# Incorporating Social Context and Domain Knowledge for Entity Recognition

Jie Tang, Zhanpeng Fang

Department of Computer Science, Tsinghua University

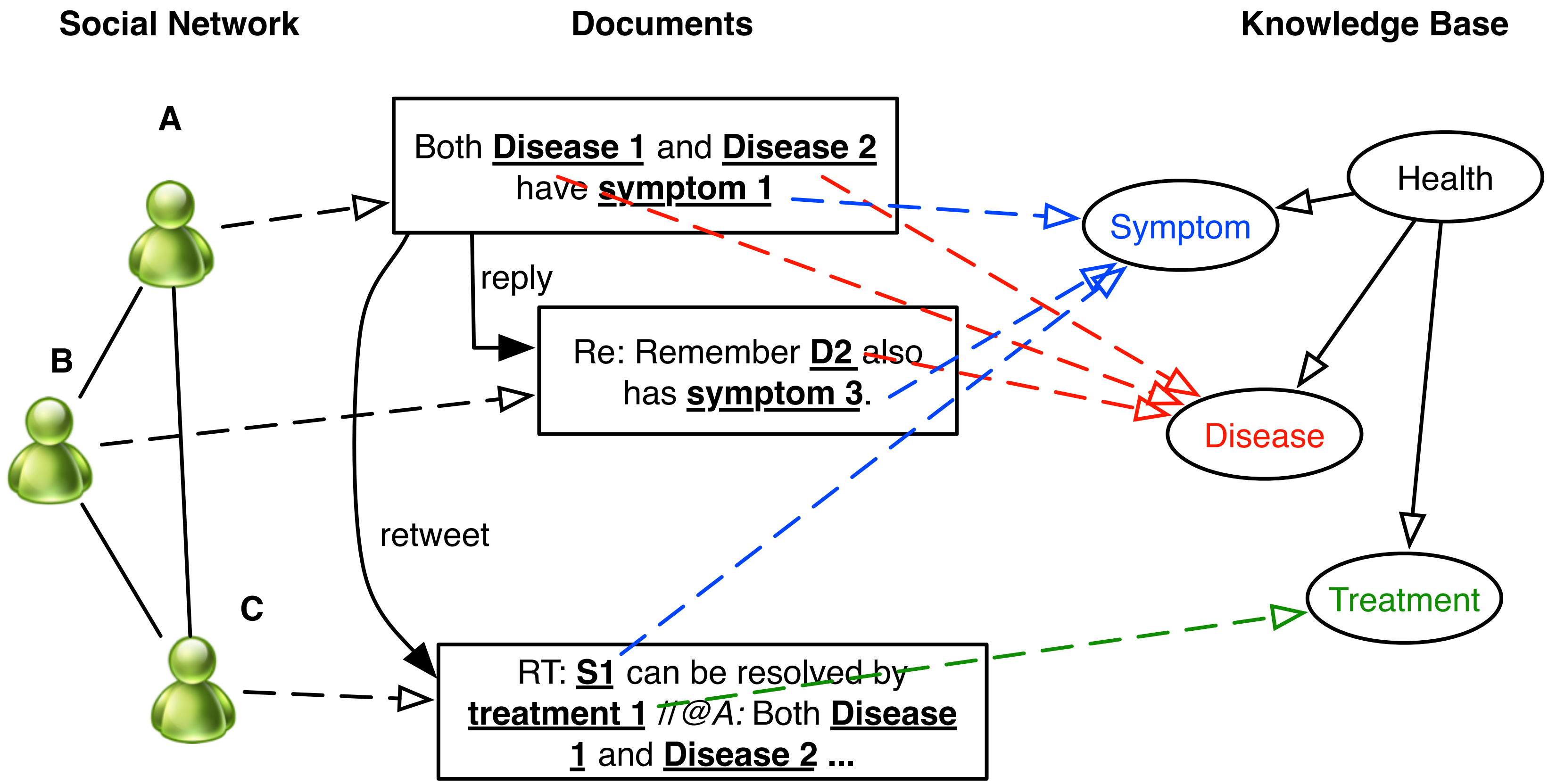
Jimeng Sun

College of Computing, Georgia Institute of Technology

# Entity Recognition in Social Media

- People use blogs, forums, and review sites to share opinions on politicians or products.
- One fundamental analytic issue is to recognize entity instances from the UGC short documents. However, the problem is very challenging
  - "S4" vs. "Samsung Galaxy S4"
  - "Fruit company" vs. "Apple Inc."
  - "Peace West King" vs. "Xilai Bo" (a sensitive Chinese politician)
  - …

# A Concrete Example




**Challenges:** short text + social networks + domain knowledge = ?

# Related Work

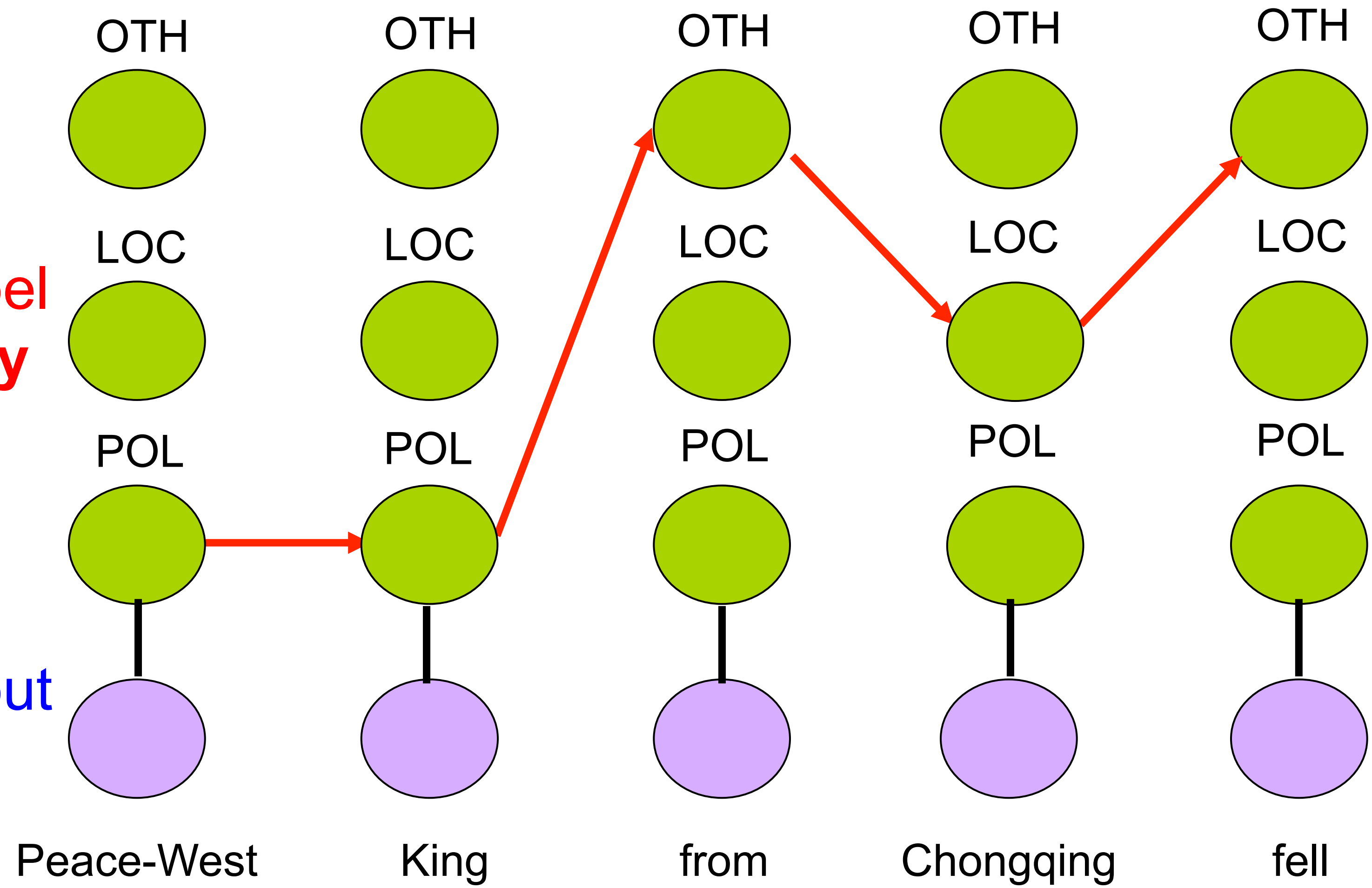
- Entity recognition
  - Modeling as a ranking problem based on boosting and voted perceptron (Collins [9])
  - Incorporating long-distance dependency (Finkel et al. [13])
  - Use Labeled LDA [26] to exploit Freebase to help extraction (Ritter et al. [27])
  - Entity morph (Huang et al. [17])
- Entity resolution
  - A collective method for entity resolution in relational data (Bhattacharya and Getoor [4])
  - A hierarchical topic model for resolving name ambiguity (Kataria et al. [18])
  - Name disambiguation in digital libraries (Tang et al. [32])

# Approach Framework
# —SOCINST

# Preliminary: Sequential Labeling

The label results **y**

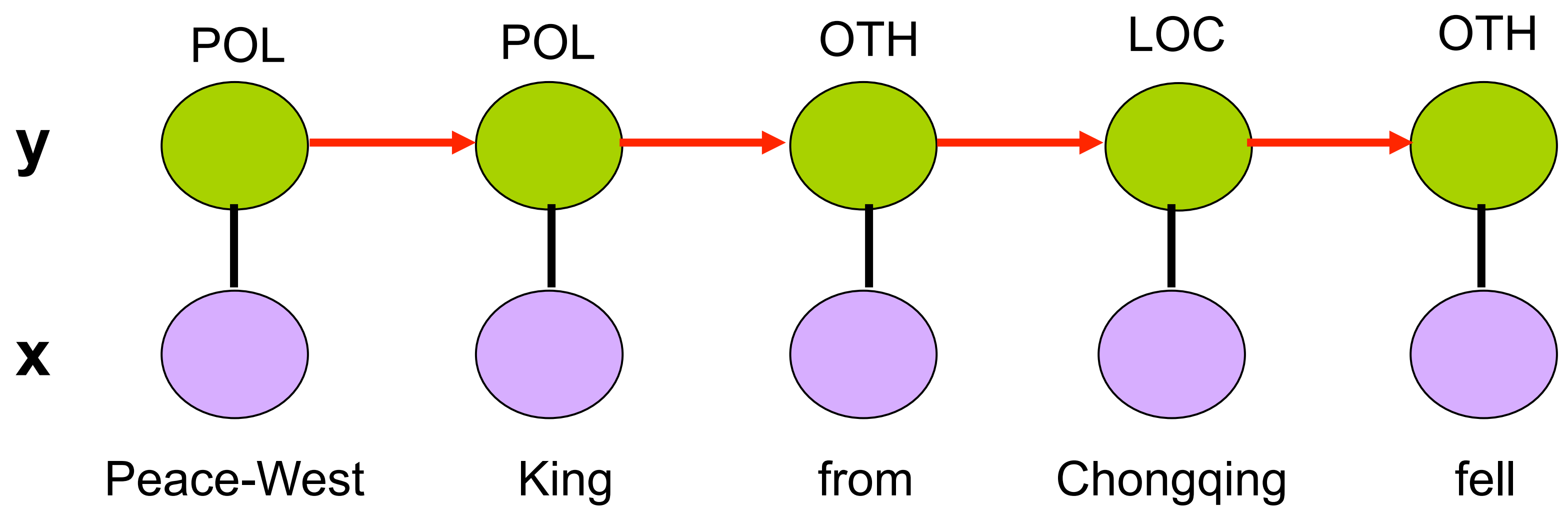
The input text **x**




$$\boxed{\mathbf{y}^* = \max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x}; f, \Theta)}$$

where $f$ represents features and $\Theta$ are model parameters.

# Sequential Labeling with CRFs



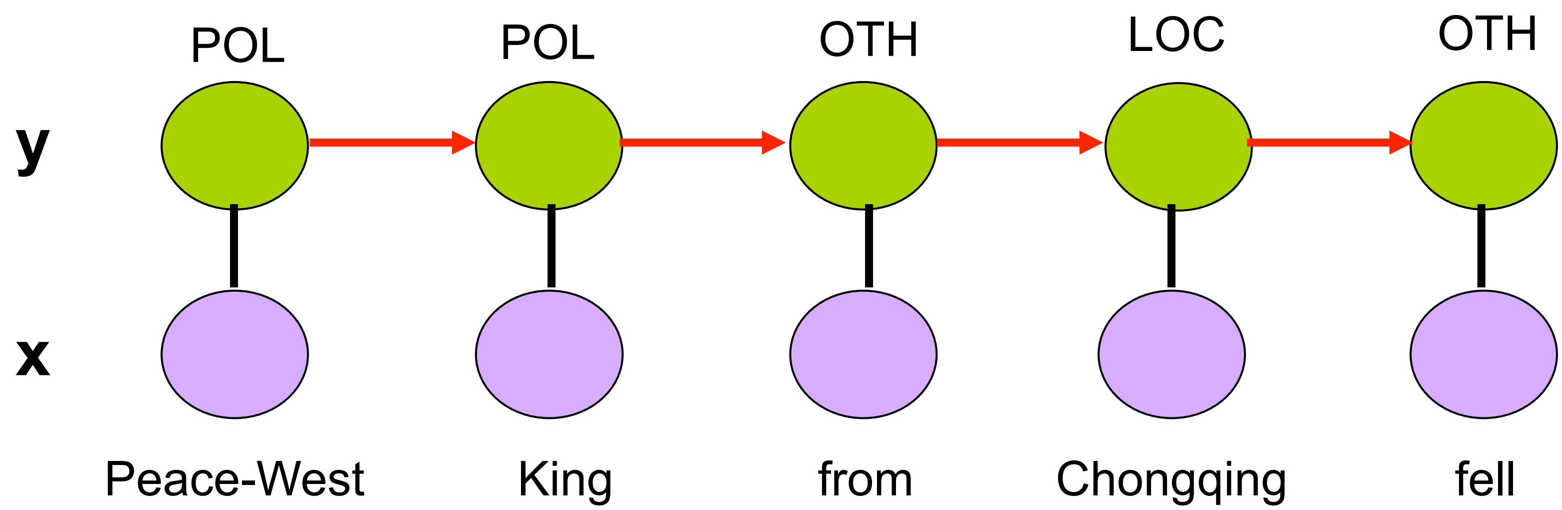

$\mu$ and $\lambda$ are parameters to be learned from the training data.

$$p(\mathbf{y} \mid \mathbf{x}, \lambda, \mu) = \frac{1}{Z} \exp\Big(\sum_i \sum_k \boxed{\lambda_k} \boxed{f_k(x_i, y_i)} + \sum_i \sum_j \boxed{\mu_j} \boxed{f_j(\mathbf{x}, y_i, y_{i+1})}\Big)$$

$f_k$ denotes the $k$-th feature defined for token $x_i$

$f_j$ denotes the $j$-th feature defined for two consecutive tokens $x_i$; and $x_j$;

# Sequential Labeling with CRFs

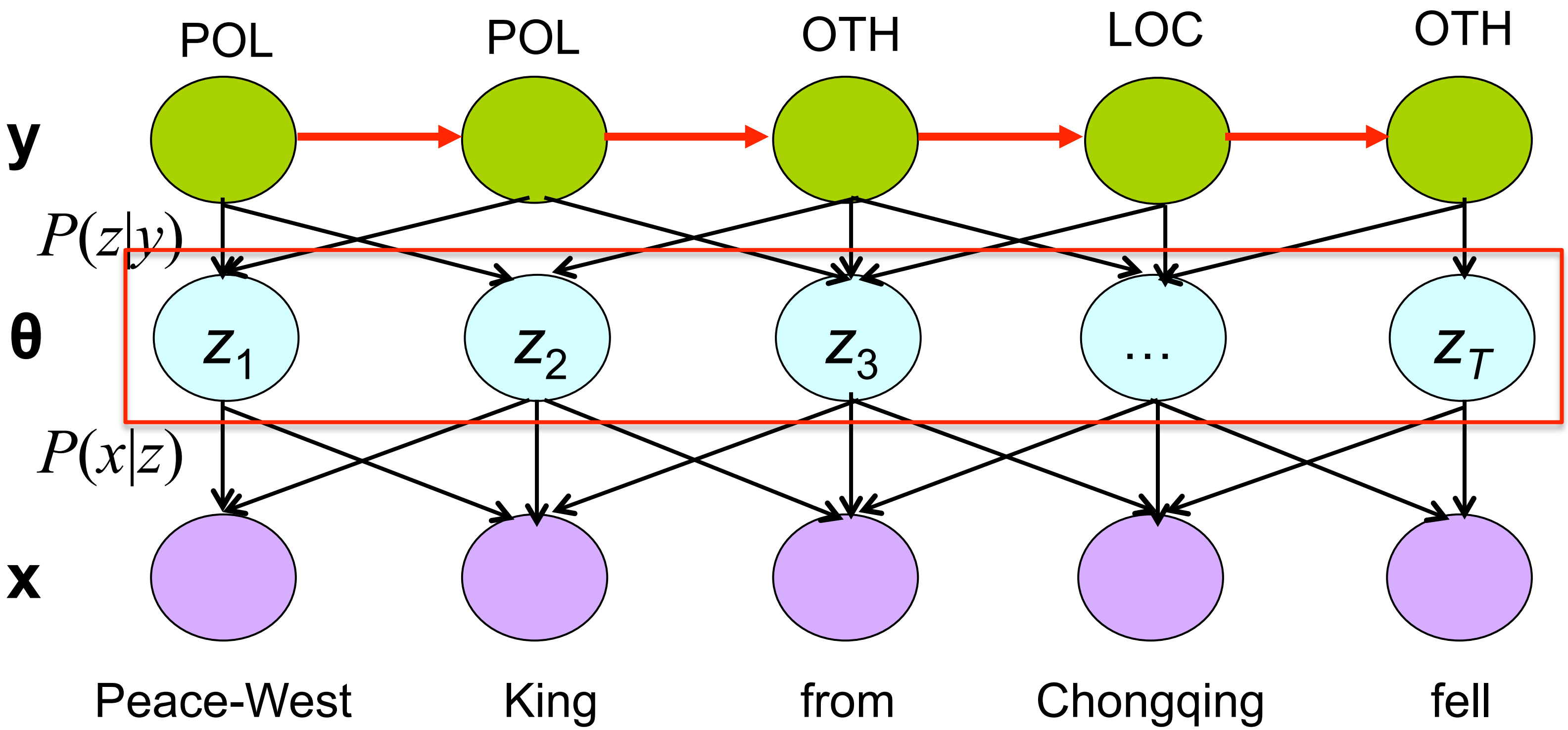



**Performance of the model will be bad when dealing with short-text due to sparsity**

$f_k$ denotes the $k$-th feature defined for token $x_i$

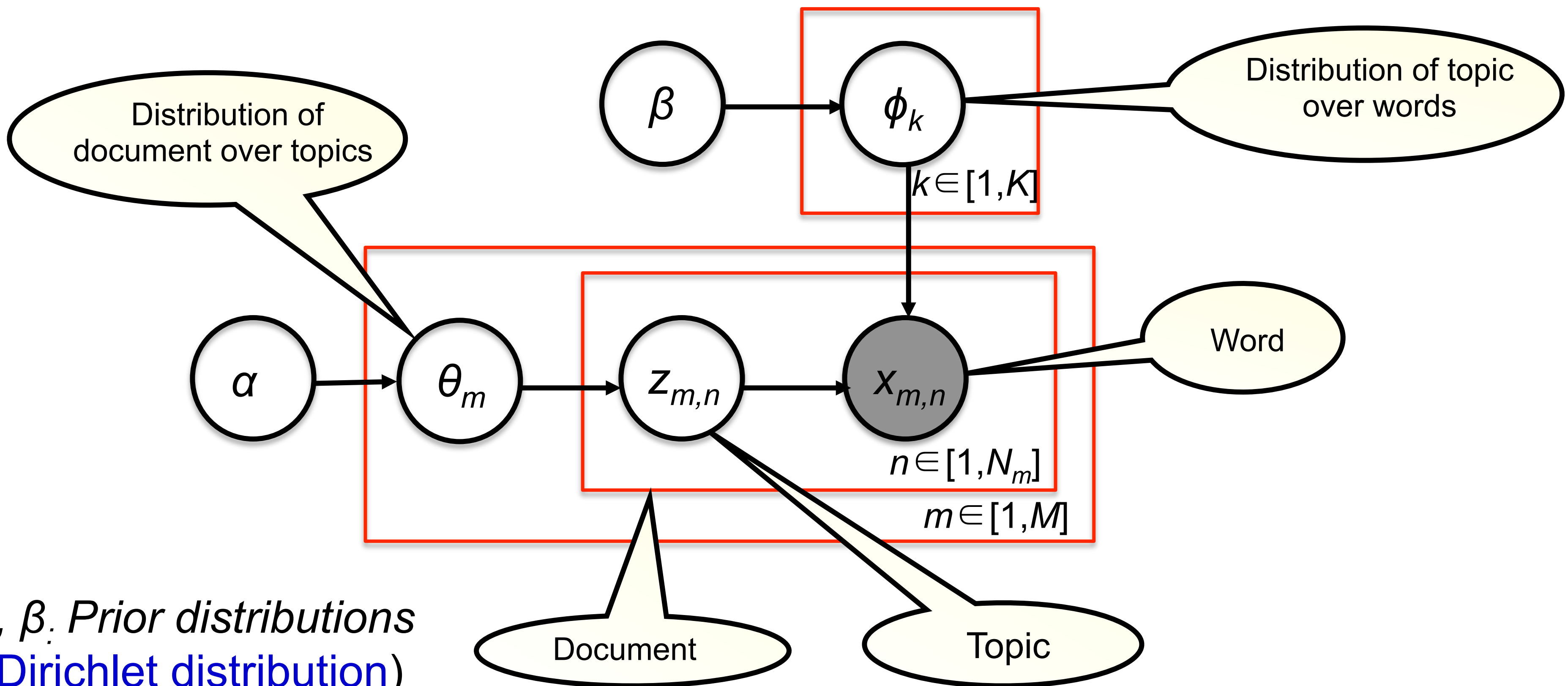
$f_j$ denotes the $j$-th feature defined for two consecutive tokens $x_i$; and $x_j$;

# Sequential Labeling Incorporating Topics




$$p(\mathbf{y} \mid \mathbf{x}, \theta, \lambda, \mu) = \frac{1}{Z} \exp\left(\sum_i \sum_k \lambda_k f_k(x_i, \theta_i, y_i) + \sum_i \sum_j \mu_j f_j(\mathbf{x}, \theta, y_i, y_{i+1})\right)$$

# Latent Dirichlet Allocation



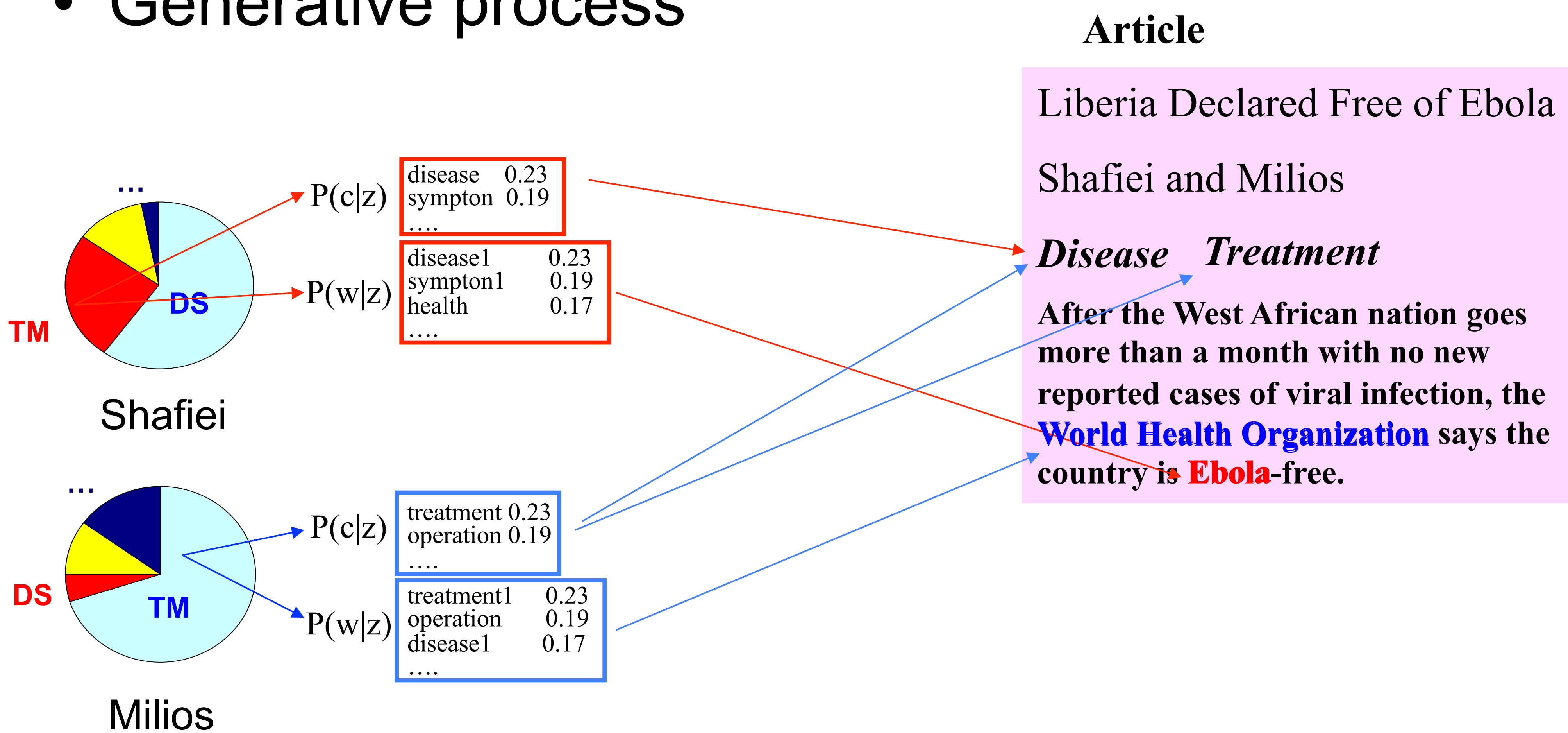

*α, β: Prior distributions*
(Dirichlet distribution)

$$p(\mathbf{x},\mathbf{z},\theta,\phi \mid \alpha,\beta) = \prod_{z=1}^{K} p(\phi_z \mid \beta) \prod_{d=1}^{M} p(\theta_d \mid \alpha) \prod_{i=1}^{N_d} p(x_i \mid \phi_z) p(z \mid \theta_d)$$


[5] D. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. JMLR, 3:993–1022, 2003.

# Extend to Model Authorship and Categories

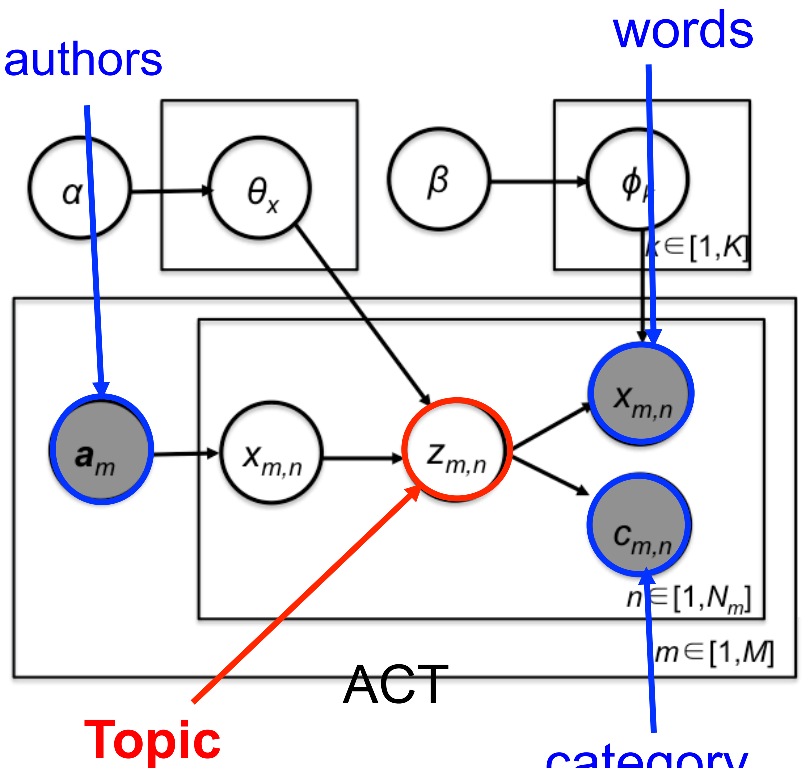
- Generative process





[35] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In KDD'08, pages 990–998, 2008

# ACT Model




Generative process:

1. For each topic $z$, draw $\phi_z$ and $\psi_z$ respectively from Dirichlet priors $\beta_z$ and $\mu_z$;

2. For each word $w_{di}$ in document $d$:

- draw an author $x_{di}$ from $\mathbf{a}_d$ uniformly;
- draw a topic $z_{di}$ from a multinomial distribution $\theta_{x_{di}}$ specific to author $x_{di}$, where $\theta$ is generated from a Dirichlet prior $\alpha$;
- draw a word $w_{di}$ from multinomial $\phi_{z_{di}}$;
- draw a category tag $c_{di}$ from multinomial $\psi_{z_{di}}$.

$$P(z_{di}, x_{di}|\mathbf{z}_{-di}, \mathbf{x}_{-di}, \mathbf{w}, \mathbf{c}, \alpha, \beta, \mu) \propto$$
$$\frac{m^{-di}_{x_{di}z_{di}} + \alpha_{z_{di}}}{\sum_z (m^{-di}_{x_{di}z} + \alpha_z)} \frac{n^{-di}_{z_{di}w_{di}} + \beta_{w_{di}}}{\sum_v (n^{-di}_{z_{di}v} + \beta_v)} \frac{n^{-d}_{z_{di}c_d} + \mu_{c_d}}{\sum_c (n^{-d}_{z_{di}c} + \mu_c)}$$

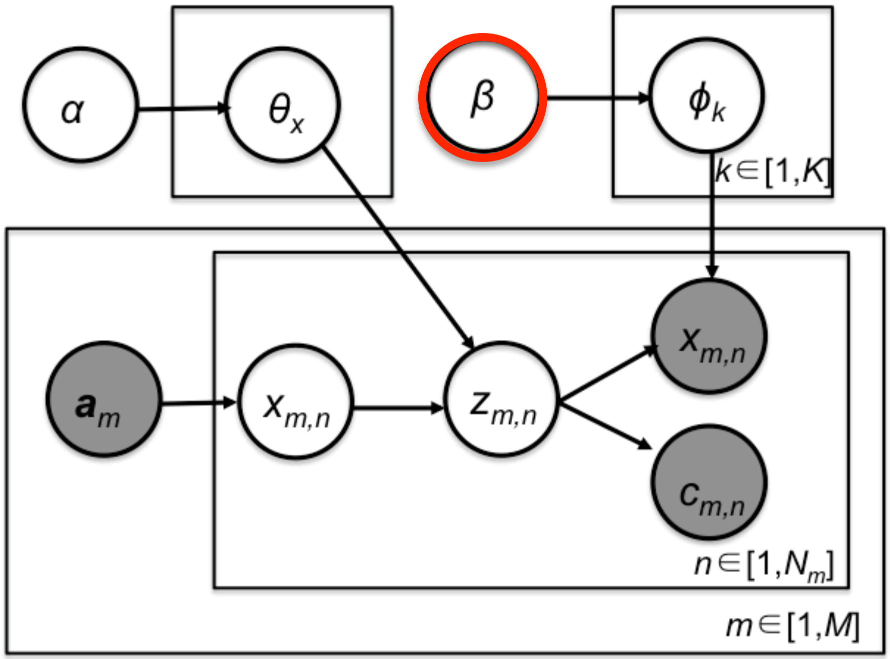

[35] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In KDD'08, pages 990–998, 2008

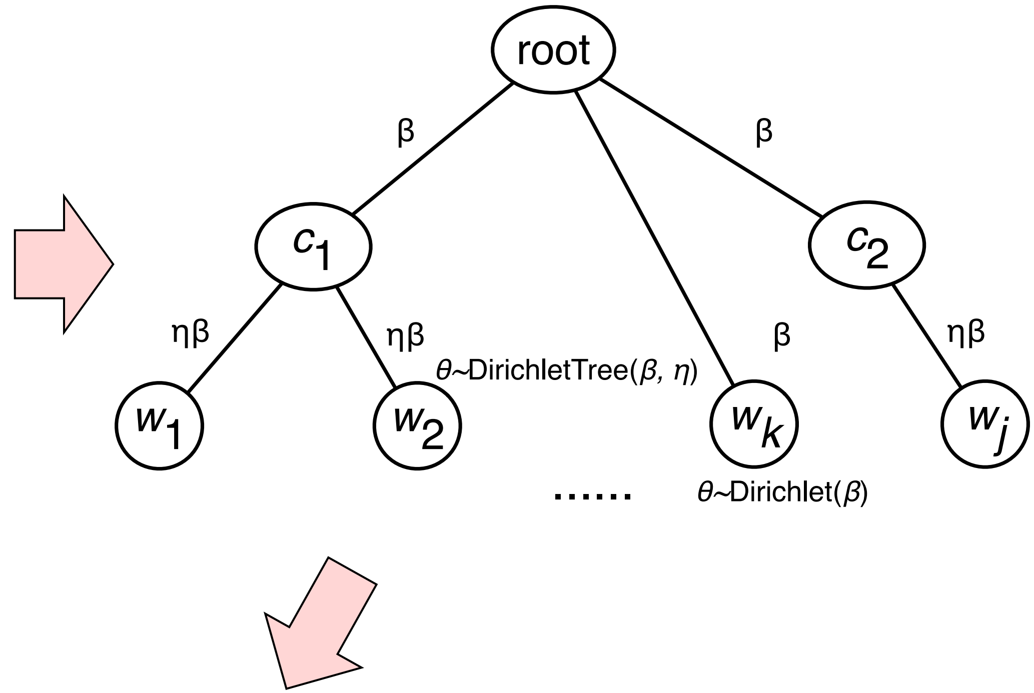
# Still challenges

However, we still cannot model domain knowledge and social context!

SOCINST: Modeling Domain Knowledge and Social Context Simultaneously

# Modeling Domain Knowledge



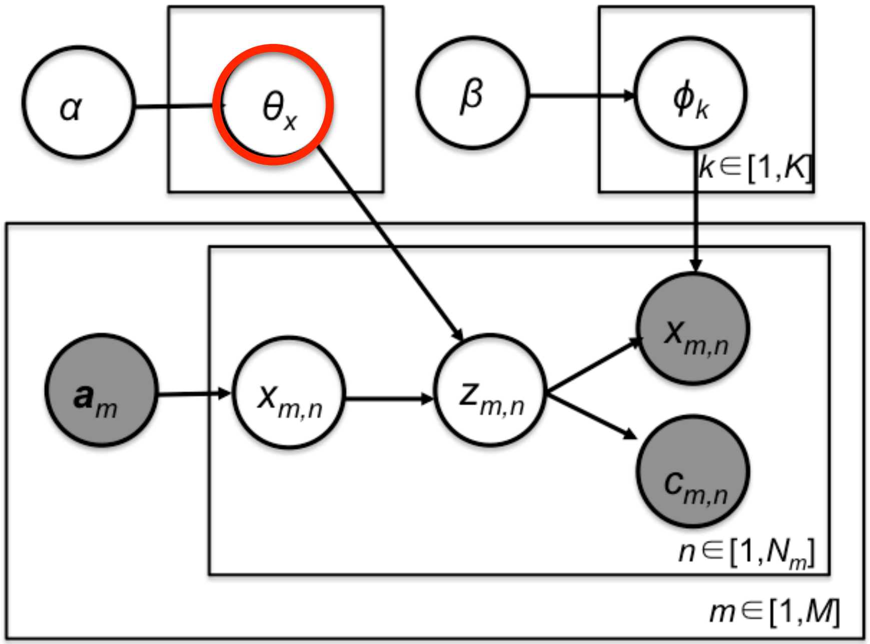



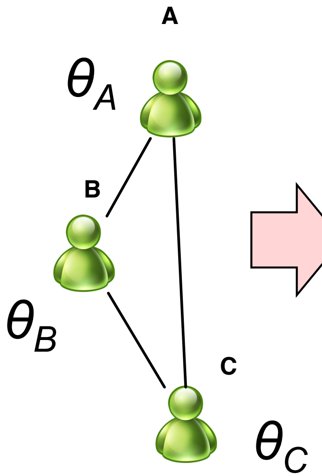

$$\mathrm{DirichletTree}(\beta, \eta) = \left( \prod_{i=1}^{W} \phi_{zw_i}^{\eta w_i} \right) \times$$

$$\left( \prod_{j, c_j \in C} \frac{\Gamma\left(\sum_{k, w_k \in W(c)} \eta_{w_k}\right)}{\prod_{k, w_k \in W(c)} \Gamma(\eta_{w_k})} \left( \sum_{k, w_k \in W(c)} \phi_{zw_i}^{\eta w_i} \right)^{\Delta(s)} \right)$$


[1] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In ICML'09, pages 25–32, 2009.

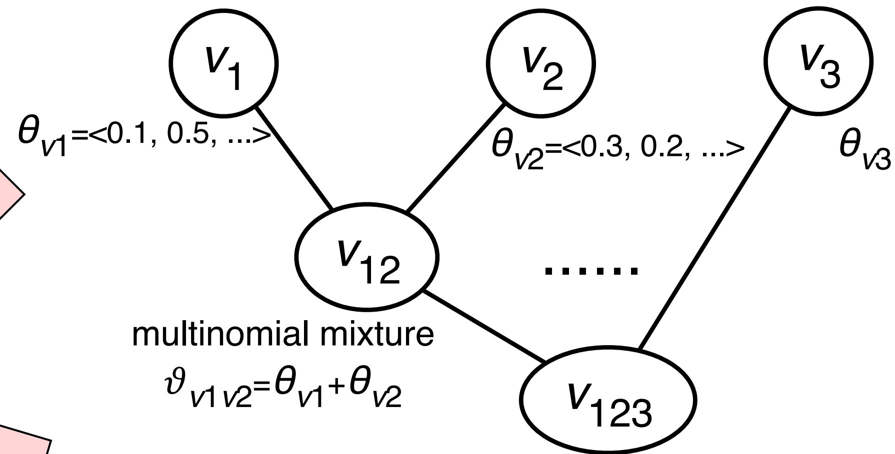
# Modeling Social Context




User A's Social context is defined as a mixture of topic distributions of neighbors, i.e.

$$\sum_{j\in NB(v_i)} \gamma_j \theta_j$$



# Theoretical Basis

- Aggregation property of Dirichlet distribution

If

$$(\theta_1,\ldots,\theta_i,\theta_{i+1},\ldots,\theta_K) \sim \text{Dirichlet}(\alpha_1,\ldots,\alpha_i,\alpha_{i+1},\ldots,\alpha_K)$$

then

$$(\theta_1,\ldots,\theta_i+\theta_{i+1},\ldots,\theta_K) \sim \text{Dirichlet}(\alpha_1,\ldots,\alpha_i+\alpha_{i+1},\ldots,\alpha_K)$$

- Inverse of the aggregation property

If

$$(\theta_1,\ldots,\theta_K) \sim \text{Dirichlet}(\alpha_1,\ldots,\alpha_K)$$

then

$$(\theta_1,\ldots,\tau\theta_i,(1-\tau)\theta_i,\ldots,\theta_K) \sim \text{Dirichlet}(\alpha_1,\ldots,\tau\alpha_i,(1-\tau)\alpha_i,\ldots,\alpha_K)$$

# Model Learning

**Input**: a social network $G$, a document set $D$, a knowledge base $KB$;
**Output**: estimated parameters $\theta$,$\phi$

For each author $v$, draw $\theta_v$ from Dirichlet prior $\alpha$;
For each topic $z$, draw $\phi_z$ from Dirichlet prior $\beta$;
**foreach** *document* $d$ **do**
  **if** $v_d$ *does not have relationship with others* **then**
    **foreach** *word* $w_{di} \in \mathbf{w}_d$ **do**
      Draw a topic $z_{di} \sim multi(\theta_v)$ from the topic model of user $v$;
      Call SamplingWord($z_{di}$, $w_{di}$);
    **end**
  **end**
  **else if** $v_d$ *have relationship with* $v'$ **then**
    Construct a multinomial mixture $\vartheta_{v_d v'}$ by combining topics distributions specific to users $v_d$ and $v'$;
    **foreach** *word* $w_{di} \in \mathbf{w}_d$ **do**
      Draw a topic $z_{di} \sim multi(\vartheta)$ from the distribution specific to the pair;
      Call SamplingWord($z_{di}$, $w_{di}$);
    **end**
  **end**
**end**

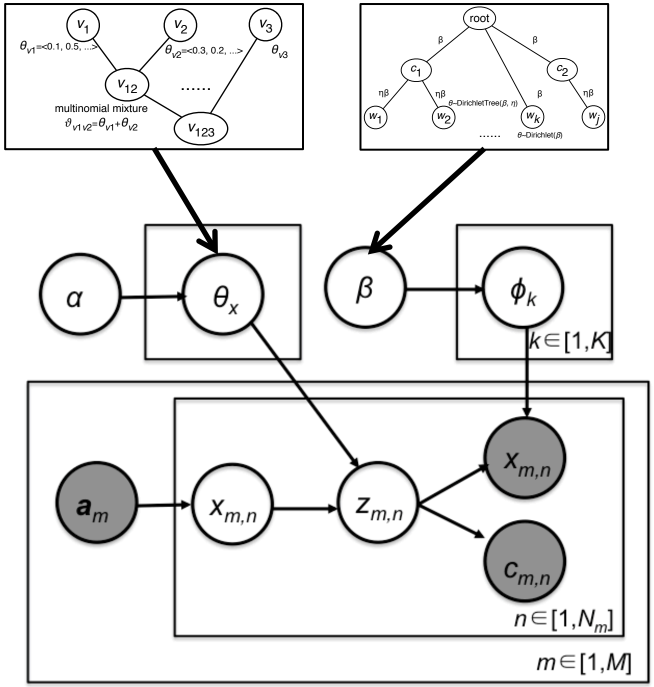
`SamplingWord(`$z_{di}$, $w_{di}$`)`
**if** $w_{di}$ *is an instance of a concept* $c \in KB$ **then**
  Draw a concept path $\{c_k\}_k \sim multi(\pi)$ from a topic-specific concept path distribution;
  Draw word $w_{di} \sim multi(\psi_c)$ from a concept-specific multinomial distribution;
**end**
**else**
  Draw word $w_{di} \sim multi(\phi_{z_{di}})$ directly from a topic-specific multinomial distribution;
**end**

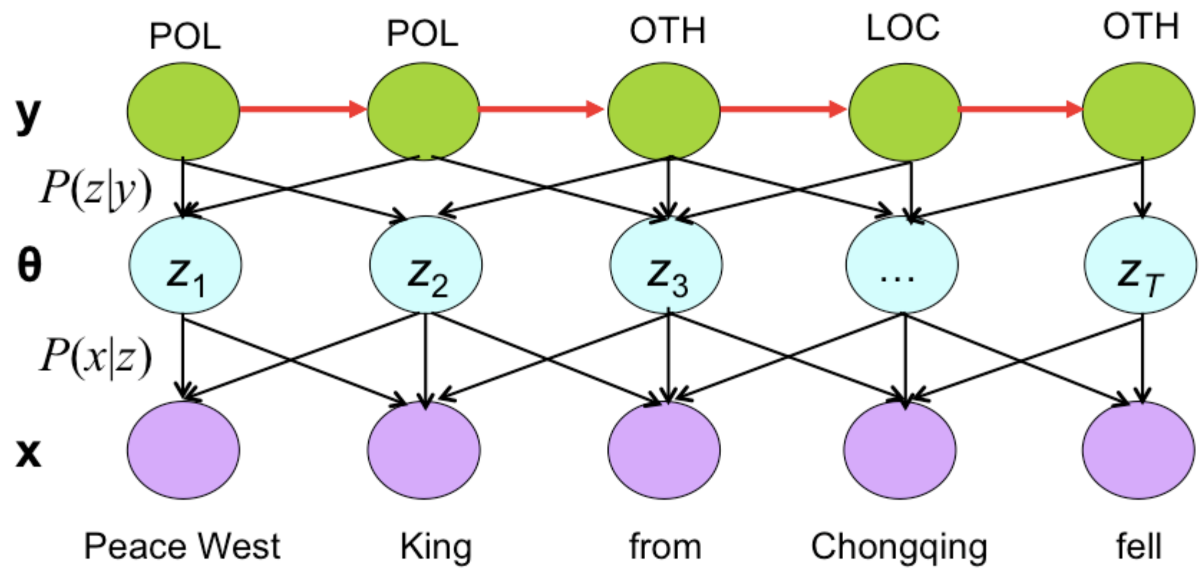
$$P(z_{di}|\mathbf{z}_{-di},\mathbf{w},\cdot) = \boxed{\frac{n_{vz_{di}}^{-di} + \gamma n_{v'z_{di}} + \alpha}{\sum_z (n_{vz}^{-di} + \gamma n_{v'z}) + W\alpha}} \times \prod_{k=1}^{T} \frac{m_{z_{di}c_{di}^k}^{-di} + W_{c_{di}^k}\beta}{\sum_{c_s}(m_{z_{di}c_s}^{-di} + W_{c_s^k}\beta)} \times \frac{m_{c_{di}w_{di}}^{-di} + \eta}{\sum_w m_{c_{di}w}^{-di} + W_c\eta}$$

$$P(z_{di}|\mathbf{z}_{-di},\mathbf{w},\cdot) = \frac{n_{vz_{di}}^{-di} + \alpha}{\sum_z n_{vz}^{-di} + W\alpha} \boxed{\times \prod_{k=1}^{T} \frac{m_{z_{di}c_{di}^k}^{-di} + W_{c_{di}^k}\beta}{\sum_{c_s}(m_{z_{di}c_s}^{-di} + W_{c_s^k}\beta)}} \times \frac{m_{c_{di}w_{di}}^{-di} + \eta}{\sum_w m_{c_{di}w}^{-di} + W_c\eta}$$

$$P(z_{di}|\mathbf{z}_{-di},\mathbf{w},\cdot) = \frac{n_{vz_{di}}^{-di} + \alpha}{\sum_z n_{vz}^{-di} + K\alpha} \times \frac{m_{z_{di}w_{di}}^{-di} + \beta}{\sum_w m_{z_{di}w}^{-di} + W\beta}$$

# Sequential Labeling Incorporating Topics




$$p(\mathbf{y} \mid \mathbf{x}, \theta, \lambda, \mu) = \frac{1}{Z} \exp\left(\sum_i \sum_k \lambda_k f_k(x_i, \theta_i, y_i) + \sum_i \sum_j \mu_j f_j(\mathbf{x}, \theta, y_i, y_{i+1})\right)$$

# Experiments

# Data Sets

- All codes and datasets can be downloaded here http://aminer.org/socinst/
- Dataset

| Domain | #documents | #instances | #relationships |
|---|---|---|---|
| Weibo | 1,800 | 545 | 10,763 |
| I2B2 | 899 | 2,400 | 27,175 |
| ICDM'12 Contest | 2,110 | 565 | NA |

- Goal:
  - **Weibo**: Our goal is to extract real morph instances in the dataset.
  - **I2B2**: Our goal here is to extract private health information instances in the dataset.
  - **ICDM'12 Contest**: Our goal is to recognize product mentions in the dataset.

# I2B2

HISTORY OF PRESENT ILLNESS :
Mr. Blind is a 79-year-old white male with a history of diabetes mellitus , inferior myocardial infarction , who underwent open repair of his increased diverticulum November 13th at Sephsandpot Center. The patient developed hematemesis November 15th and was intubated for respiratory distress.
He was transferred to the Valtawnprinceel Community Memorial Hospital for endoscopy and esophagoscopy on the 16th of November which showed a 2 cm linear tear of the esophagus at 30 to 32 cm .

Patient

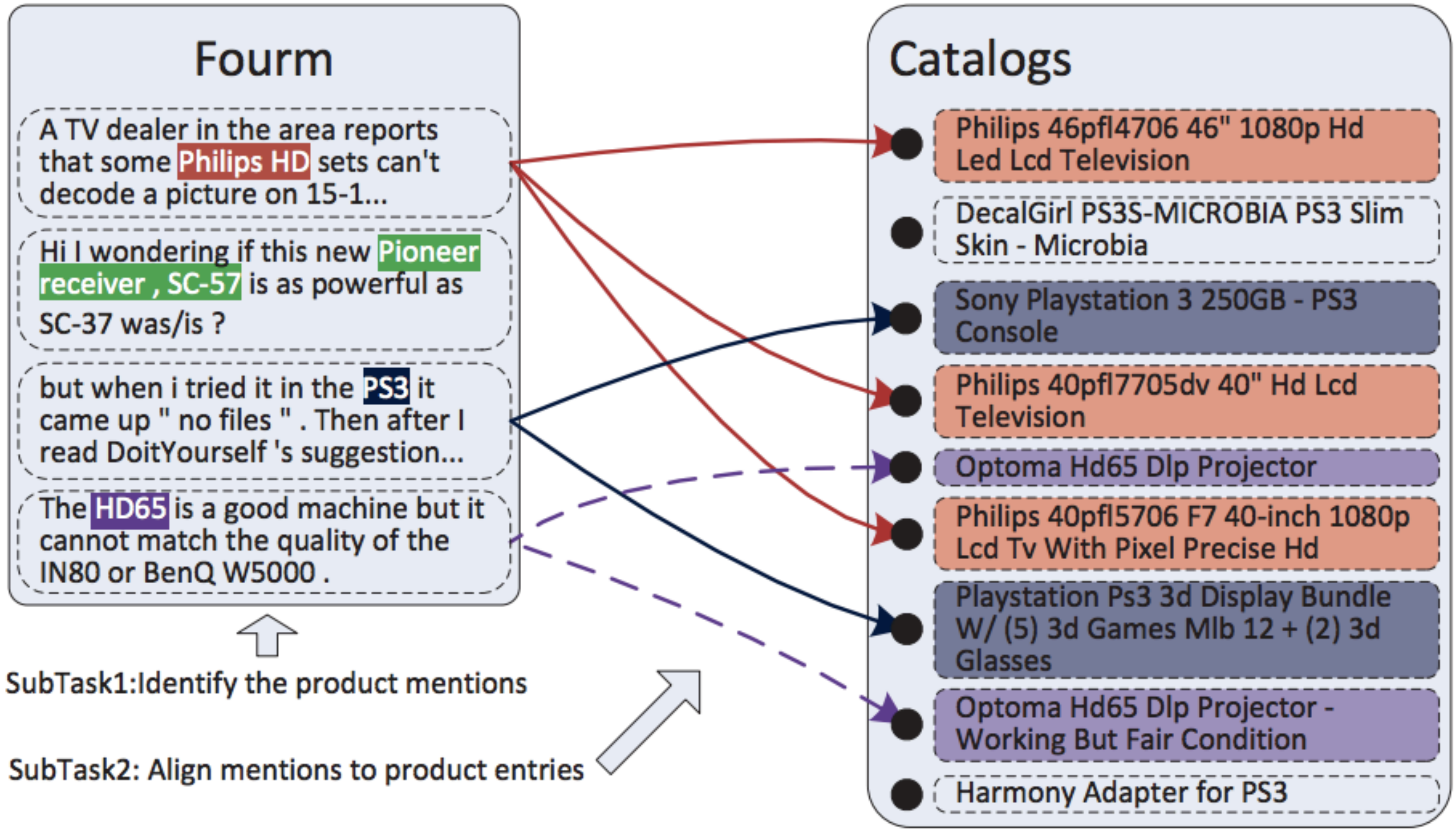
Doctor

Date

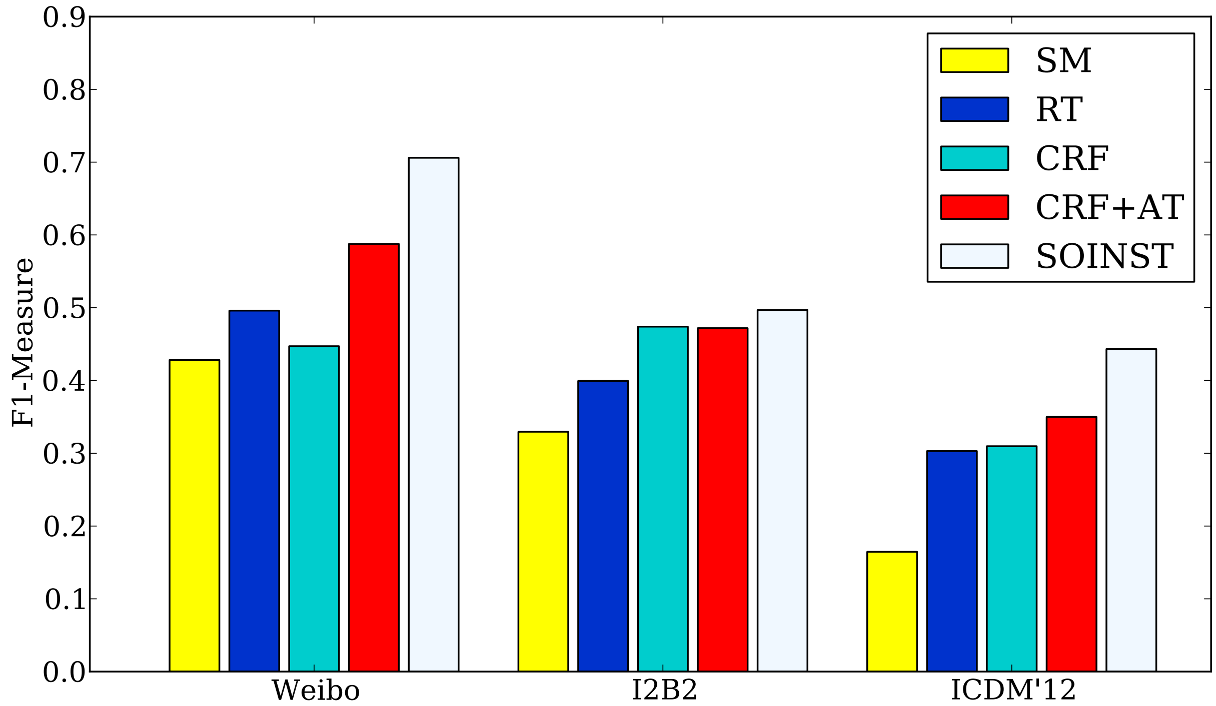
Location

Hospital

# ICDM'12 Contest



# Results




- **SM:** Simply extracts all the terms/symbols that are annotated
- **RT:** Recognizes target instances from the test data by a set of rule templates
- **CRF:** Trains a CRF model using features associated with each token
- **CRF+AT:** Uses Author-Topic (AT) [30] to train a model and then it use the learned topics as features for CRF for instance recognition
- **SOCINST:** Our proposed model

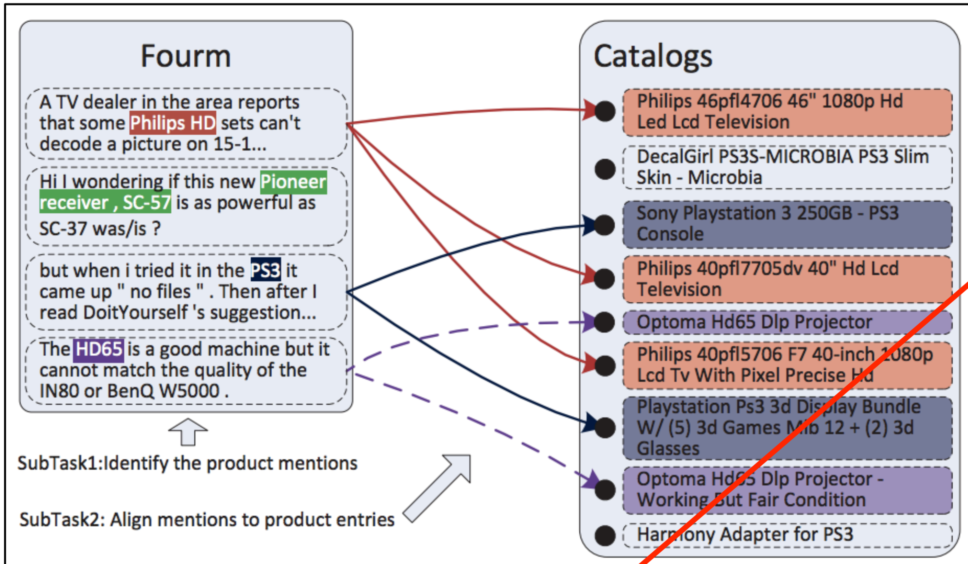
# Results

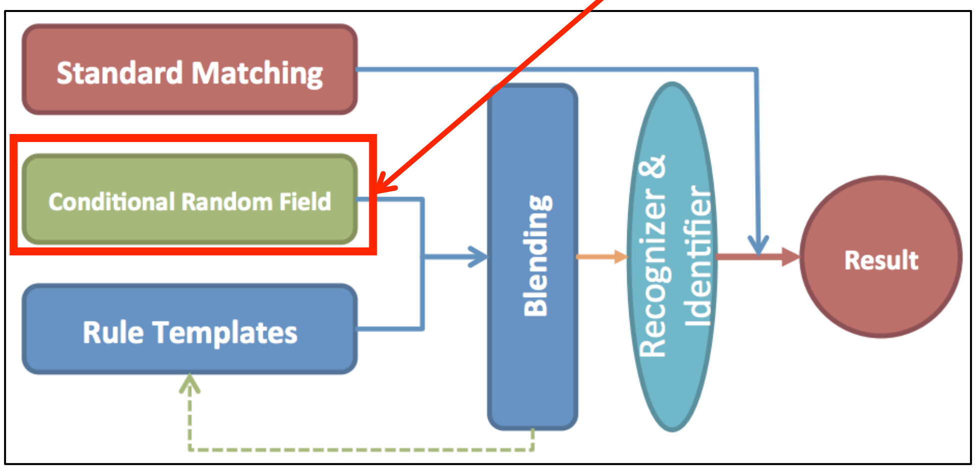
- **SM:** Simply extracts all the terms/symbols that are annotated
- **RT:** Recognizes target instances from the test data by a set of rule templates.
- **CRF:** Trains a CRF model using features associated with each token
- **CRF+AT:** Uses Author-Topic (AT) [30] to train a model and then it use the learned topics as features for CRF for instance recognition
- **SOCINST:** Our proposed model

| Data | Method | Recall | Precision | F1-Measure |
|---|---|---|---|---|
| Weibo | SM | 55.34 | 34.92 | 42.82 |
| | RT | 39.62 | 66.31 | 49.60 |
| | CRF | 29.24 | 94.89 | 44.71 |
| | CRF+AT | 43.71 | 89.67 | 58.77 |
| | SOCINST | **65.72** | 76.27 | **70.60** |
| I2B2 | SM | 39.58 | 28.24 | 32.96 |
| | RT | 39.60 | 40.29 | 39.94 |
| | CRF | 40.99 | 56.19 | 47.40 |
| | CRF+AT | 41.37 | 54.92 | 47.19 |
| | SOCINST | **43.94** | **57.18** | **49.69** |
| ICDM'12 Contest | SM | 9.47 | 62.50 | 16.46 |
| | RT | 23.69 | 42.01 | 30.30 |
| | CRF | 21.80 | 53.48 | 30.97 |
| | CRF+AT | 26.54 | 51.37 | 35.00 |
| | SOCINST | **37.91** | 53.33 | **44.32** |

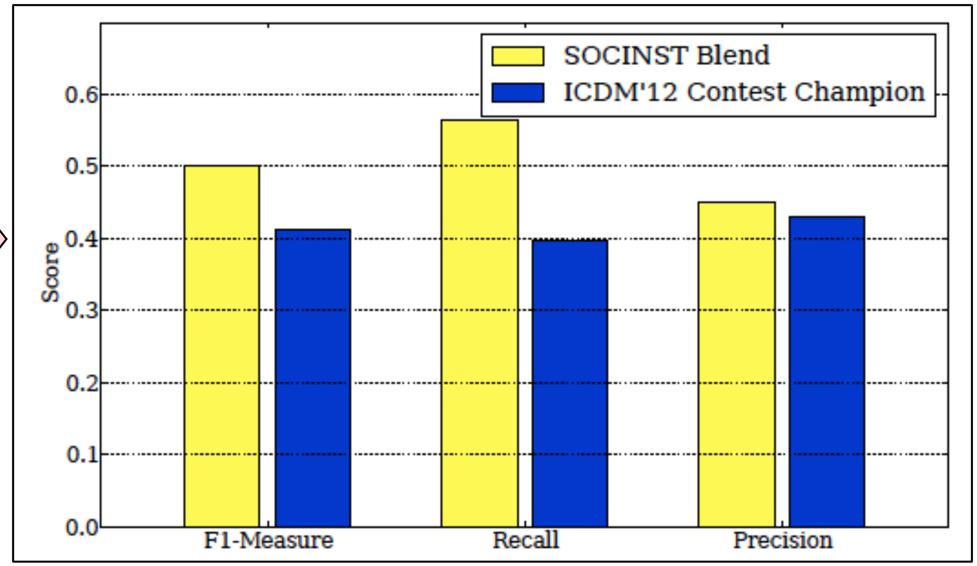
# More Results—ICDM'12 Contest

Performance comparison of SOCINST and the first place [38] in ICDM'12 Contest.



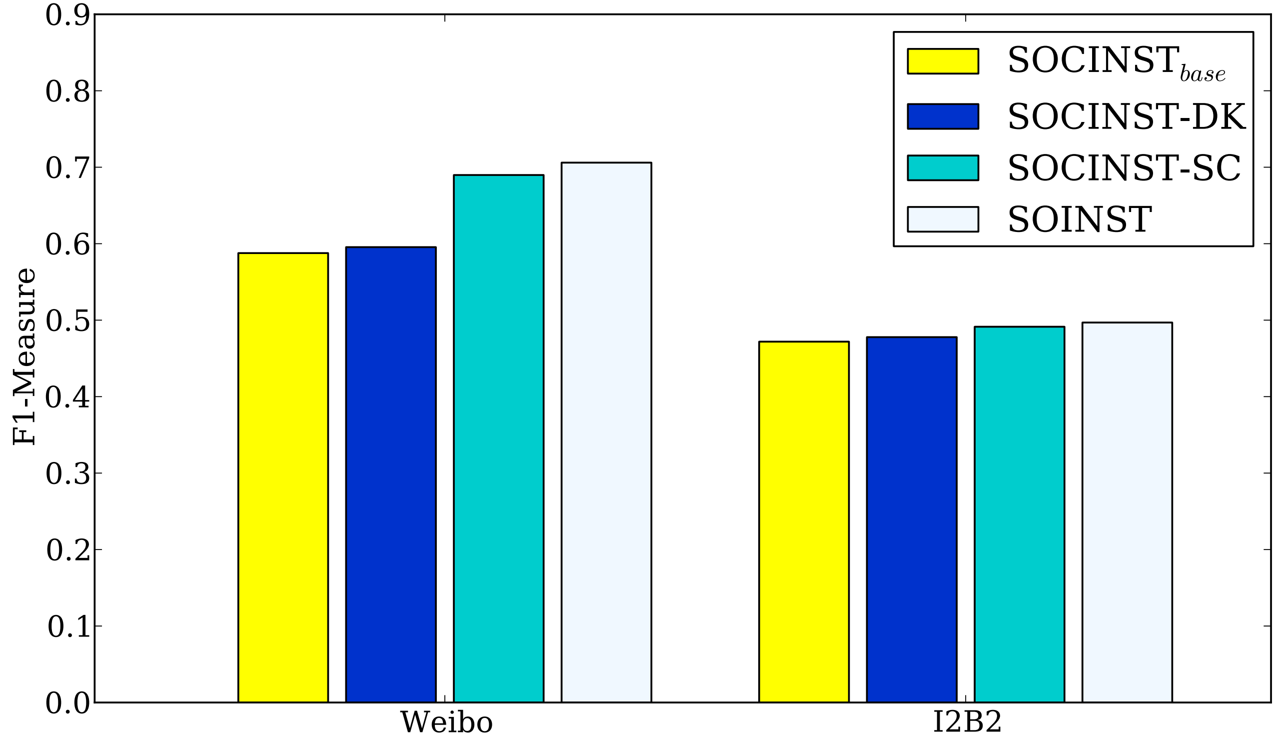




**By incorporating the modeling results into the CRF model**



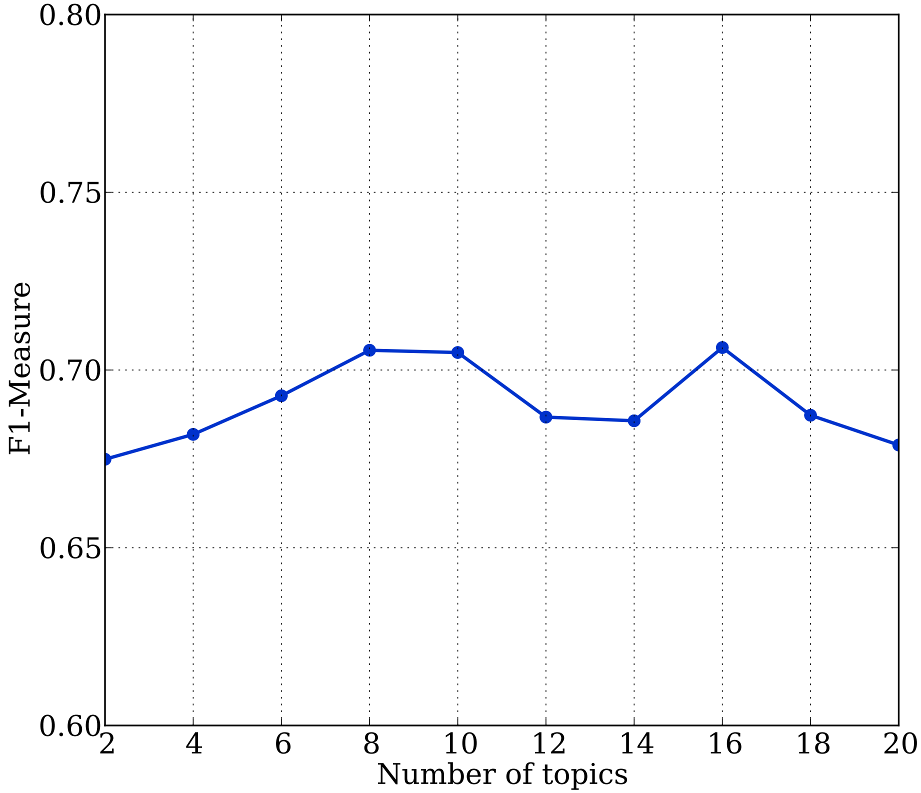

[38] S. Wu, Z. Fang, and J. Tang. Accurate product name recognition from user generated content. In ICDM'12 Contest.

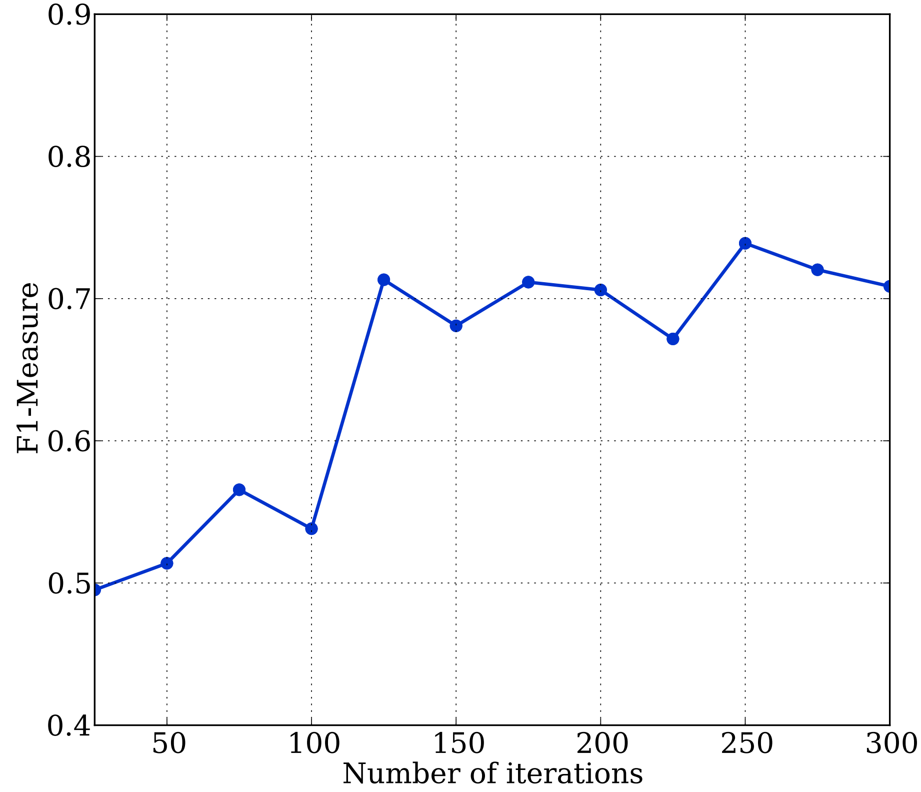
# Effects of Social Context and Domain Knowledge



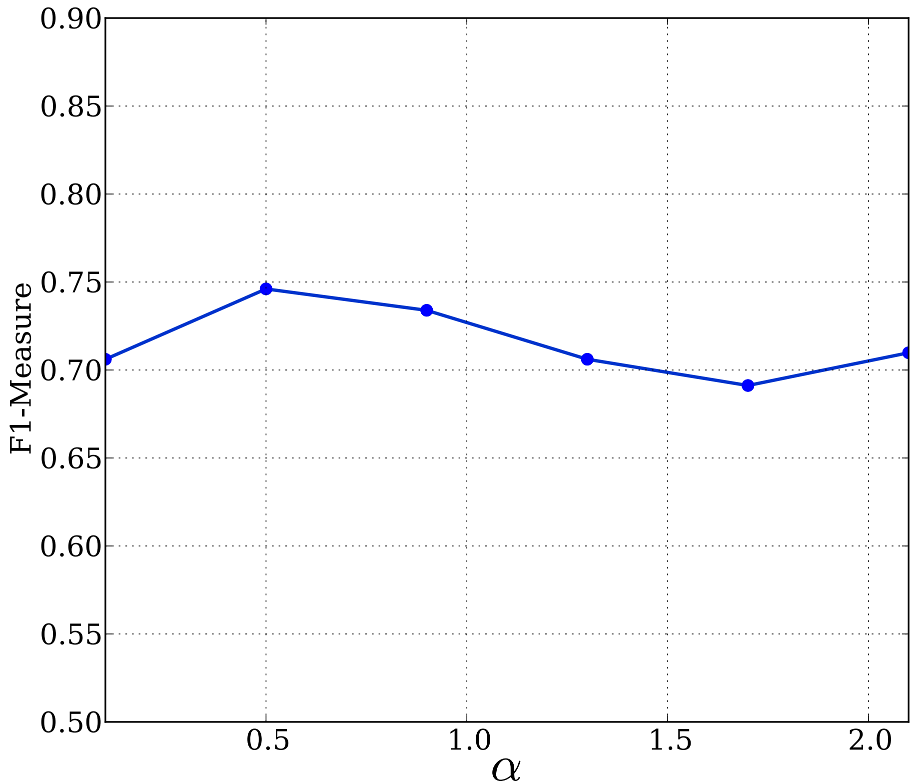

**SOCINST$_{base}$**— we removed both social context and domain knowledge from our method;
**SOCINST-SC**— we removed social context from our method;
**SOCINST-DK**— we removed domain knowledge from our method;

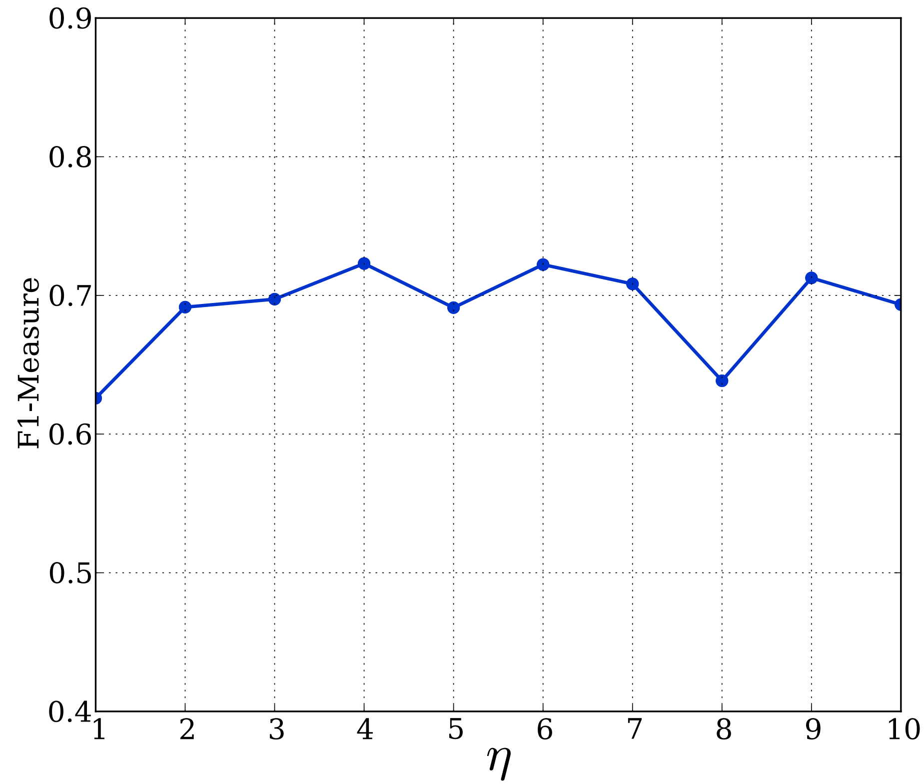
# Parameter Analysis



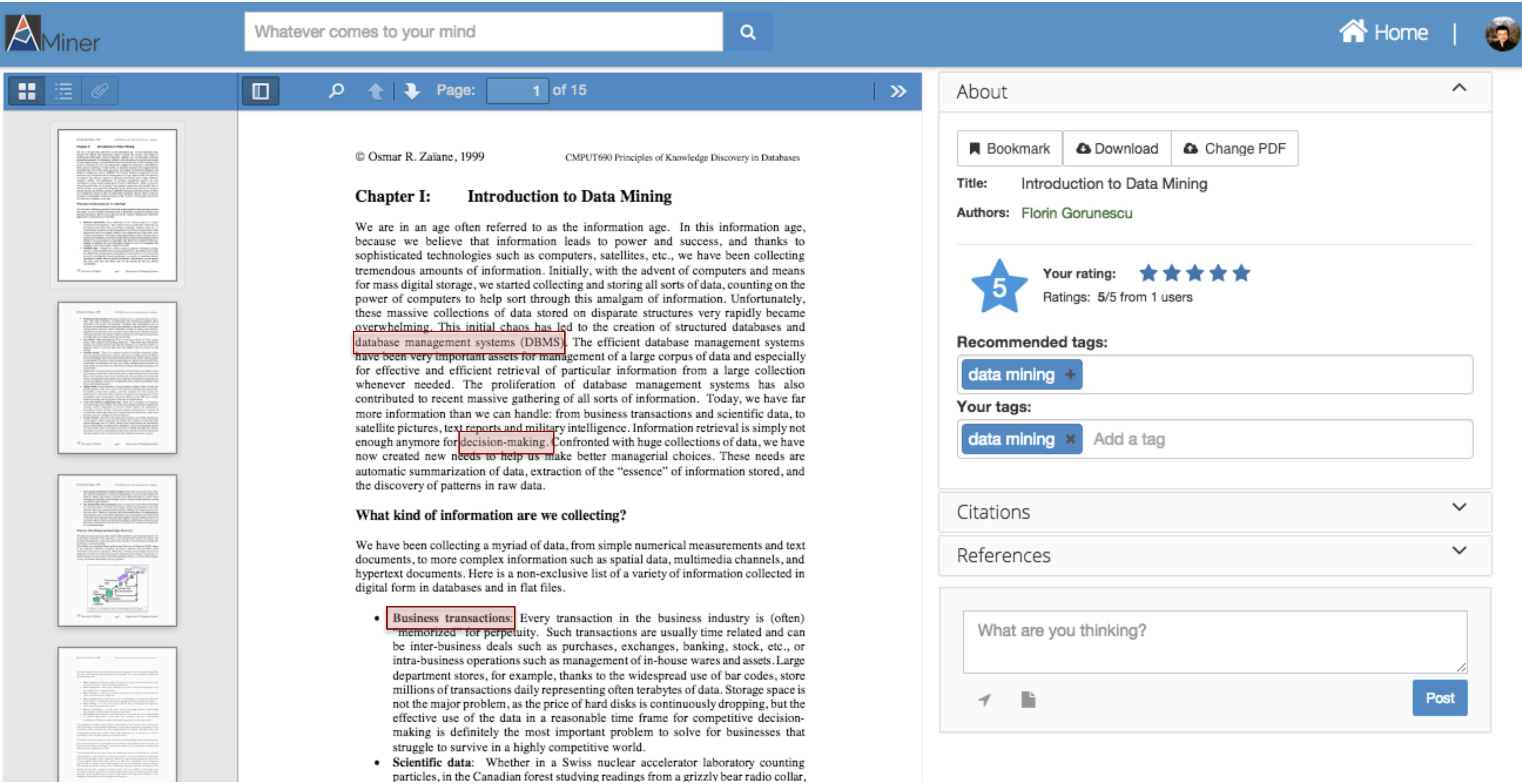
# Parameter Analysis (cont.)




*** All the other hyperparameters fixed**
**The number of topics is set to K = 15**

# AMiner
(http://aminer.org)



# Conclusion

- Study the problem of instance recognition by incorporating social context and domain knowledge
- Propose a topic modeling approach to learn topics by considering social relationships between users and context information from a domain knowledge base
- Experimental results on three different datasets validate the effectiveness and the efficiency of the proposed method.

# Future work

- The general idea of incorporating social context and domain knowledge for entity recognition represents a new research direction
- Combining the sequential labeling model and the proposed SOCINST into a unified model should be beneficial
- Further incorporating other social interactions, such as social influence, to help instance recognition is an intriguing direction

# Thank you！

**Collaborators:**

Jimeng Sun (**Georgia Tech**)

Zhanpeng Fang (**THU**)

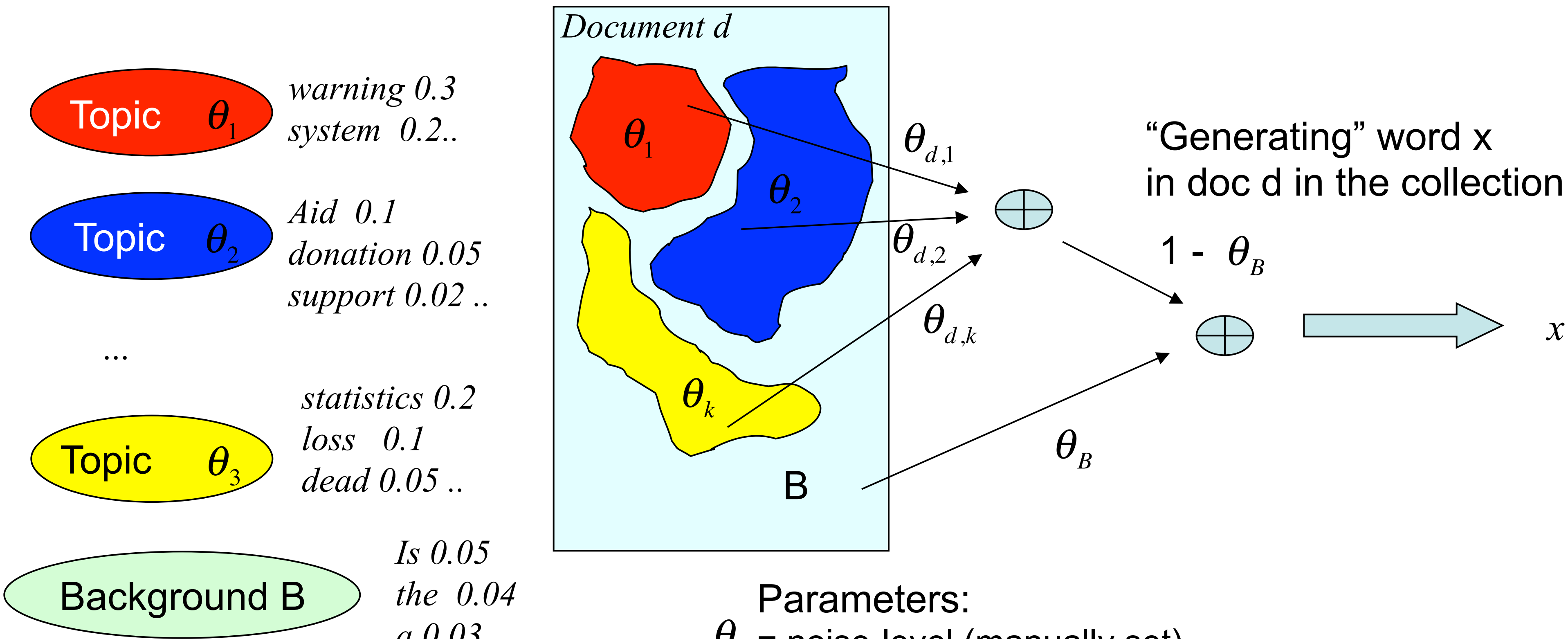
Jie Tang, KEG, Tsinghua U, http://keg.cs.tsinghua.edu.cn/jietang

**Download all data & Codes,** http://aminer.org/socinst

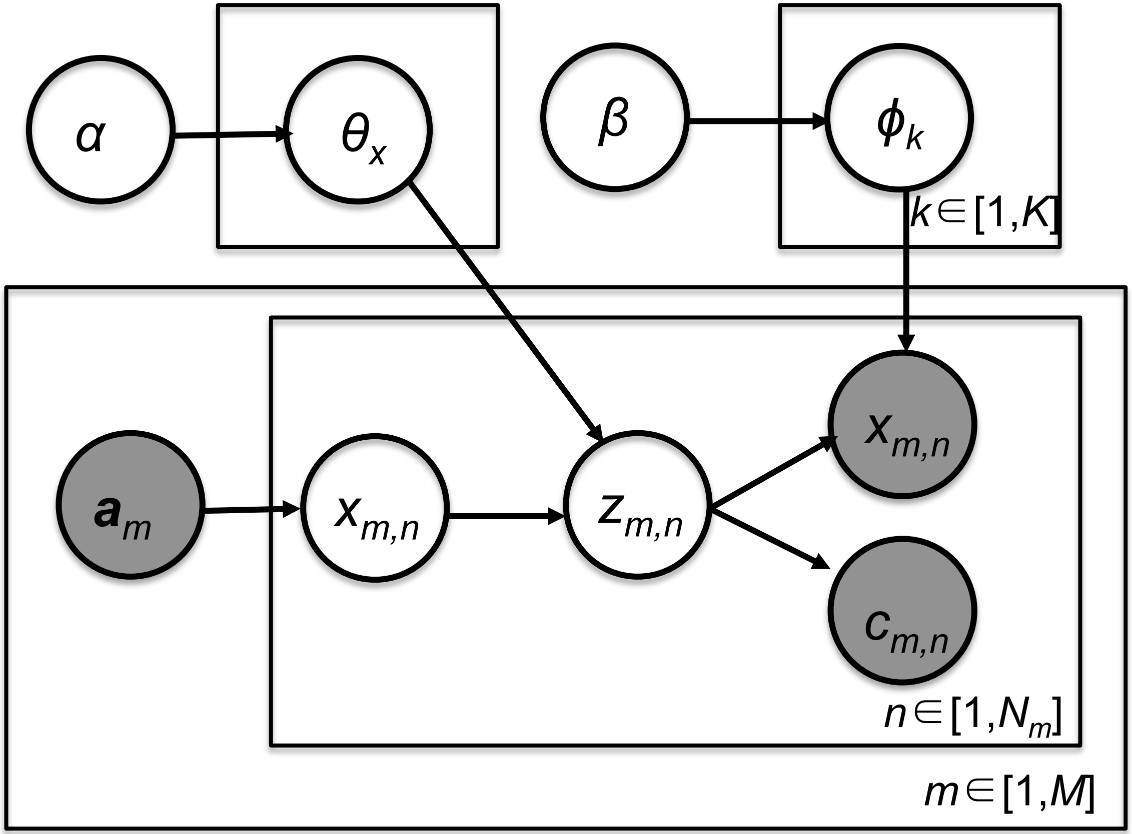
# Modeling Short Text with Topics

$$p_d(x) = \lambda_B p(x \mid \theta_B) + (1-\lambda)\sum_{k=1}^{K} \pi_{d,k} p(x \mid \theta_k)$$

$$\log p(d) = \sum_{x \in V} n(x,d) \log[\lambda_B p(x \mid \theta_B) + (1-\lambda)\sum_{k=1}^{K} \pi_{d,k} p(x \mid \theta_k)]$$




Parameters:
$\theta_B$ = noise-level (manually set)
$\theta_1$ and $\pi$ are estimated with Maximum Likelihood



α
θ_x
β
ϕ_k
k∈[1,K]
a_m
x_m,n
z_m,n
x_m,n
c_m,n
n∈[1,N_m]
m∈[1,M]