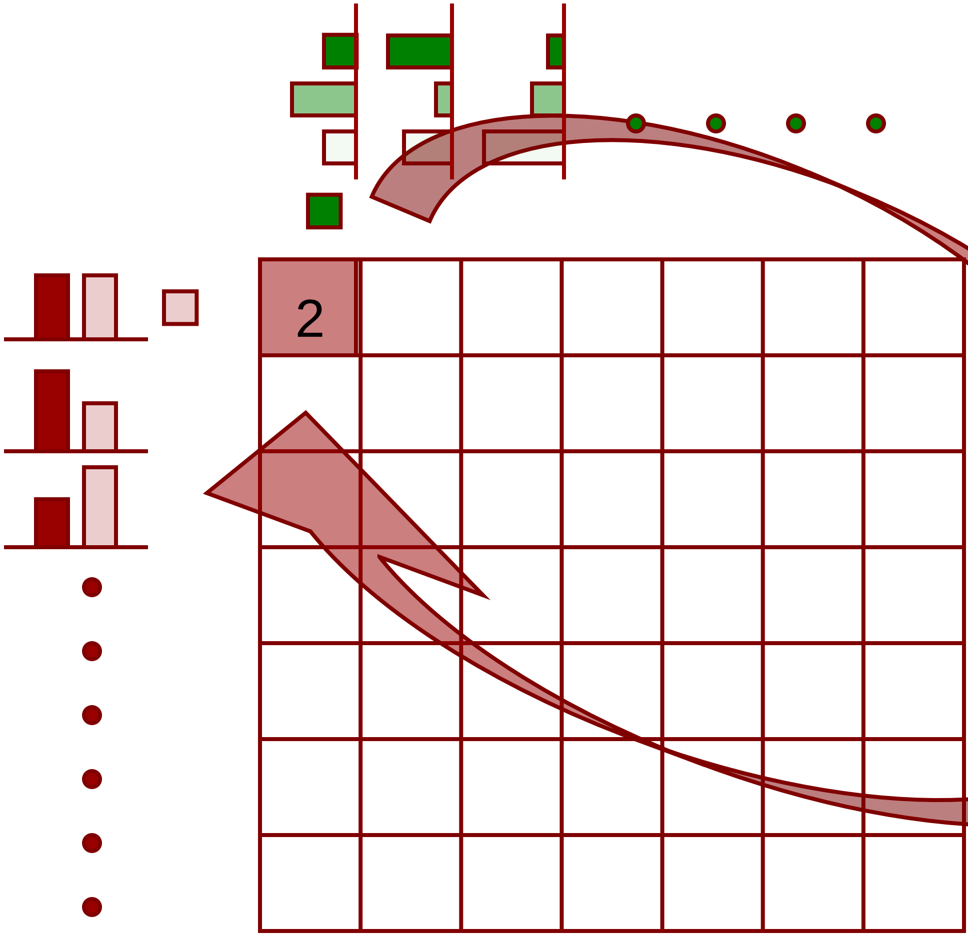
Probabilistic Co-clustering



Probabilistic Co-clustering

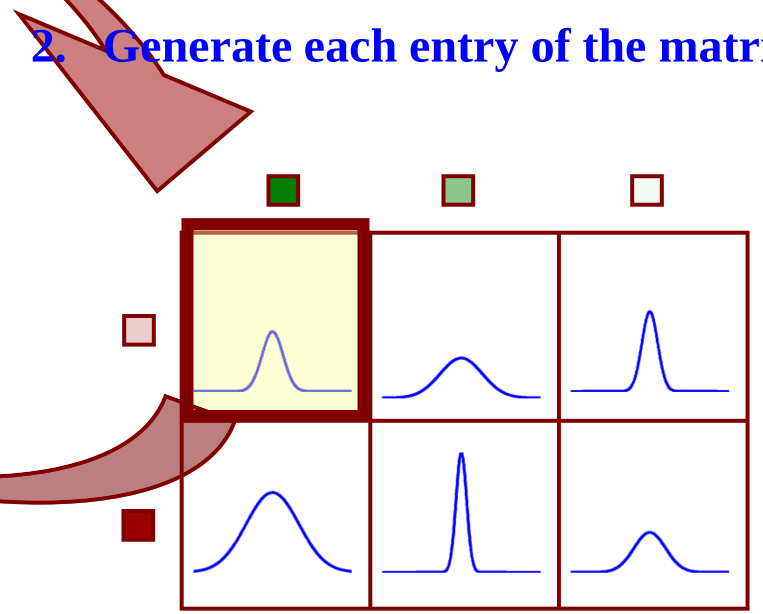


Generative Process

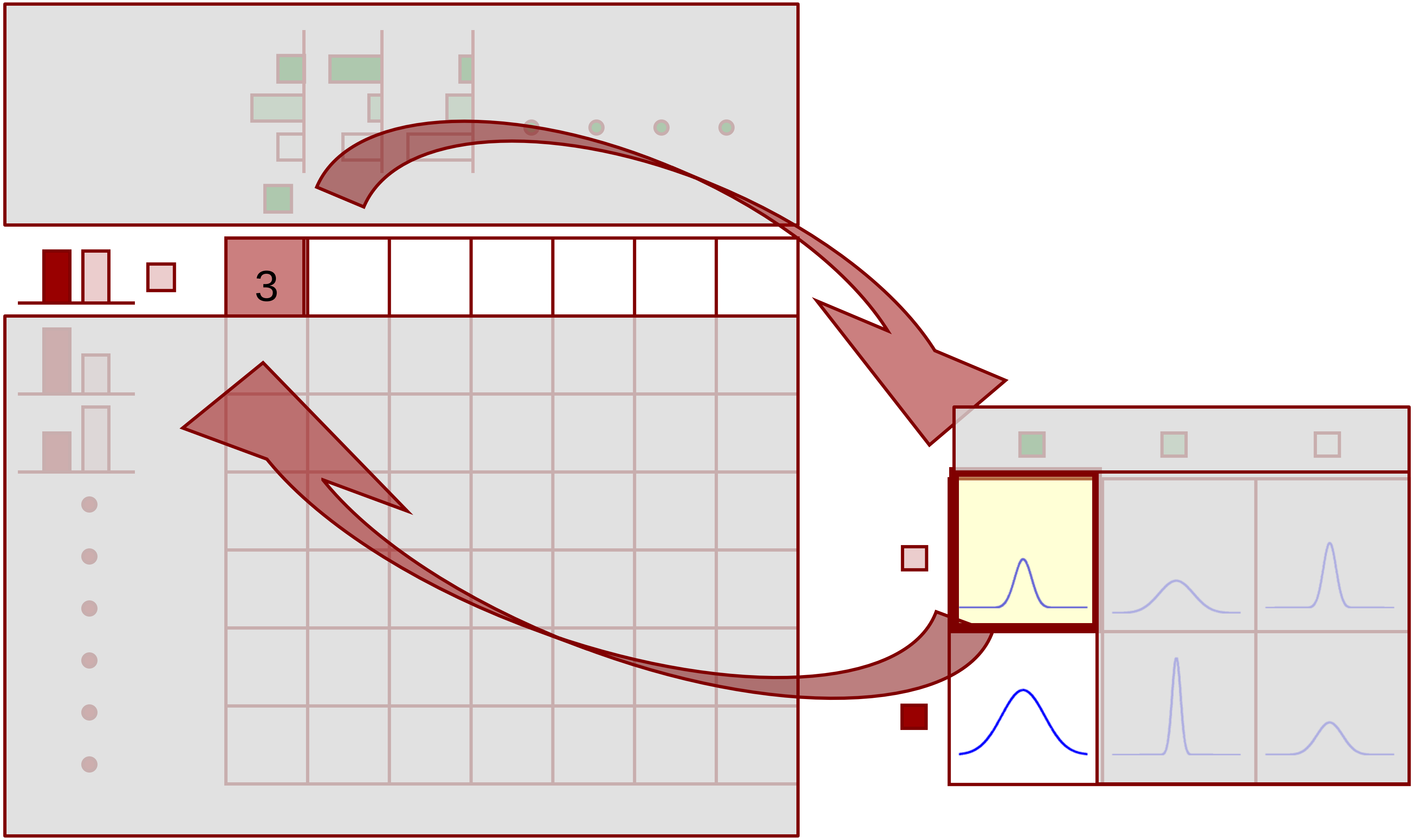


- **Assume a mixed membership for each row and column**
- **Assume a Gaussian for each co-cluster**

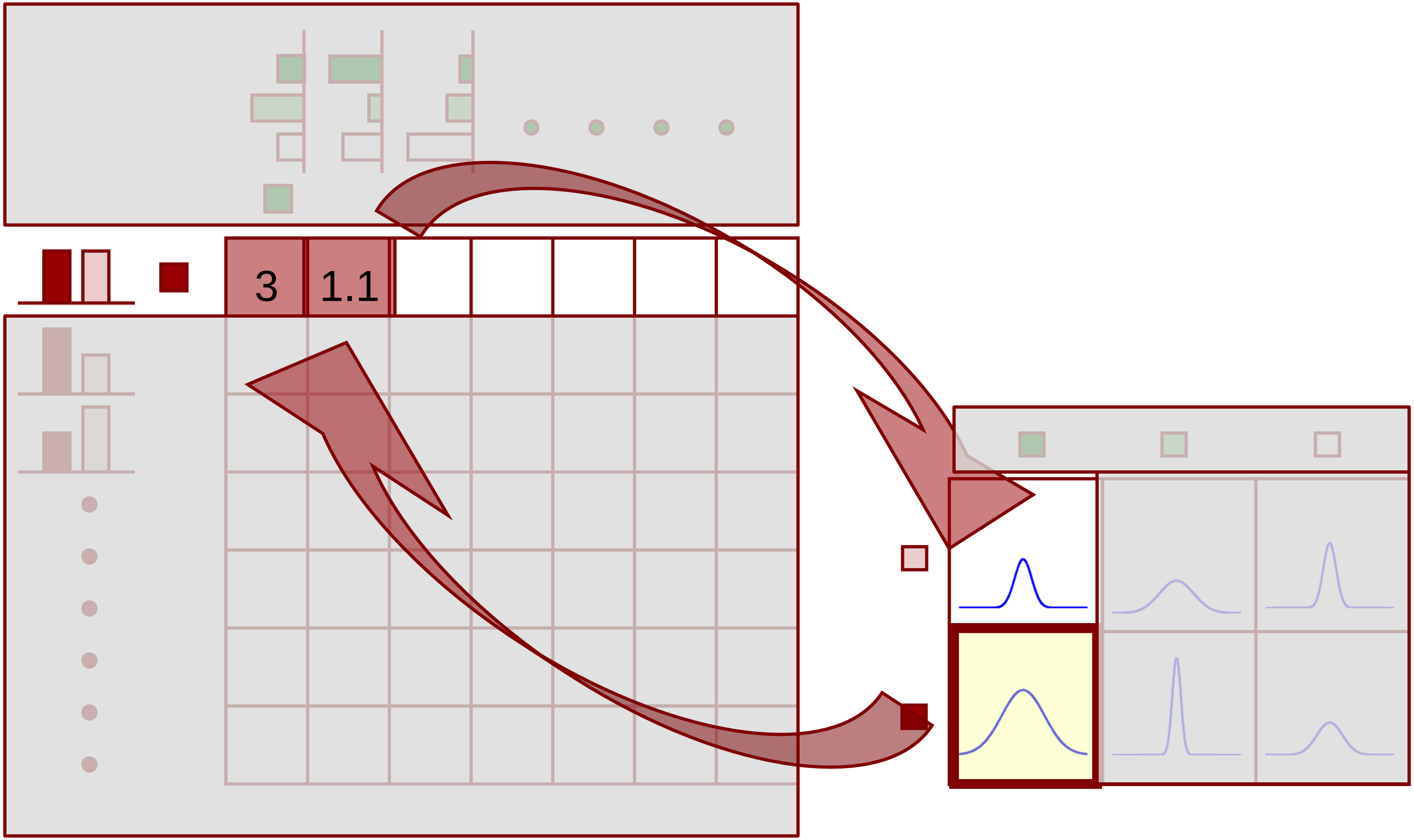
1. **Pick row/column clusters**
2. **Generate each entry of the matrix**



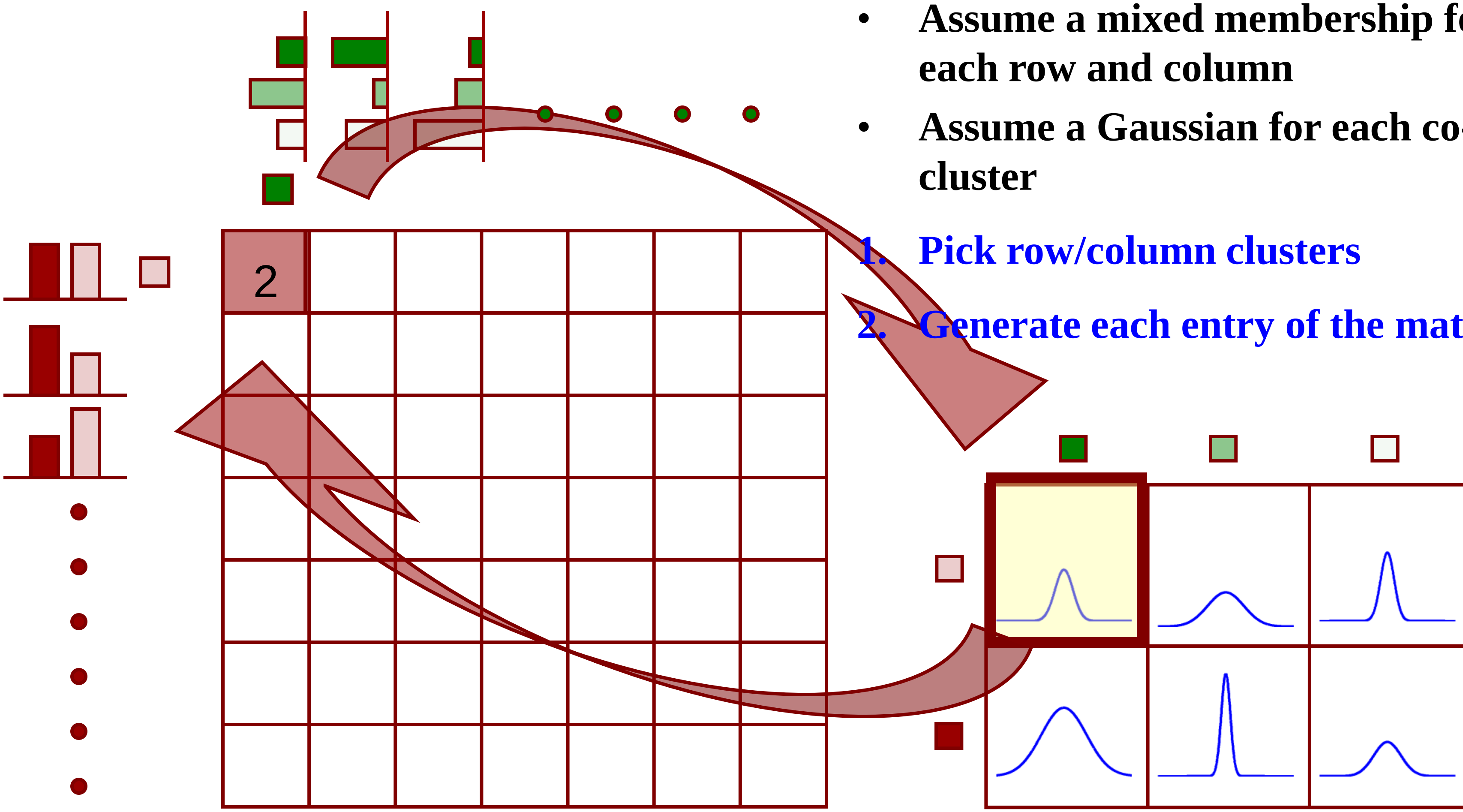
Reduction to Mixture Models



Reduction to Mixture Models



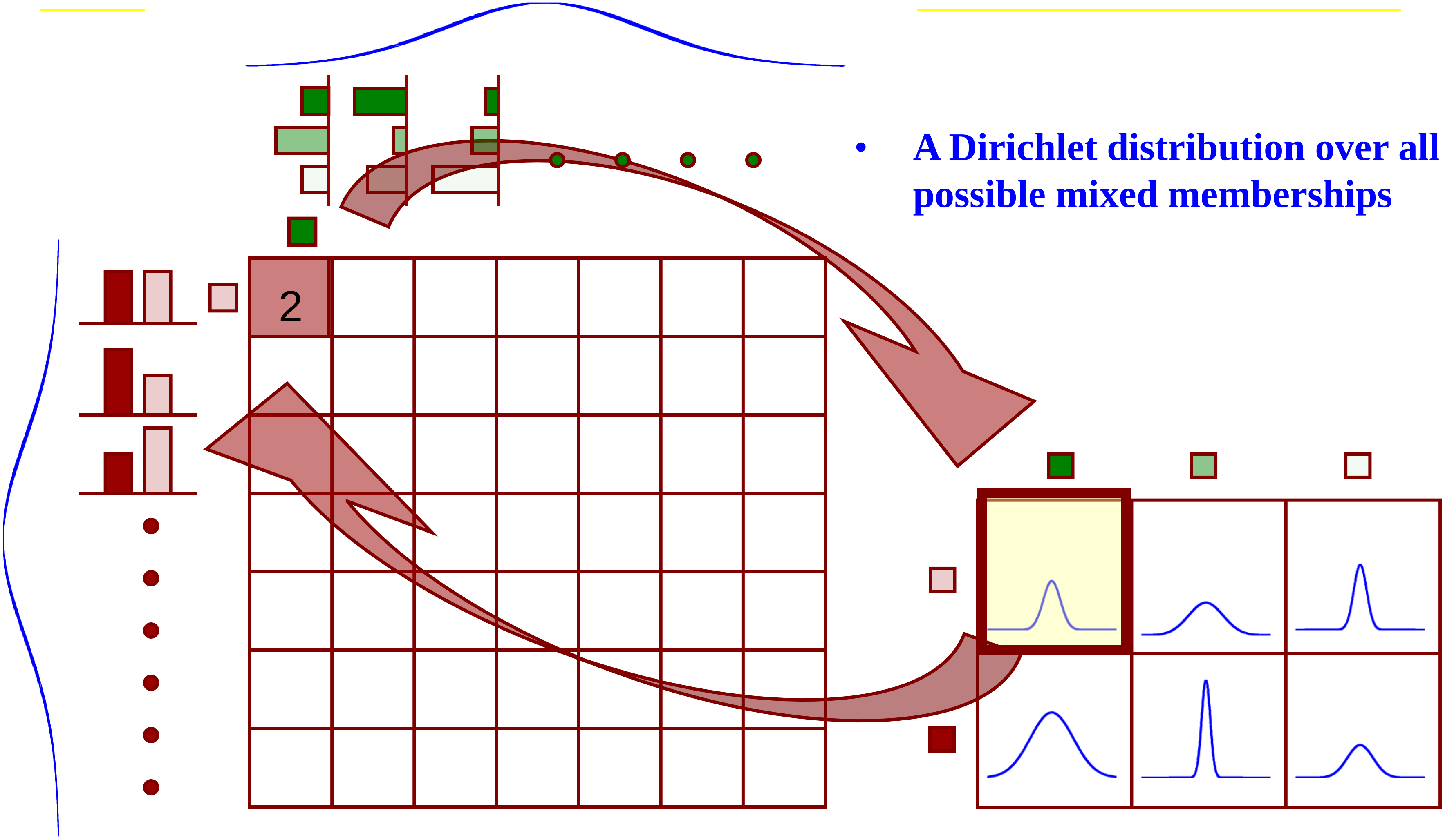
Generative Process



- **Assume a mixed membership for each row and column**
- **Assume a Gaussian for each co-cluster**

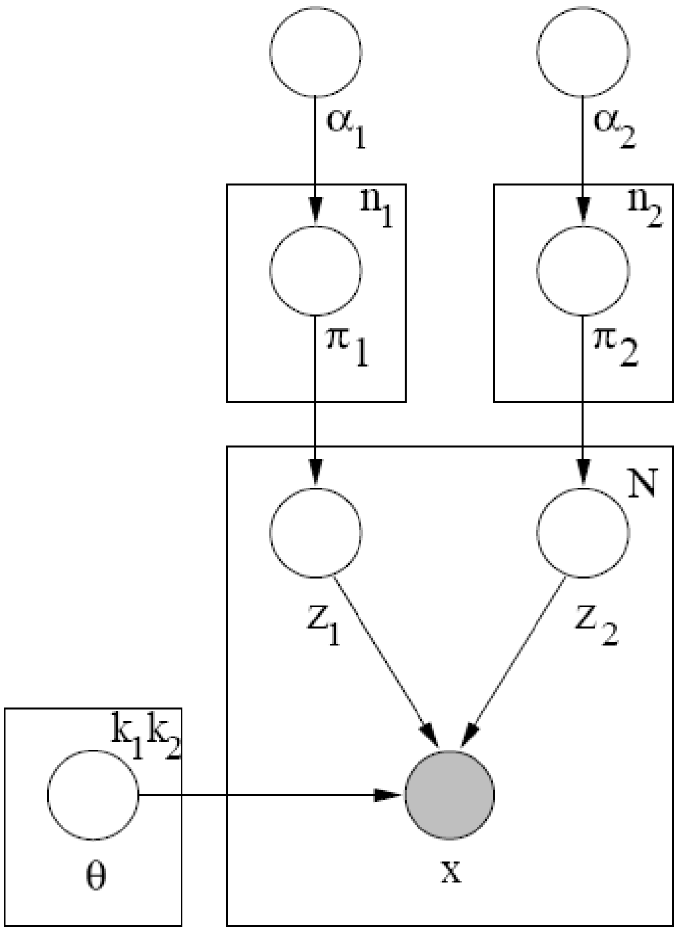
1. **Pick row/column clusters**
2. **Generate each entry of the matrix**

Bayesian Co-clustering (BCC)



- **A Dirichlet distribution over all possible mixed memberships**

Bayesian Co-clustering (BCC)



1. For each row $u, [u]_1^{n_1}$, choose $\pi_{1u} \sim \text{Dir}(\alpha_1)$.
2. For each column $v, [v]_1^{n2}$, choose $\pi_{2v} \sim \text{Dir}(\alpha_2)$.
3. For each non-missing entry in row u and column v:
   - (a) Choose $z_1 \sim \text{Discrete}(\pi_{1u})$.
   - (b) Choose $z_2 \sim \text{Discrete}(\pi_{2v})$.
   - (c) Choose $x_{uv} \sim p_(x|\theta_{z_1 z_2})$.

$$\boxed{\log p(X|\alpha_1, \alpha_2, \Theta) \neq \sum_{n=1}^{N} \log p(x_n|\alpha_1, \alpha_2, \Theta)}$$

Learning: Inference and Estimation

- Learning
  - Estimate model parameters $(\alpha_1, \alpha_2, \theta)$
  - Infer 'mixed memberships' of individual rows and columns
- Expectation Maximization
  - E-step: Calculate posterior probability $p(\pi_1, \pi_2, \mathbf{z}_1, \mathbf{z}_2 | \alpha_1, \alpha_2, \Theta, X)$ to obtain log-likelihood $L(\alpha, \Theta)$.
  - M-step: Maximize $L(\alpha, \Theta)$ w.r.t α, Θ.
- Issues
  - Posterior probability cannot be obtained in closed form
  - Parameter estimation cannot be done directly
- Approach: Approximate inference
  - Variational Inference
  - Collapsed Gibbs Sampling, Collapsed Variational Inference

Variational EM

- Introduce a variational distribution $q(\pi_1, \pi_2, \mathbf{z}_1, \mathbf{z}_2 | \gamma_1, \gamma_2, \phi_1, \phi_2)$ to approximate $p(\pi_1, \pi_2, \mathbf{z}_1, \mathbf{z}_2 | \alpha_1, \alpha_2, \Theta, X)$
- Use Jensen's inequality to get a tractable lower bound

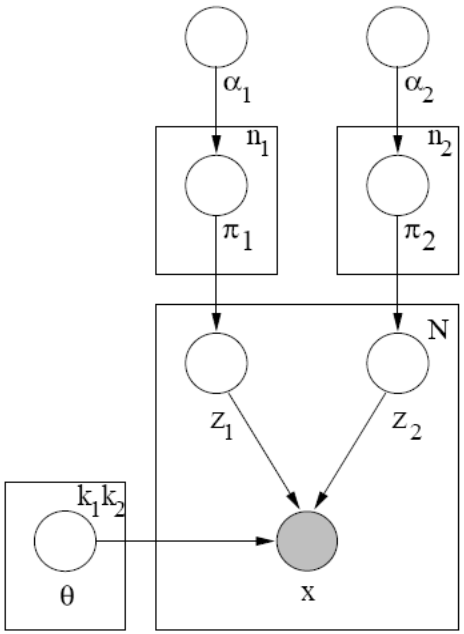
$$\log p(X | \alpha_1, \alpha_2, \Theta) \geq E_q[\log p(X, \mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2 | \alpha_1, \alpha_2, \Theta)] + H(q(\mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2))$$

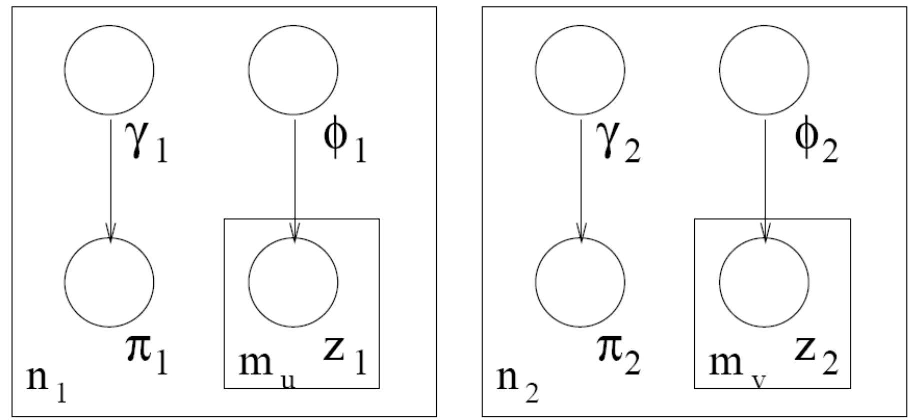
- Maximize the lower bound w.r.t $(\phi_1, \gamma_1, \phi_2, \gamma_2)$
  - Alternatively minimize the KL divergence between $q(\pi_1, \pi_2, \mathbf{z}_1, \mathbf{z}_2 | \gamma_1, \gamma_2, \phi_1, \phi_2)$ and $p(\pi_1, \pi_2, \mathbf{z}_1, \mathbf{z}_2 | \alpha_1, \alpha_2, \Theta, X)$

- Maximize the lower bound w.r.t. $(\alpha_1, \alpha_2, \Theta)$

Variational Distribution

- $\mathrm{Dir}(\gamma_1), \mathrm{Disc}(\phi_1)$ for each row, $\mathrm{Dir}(\gamma_2), \mathrm{Disc}(\phi_2)$ for each column





(a) row (b) column

$$q(\pi_1, \pi_2, \mathbf{z}_1, \mathbf{z}_2 | \gamma_1, \gamma_2, \phi_1, \phi_2) = \left(\prod_{u=1}^{n_1} q(\pi_{1u} | \gamma_{1u}) \right) \times \left(\prod_{v=1}^{n_2} q(\pi_{2v} | \gamma_{2v}) \right)$$

$$\times \left(\prod_{u=1}^{n_1} \prod_{v=1}^{n_2} q(z_{1uv} | \phi_{1u}) q(z_{2uv} | \phi_{2v}) \right)$$

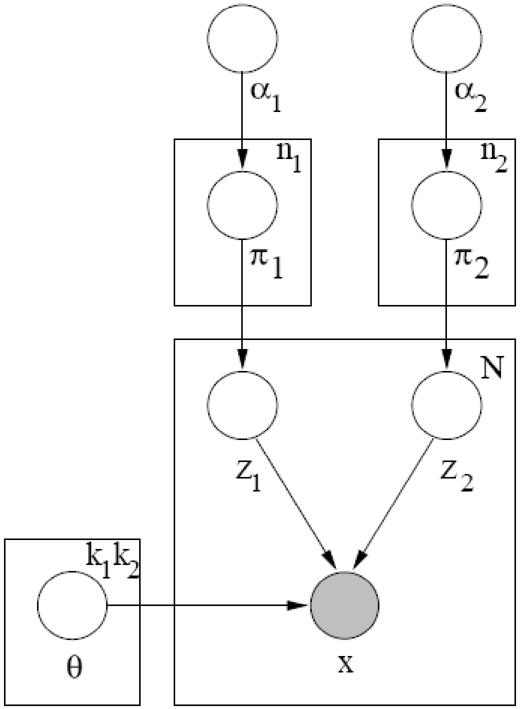
Collapsed Inference

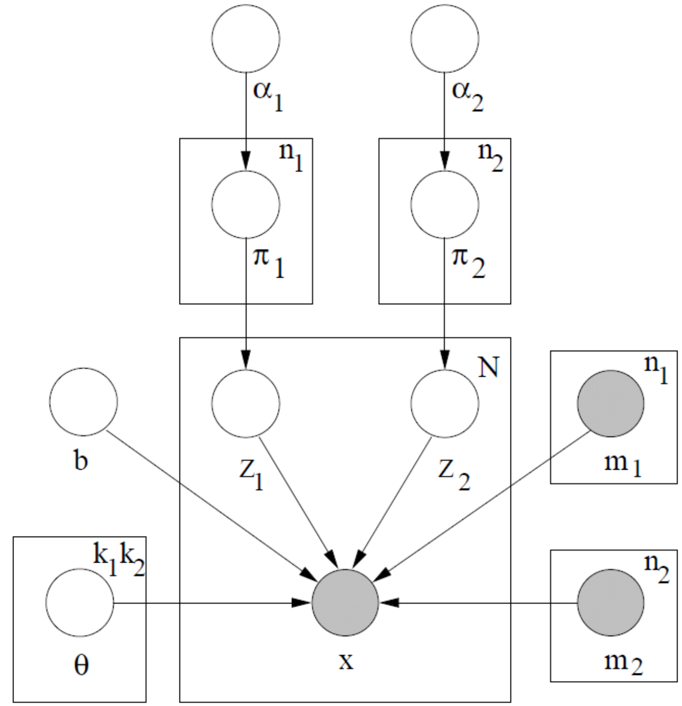
- Latent distribution can be exactly marginalized over (π_1, π_2)
  - Obtain $p(X, z_1, z_2 | \alpha_1, \alpha_2, \beta)$ in closed form
  - Analysis assumes discrete/categorical entries
  - Can be generalized to exponential family distributions
- Collapsed Gibbs Sampling
  - Conditional distribution of (z1uv,z2uv) in closed form

$$P(z_1^{uv}=i, z_2^{uv}=j \mid X, z_1^{-uv}, z_{2\text{-}uv}, \alpha_1, \alpha_2, \beta)$$

  - Sample states, run sampler till convergence
- Collapsed Variational Bayes
  - Variational distribution $q(z_1, z_2 | \gamma) = \prod_{u,v} q(z_1^{uv}, z_2^{uv} | \gamma^{uv})$
  - Gaussian and Taylor approximation to obtain updates for γ^{uv}

Residual Bayesian Co-clustering (RBC)





$$x_{uv} \sim N(x|\mu_{z_1 z_2}, \sigma^2_{z_1 z_2})$$

$$x_{uv} \sim N(x|\mu_{z_1 z_2} + bm_{1u} + bm_{2v}, \sigma^2_{z_1 z_2})$$

- *(z1,z2)* determines the distribution
- Users/movies may have bias

- *(m1,m2)*: row/column means
- *(bm1,bm2)*: row/ column bias

Results: Datasets

- Movielens: Movie recommendation data
  - 100,000 ratings (1-5) for 1682 movies from 943 users (6.3%)
  - Binarize: 0 (1-3), 1(4-5).
  - Discrete (original), Bernoulli (binary), Real (z-scored)
- Foodmart: Transaction data
  - 164,558 sales records for 7803 customers and 1559 products (1.35%)
  - Binarize: 0 (less than median), 1(higher than median)
  - Poisson (original), Bernoulli (binary), Real (z-scored)
- Jester: Joke rating data
  - 100,000 ratings (-10.00,+10.00) for 100 jokes from 1000 users (100%)
  - Binarize: 0 (lower than 0), 1 (higher than 0)
  - Gaussian (original), Bernoulli (binary), Real (z-scored)

Perplexity Comparison with 10 Clusters

Training Set

	MMNB	BCC	LDA
Jester	**1.7883**	1.8186	98.3742
Movielens	**1.6994**	1.9831	439.6361
Foodmart	**1.8691**	1.9545	1461.7463

Test Set

	MMNB	BCC	LDA
Jester	4.0237	**2.5498**	98.9964
Movielens	3.9320	**2.8620**	1557.0032
Foodmart	6.4751	**2.1143**	6542.9920

On Binary Data

Training Set

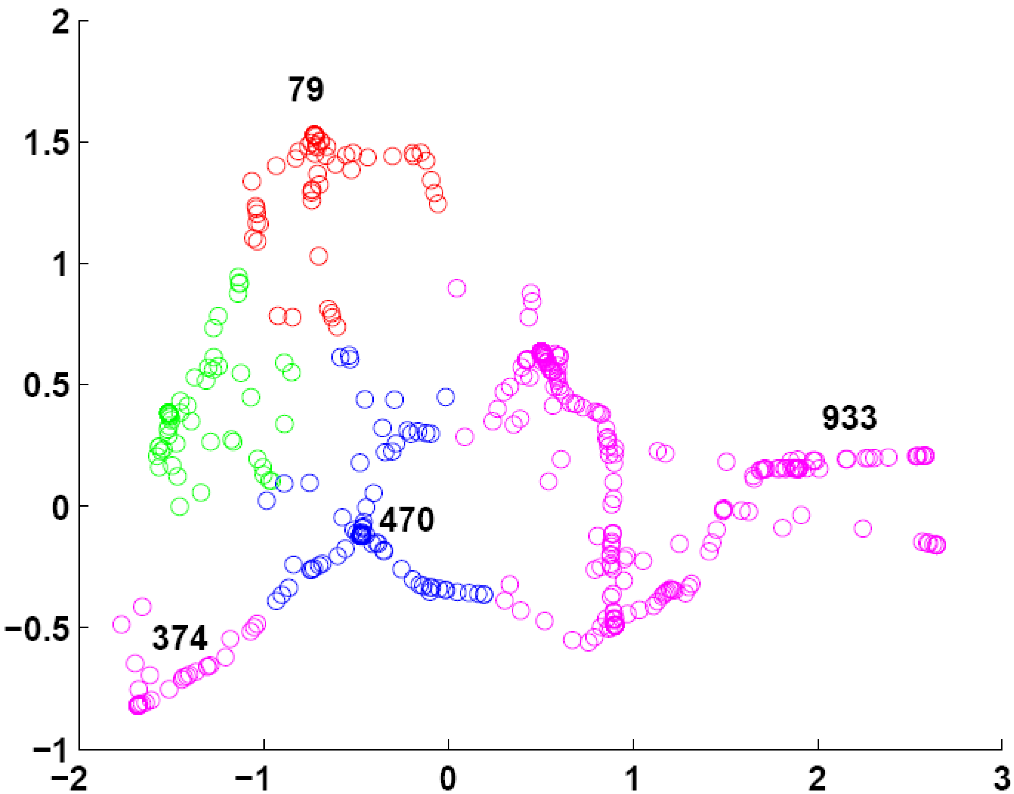
	MMNB	BCC
Jester	**15.4620**	18.2495
Movielens	3.1495	**0.8068**
Foodmart	**4.5901**	4.5938

Test Set

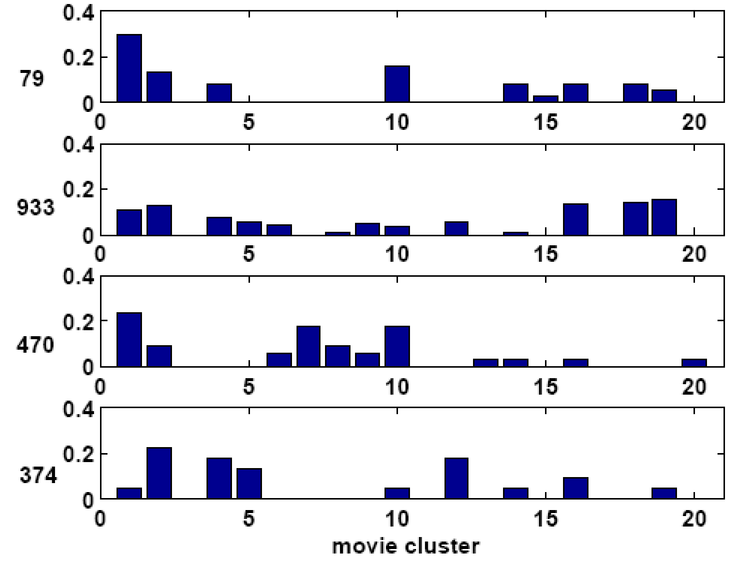
	MMNB	BCC
Jester	39.9395	**24.8239**
Movielens	38.2377	**1.0265**
Foodmart	4.6681	**4.5964**

On Original Data

Co-embedding: Users



Users

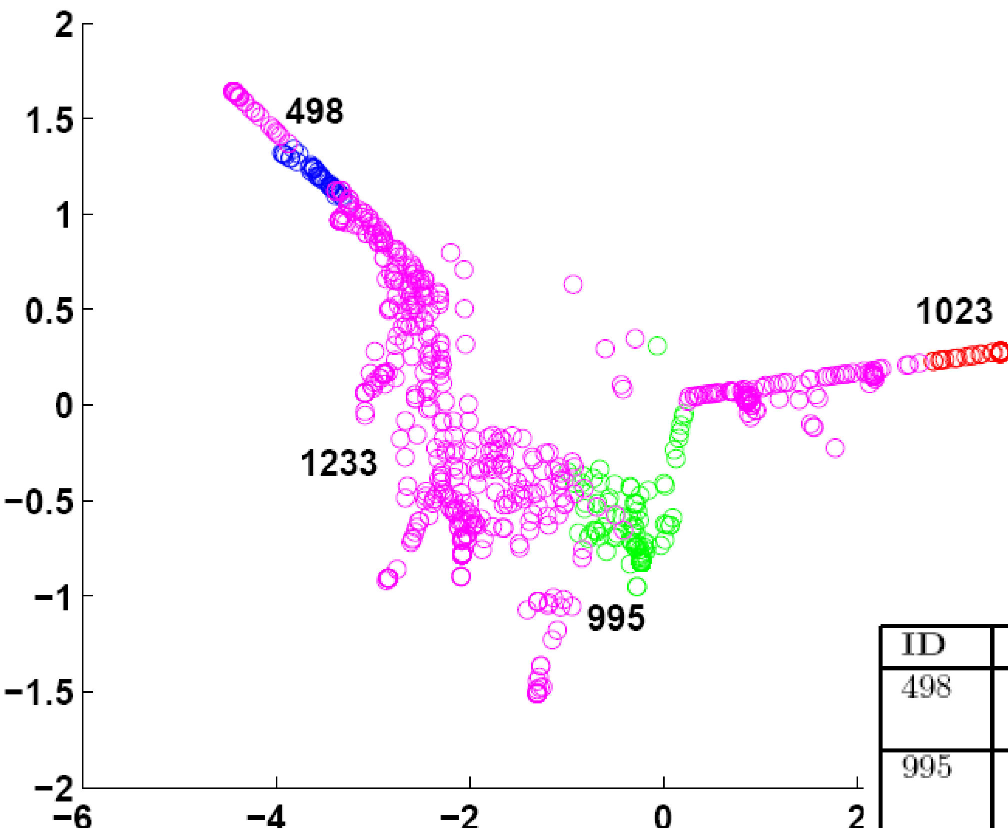


User signatures

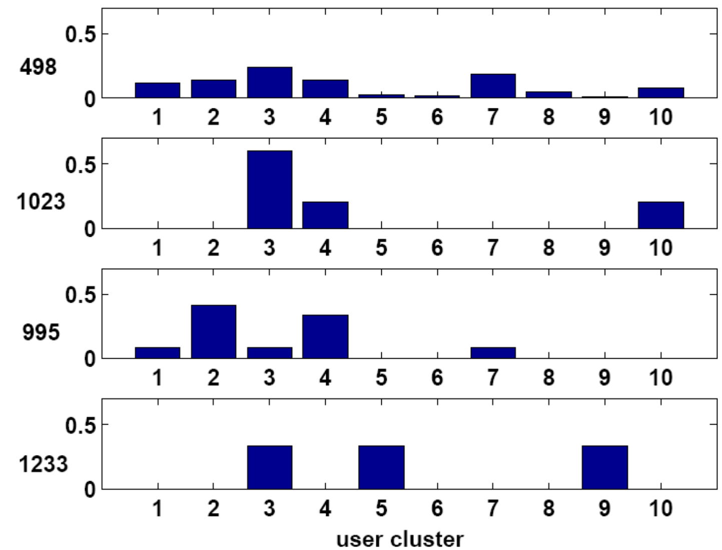
ID	Age	Sex	Occupation
79	39	F	administrator
374	36	M	executive
470	24	M	programmer
933	28	M	student

User profiles.

Co-embedding: Movies



Movies

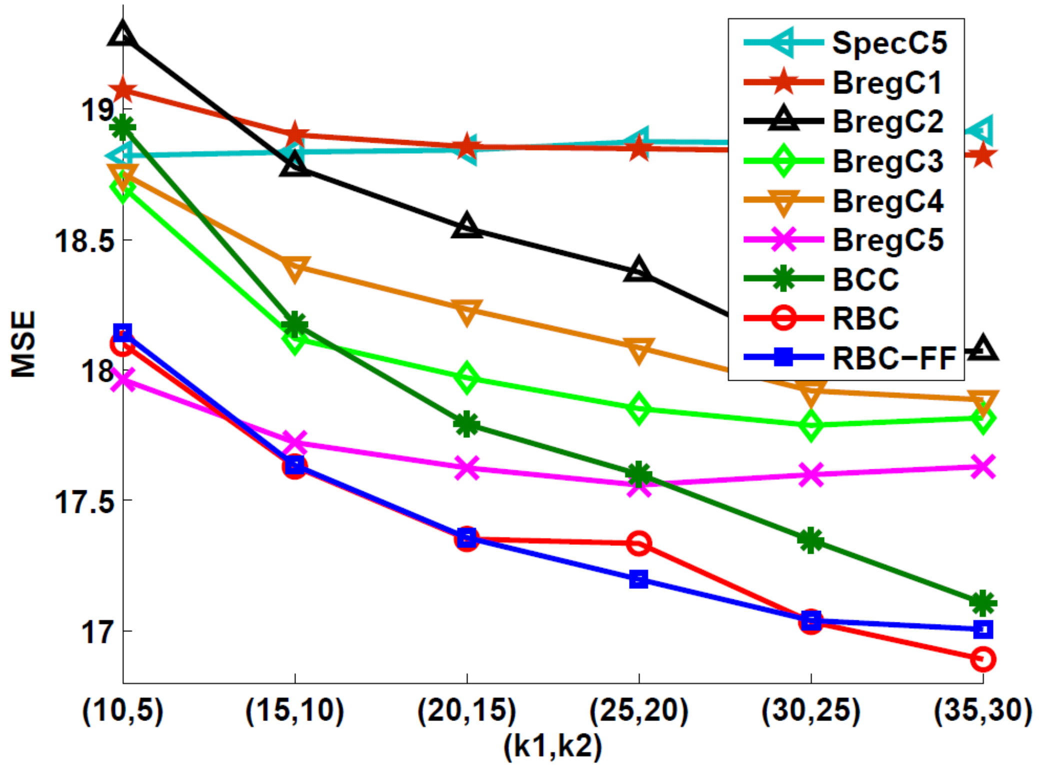


Movie signatures

ID	Movie	Keywords
498	The African Queen	American Expatriate, Boat, Mission, African Tribe
995	Kiss Me, Guido	Italian Food, Homosexual, Pizza, Gay Interest
1023	Fathers' Day	Seduction, Con, Box Office Flop, Friendship
1233	Nënette et Boni	Brother Sister Relationship, Teen, Pregnancy, Teenage Pregnancy

Movie names and keywords.

RBC vs. other co-clustering algorithms



Jester

- RBC and RBC-FF perform better than BCC
- RBC and RBC-FF are also the best among others

RBC vs. other co-clustering algorithms

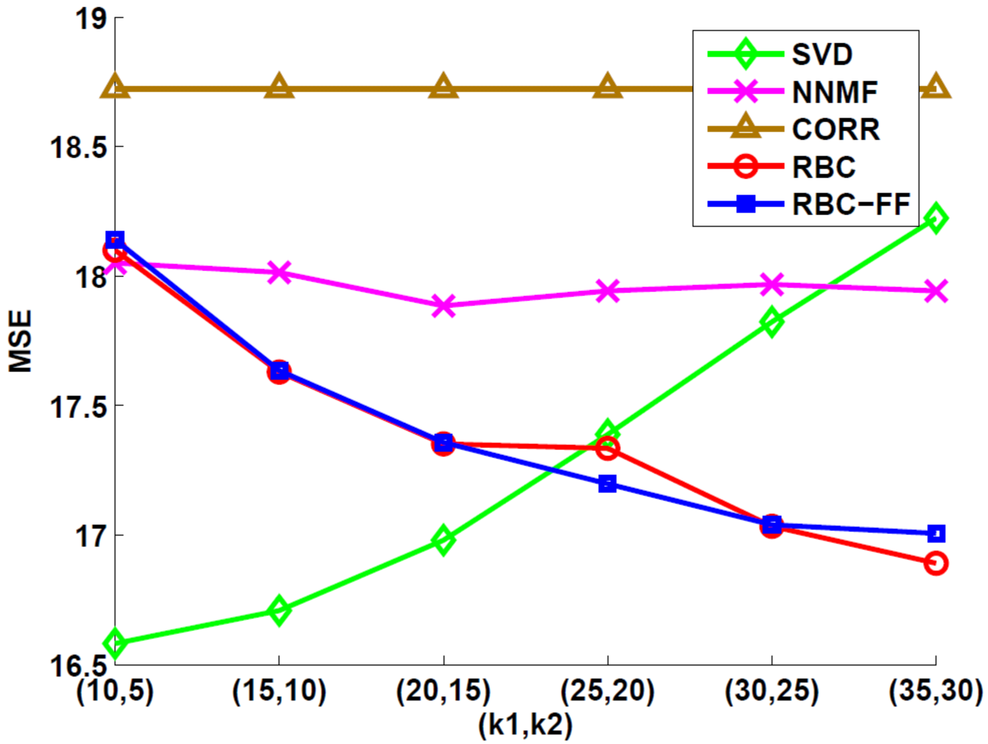
$k1, k2$	SpecC2	SpecC5	BregC1	BregC2	BregC3	BregC4	BregC5	BregC6	BCC	RBC	RBC-FF
5,10	0.1175 ±0.0019	0.0979 ±0.0013	0.0956 ±0.0015	0.1073 ±0.0026	0.0949 ±0.0015	0.1201 ±0.0033	0.1073 ±0.0026	0.1715 ±0.0080	0.0957 ±0.0012	**0.0943 ±0.0012**	**0.0943 ±0.0010**
10,15	0.1141 ±0.0016	0.0963 ±0.0013	0.0948 ±0.0013	0.0959 ±0.0013	0.0942 ±0.0012	0.1173 ±0.0040	0.1090 ±0.0037	0.2603 ±0.0084	0.0953 ±0.0011	**0.0935 ±0.0010**	**0.0935 ±0.0011**
15,20	0.1136 ±0.0014	0.0960 ±0.0009	0.0944 ±0.0010	0.1100 ±0.0040	0.0954 ±0.0012	0.1178 ±0.0048	0.1100 ±0.0040	0.3399 ±0.1112	0.0952 ±0.0013	**0.0931 ±0.0013**	**0.0931 ±0.0013**

Movielens

$k1, k2$	SpecC2	SpecC5	BregC1	BregC2	BregC3	BregC4	BregC5	BregC6	BCC	RBC	RBC-FF
10,5	0.9758 ±0.0221	0.9159 ±0.0199	0.9123 ±0.0194	0.9765 ±0.0212	0.9819 ±0.0217	0.9855 ±0.0221	0.9415 ±0.0154	1.4148 ±0.0168	0.9591 ±0.0212	**0.9119 ±0.0196**	0.9136 ±0.0197
15,10	0.9767 ±0.0214	0.9170 ±0.0191	0.9126 ±0.0200	1.0178 ±0.0236	1.0206 ±0.0237	1.0269 ±0.0239	0.9875 ±0.0229	2.0442 ±0.0262	0.9582 ±0.0217	**0.9111 ±0.0202**	0.9113 ±0.0204
20,15	0.9785 ±0.0209	0.9187 ±0.0192	0.9129 ±0.0196	1.0561 ±0.0227	1.0648 ±0.0222	1.0670 ±0.0164	1.0304 ±0.0217	2.9876 ±0.0505	0.9580 ±0.0215	**0.9106 ±0.0198**	0.9112 ±0.0217

Foodmart

RBC vs. SVD, NNMF, and CORR



Jester

- RBC and RBC-FF are competitive with other algorithms

RBC vs. SVD, NNMF and CORR

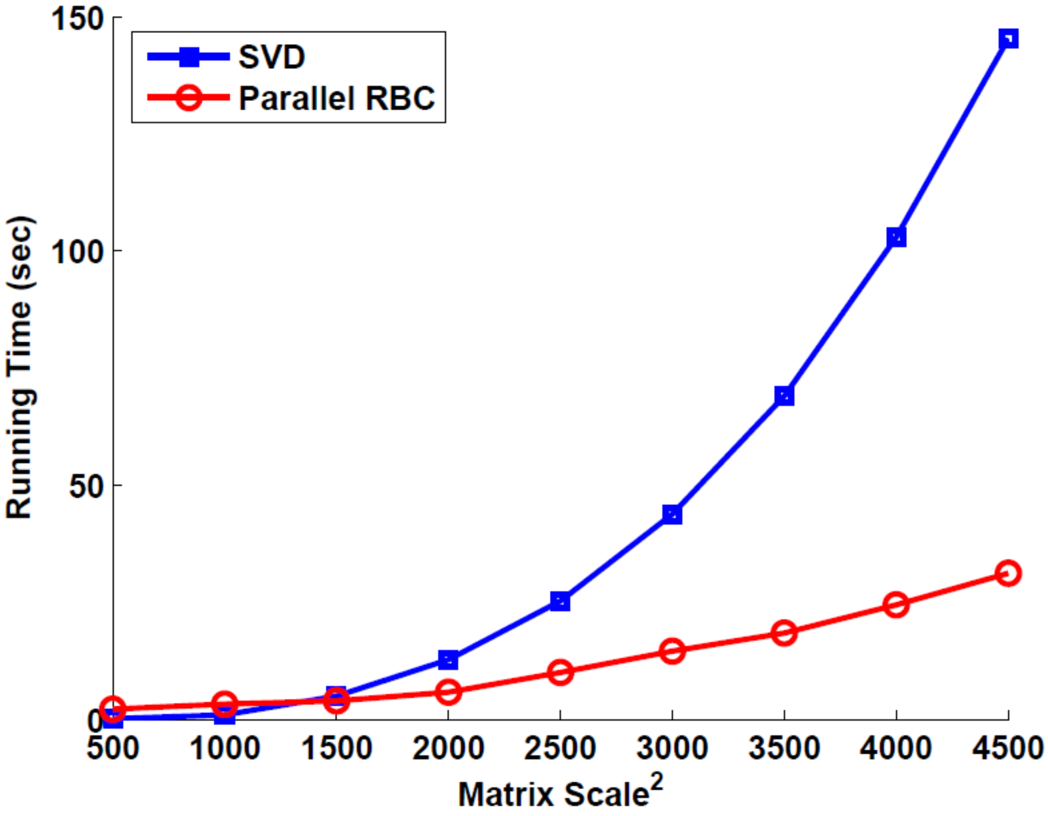
$k1, k2$	SVD	NNMF	CORR	RBC	RBC-FF
5,10	0.0986 $\pm$0.0012	0.1086 $\pm$0.0012	0.4118 $\pm$0.0061	**0.0943 $\pm$0.0012**	**0.0943 $\pm$0.0010**
10,15	0.0988 $\pm$0.0011	0.1078 $\pm$0.0013	0.4118 $\pm$0.0061	**0.0935 $\pm$0.0010**	**0.0935 $\pm$0.0011**
15,20	0.0991 $\pm$0.0011	0.1080 $\pm$0.0012	0.4118 $\pm$0.0061	**0.0931 $\pm$0.0013**	**0.0931 $\pm$0.0013**

Movielens

$k1, k2$	SVD	NNMF	CORR	RBC	RBC-FF
10,5	**0.8998 $\pm$0.0210**	0.9197 $\pm$0.0212	1.4528 $\pm$0.0281	0.9119 $\pm$0.0196	0.9136 $\pm$0.0197
15,10	**0.8995 $\pm$0.0208**	0.9216 $\pm$0.0207	1.4528 $\pm$0.0281	0.9111 $\pm$0.0202	0.9113 $\pm$0.0204
20,15	**0.9021 $\pm$0.0211**	0.9202 $\pm$0.0208	1.4528 $\pm$0.0281	0.9106 $\pm$0.0198	0.9112 $\pm$0.0217

Foodmart

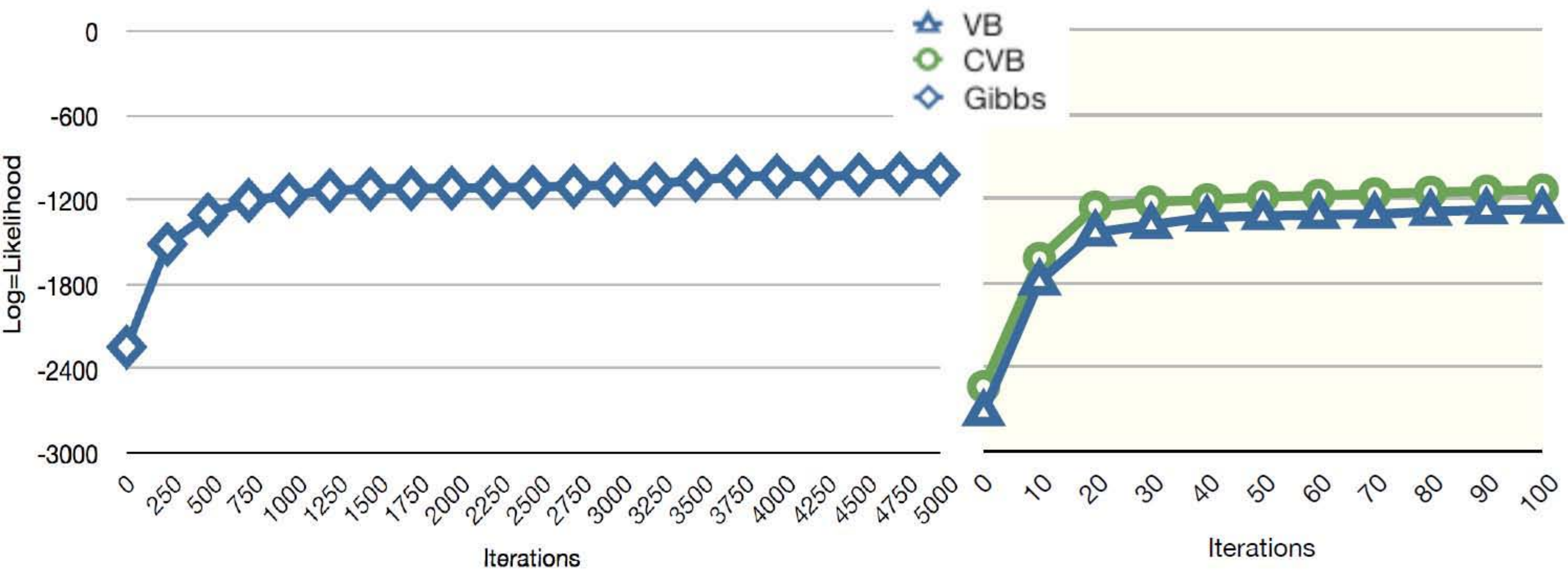
SVD vs. Parallel RBC



Parallel RBC scales well to large matrices

Inference Methods: VB, CVB, Gibbs

	Gibbs	CVB	VB
MovieLens	3.247	4.553	5.849
Binarized Jester	2.954	3.216	4.023

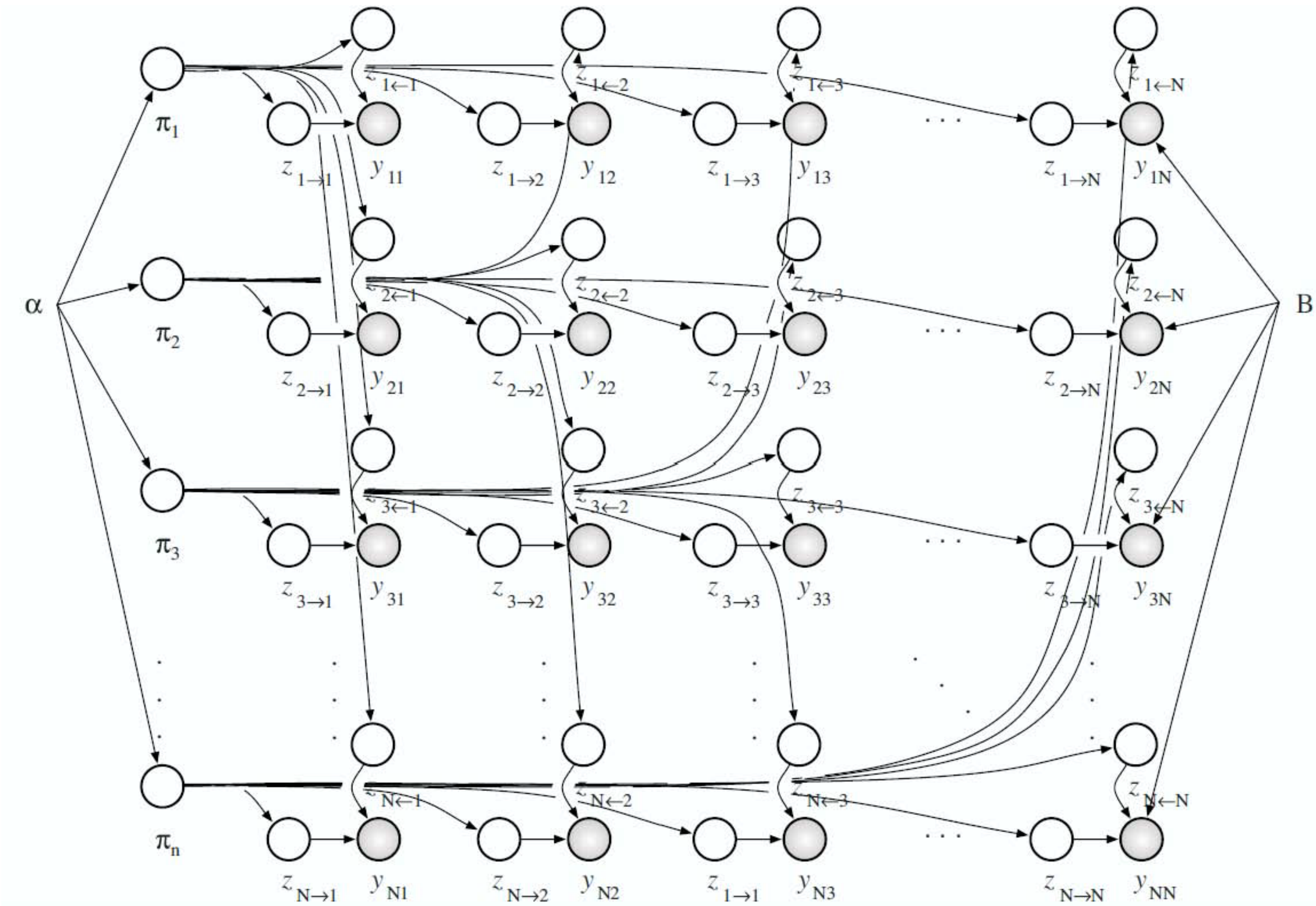


Mixed Membership Stochastic Block Models

- Network data analysis
  - Relational View: Rows and Columns are the same entity
  - Example: Social networks, Biological networks
  - Graph View: (Binary) adjacency matrix
- Model

- For each node $p \in \mathcal{N}$:
  - Draw a K dimensional mixed membership vector $\vec{\pi}_p \sim \text{Dirichlet}\,(\,\vec{\alpha}\,)$.
- 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, $Y(p, q) \sim \text{Bernoulli}\,(\,\vec{z}_{p \to q}^{\top} B\, \vec{z}_{p \leftarrow q}\,)$.

MMB Graphical Model



Variational Inference

- Variational lower bound

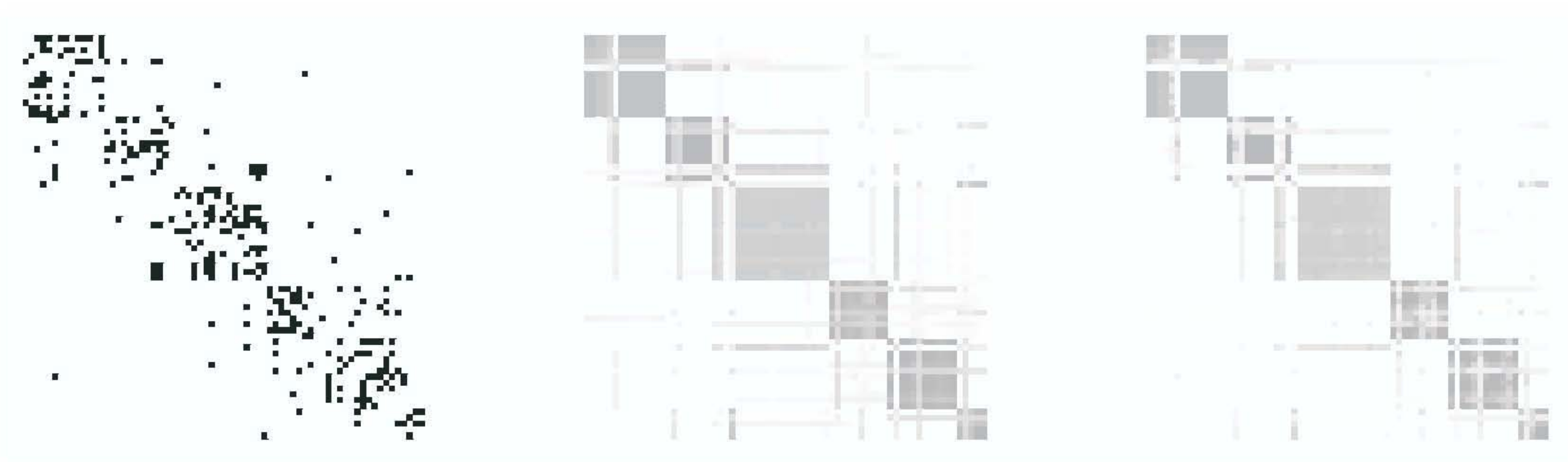
$$\log p(Y \mid \alpha, B) \geq \mathbb{E}_q \left[\log p(Y, \vec{\pi}_{1:N}, Z_{\rightarrow}, Z_{\leftarrow} | \alpha, B) \right] - \mathbb{E}_q \left[\log q(\vec{\pi}_{1:N}, Z_{\rightarrow}, Z_{\leftarrow}) \right]$$

- Fully factorized variational distribution

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

- Variational EM
  - E-step: Update variational parameters (γ,φ)
  - M-step: Update model parameters (α,B)

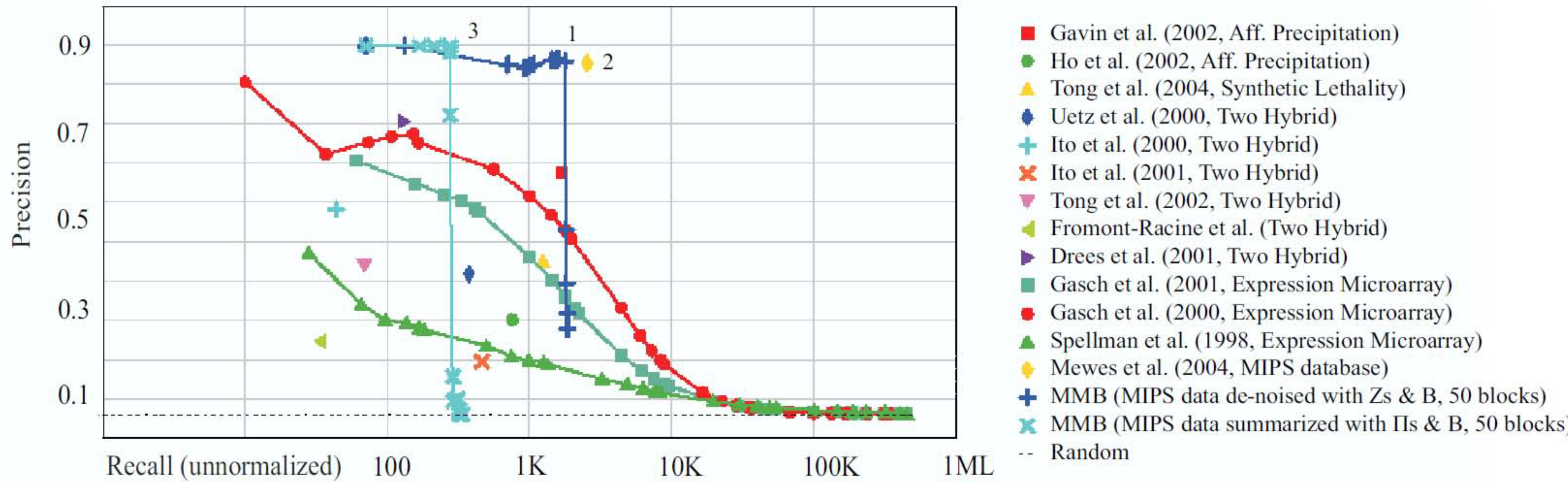
Results: Inferring Communities



Original friendship matrix

Friendships inferred from the posterior, respectively based on thresholding $\pi_p^T B \pi_q$ and $\varphi_p^T B \varphi_q$

Results: Protein Interaction Analysis



“Ground truth”: MIPS collection of protein interactions (yellow diamond)

Comparison with other models based on protein interactions and microarray expression analysis

Non-parametric Bayes

Dirichlet Process Mixtures

Gaussian Processes

Hierarchical Dirichlet Processes

Chinese Restaurant Processes

Pittman-Yor Processes

Mondrain Processes

Indian Buffet Processes

References: Graphical Models

- S. Russell & P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 2009.
- D. Koller & N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, 2009.
- C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2007.
- D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, 2010.
- M. I. Jordan (Ed), *Learning in Graphical Models*, MIT Press, 1998.
- S. L. Lauritzen, *Graphical Models*, Oxford University Press, 1996.
- J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.

References: Inference

- F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, “Factor graphs and the
- sum-product algorithm,” *IEEE Transactions on Information Theory*, *vol.*47, no. 2, 498–519, 2001.
- S. M. Aji and R. J. McEliece, “The generalized distributive law,” *IEEE Transactions on Information Theory*, 46, 325–343, 2000.
- M. J. Wainwright and M. I. Jordan, “Graphical models, exponential families, and variational inference,” *Foundations and Trends in Machine Learning*, vol. 1, no. 1-2, 1-305, December 2008.
- C. Andrieu, N. De Freitas, A. Doucet, M. I. Jordan, “An Introduction to MCMC for Machine Learning,” *Machine Learning*, 50, 5-43, 2003.
- J. S. Yedidia, W. T. Freeman, and Y. Weiss, “Constructing free energy approximations and generalized belief propagation algorithms,” *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2282–2312, 2005.

References: Mixed-Membership Models

- S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. "Indexing by latent semantic analysis," *Journal of the Society for Information Science*, 41(6):391–407, 1990.
- T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, 42(1):177–196, 2001.
- D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003.
- T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, 101(Suppl 1): 5228–5235, 2004.
- Y. W. Teh, D. Newman, and M. Welling. "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation, " *Neural Information Processing Systems* (*NIPS)*, 2007.
- A. Asuncion, P. Smyth, M. Welling, Y.W. Teh, "On Smoothing and Inference for Topic Models," *Uncertainty in Artificial Intelligence (UAI)*, 2009.
- H. Shan, A. Banerjee, and N. Oza, "Discriminative Mixed-membership Models," *IEEE Conference on Data Mining (ICDM)*, 2009.

References: Matrix Factorization

- S. Funk, "Netflix update: Try this at home," http://sifter.org/~simon/journal/20061211.html
- R. Salakhutdinov and A. Mnih. "Probabilistic matrix factorization," *Neural Information Processing Systems* (*NIPS), *2008.
- R. Salakhutdinov and A. Mnih. "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo," *International Conference on Machine Learning (ICML),* 2008.
- I. Porteous, A. Asuncion, and M. Welling, "Bayesian matrix factorization with side information and Dirichlet process mixtures," *Conference on Artificial Intelligence* (*AAAI), 2010.*
- I. Sutskever, R. Salakhutdinov, and J. Tenenbaum. "Modelling relational data using Bayesian clustered tensor facotrization," *Neural Information Processing Systems (NIPS), 2009.*
- A. Singh and G. Gordon, "A Bayesian matrix factorization model for relational data," *Uncertainty in Artificial Intelligence (UAI),* 2010.

References: Co-clustering, Block Structures

- A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, D. Modha., “A Generalized Maximum Entropy Approach to Bregman Co-clustering and Matrix Approximation,” *Journal of Machine Learning Research (JMLR)*, 2007.
- M. M. Shafiei and E. E. Milios, “Latent Dirichlet Co-Clustering,” *IEEE Conference on Data Mining (ICDM)*, 2006.
- H. Shan and A. Banerjee, “Bayesian co-clustering,” *IEEE International Conference on Data Mining (ICDM)*, 2008.
- P. Wang, C. Domeniconi, and K. B. Laskey, “Latent Dirichlet Bayesian Co-Clustering,” *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, 2009.
- H. Shan and A. Banerjee, “Residual Bayesian Co-clustering for Matrix Approximation,” *SIAM International Conference on Data Mining (SDM)*, 2010.
- T. A. B. Snijders and K. Nowicki, “Estimation and prediction for stochastic blockmodels for graphs with latent block structure,” *Journal of Classification*, 14:75–100, 1997.
- E.M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, “Mixed-membership stochastic blockmodels,” *Journal of Machine Learning Research (JMLR), 9*, 1981-2014, 2008.

Acknowledgements

Hanhuai Shan

Amrudin Agovic

Thank you!