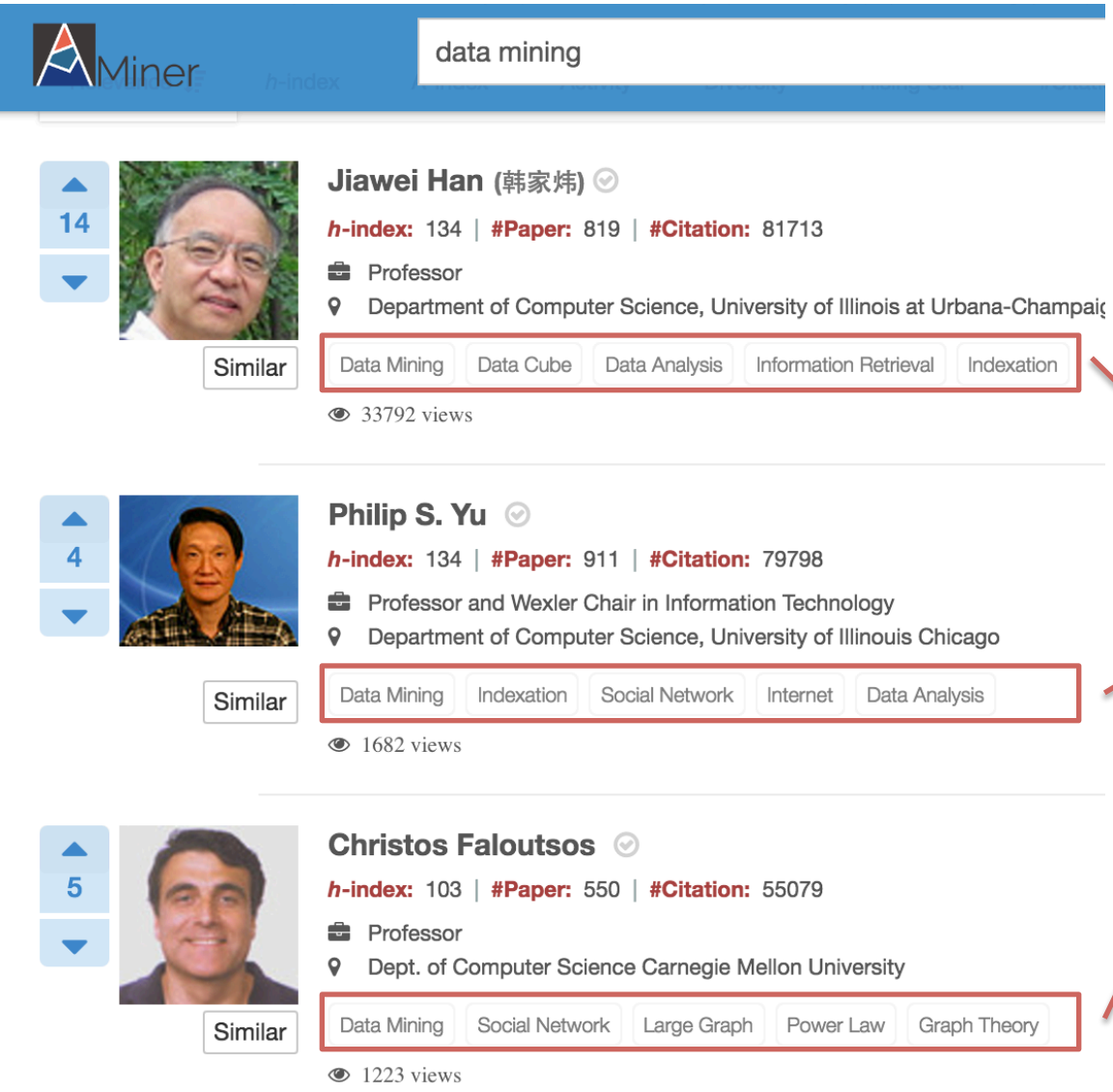
Multi-Modal Bayesian Embeddings for Learning Social Knowledge Graphs

Zhilin Yang[12], Jie Tang[1], William W. Cohen[2]

[1]Tsinghua University

[2]Carnegie Mellon University

AMiner: academic social network



Research interests

Text-Based Approach

818

Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions

Wei Shen, Jianyong Wang, **Jiawei Han**

Knowledge and Data Engineering, IEEE Transactions (2015)

Bibtex http://dx.doi.org/10.1109/TKDE.2014.2327028

817

Power-Based Diagnosis of Node Silence in Remote High-End Sensing Sy

Yong Yang, Lu Su, Mohammad Maifi Hasan Khan, Michael LeMay, Tarek F. Abdelzaher, Jia

TOSN (2015)

Bibtex http://doi.acm.org/10.1145/2661639

816

A Framework of Mining Trajectories from Untrustworthy Data in Cyber-P

Lu An Tang, Xiao Yu, Quanquan Gu, Jiawei Han, Guofei Jiang, Alice Leung, Thomas F. La

List of publications

Infer

Similar

Jiawei Han (韩家炜)

H-Index: 133 | **#Paper:** 818 | **#Citation:** 111164

Department of Computer Science, University of Illinois at Urbana-

Professor

Data Mining | Information Extraction | Data Analysis | Machine Learning

Research interests

Text-Based Approach

818

Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions

Wei Shen, Jianyong Wang, **Jiawei Han**

Knowledge and Data Engineering, IEEE Transactions (2015)

Bibtex http://dx.doi.org/10.1109/TKDE.2014.2327028

817

Power-Based Diagnosis of Node Silence in Remote High-End Sensing Sy

Yong Yang, Lu Su, Mohammad Maifi Hasan Khan, Michael LeMay, Tarek F. Abdelzaher, Jia

TOSN (2015)

Bibtex http://doi.acm.org/10.1145/2661639

816

A Framework of Mining Trajectories from Untrustworthy Data in Cyber-P

Lu An Tang, Xiao Yu, Quanquan Gu, Jiawei Han, Guofei Jiang, Alice Leung, Thomas F La

Similar

Jiawei Han (韩家炜)

H-Index: 133 | **#Paper:** 818 | **#Citation:** 111164

Department of Computer Science, University of Illinois at Urbana-

Professor

Data Mining | Information Extraction | Data Analysis | Machine Learning

Term Frequency => “challenging problem”

TF-IDF => “line drawing”

Knowledge-Driven Approach

818

Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions

Wei Shen, Jianyong Wang, **Jiawei Han**

Knowledge and Data Engineering, IEEE Transactions (2015)

Bibtex http://dx.doi.org/10.1109/TKDE.2014.2327028

817

Power-Based Diagnosis of Node Silence in Remote High-End Sensing Sy

Yong Yang, Lu Su, Mohammad Maifi Hasan Khan, Michael LeMay, Tarek F. Abdelzaher, Jia

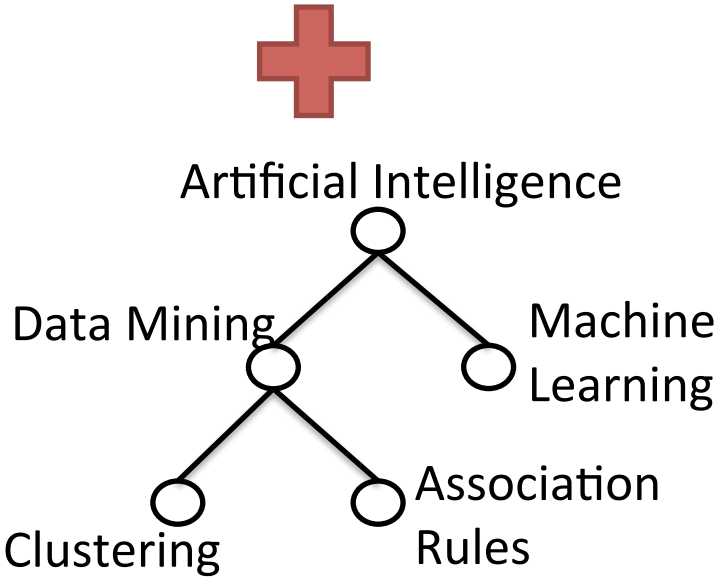
TOSN (2015)

Bibtex http://doi.acm.org/10.1145/2661639

816

A Framework of Mining Trajectories from Untrustworthy Data in Cyber-P

List of publications



Knowledge bases

Infer

Jiawei Han (韩家炜)

H-Index: 133 | **#Paper:** 818 | **#Citation:** 111164

Department of Computer Science, University of Illinois at Urbana-

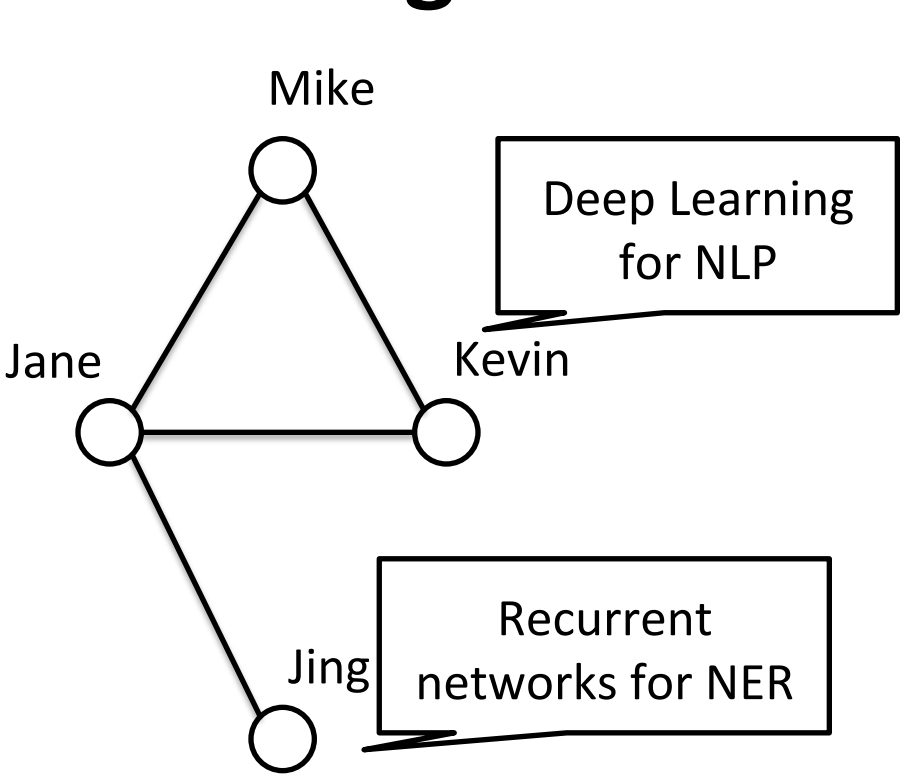
Professor

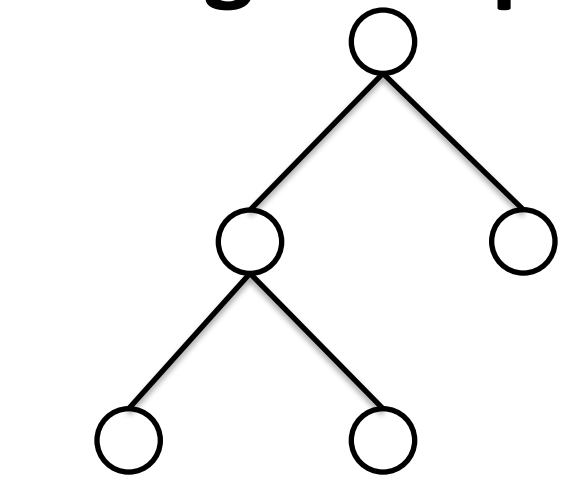
Similar

Data Mining | Information Extraction | Data Analysis | Machine Learning

Research interests

Problem:
Learning Social Knowledge Graphs

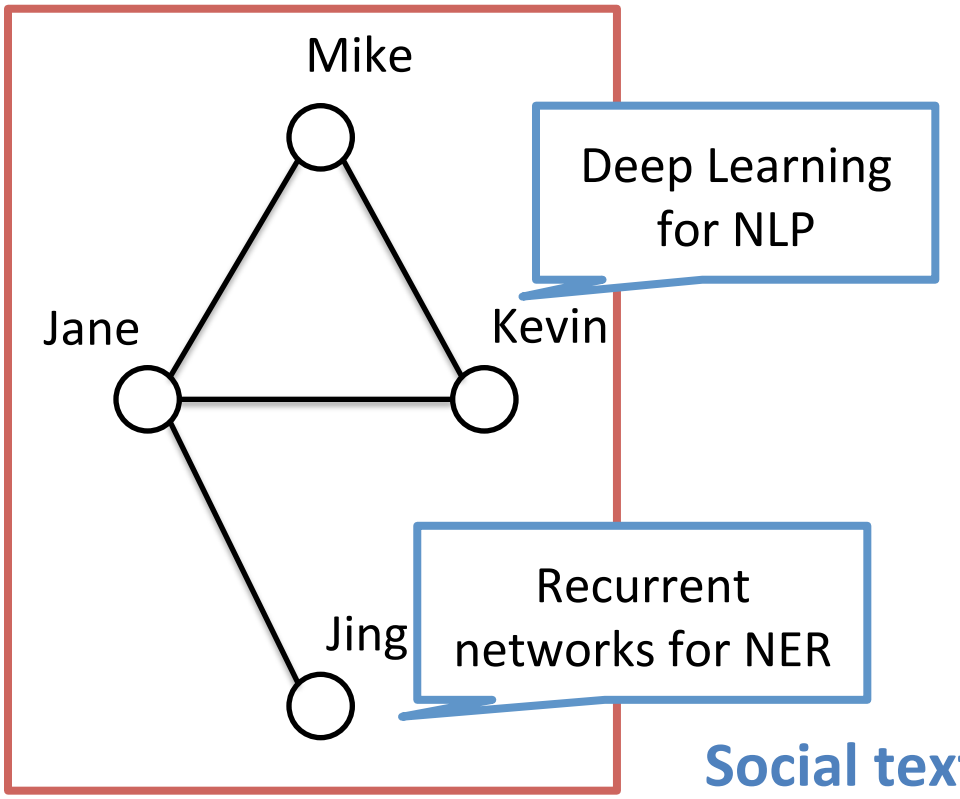




Deep Learning

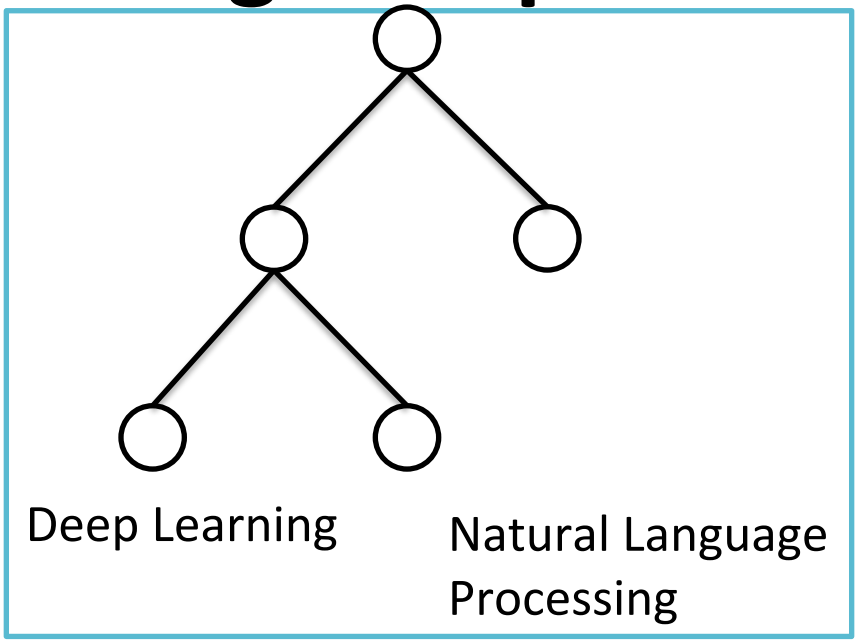
Natural Language Processing

Problem:
Learning Social Knowledge Graphs



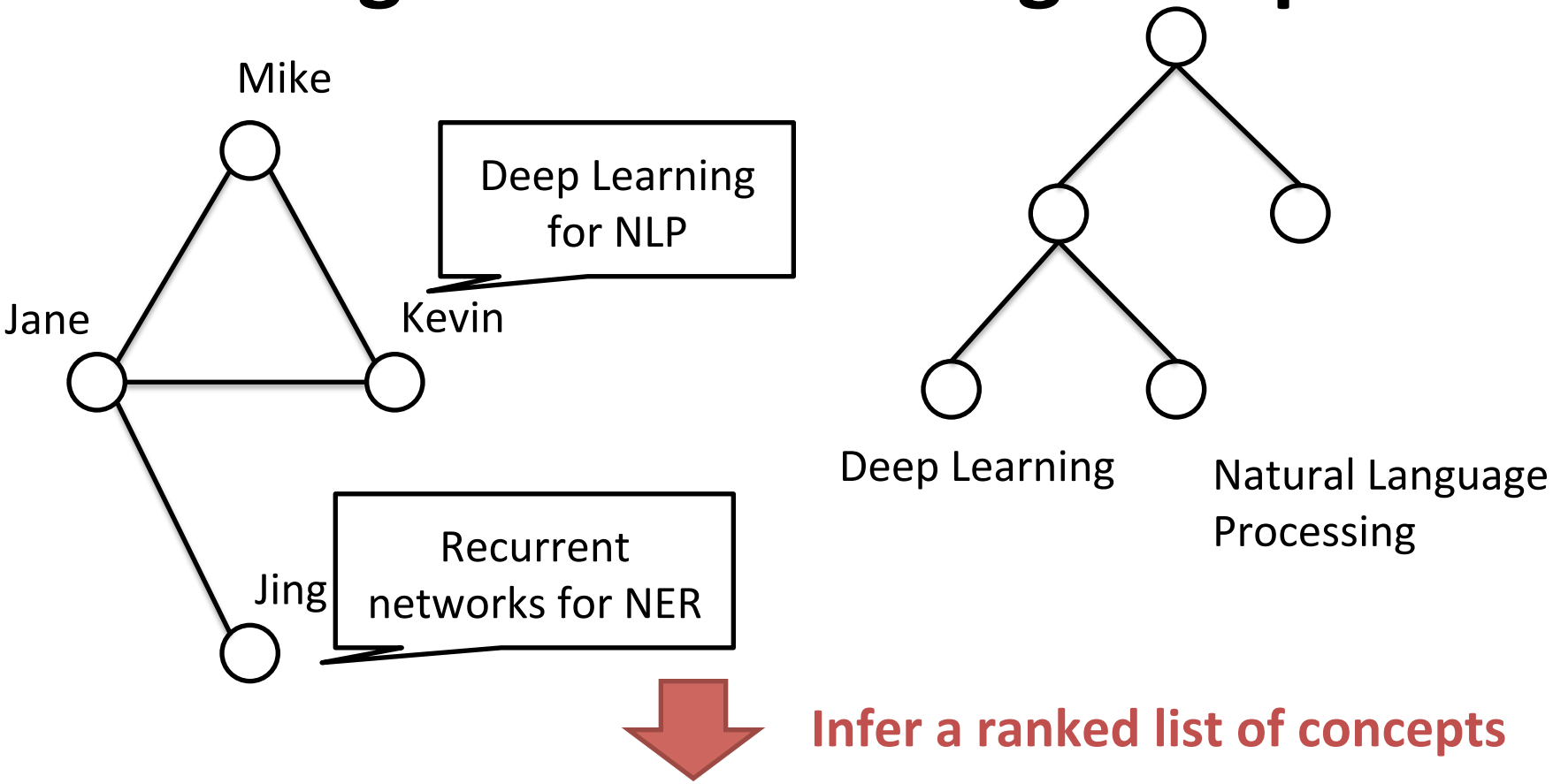
Social network structure

Social text



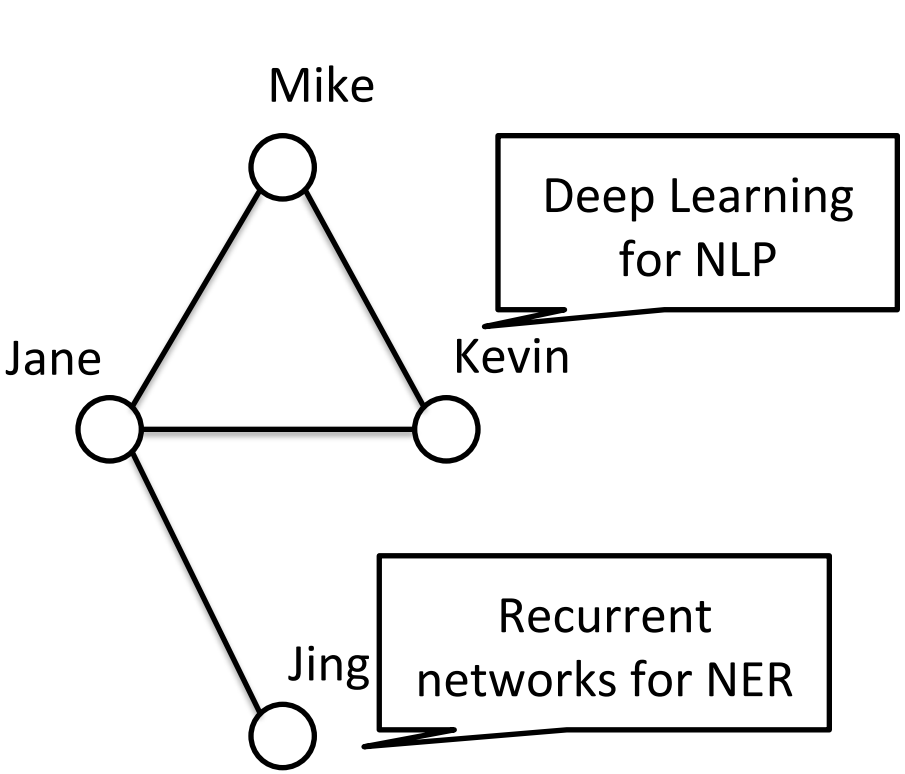
Knowledge base

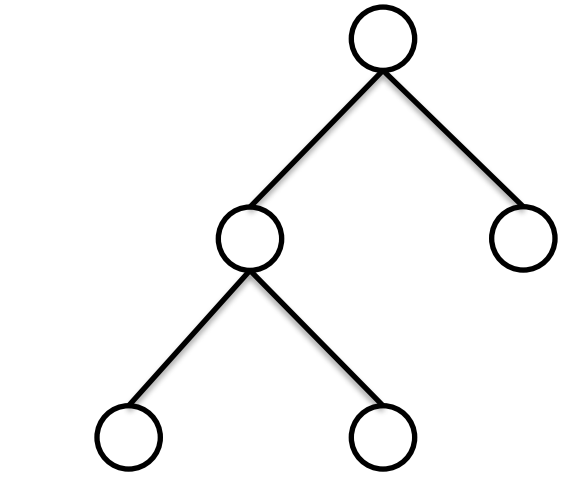
Problem:
Learning Social Knowledge Graphs



Kevin: Deep Learning, Natural Language Processing
Jing: Recurrent Networks, Named Entity Recognition

Challenges



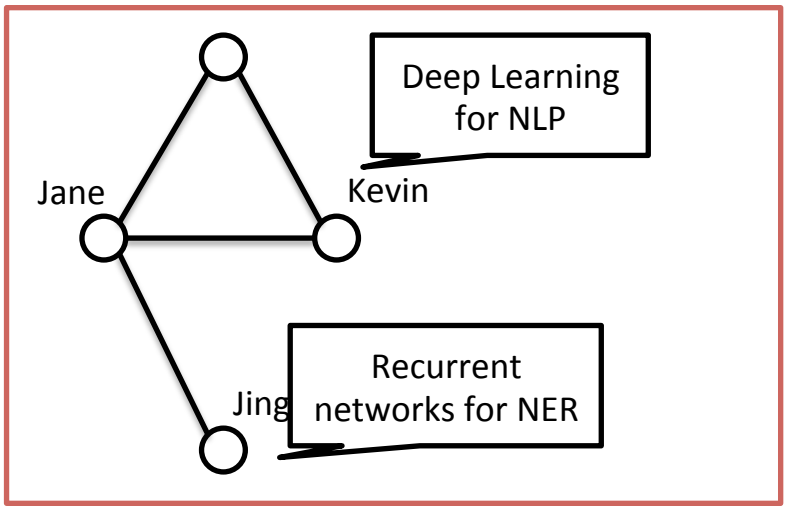


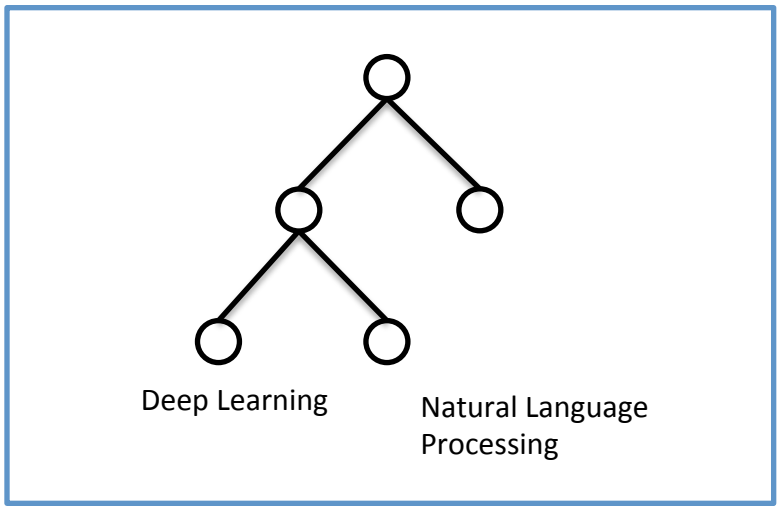
Two modalities – users and concepts

How to leverage information from both modalities?

How to connect these two modalities?

Approach





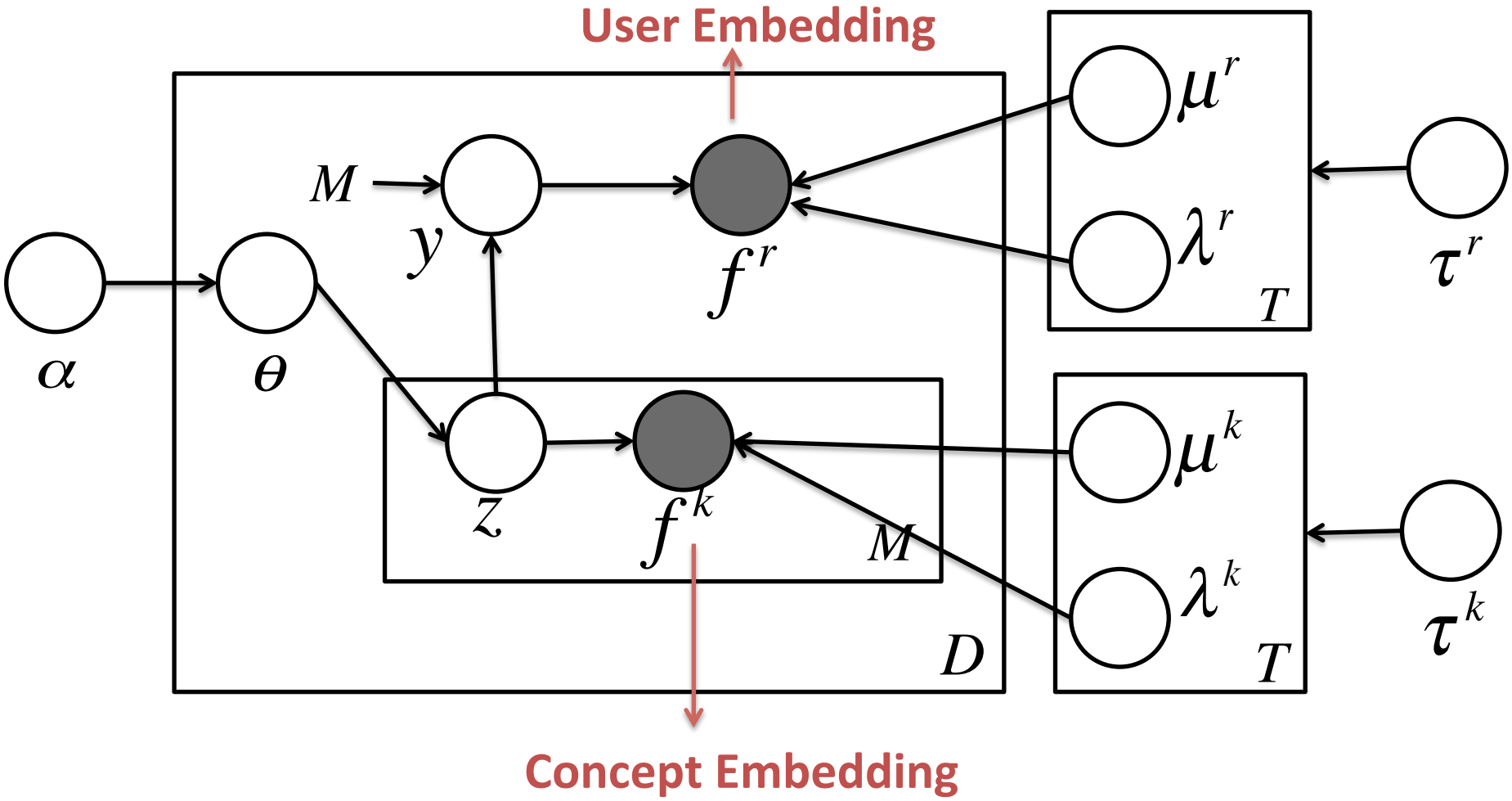
Learn user embeddings

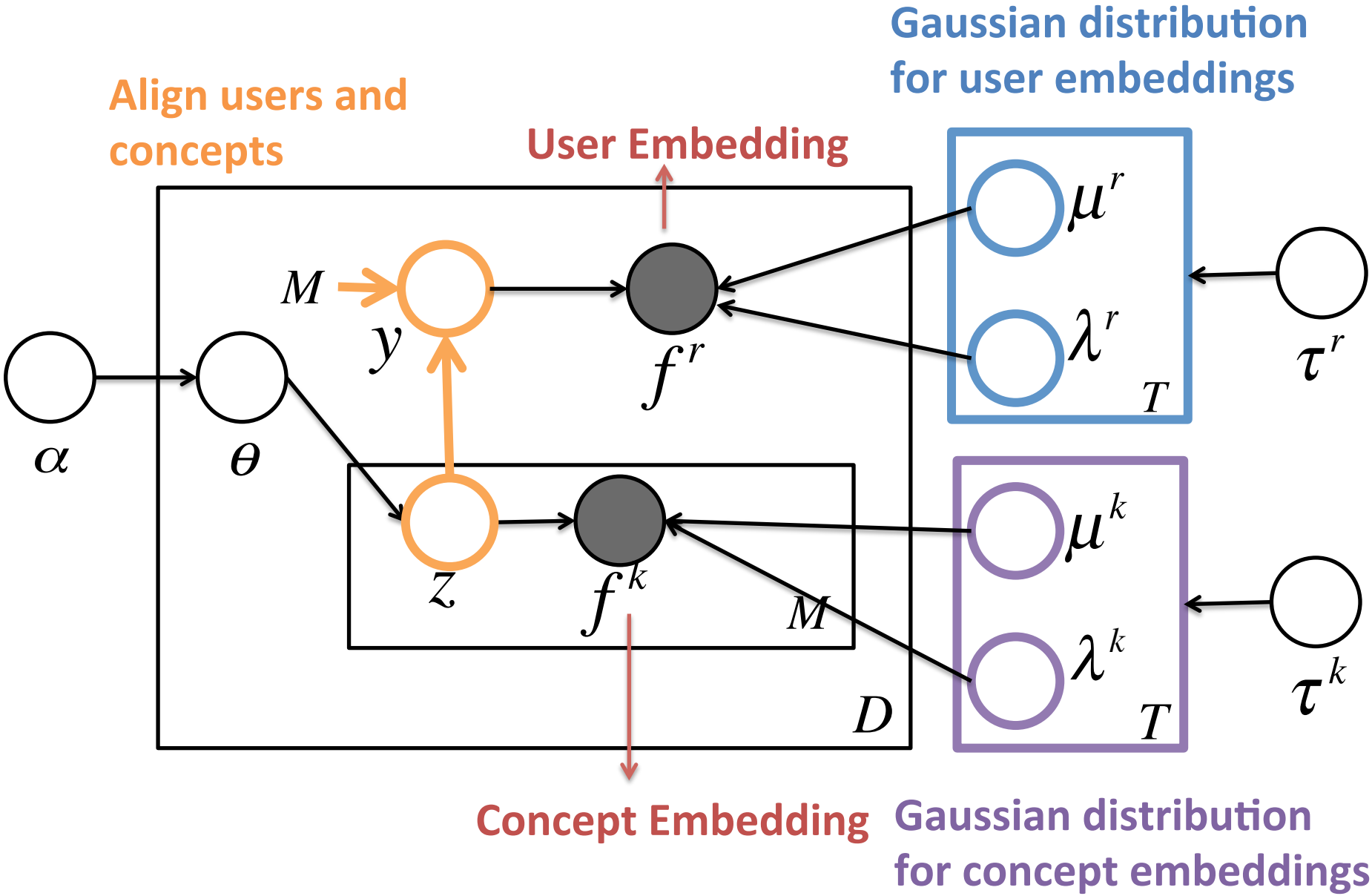
Learn concept embeddings

Model

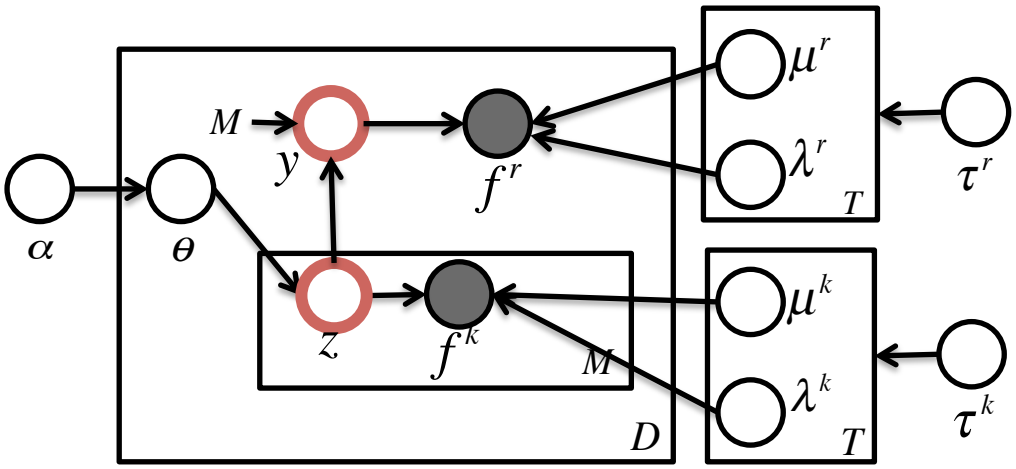
Social KG

Model


User Embedding
μ^r
λ^r
τ^r
T
M
y
f^r
α
θ
z
f^k
M
μ^k
λ^k
τ^k
T
D
Concept Embedding


Gaussian distribution
for user embeddings
Align users and
concepts
User Embedding
μ^r
λ^r
T
τ^r
M
y
f^r
α
θ
z
f^k
M
μ^k
λ^k
T
τ^k
D
Concept Embedding
Gaussian distribution
for concept embeddings

Inference and Learning

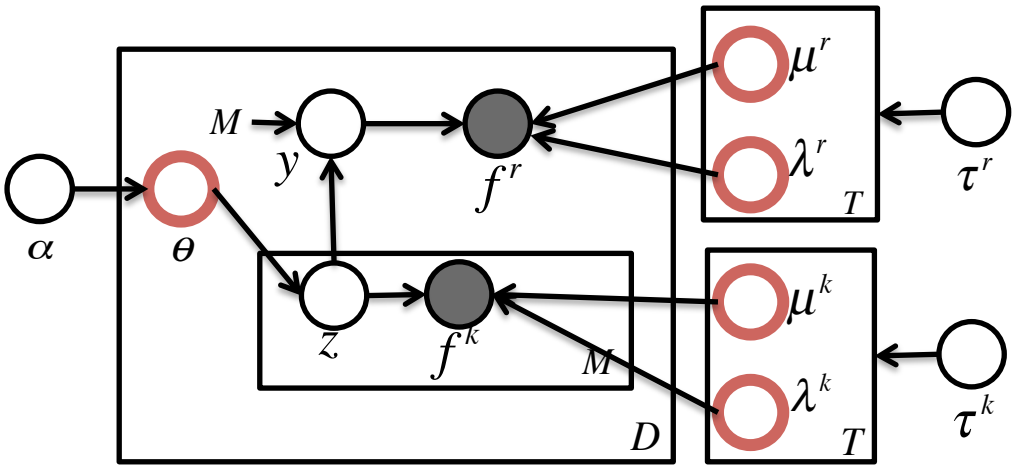


Collapsed Gibbs sampling

Iterate between:

1. **Sample latent variables**

Inference and Learning

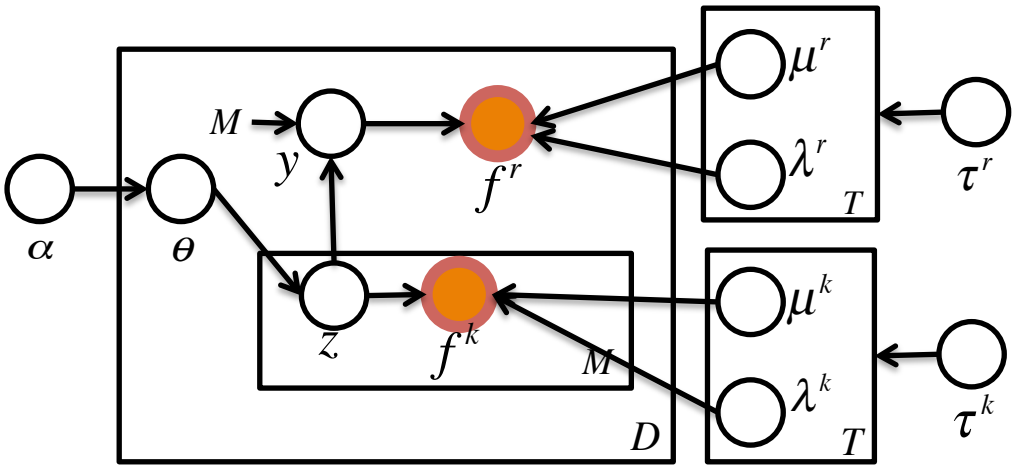


Collapsed Gibbs sampling

Iterate between:

1. Sample latent variables
2. **Update parameters**

Inference and Learning



Collapsed Gibbs sampling

Iterate between:

1. Sample latent variables
2. Update parameters
3. **Update embeddings**

AMiner Research Interest Dataset

- 644,985 researchers
- Terms in these researchers' publications
  - Filtered with Wikipedia
- Evaluation
  - Homepage matching
    - 1,874 researchers
    - Using homepages as ground truth
  - LinkedIn matching
    - 113 researchers
    - Using LinkedIn skills as ground truth

Code and data available:

https://github.com/kimiyoung/genvector

Homepage Matching

Using homepages as ground truth.

Method	Precision@5
GenVector	**78.1003%**
GenVector-E	77.8548%
Sys-Base	73.8189%
Author-Topic	74.4397%
NTN	65.8911%
CountKG	54.4823%

GenVector	Our model
GenVector-E	Our model w/o embedding update
Sys-Base	AMiner baseline: key term extraction
CountKG	Rank by frequency
Author-topic	Classic topic models
NTN	Neural tensor networks

LinkedIn Matching

Using LinkedIn skills as ground truth.

Method	Precision@5
GenVector	**50.4424%**
GenVector-E	49.9145%
Author-Topic	47.6106%
NTN	42.0512%
CountKG	46.8376%

GenVector Our model
GenVector-E Our model w/o embedding update

CountKG Rank by frequency
Author-topic Classic topic models
NTN Neural tensor networks

Error Rate of Irrelevant Cases

Manually label terms that are clearly NOT research interests, e.g., challenging problem.

Method	Error rate
GenVector	**1.2%**
Sys-Base	18.8%
Author-Topic	1.6%
NTN	7.2%

GenVector Our model
Sys-Base AMiner baseline: key term extraction
Author-topic Classic topic models
NTN Neural tensor networks

Qualitative Study: Top Concepts within Topics

GenVector	Author-Topic
Query expansion	Speech recognition
Concept mining	Natural language
Language modeling	***Integrated circuits**
Information extraction	Document retrieval
Knowledge extraction	Language models
Entity linking	Language model
Language models	***Microphone array**
Named entity recognition	Computational linguistics
Document clustering	***Semidefinite programming**
Latent semantic indexing	Active learning

Qualitative Study: Top Concepts within Topics

GenVector	Author-Topic
Image processing	Face recognition
Face recognition	***Food intake**
Feature extraction	Face detection
Computer vision	Image recognition
Image segmentation	***Atmospheric chemistry**
Image analysis	Feature extraction
Feature detection	Statistical learning
Digital image processing	Discriminant analysis
Machine learning algorithms	Object tracking
Machine vision	***Human factors**

Qualitative Study: Research Interests

GenVector

Feature extraction
Image segmentation
Image matching
Image classification
Face recognition

Sys-Base

Face recognition
Face image
***Novel approach**
***Line drawing**
Discriminant analysis

Qualitative Study: Research Interests

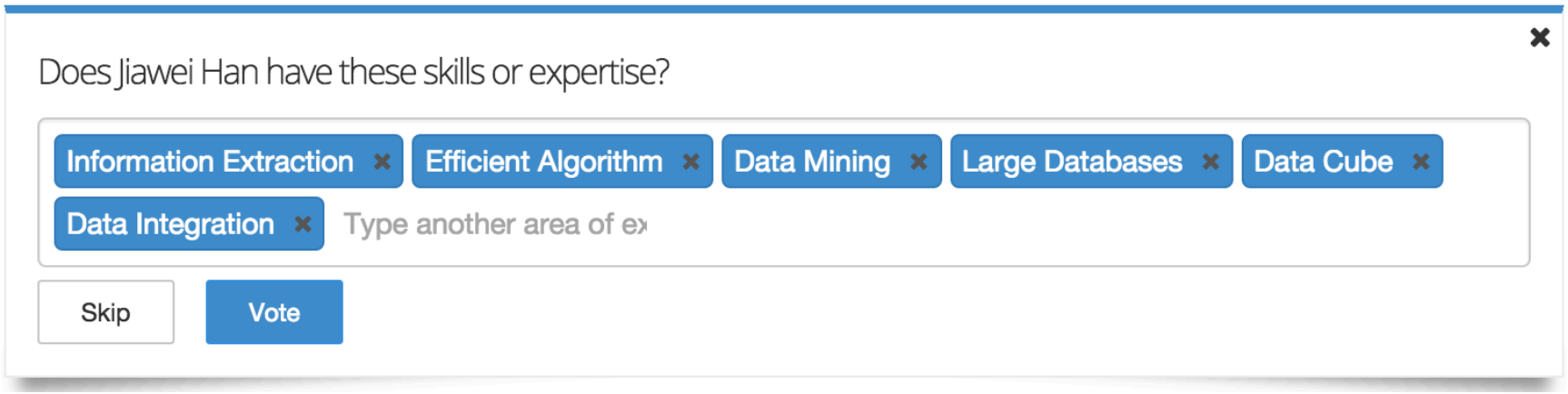
GenVector

Unsupervised learning
Feature learning
Bayesian networks
Reinforcement learning
Dimensionality reduction

Sys-Base

***Challenging problem**
Reinforcement learning
***Autonomous helicopter**
***Autonomous helicopter flight**
Near-optimal planning

Online Test

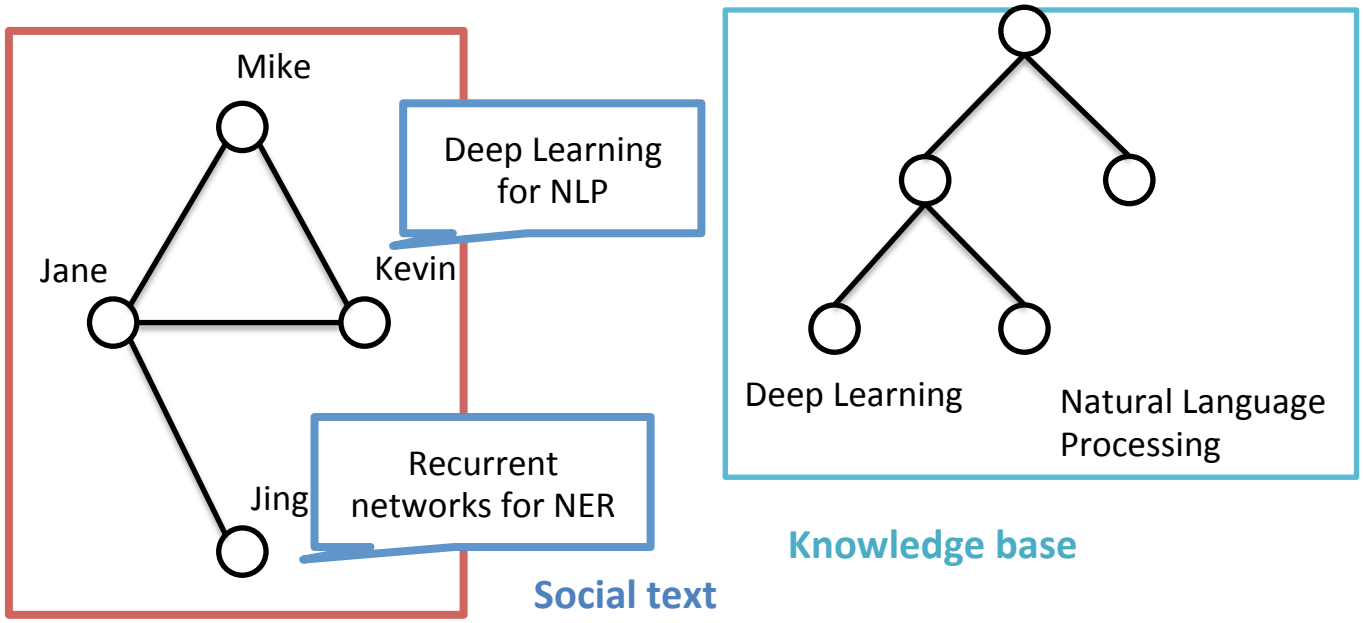


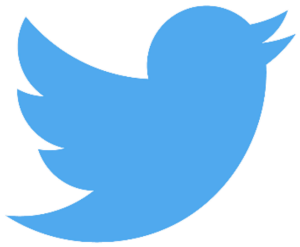
A/B test with live users

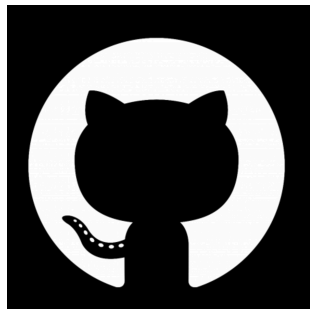
- Mixing the results with Sys-Base

Method	Error rate
GenVector	**3.33%**
Sys-Base	10.00%

Other Social Networks?







Conclusion

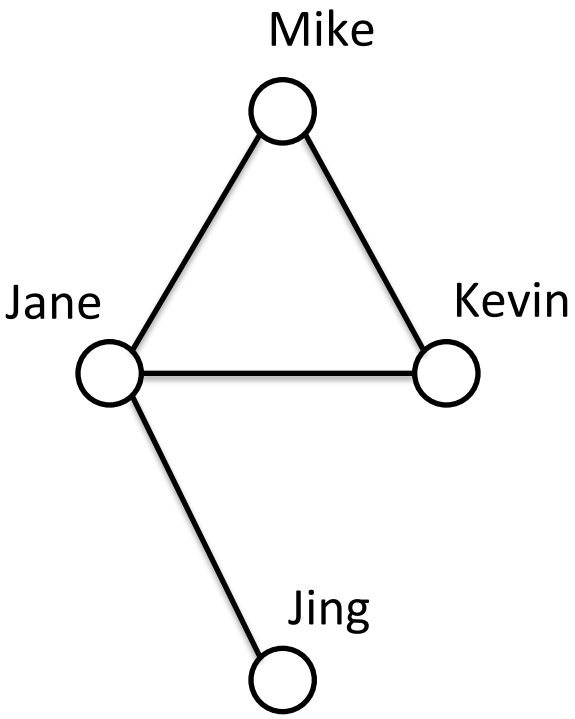
- Study a novel problem
  - Learning social knowledge graphs
- Propose a model
  - Multi-modal Bayesian embedding
  - Integrate embeddings into graphical models
- AMiner research interest dataset
  - 644,985 researchers
  - Homepage and LinkedIn matching as ground truth
- Online deployment on AMiner

Thanks!

Code and data:

https://github.com/kimiyoung/genvector

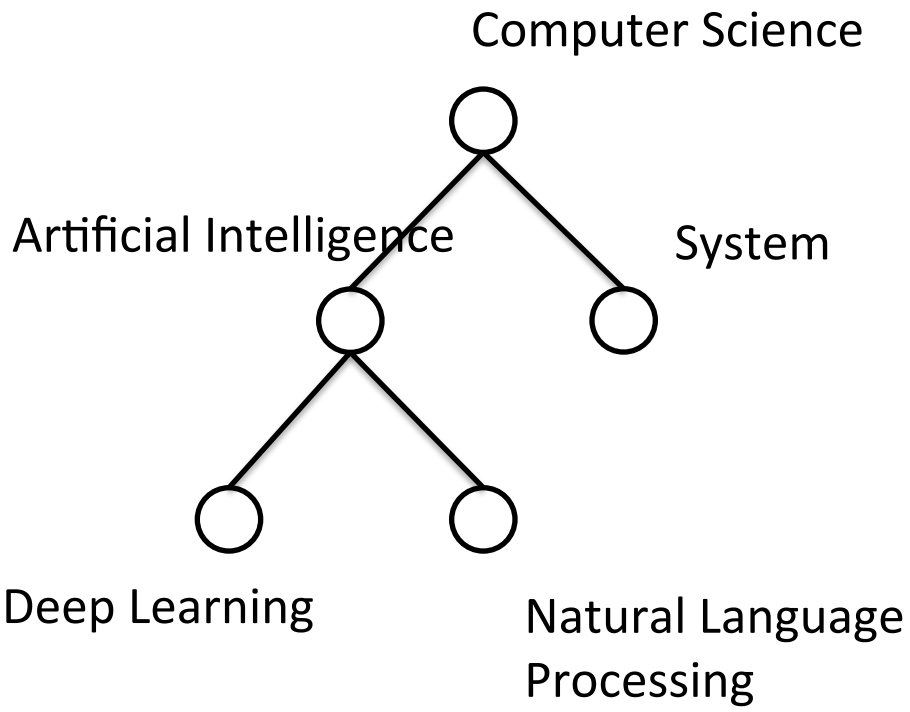
Social Networks



AMiner, Facebook, Twitter…

Huge amounts of information

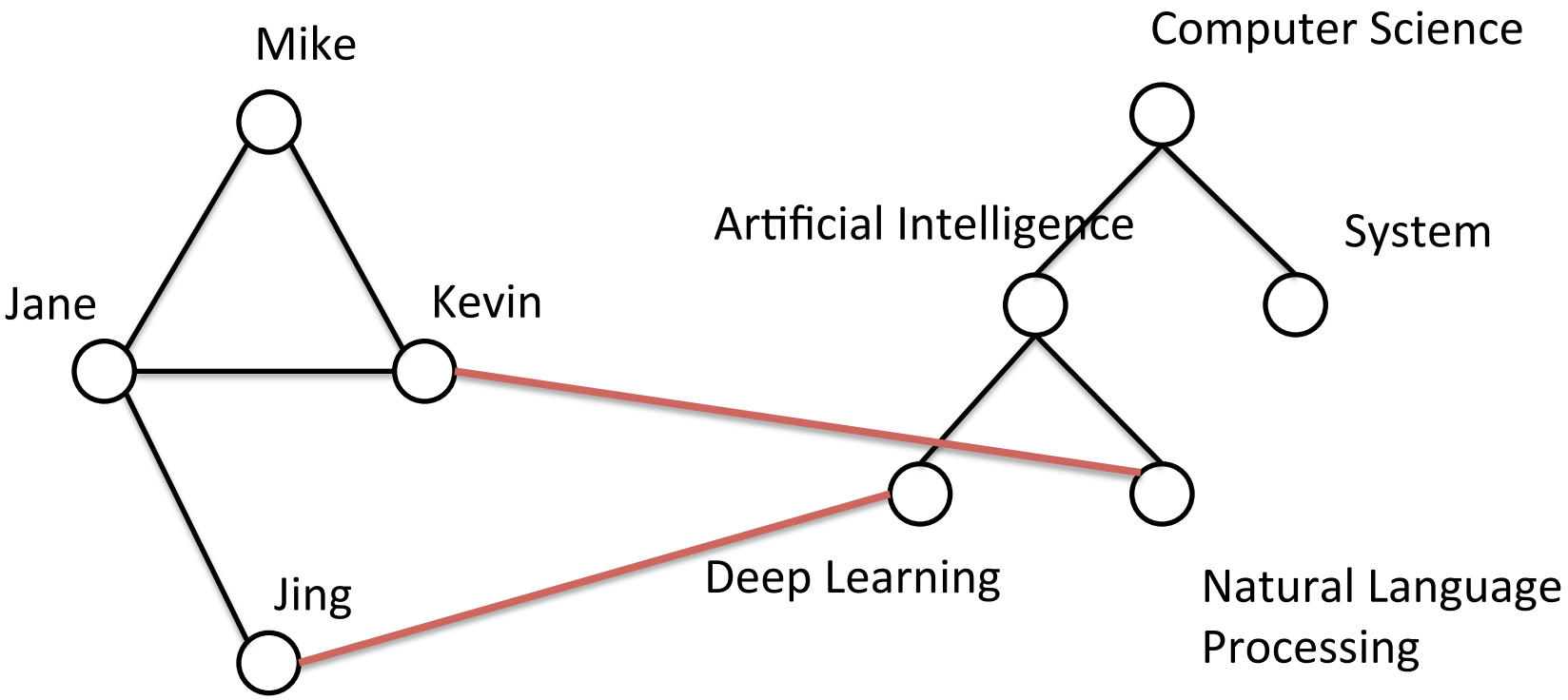
Knowledge Bases



Wikipedia, Freebase, Yago, NELL…

Huge amounts of knowledge

Bridge the Gap

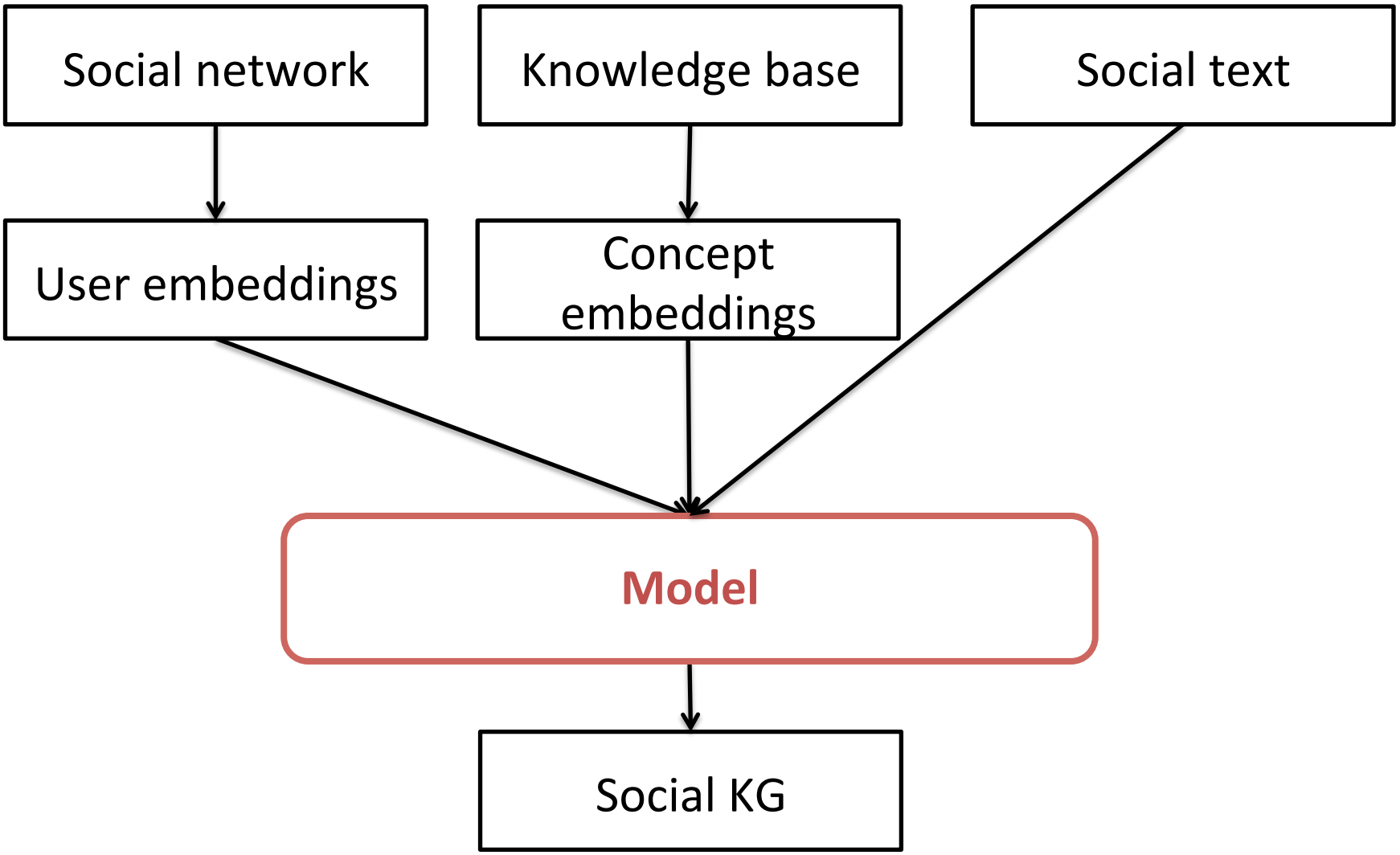


Better user understanding

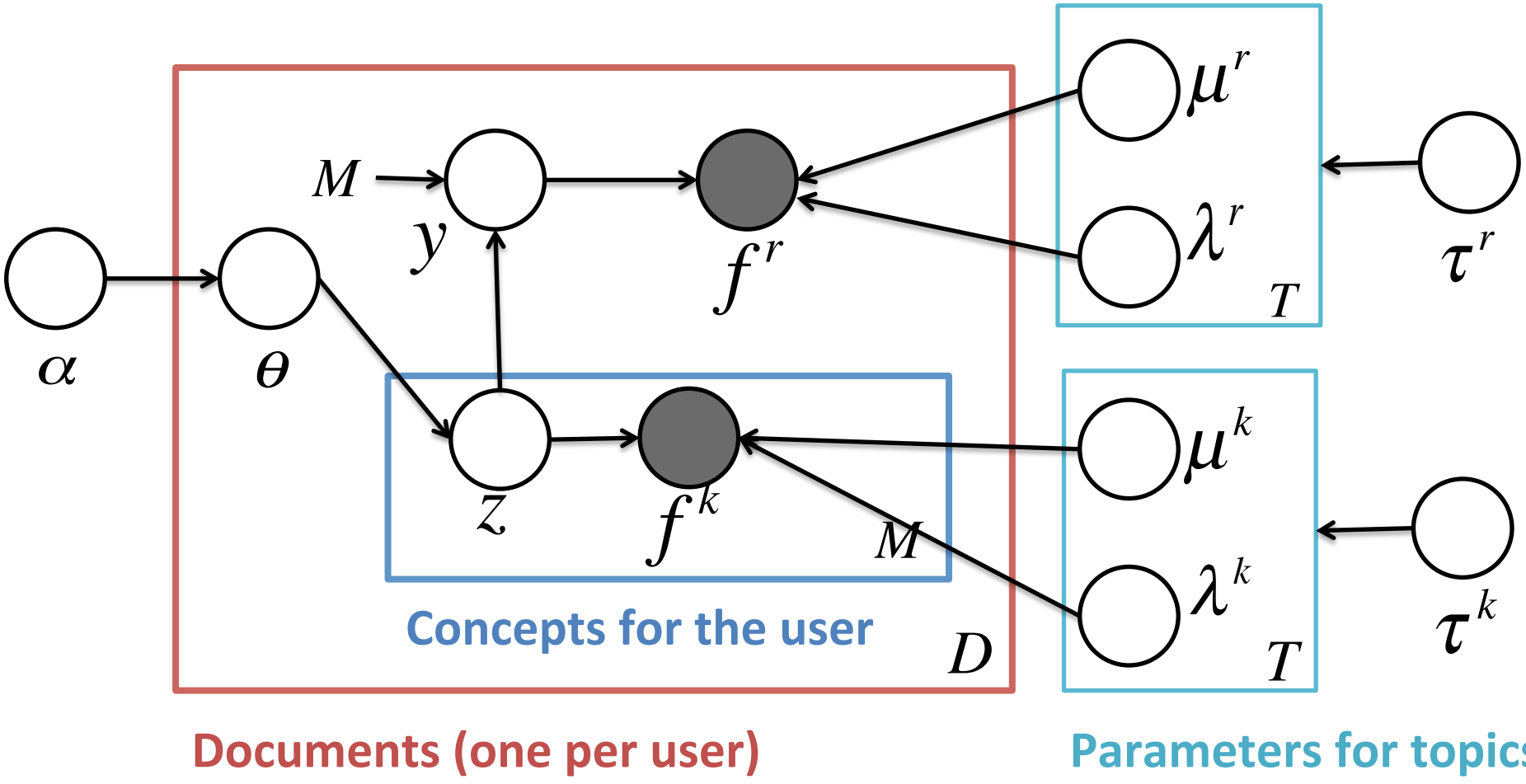
e.g. mine research interests on AMiner

Approach

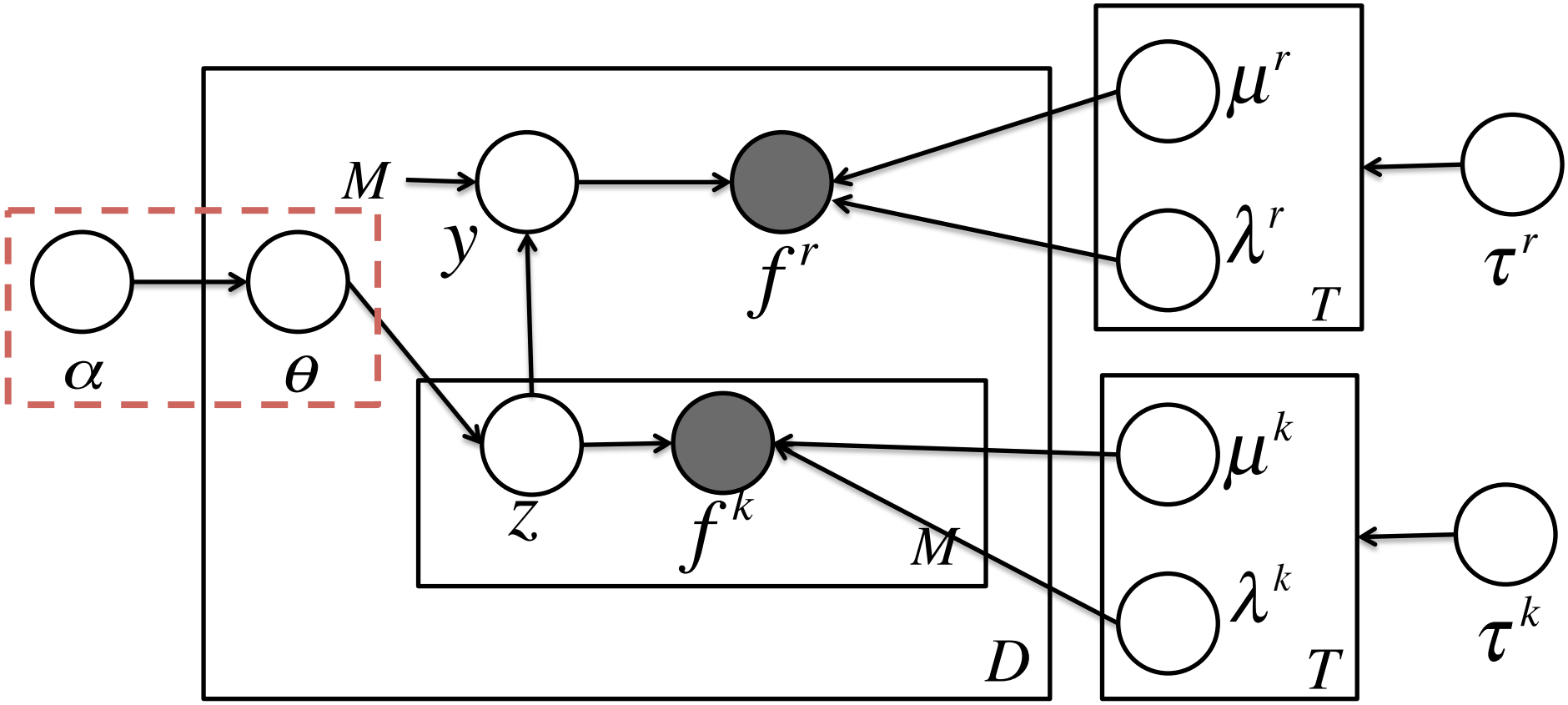
Copy picture



Model

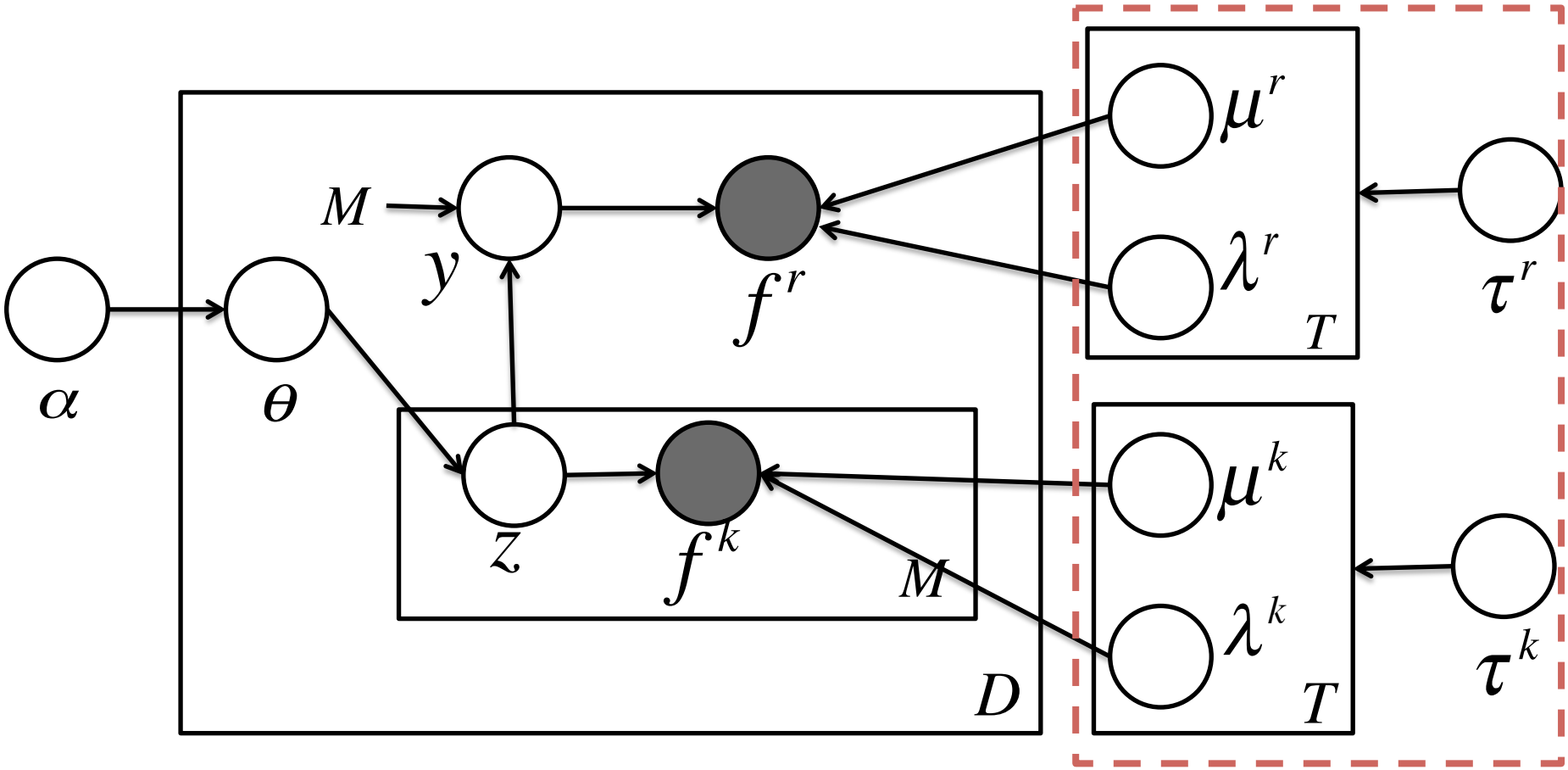

μ^r
λ^r
T
τ^r
M
y
f^r
α
θ
z
f^k
M
μ^k
λ^k
T
τ^k
Concepts for the user
D
Documents (one per user)
Parameters for topics

Model



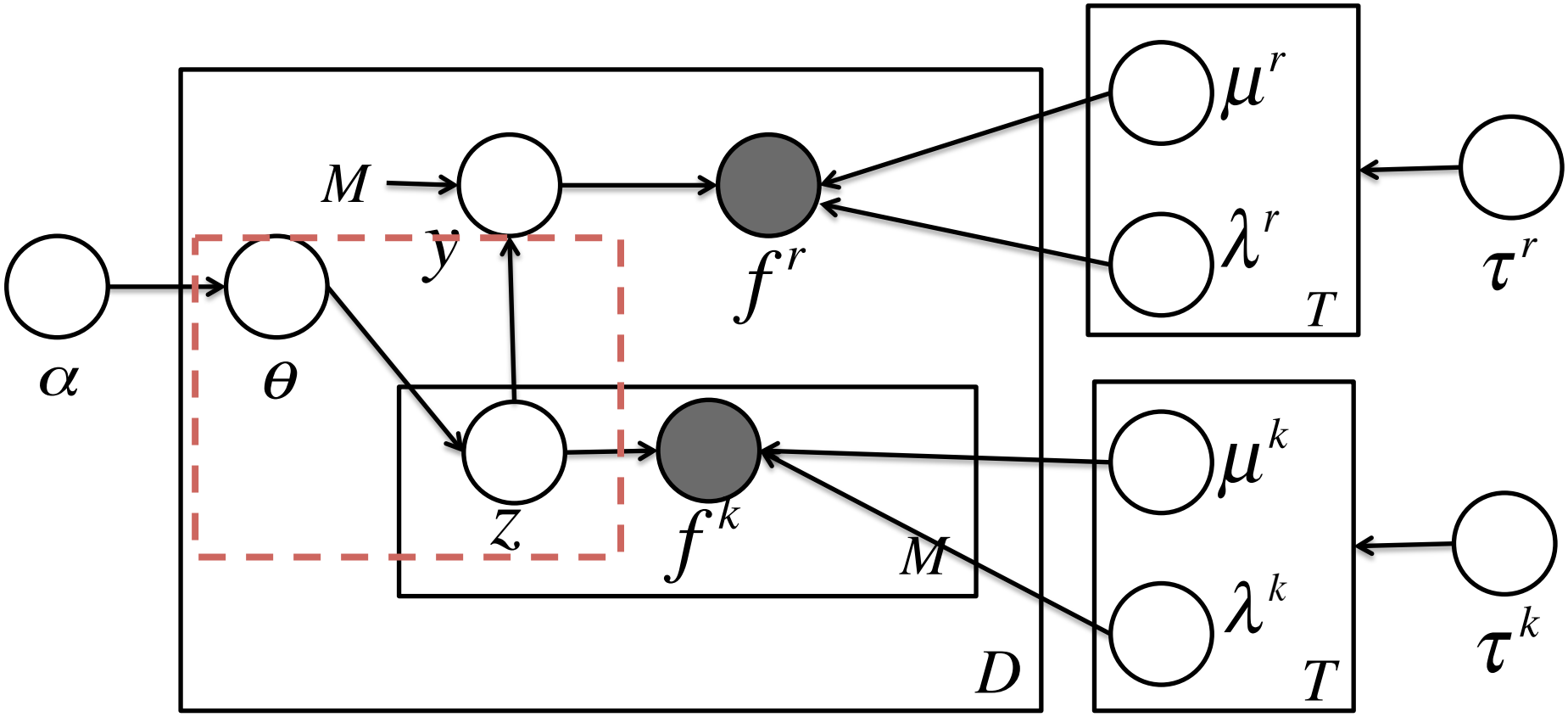
Generate a topic distribution for each document (from a Dirichlet)

Model



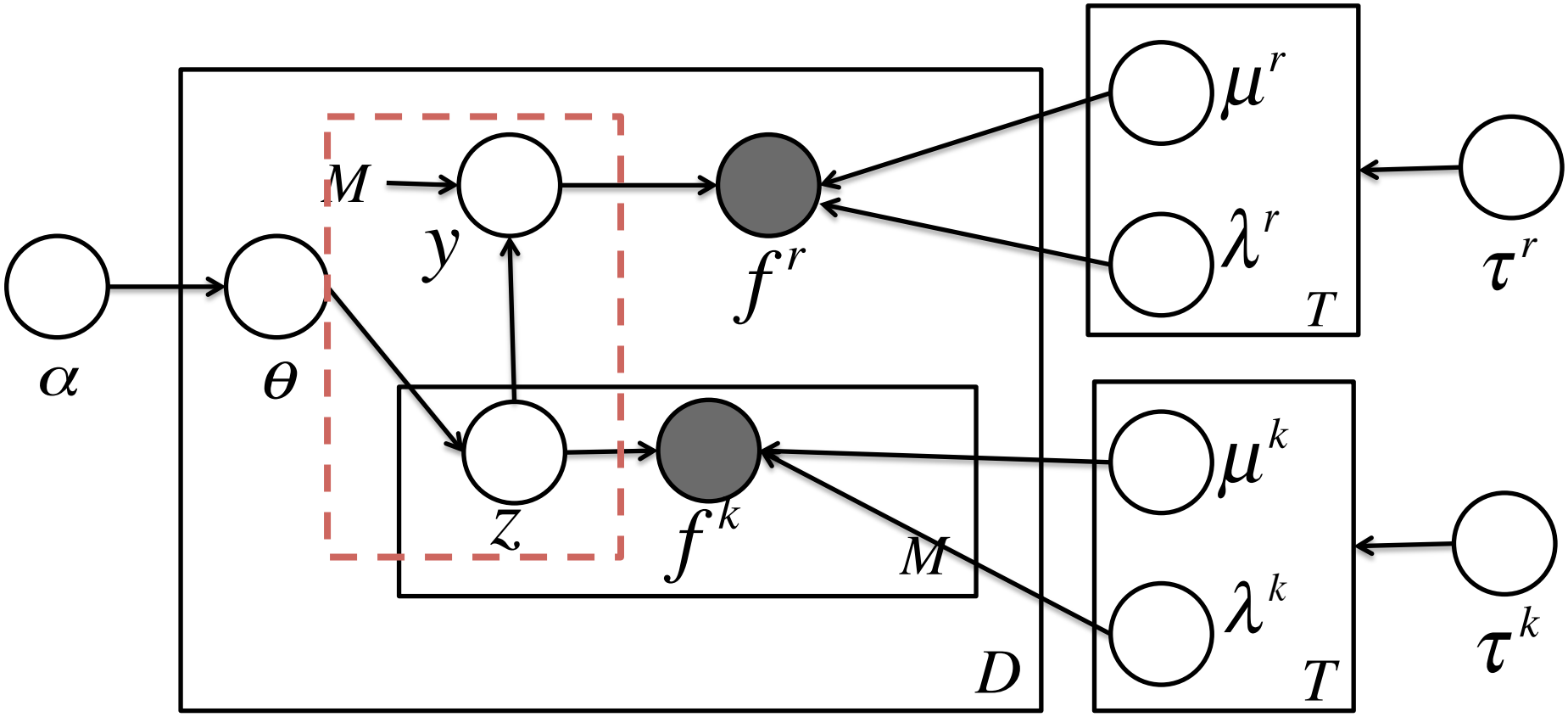
Generate Gaussian distribution for each embedding space (from a Normal Gamma)

Model



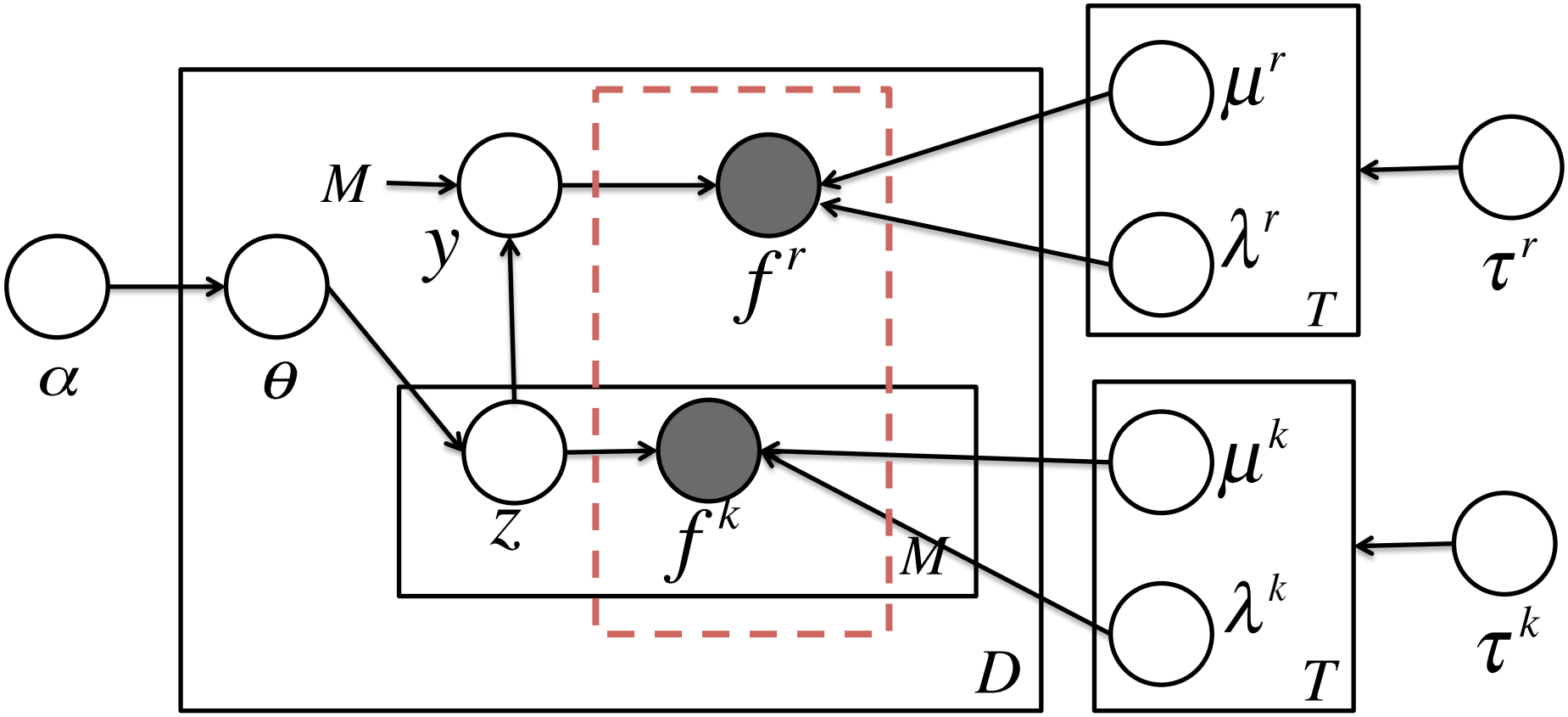
Generate the topic for each concept (from a Multinomial)

Model



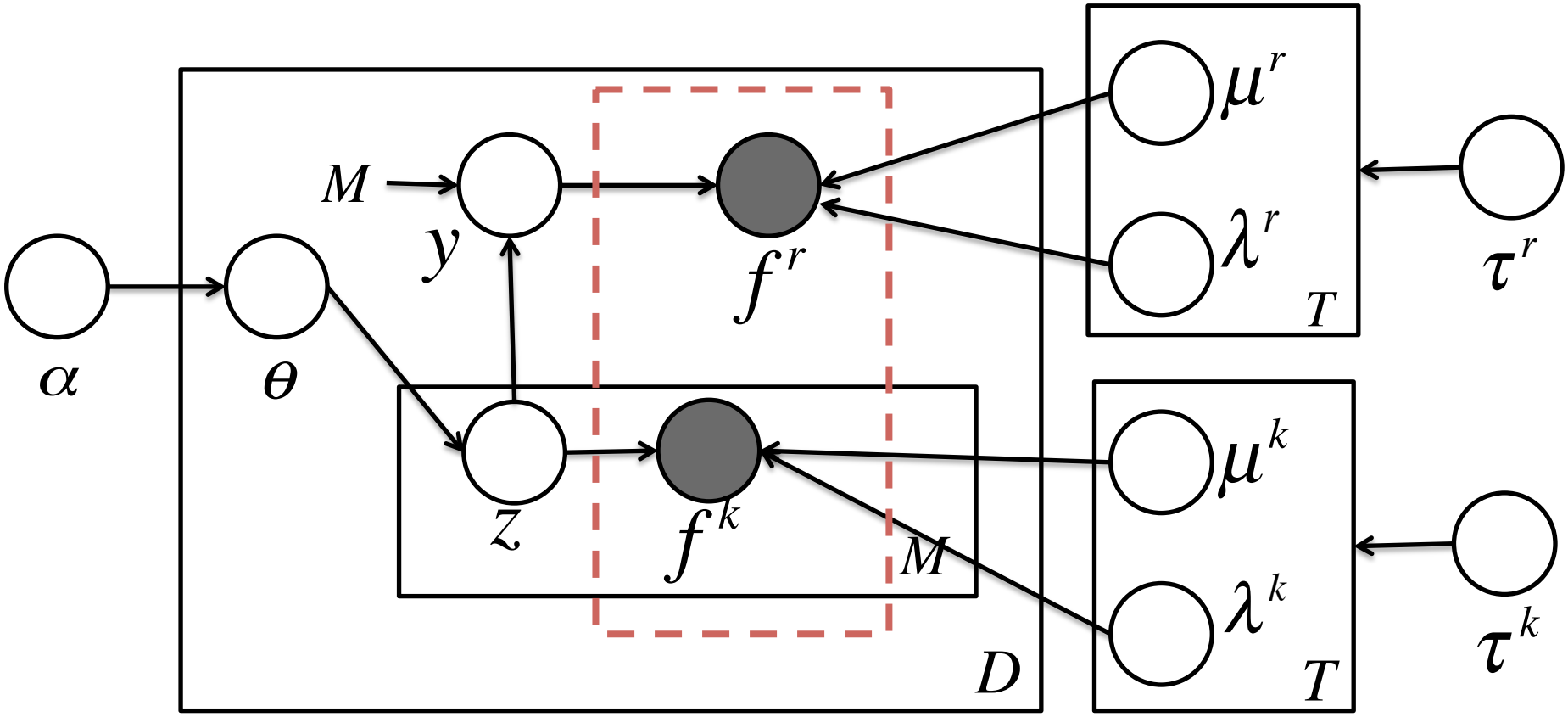
Generate the topic for each user (from a Uniform)

Model



Generate embeddings for users and concepts (from a Gaussian)

Model



General

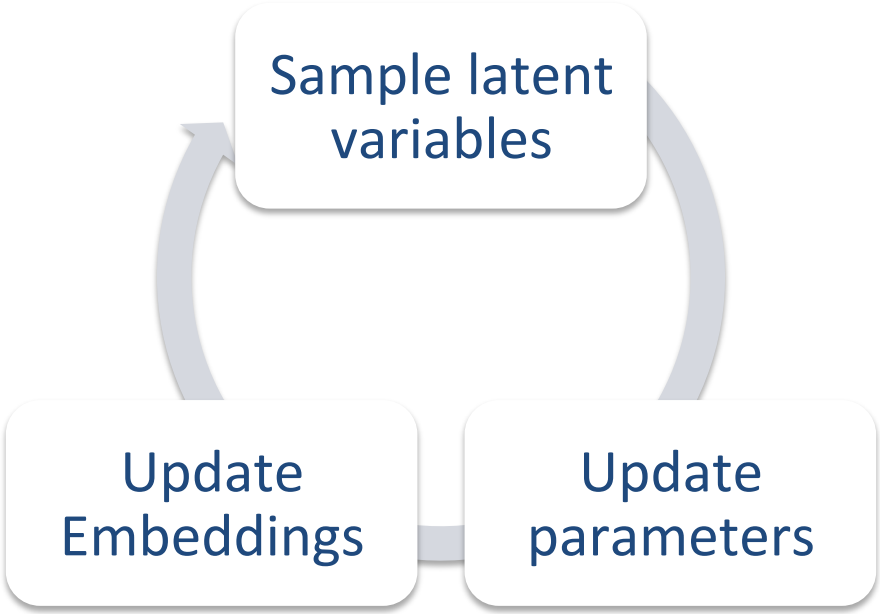
Inference and Learning

Collapsed Gibbs sampling for inference

Add picture

Update the embedding during learning

Different from LDAs with discrete observed variables



Methods for Comparison

Method	Description
GenVector	Our model
GenVector-E	Our model w/o embedding update
Sys-Base	AMiner baseline: key term extraction
CountKG	Rank by frequency
Author-topic	Classic topic models
NTN	Neural tensor networks