Joint Modeling of Topics, Citations, and Topical Authority in Academic Corpora

Jooyeon Kim KAIST
Dongwoo Kim Australian National University
Alice Oh KAIST

Research Overview

Introduction

Michael I. Jordan

Citation Indices

Citations: 122,672
h-index: 138

Reinforcement Learning	Statistical Inference	Image Processing	Information Retrieval
2.76	10.29	0.14	0.10

Measure topical authority of academic researchers

Proposed Model

- Latent Topical-Authority Indexing (LTAI)
  - A Bayesian probabilistic model that discovers topic-specific authority of scholars
  - Uses paper contents and citation networks

- Key Idea

Paper written by authors of high topical authority values → High probability of being cited by papers of similar topic

Topic	RL	Statistical Inference	Image Processing	IR
Authority	2.76	10.29	0.14	0.10

highly cited

less cited

Possible Applications

- Recommending fine-grained authoritative researcher
  - ex) Newly entering graduate student

Interested in Statistical Learning → LTAI recommends Michael I. Jordan

Interested in Natural Language Processing → LTAI recommends Christopher D. Manning

- Finding academic papers from given topical interests
- Discovering research topics from academic corpus

LTAI: Input & Output

- Input
  - Paper content (text)
  - Authorship (author-paper link)
  - Citation network (paper-paper link)
- Output
  - Academic Topics (global)
  - Topic distribution (per paper)
  - Topical authority (per author)

Topic No.	Hand-Tagged Labels	Top Frequent Words
Topic 1	Information Retrieval	information user document text retrieval web system content collection using
Topic 2	Image Processing	image object visual motion recognition model feature shape vision face
Topic 3	Distributed System	distributed system protocol group failure message fault recovery process asynchronous
Topic 4	Database	query database data transaction system rule view processing paper relational
…	…	…

Global Academic Topics

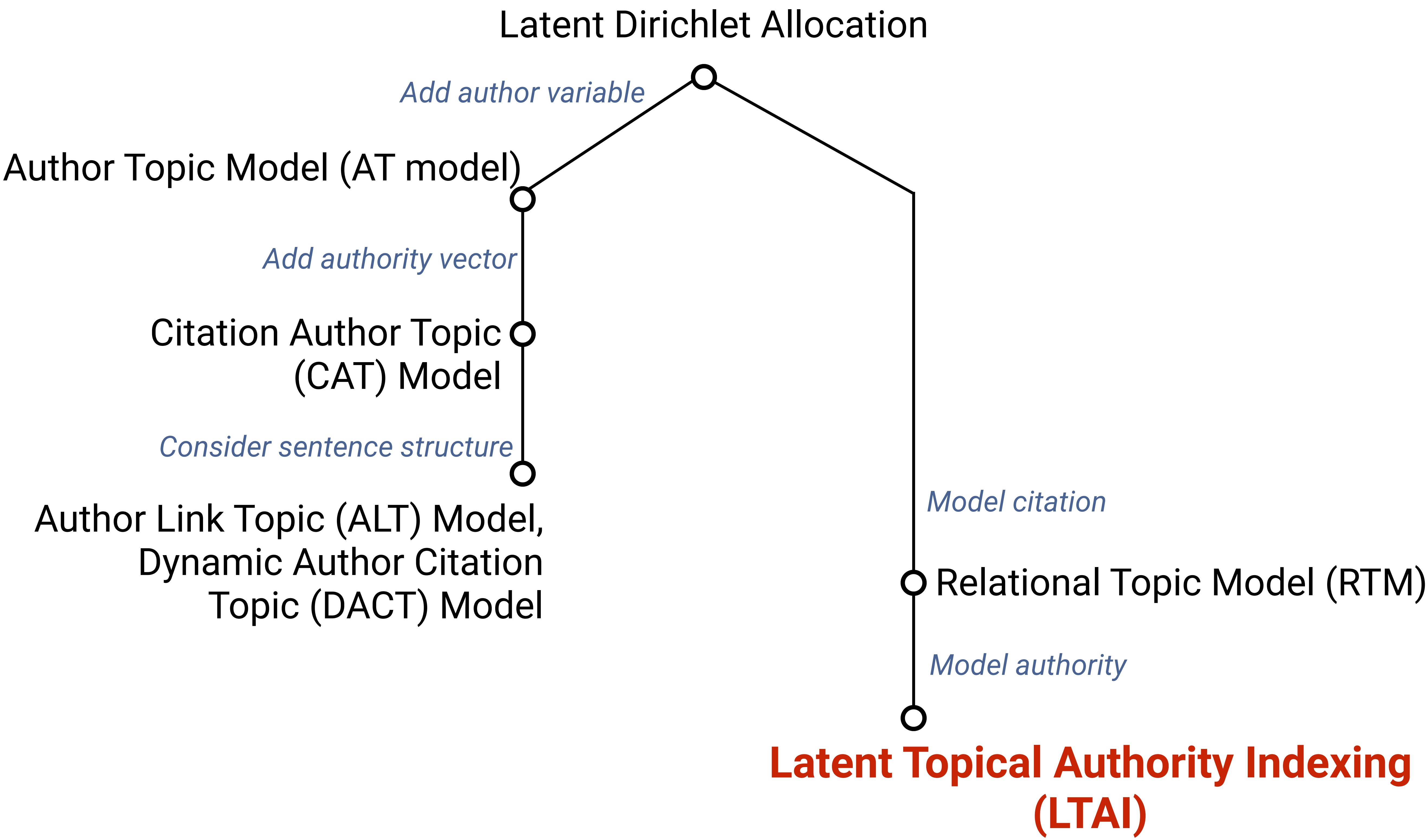
Topic No.	Topic 1	Topic 2	Topic 3	Topic 4	…
Authority	3.82	9.58	1.27	0.32	…

Per-Author Topical Authority

Topic No.	Topic 1	Topic 2	Topic 3	Topic 4	…
Weight	0.40	0.12	0.01	0.08	…

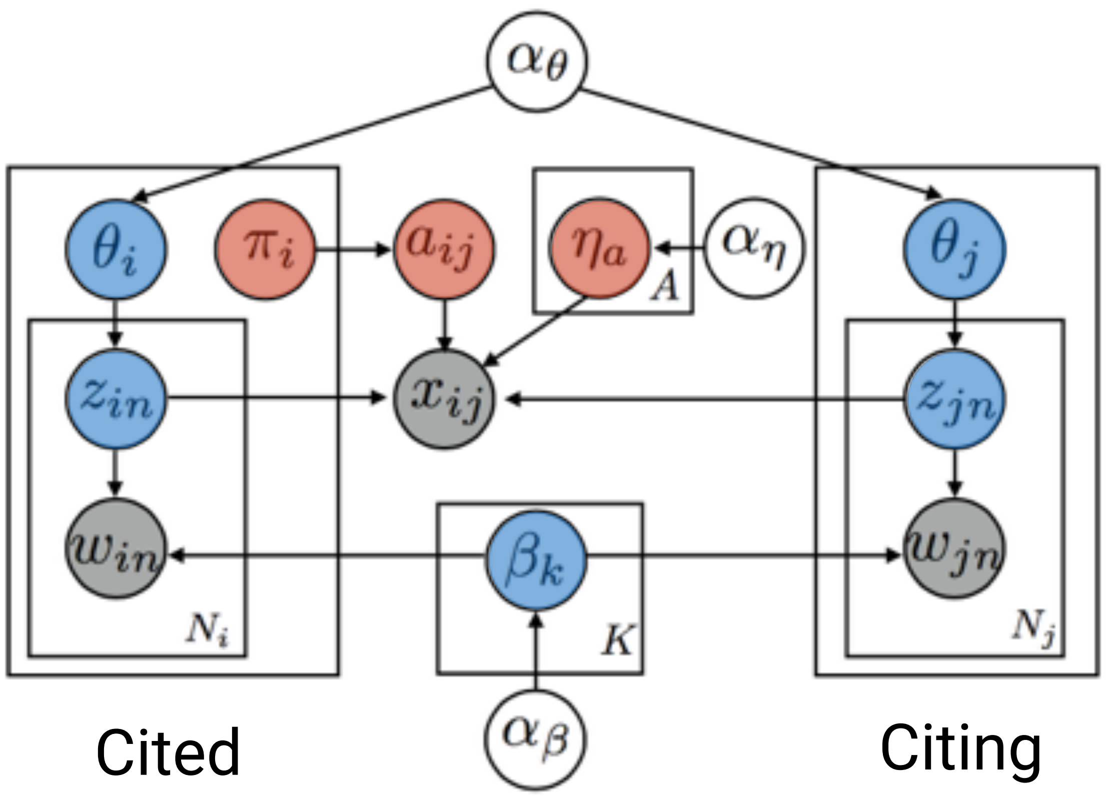
Per-Paper Topic Distribution

Related Work



Model Description

Model Description: Full Graphical Model



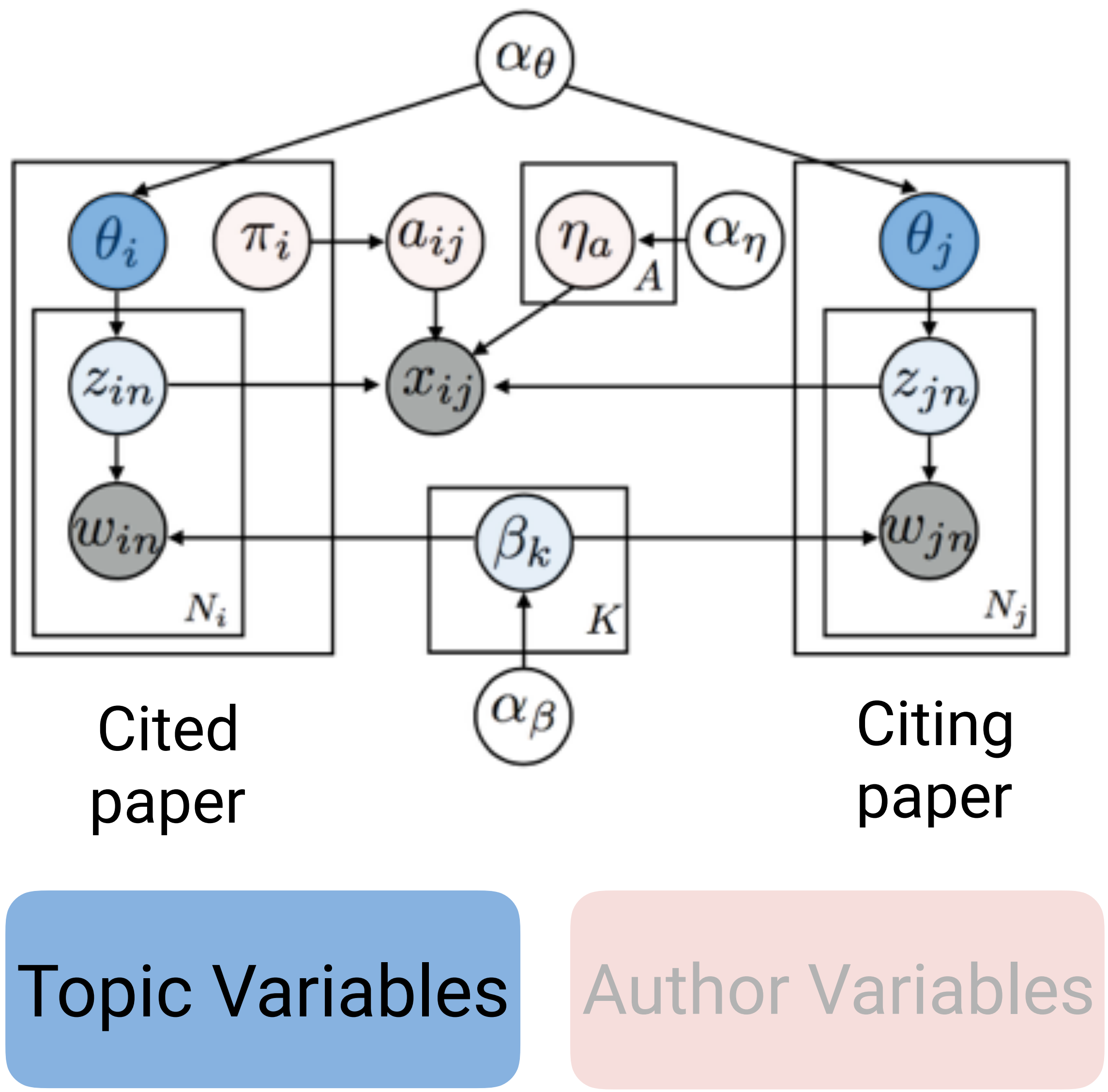
Cited paper

Citing paper

Topic Variables

Author Variables

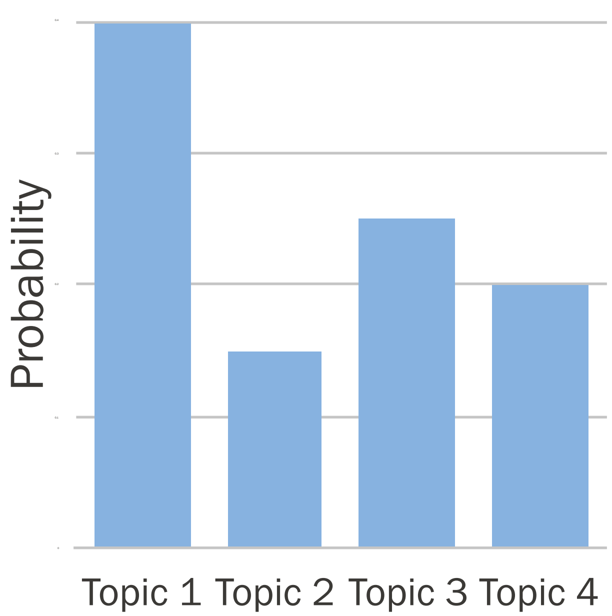
Model Description: Topic Variables



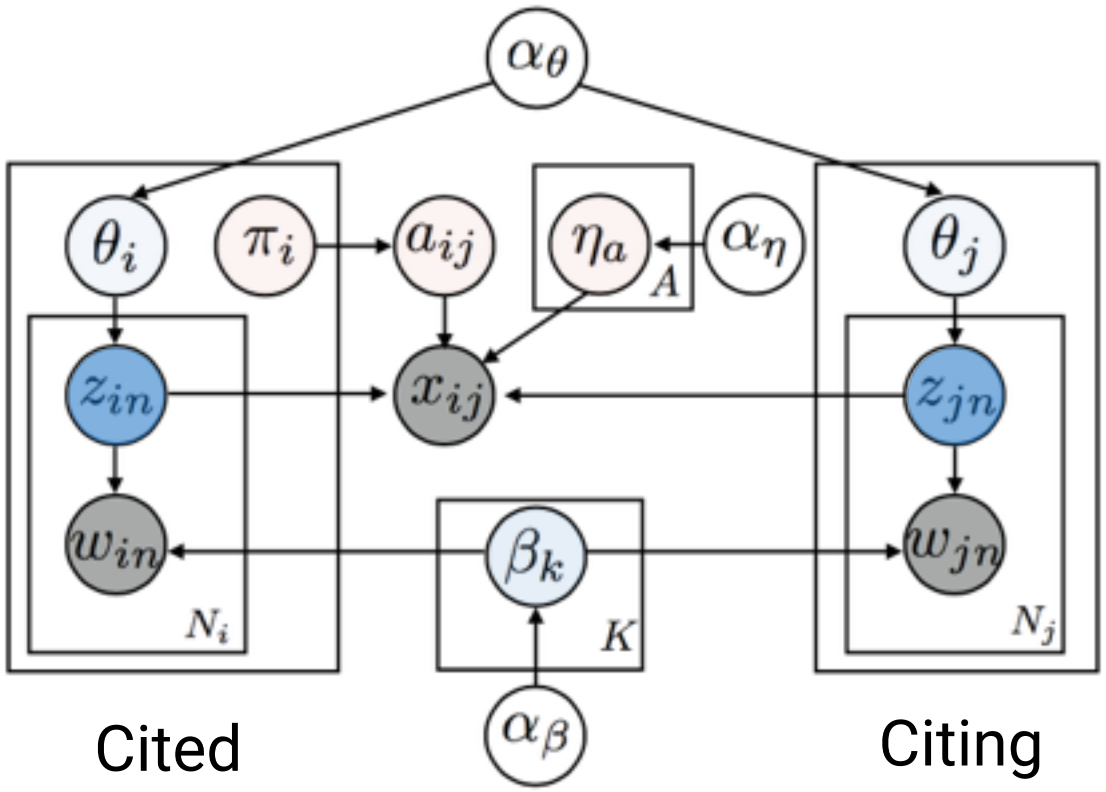
θ **Document-Topic distribution**
K (# topics)-dimensional vector

z **Per-word topic indicator variable**

β_k **Topic-Word distribution**
V (vocab. size)-dimensional vector



Model Description: Topic Variables



Cited paper

Citing paper

Topic Variables

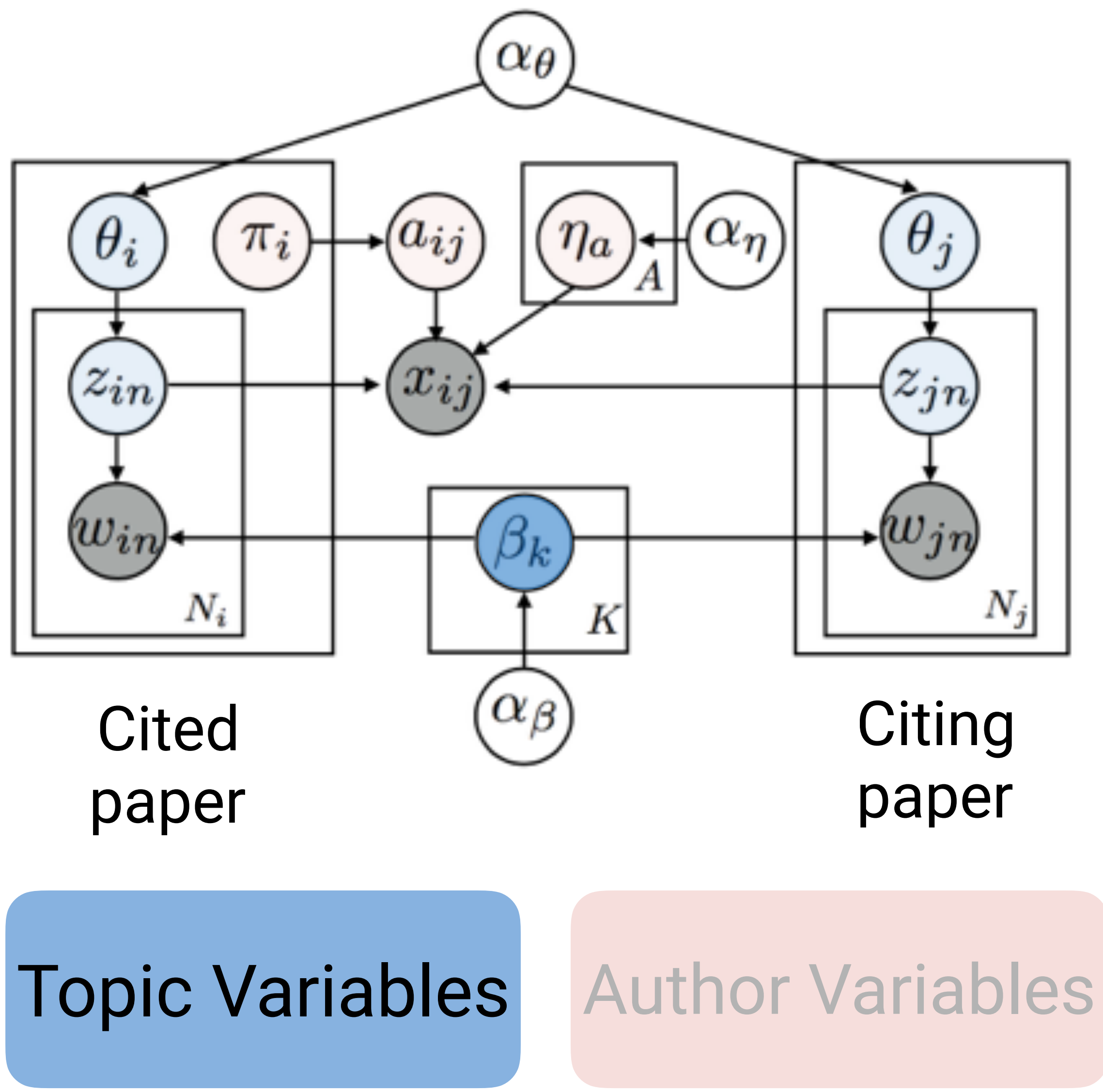
Author Variables

θ **Document-Topic distribution**
K (# topics)-dimensional vector

z **Per-word topic indicator variable**

β_k **Topic-Word distribution**
V (vocab. size)-dimensional vector

Model Description: Topic Variables



θ **Document-Topic distribution**
K (# topics)-dimensional vector

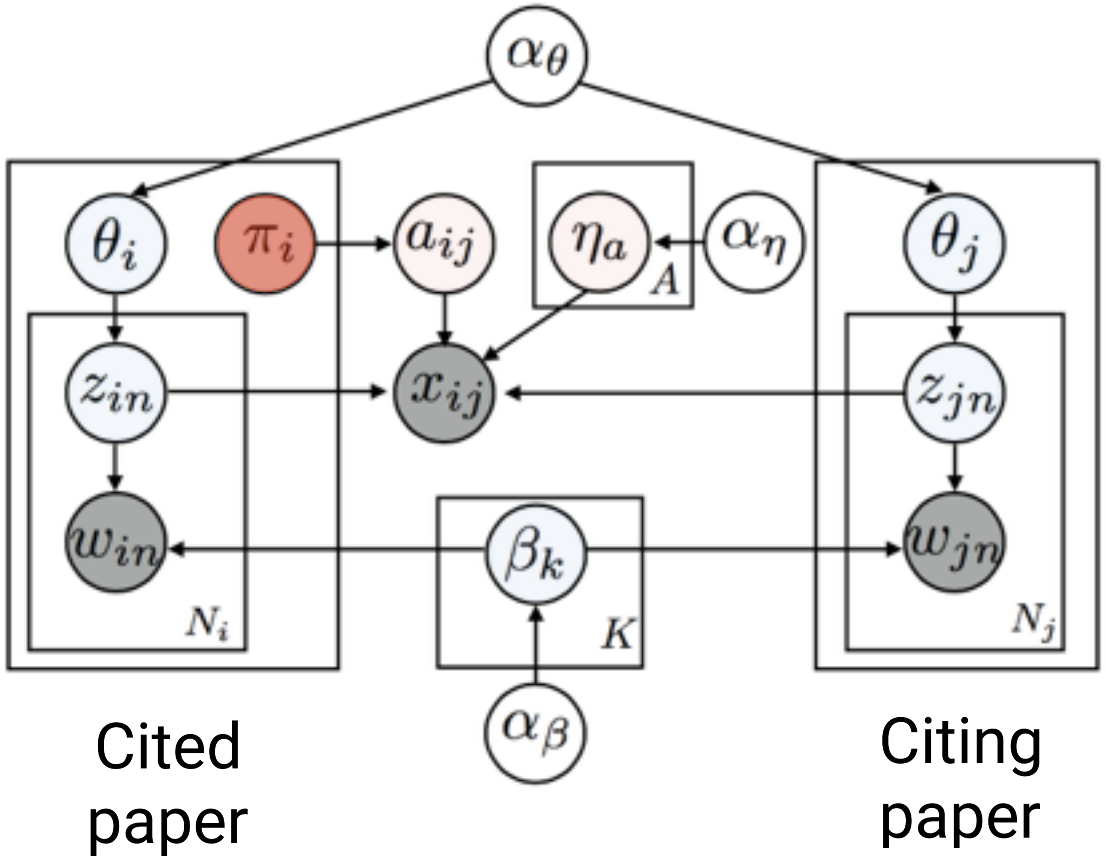
z **Per-word topic indicator variable**

β_k **Topic-Word distribution**
V (vocab. size)-dimensional vector

Topic 1	Topic 2
machine 0.05	image 0.07
learning 0.04	pattern 0.02
probability 0.02	recognition 0.02
distribution 0.01	pixel 0.01
...	...

...

Model Description: Author Variables



Topic Variables

Author Variables

π_i **Mixture weight over authors of publication i**

Given to each cited paper

a_{ij} **Selected author for cited paper i regarding citing paper j**

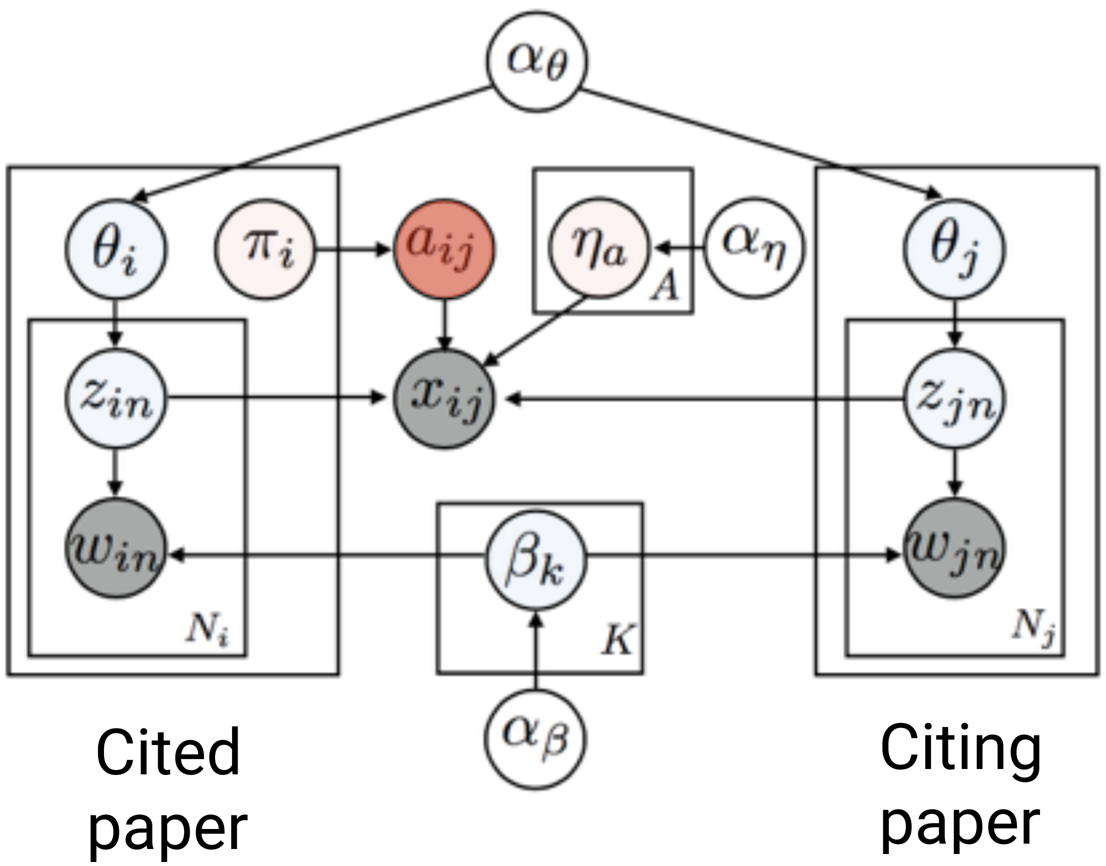
depends on the mixture weight π_i

η_a **Authority variable**

Given to each author

K-dimensional vector

Model Description: Author Variables



Topic Variables

Author Variables

π_i **Mixture weight over authors of publication i**

Given to each cited paper

a_{ij} **Selected author for cited paper i regarding citing paper j**

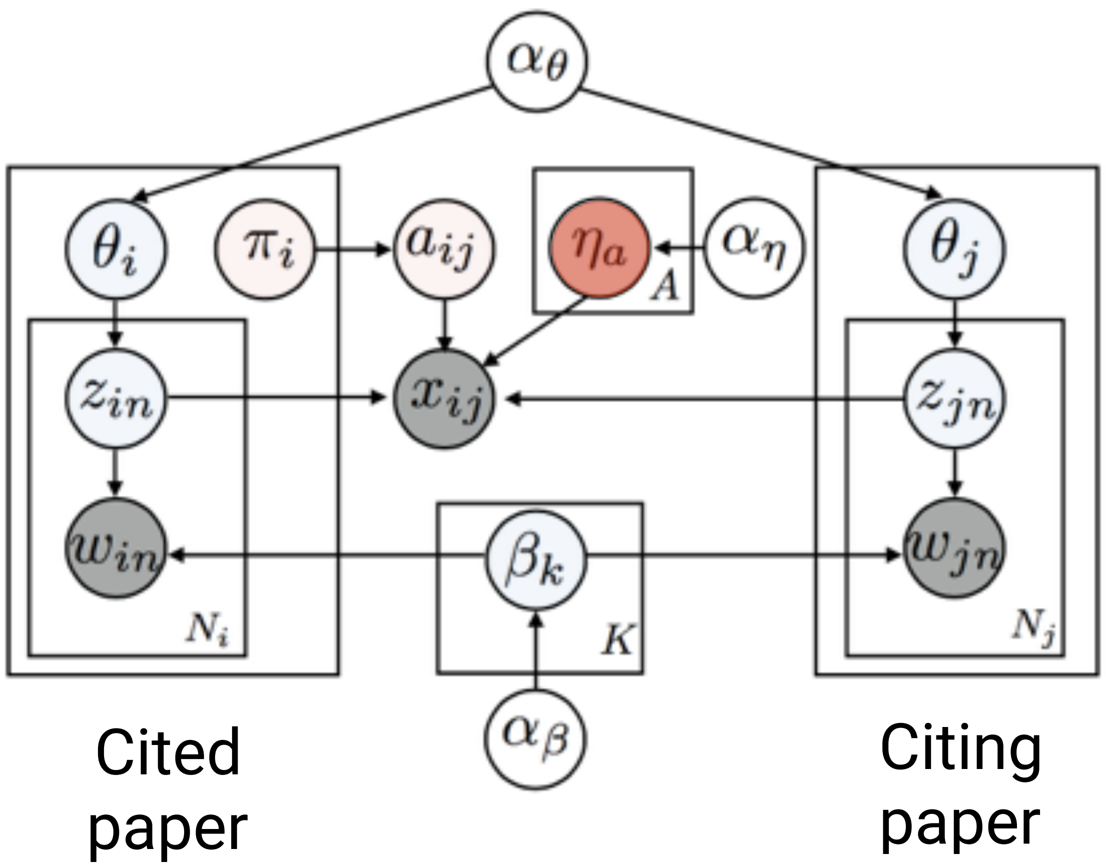
depends on the mixture weight π_i

η_a **Authority variable**

Given to each author

K-dimensional vector

Model Description: Author Variables



Topic Variables

Author Variables

π_i **Mixture weight over authors of publication i**

Given to each cited paper

a_{ij} **Selected author for cited paper i regarding citing paper j**

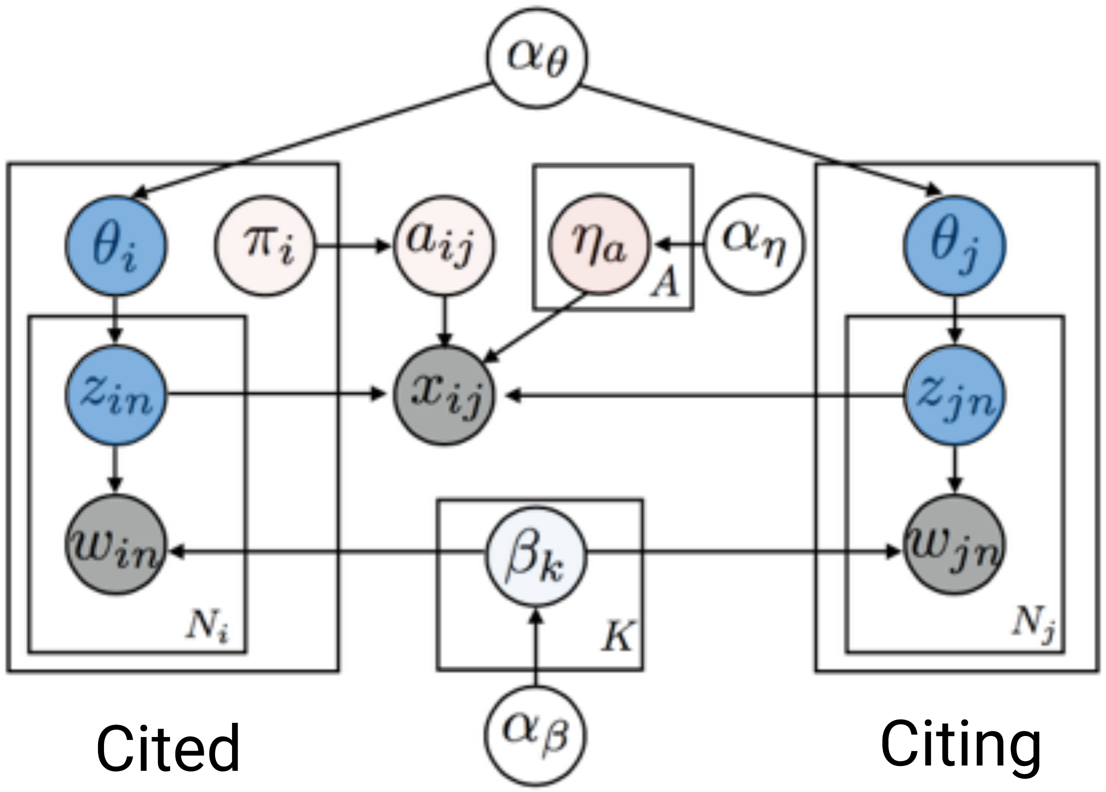
depends on the mixture weight π_i

η_a **Authority variable**

Given to each author

K-dimensional vector

Model Description: Citation Modeling



Topic Variables

Author Variables

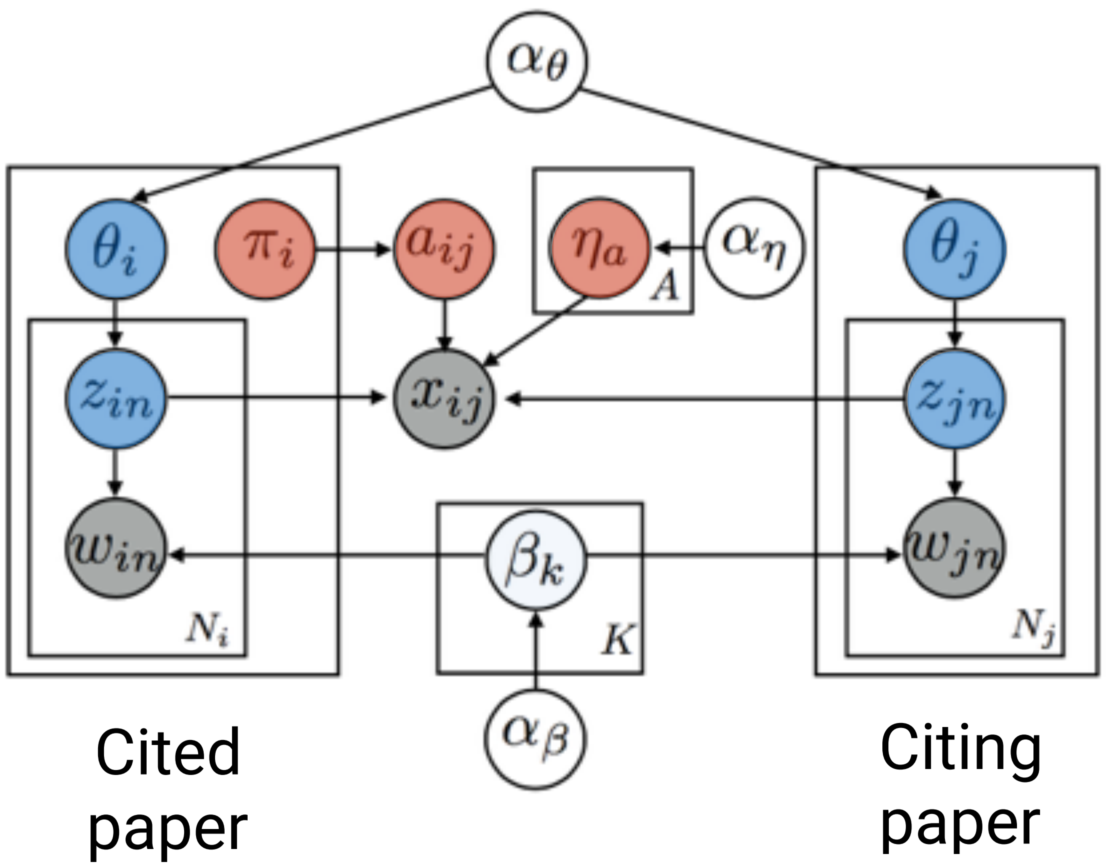
$$p(x_{ij} = 1) =$$

$$p(i \leftarrow j = 1) \propto$$

$$\overline{z_i}^\intercal \overline{z_j}$$

Topic Similarity for Citation

Model Description: Citation Modeling



Topic Variables

Author Variables

$$p(x_{ij} = 1) =$$

$$p(i \leftarrow j = 1) \propto$$

$$\overline{z_i}^\intercal \overline{z_j}$$

Topic Similarity for Citation

$$\overline{z_i}^\intercal diag(\eta_{a_{i \leftarrow j}}) \overline{z_j}$$

Adding Authority Variable

Model Description: Generative Process

1. For each topic k $\in${1, 2, ..., K}, draw topic distribution β_k~ Dirichlet(α_β)

2. For each document i $\in${1, 2, ..., D}:

   (a) Draw topic proportion: θ_i ~ Dirichlet(α_θ)

   (b) For each word token: $n \in$ {1, 2, ... ,N_i}:

      i. Draw topic assignment: $z_{in} \sim Mult(\theta_d)$

      ii. Draw word token: $w_{in} \sim Mult(\beta_{z_{in}})$

3. For each author *a and topic k* :

   (a) Draw authority index of author a: $\eta_{ak} \sim \mathcal{N}(0, \alpha_\eta^{-1} I)$

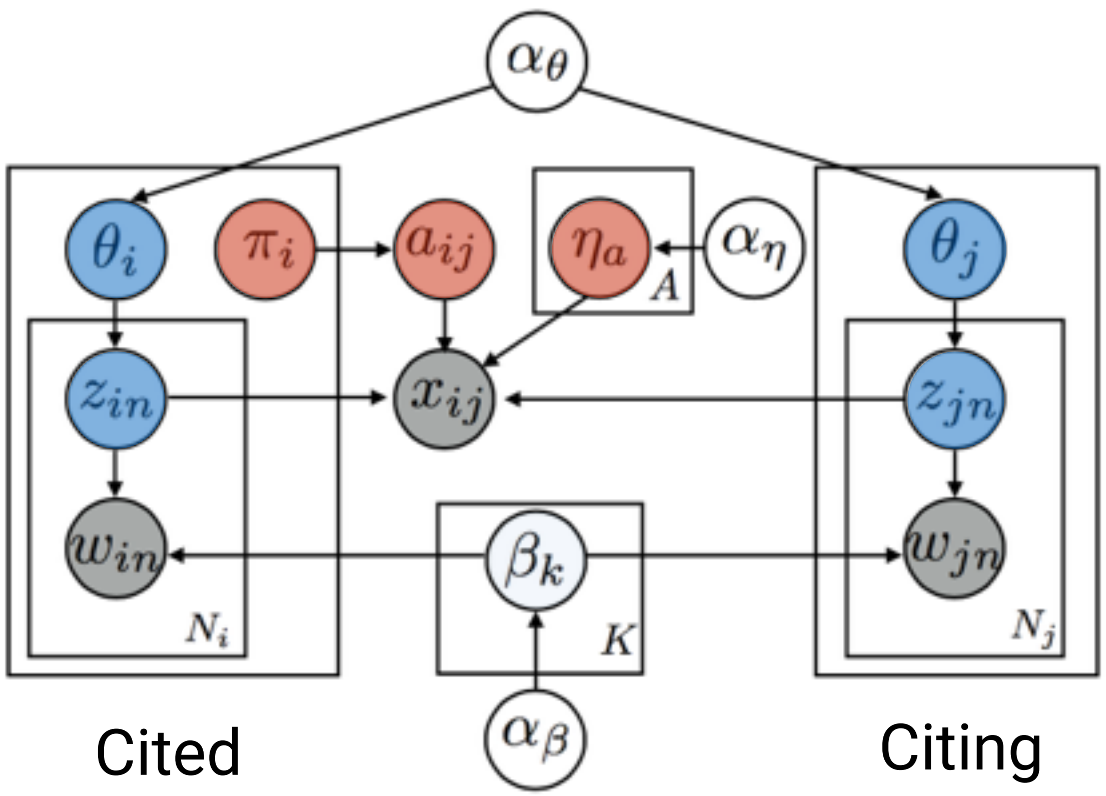
4. For each ordered document pair *i* and *j* :

   (a) Draw influence proportion parameter: ~ Dirichlet(π_i)

   (b) Draw one author from a set of authors of cited document *i* : $a_{i \leftarrow j} \sim Mult(\pi_{i \leftarrow j}$)

   (c) Draw link from document *j* to document *i* : $x_{i \leftarrow j} \sim \mathcal{N}(\overline{z_i}^\intercal diag(\eta_{a_{i \leftarrow j}})\overline{z_j}, c_{i \leftarrow j}^{-1})$

Model Inference



Cited paper

Citing paper

Topic Variables

Author Variables

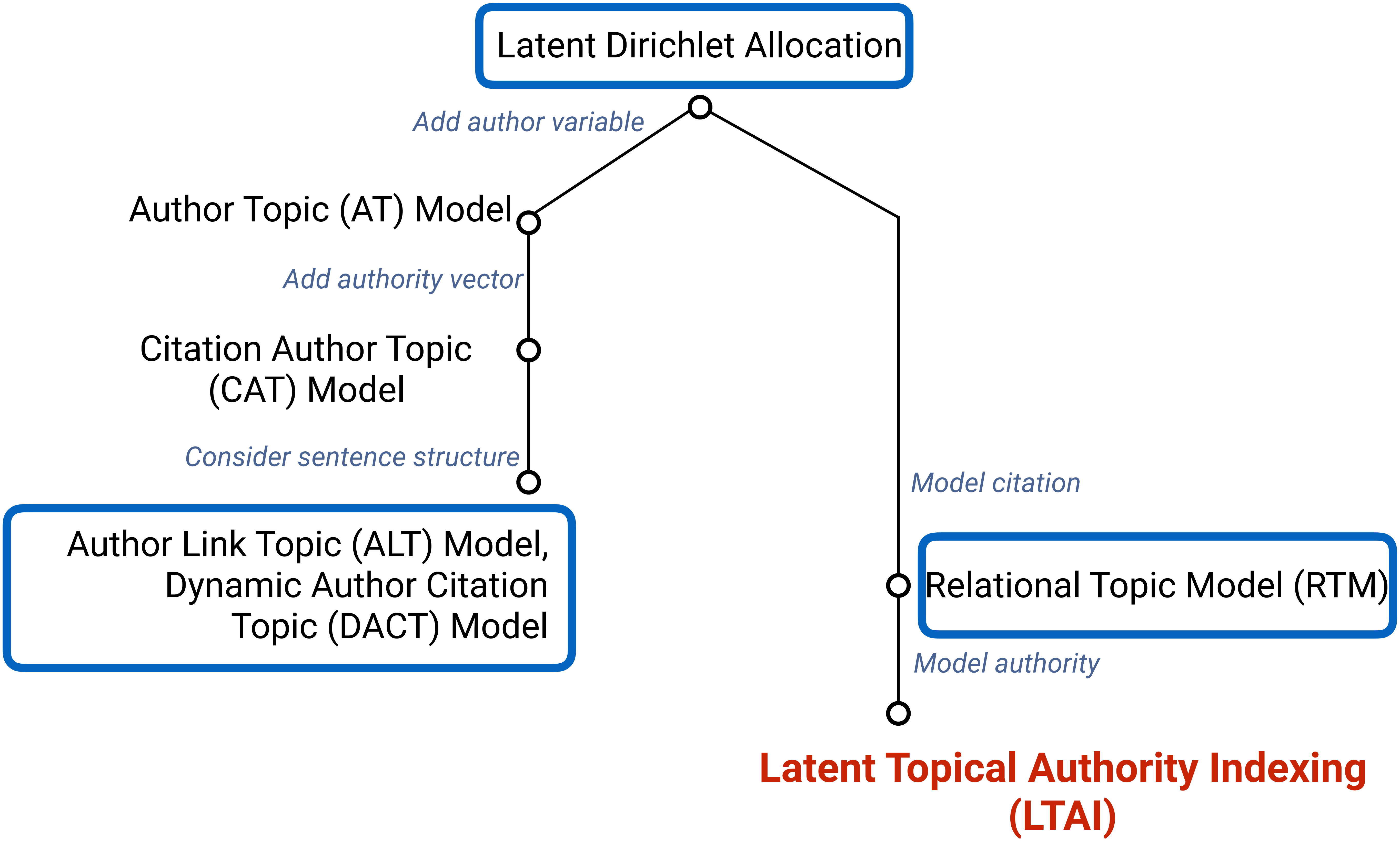
- **Topic Variables**
  - Tracking exact posterior distribution: Infeasible
  - Stochastic Variational Inference
    - Create approximate posterior distribution
- **Author Variables**
  - EM approach
    - Fix topic variables
    - Take gradient of the model likelihood with respect to eta
    - Fix eta and reassign values with respect to pi
    - Subsample the negative links

Experiments

Experiments: Dataset

	# Tokens	# Docs	# Authors	Avg. citation /Doc	Avg. citation /Author
CORA	17,059	13,147	12,111	3.46	12.17
arXiv-Physics	49,807	27,770	10,950	12.70	67.93
PNAS	39,664	31,054	9,862	1.57	13.18
Citeseer	21,223	4,255	6,384	1.24	4.38

Experiments: Related Models



Experiments: Related Models

Latent Topical Authority Indexing (LTAI)

- LTAI - *n*%
  Remove n % of citations,
  displace with random connections
- LTAI - *C*
  Set the precision parameter constant
- LTAI - *SEP*
  Learn the topic and the authority
  variables in separate manner

Experiments: Model Comparison

	Unified Model	Authority	Using cited contents	No sentence structure required
RTM	o	x	o	o
ALTM	o	o	x	o
DACTM	o	o	x	x
LTAI-n%	o	o	o	o
LTAI-C	o	o	o	o
LTAI-SEP	x	o	o	o
LTAI	**o**	**o**	**o**	**o**

Experiments: Evaluation Metric

- Mean Reciprocal Rank (MRR) = 1/(harmonic rank)

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

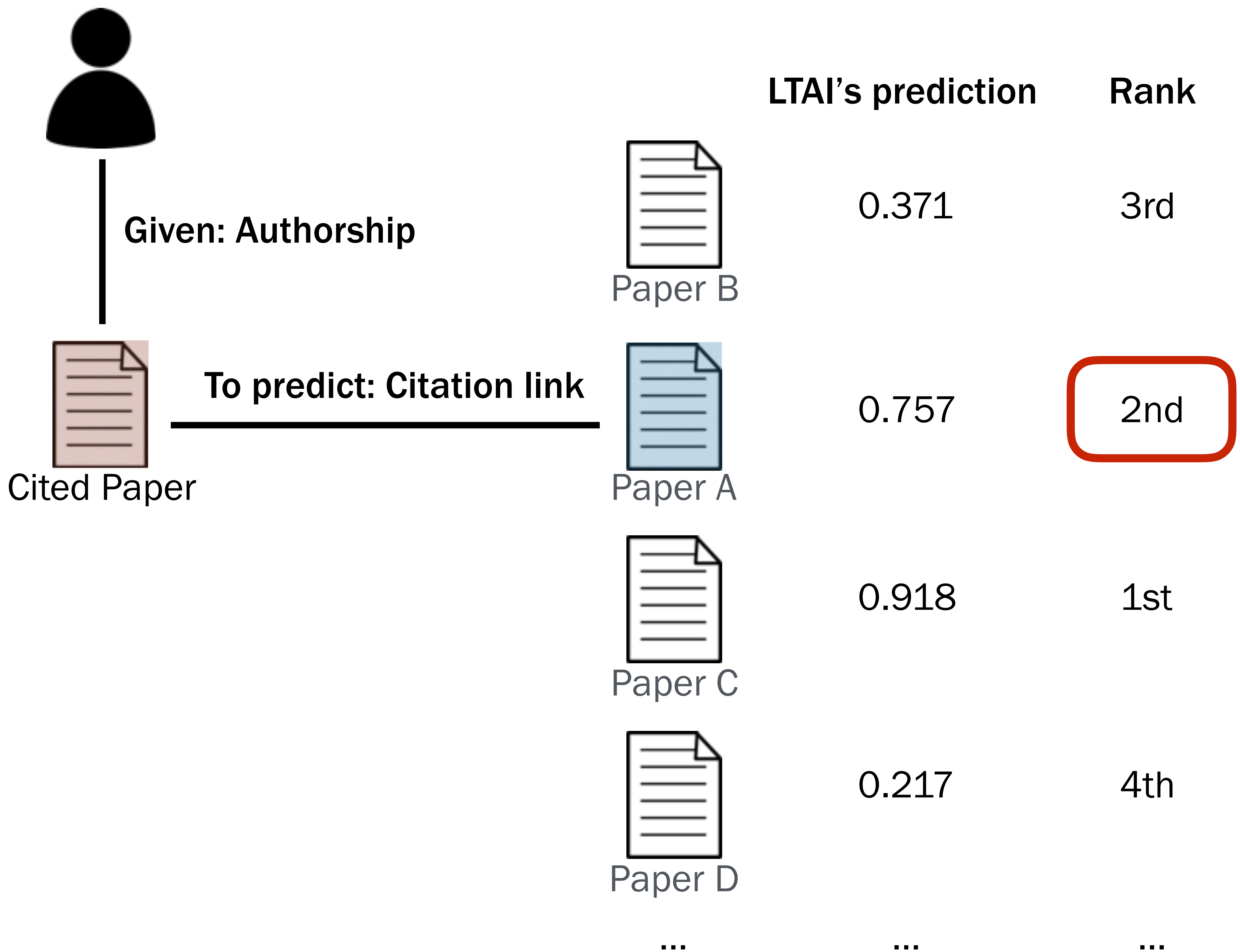
  - For author, citation prediction

- Word predictive probability

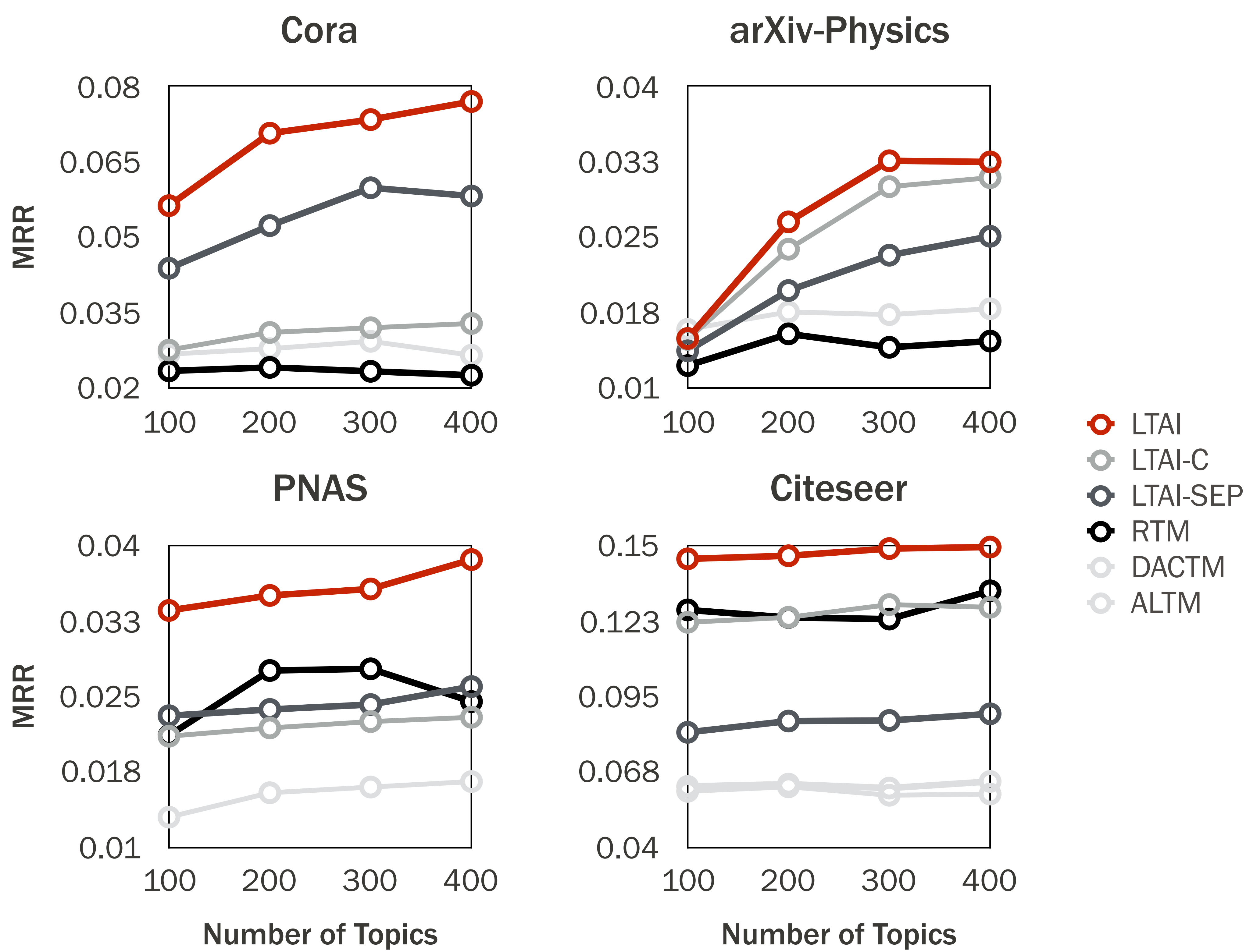
$$p(w_{new} | \mathcal{D}_{train}, w_{obs}) = \sum_{k=1}^{K} \mathbb{E}_q[\theta_k] \mathbb{E}_q[\beta_{k,w_{new}}]$$

  - For word prediction

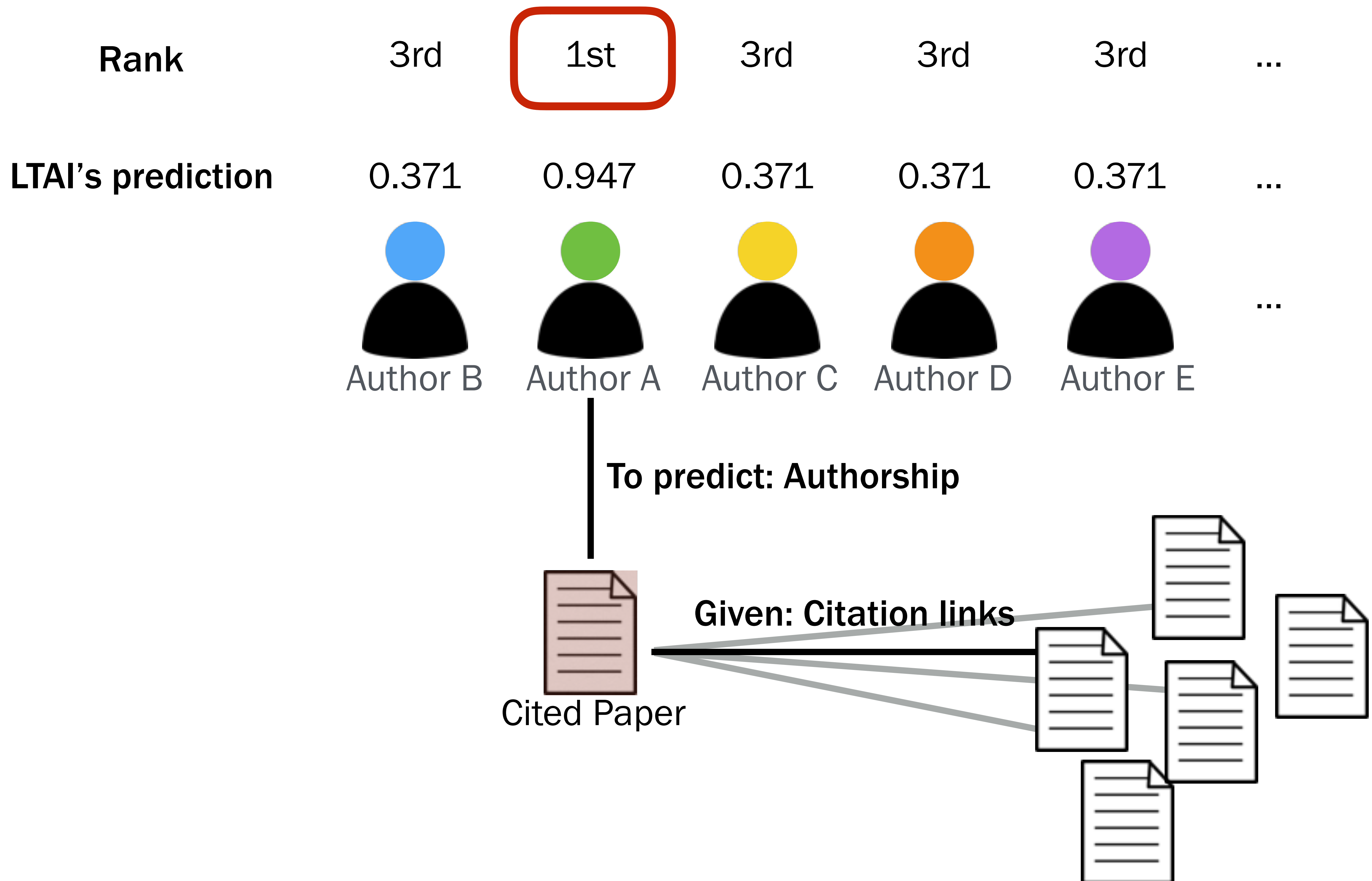
Experiments: Citation Prediction



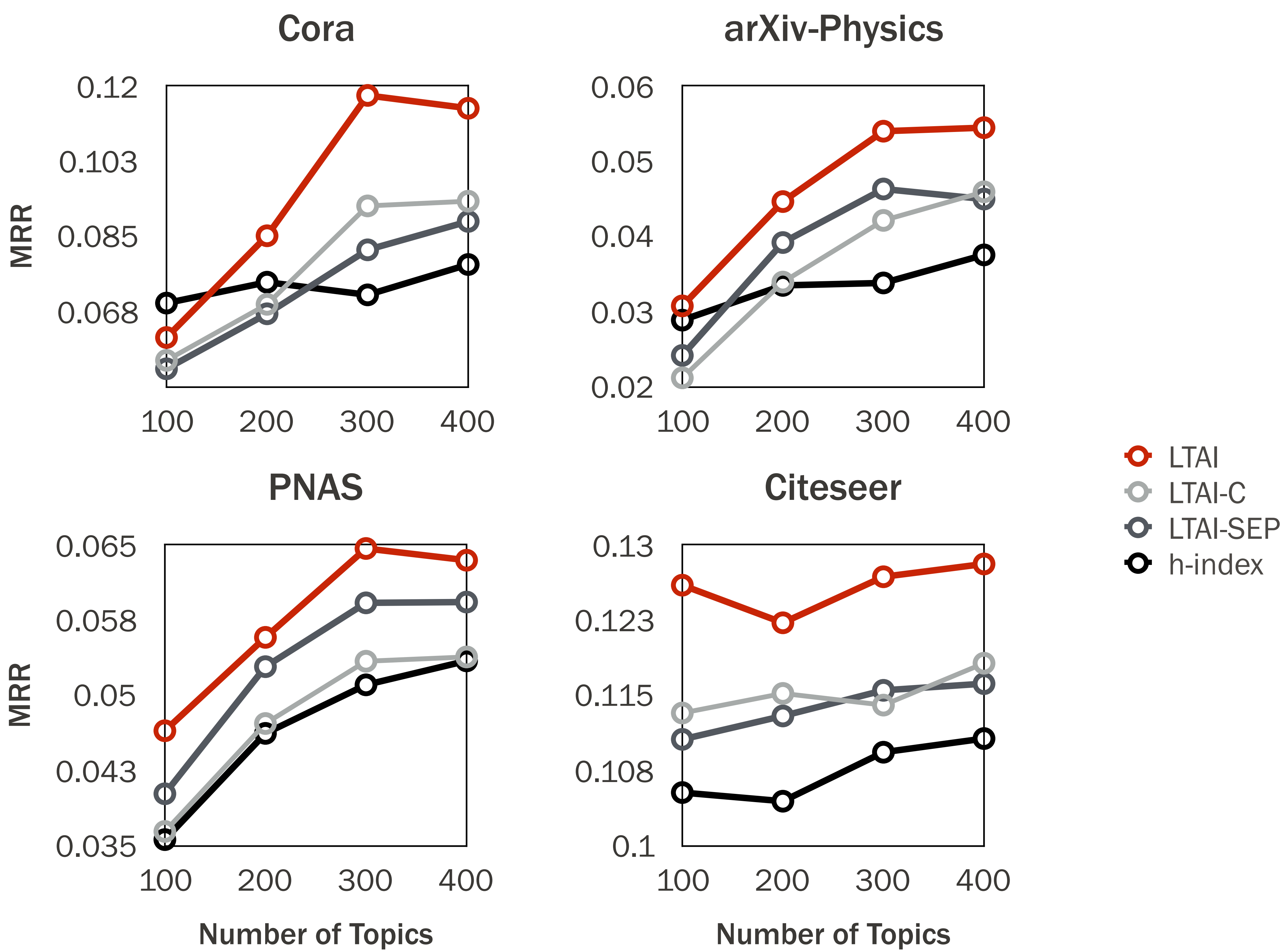
Experimental Results: Citation Prediction



Experiments: Author Prediction



Experimental Results: Author Prediction



Experimental Results: Word Prediction

Log Predictive Probability

	LTAI	LTAI-10%	LTAI-20%	LTAI-30%	LDA
CORA	**-7.624**	-7.672	-7.711	-7.754	-7.740
arXiv-Physics	**-7.724**	-7.744	-7.761	-7.813	-7.805
PNAS	**-8.214**	-8.262	-8.298	-8.321	-8.280
Citeseer	**-7.808**	-7.850	-7.863	-7.875	-7.866

Qualitative Analysis

Qualitative Analysis

T 27: *approximation, intelligence, artificial, correlation, support, recognition, model, representation*

Author	Authority Score of T27	h-index	# cite	# papers
M Jordan	9.85	9	245	28
F Girosi	4.76	4	117	13
T Poggio	4.67	6	176	28
M Jones	2.83	7	135	20

- Lists famous researchers with their topical authority score, h-index, number of citations, and number of papers

Qualitative Analysis: Researcher with Focused Research Domain

T 27: *approximation, intelligence, artificial, correlation, support, recognition, model, representation*

Author	Authority Score of T27	h-index	# cite	# papers
M Jordan	9.85	9	245	28
F Girosi	4.76	4	117	13
T Poggio	4.67	6	176	28
M Jones	2.83	7	135	20

- High concentration on statistical learning
  - ***Training support vector machines: an application to face detection***
  - ***An improved training algorithm for support vector machines***
  - ***Regularization theory and neural networks architectures***
- Relatively low number of h-index, #cite, #papers, but high topical authority score

Qualitative Analysis: Researcher with Broader Research Domain

T 27: *approximation, intelligence, artificial, correlation, support, recognition, model, representation*

Author	Authority Score of T27	h-index	# cite	# papers
M Jordan	9.85	9	245	28
F Girosi	4.76	4	117	13
T Poggio	4.67	6	176	28
M Jones	2.83	7	135	20

- Has broader academic interest than Federico Girosi: Statistical learning + computer vision
  - ***Face recognition: Features versus templates***
  - ***Hierarchical models of object recognition in cortex***
  - ***Example-based learning for view-based human face detection***
- Coauthored all papers written by Federico Girosi

Qualitative Analysis: Researcher with Different Research Domain

T 27: *approximation, intelligence, artificial, correlation, support, recognition, model, representation*

Author	Authority Score of T27	h-index	# cite	# papers
M Jordan	9.85	9	245	28
F Girosi	4.76	4	117	13
T Poggio	4.67	6	176	28
M Jones	2.83	7	135	20

- Research interests:
  - Programming language design, implementation, and application
- Main topic extracted by LTAI:
  - Language, type, programming, higher order
- Algorithms and inference techniques often used in the paper

Qualitative Analysis: Researcher with Different Research Domain

T 27: *approximation, intelligence, artificial, correlation, support, recognition, model, representation*

Author	Authority Score of T27	h-index	# cite	# papers
M Jordan	9.85	9	245	28
F Girosi	4.76	4	117	13
T Poggio	4.67	6	176	28
M Jones	2.83	7	135	20

The efficiency of a parallel implementation of the conjugate gradient method preconditioned by an incomplete Cholesky factorization can vary dramatically depending on the column ordering chosen. One method to minimize the number of major parallel steps is to choose an ordering based on a coloring of the symmetric graph representing the nonzero adjacency structure of the matrix. In this paper, we compare the performance of the preconditioned conjugate gradient method using these coloring orderings with a number of standard orderings on matrices arising from applications in structural engineering. Because optimal colorings for these systems may not be a priori known, we employ several graph coloring heuristics to obtain consistent colorings. Based on lower bounds obtained from the local structure of these systems, we find that the colorings determined by these heuristics are nearly optimal.

- Research ir
  - Programm … pplication
- Main topic
  - Language
- Algorithms … the paper

Conclusion

- LTAI models topical authority of academic researchers
  - Input: Text data, link data (authorship, citation)
  - Output: Research topic, topic distribution (paper), topical authority (author)
- LTAI outperforms related citation models
  - 3 Experiments: Predicting citation/authorship links and words
- LTAI's possible applications
  - Finding authoritative researchers from given topical interests
  - Finding academic papers from given topical interests
  - Discovering research topics from academic corpus

References

1. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. JMLR, 2003.

2. J. Chang and D. M. Blei. Relational topic models for document networks. In AISTATS, 2010.

3. M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In UAI, 2004.

4. Y. Tu, N. Johri, D. Roth, and J. Hockenmaier. Citation author topic model in expert search. In ICCL, 2010.

5. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link lda: joint models of topic and author community. In ICML, 2009.

Appendix: Variational Inference

- Variational distributions over the topic-related latent variables

$$q(\theta, \beta, z) = \prod_i q(\theta_i) \prod_{N_i} q(z_{in}) \prod_k q(\beta_k)$$

- where

$$q(z_{in}) = \text{Multinomial}(z_{in}|\phi_{in})$$
$$q(\theta_i) = \text{Dirichlet}(\theta_i|\gamma_i)$$
$$q(\beta_k) = \text{Dirichlet}(\phi_k|\lambda_k).$$

- Then ELBO of log-likelihood of the variational distribution becomes

$$\begin{aligned}\mathcal{L}_{[q]} = \mathbb{E}_q[&\sum_k \log p(\beta_k|\alpha_\beta) + \sum_i \log p(\theta_i|\alpha_\theta) \\ &+ \sum_i \sum_{N_i} \log p(z_{in}|\theta_d) + \log p(w_{in}|\beta_{z_{in}}) \\ &+ \sum_{i,j} \log p(x_{ij}|z_i, z_j, \pi_i)] - \mathcal{H}[q],\end{aligned}$$

Appendix: Variational Inference

- Taking gradient w.r.t gamma and lambda leads to

$$\gamma_{ik} = \alpha_\theta + \sum_{N_i} \phi_{ink} \qquad \text{and} \qquad \lambda_{kw} = \alpha_\beta + \sum_i \sum_{N_i} \phi_{ink} \delta(w_{in} = w)$$

- Taking gradient w.r.t. phi leads to

$$\phi_{ink} \propto \exp\left\{ \frac{\sum_j \partial \mathbb{E}_q[\log p(x_{ij}|\bar{z}_i, \bar{z}_j, \pi_i, \eta)]}{\partial \phi_{ink}} + \frac{\sum_j \partial \mathbb{E}_q[\log p(x_{ji}|\bar{z}_j, \bar{z}_i, \pi_j, \eta)]}{\partial \phi_{ink}} + \mathbb{E}_q[\log \theta_{ik}] + \mathbb{E}_q[\log \beta_{kw_{in}}] \right\}$$

- Where

$$\begin{aligned} &\mathbb{E}_q[\log p(x_{ij}|\bar{z}_i, \bar{z}_j, \pi_i, \eta)] \\ &\quad = \mathbb{E}_q[\log \sum_{a \in A_i} p(a_{ij} = a|\pi_i) p(x_{ij}|\bar{z}_i, \bar{z}_j, \eta_a)] \\ &\quad \geq \sum_{a \in A_i} p(a_{ij} = a|\pi_i) \mathbb{E}_q[\log p(x_{ij}|\bar{z}_i, \bar{z}_j, \eta_a)] \end{aligned}$$

using Jensen's inequality and

Appendix: Variational Inference

$$\mathbb{E}_q[\log p(x_{ij}|\bar{z}_i, \bar{z}_j, \eta_a)] = \mathcal{N}(x_{ij}|\bar{\phi}_i^\top \mathrm{diag}(\eta_a)\bar{\phi}_j, c_{ij})$$

by first-order Taylor expansion.

Thus,

$$\frac{\sum_j \partial \mathbb{E}_q[\log p(x_{ij}|\bar{z}_i, \bar{z}_j, \pi_i, \eta)]}{\partial \phi_{ink}}$$
$$\approx \sum_j \frac{\bar{\phi}_{jk} c_{ij}}{N_i} \sum_{a \in A_i} \eta_{ak}(x_{ij} - \bar{\phi}_i^\top \mathrm{diag}(\eta_a)\bar{\phi}_j) p(a_{ij} = a|\pi_i)$$