Poisson Matrix Factorisation with Side Information

He Zhao (赵贺)

Final-year Ph.D. student, Monash University, Australia

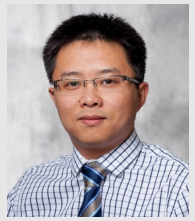
Supervisors:

Wray Buntine@Monash Lan Du@Monash

Collaborators:

Mingyuan Zhou@UT Austin Piyush Rai@IITK






Roadmap

1

- Background
  - Bayesian probabilistic models
  - Bayesian inference

2

- Poisson matrix factorisation
  - Basic framework
  - Applications

3

- Our recent work in Poisson matrix factorisation
  - Relational network analysis
  - Topic models for text analysis

4

- Future work

Background of Bayesian probabilistic models and Bayesian Inference

Bayesian probabilistic models

The *i.i.d.* assumption

- Usually $X = \{x_1, \ldots, x_n\}$ represents the data and $\boldsymbol{\theta}$ represents the model parameters.
- One usually assumes that $\{x_i\}_i$ are independent and identically distributed (*i.i.d*) conditioning on $\boldsymbol{\theta}$.
- Under the conditional *i.i.d.* assumption:
  - $P(X|\boldsymbol{\theta}) = \prod_{i=1}^{n} P(x_i|\boldsymbol{\theta})$.
  - The data in X are exchangeable, which means that $P(x_1, \ldots, x_n) = P(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$ for any random permutation σ of the data indices $1, 2, \ldots, n$.

Bayes' rule

- In equation:

$$P(\boldsymbol{\theta}|X) = \frac{P(X|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(X)} = \frac{P(X|\boldsymbol{\theta})P(\boldsymbol{\theta})}{\int P(X|\boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

  If $\boldsymbol{\theta}$ is discrete, then $\int f(\boldsymbol{\theta})d\boldsymbol{\theta}$ is replaced with $\sum f(\boldsymbol{\theta})$.
- In words:

$$\text{Posterior of } \boldsymbol{\theta} \text{ given } X = \frac{\text{Conditional Likelihood} \times \text{Prior}}{\text{Marginal Likelihood}}$$

MLSS 2015 Tutorial at UT Austin by Mingyuan Zhou
https://mingyuanzhou.github.io/Papers/MLSS_01072015.pdf

Bayesian probabilistic models

Conjugate priors

If the prior $P(\boldsymbol{\theta})$ is conjugate to the likelihood $P(X|\boldsymbol{\theta})$, then the posterior $P(\boldsymbol{\theta}|X)$ and the prior $P(\boldsymbol{\theta})$ are in the same family.

- Conjugate priors are widely used to construct hierarchical Bayesian models.
- Although conjugacy is not required for MCMC inference, it helps develop closed-form Gibbs sampling update equations.

- Example (i): beta is conjugate to Bernoulli.

$$x_i|p \sim \text{Bernoulli}(p),\ p \sim \text{Beta}(\beta_0, \beta_1)$$

  - Conditional likelihood:

$$P(x_1, \ldots, x_n|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

  - Prior:

$$P(p|\beta_0, \beta_1) = \frac{\Gamma(\beta_0+\beta_1)}{\Gamma(\beta_0)\Gamma(\beta_1)} p^{\beta_0-1}(1-p)^{\beta_1-1}$$

  - Posterior:

$$P(p|X, \beta_0, \beta_1) \propto \left\{\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}\right\}\left\{p^{\beta_0-1}(1-p)^{\beta_1-1}\right\}$$

$$(p|x_1, \ldots, x_n, \beta_0, \beta_1) \sim \text{Beta}\left(\beta_0 + \sum_{i=1}^{n} x_i,\ \beta_1 + n - \sum_{i=1}^{n} x_i\right)$$

Bayesian probabilistic models

Conjugate priors

- Example (iii): $x_i \sim \mathcal{N}(\mu, \varphi^{-1}),\ \mu \sim \mathcal{N}(\mu_0, \varphi_0^{-1})$
- Example (iv): $x_i \sim \text{Poisson}(\lambda),\ \lambda \sim \text{Gamma}(\alpha, \beta)$
- Example (v): $x_i \sim \text{NegBino}(r, p),\ p \sim \text{Beta}(\alpha_0, \alpha_1)$
- Example (vi): $x_i \sim \text{Gamma}(\alpha, \beta),\ \beta \sim \text{Gamma}(\alpha_0, \beta_0)$
- Example (vii):

$$(x_{i1}, \ldots, x_{ik}) \sim \text{Multinomial}(n_i, p_1, \ldots, p_k),$$

$$(p_1, \ldots, p_k) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k p_j^{\alpha_j - 1}$$

Hierarchical priors

- One may construct a complex prior distribution using a hierarchy of simple distributions as

$$P(\boldsymbol{\theta}) = \int \ldots \int P(\boldsymbol{\theta}|\boldsymbol{\alpha}_t) P(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}) \ldots P(\boldsymbol{\alpha}_1) d\boldsymbol{\alpha}_1 \ldots d\boldsymbol{\alpha}_t$$

- Draw $\boldsymbol{\theta}$ from $P(\boldsymbol{\theta})$ using a hierarchical model:

$$\begin{aligned} \boldsymbol{\theta}|\boldsymbol{\alpha}_t, \ldots, \boldsymbol{\alpha}_1 &\sim P(\boldsymbol{\theta}|\boldsymbol{\alpha}_t) \\ \boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \ldots, \boldsymbol{\alpha}_1 &\sim P(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}) \\ &\ldots \\ \boldsymbol{\alpha}_1 &\sim P(\boldsymbol{\alpha}_1) \end{aligned}$$

Bayesian inference

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{\int P(X|\theta)P(\theta)}$$

Tricky part

Optimisation

Maximum likelihood: maximise $\log P(X|\theta)$

↓ Prior

Maximum a posteriori: maximise $\log P(X|\theta)P(\theta)$

↓ Unobserved latent variables

Expectation maximisation:

- Compute expectation of $\log P(X|\theta)P(\theta)$ with latent variables
- Maximise $\log P(X|\theta)P(\theta)$

- Toss out the marginal likelihood
- Not fully Bayesian estimation of the posterior
- Point estimation of θ
- Sometimes good enough for downstream applications but no uncertainty in predictions

Variational inference

- $Q(\theta)$ as an approximation to $P(\theta|X)$
- Maximise $\mathrm{KL}(Q||P)$ to find good $Q(\theta)$

→ Approximated distribution to the posterior

Sampling

Markov chain Monte Carlo (MCMC) sampling

- …
- Gibbs sampling

→ Converge to the true posterior

Bayesian inference with Gibbs sampling

Inference via Gibbs sampling

- Gibbs sampling:
  - The simplest Markov chain Monte Carlo (MCMC) algorithm.
  - A special case of the Metropolis-Hastings algorithm.
  - Widely used for statistical inference.

- For a multivariate distribution $P(x_1, \ldots, x_n)$ that is difficult to sample from, if it is simpler to sample each of its variables conditioning on all the others, then we may use Gibbs sampling to obtain samples from this distribution as
  - Initialize $(x_1, \ldots, x_n)$ at some values.
  - For $s = 1 : S$
    For $i = 1 : n$
    Sample x_i conditioning on the others from
    $P(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$
    End
    End

- 😀
  - Guaranteed to converge to true posterior
  - Sampling formulas are easy to derive with conjugacy
  - No parameter tuning
  - Reasonably efficient with conjugacy
- 😞
  - Each iteration has to go through all data samples
  - May take long time to converge
  - Hard to determine when to stop
  - Not easy to parallelise
  - May be slow for big data

MLSS 2015 Tutorial at UT Austin by Mingyuan Zhou
https://mingyuanzhou.github.io/Papers/MLSS_01072015.pdf

Poisson matrix factorisation (PMF)

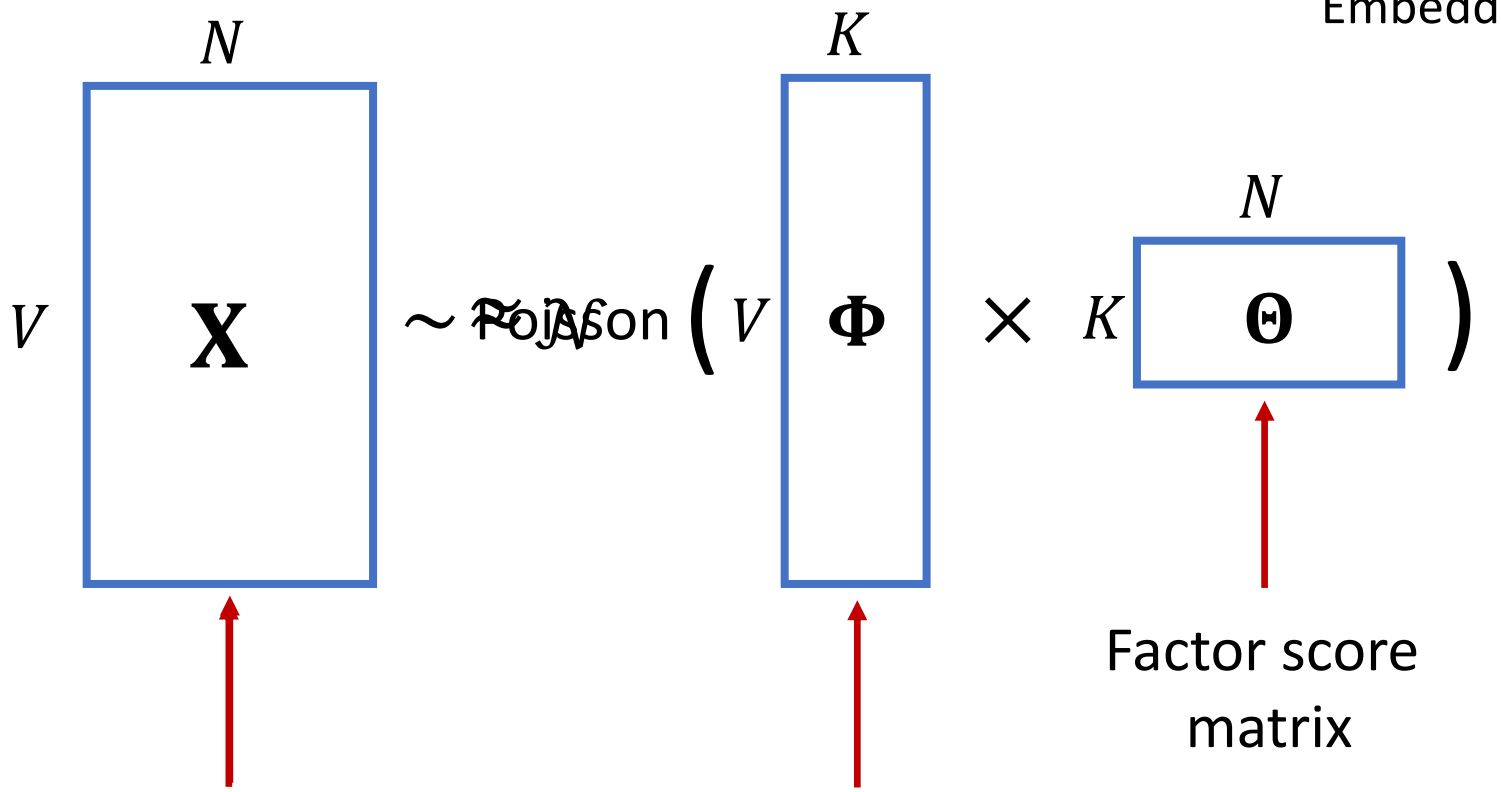
Basic PMF framework

Documents

- N: the number of documents
- V: the vocabulary size
- x_{vi}: how many words of v in document i

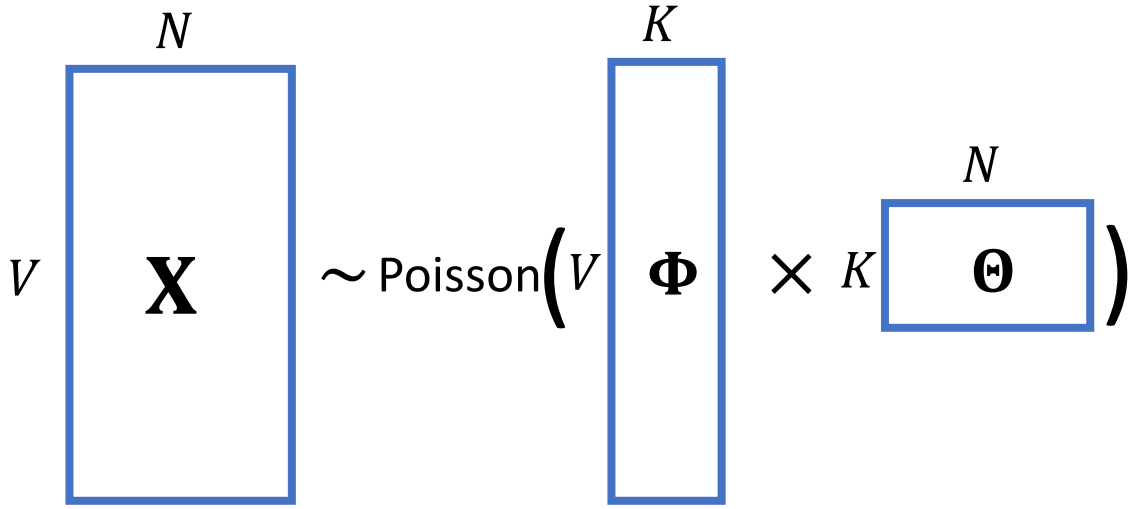
Collaborative filtering

- N: the number of users
- V: the number of songs
- x_{vi}: how many times that user i listens to song v



Probabilistic Matrix Factorization
(Salakhutdinov, et. al, NIPS, 2008)

PMF and Latent Dirichlet Allocation (LDA)



PMF

- The length of each document is drawn from Poisson
- $\boldsymbol{\theta_i}$ and $\boldsymbol{\phi_k}$ are not necessarily normalised

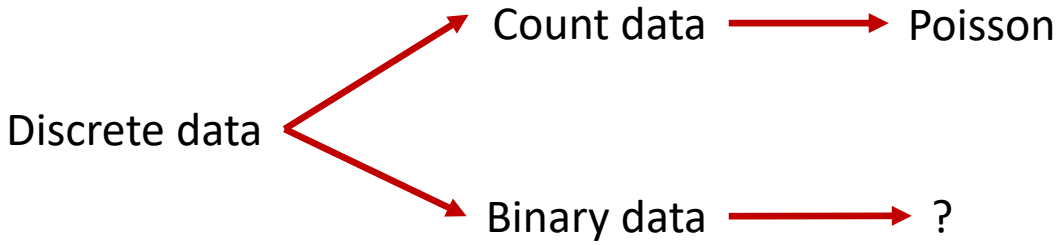
LDA

- The length of each document is fixed and known
- $\boldsymbol{\theta_i}$ is the topic distribution of document i
  $\boldsymbol{\phi_k}$ is the word distribution of topic k
- $\boldsymbol{\theta_i}$ and $\boldsymbol{\phi_k}$ are normalised

Wray's Quora answer:
https://www.quora.com/What-is-the-difference-between-NMF-and-LDA-Why-are-the-priors-of-LDA-sparse-induced

Zhou, Bayesian Analysis, 2017

PMF for (nonnegative) binary data



Bernoulli-Poisson link (BerPois)
(Zhou, AISTATS, 2015)

$$y_{vi} \sim \text{Poisson}(-)$$

$$x_{vi} = 1_{y_{vi}>0}$$

If $x_{vi} = 0$, then $y_{vi} = 0$ almost surely
If $x_{vi} = 1$, then $y_{vi} \sim Poisson_+(-)$

We only need to infer y when $x > 0$

Non-zeros of $\mathbf{X}$

Applications of PMF on discrete data

- Collaborative filtering
  - Customer-item
  - User-movie
- K: the number of latent factors

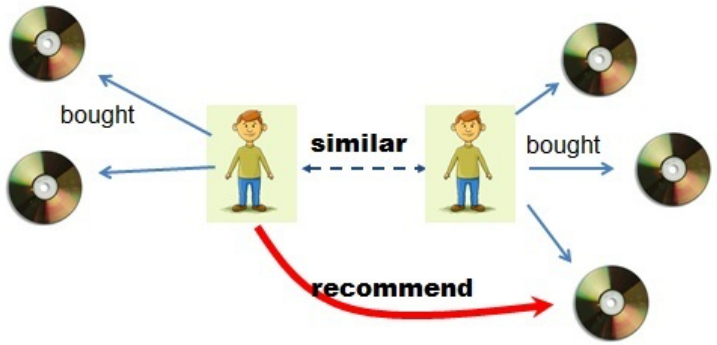


Figure source: https://spatnaik77.wordpress.com/2013/07/17/recommendation-engine/

- Relational network analysis
  - Social networks
  - Citation networks
  - Knowledge graphs
- K: the number of communities

Figure source: http://www.goldsteinepi.com/blog/socialnetworkanalysisinepidemiologypart2

- Text analysis: Topic models
- Document-word
- K: the number of topics

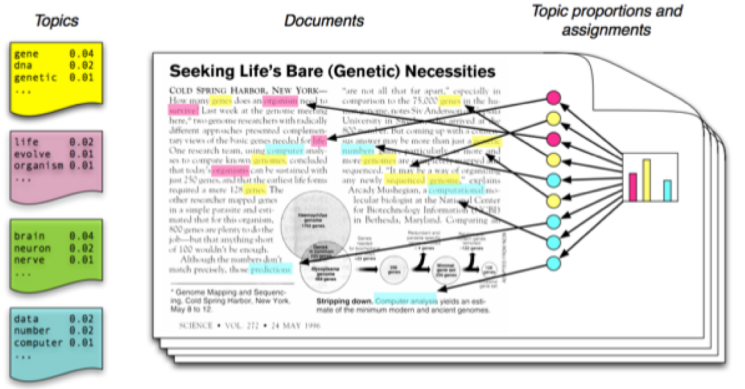
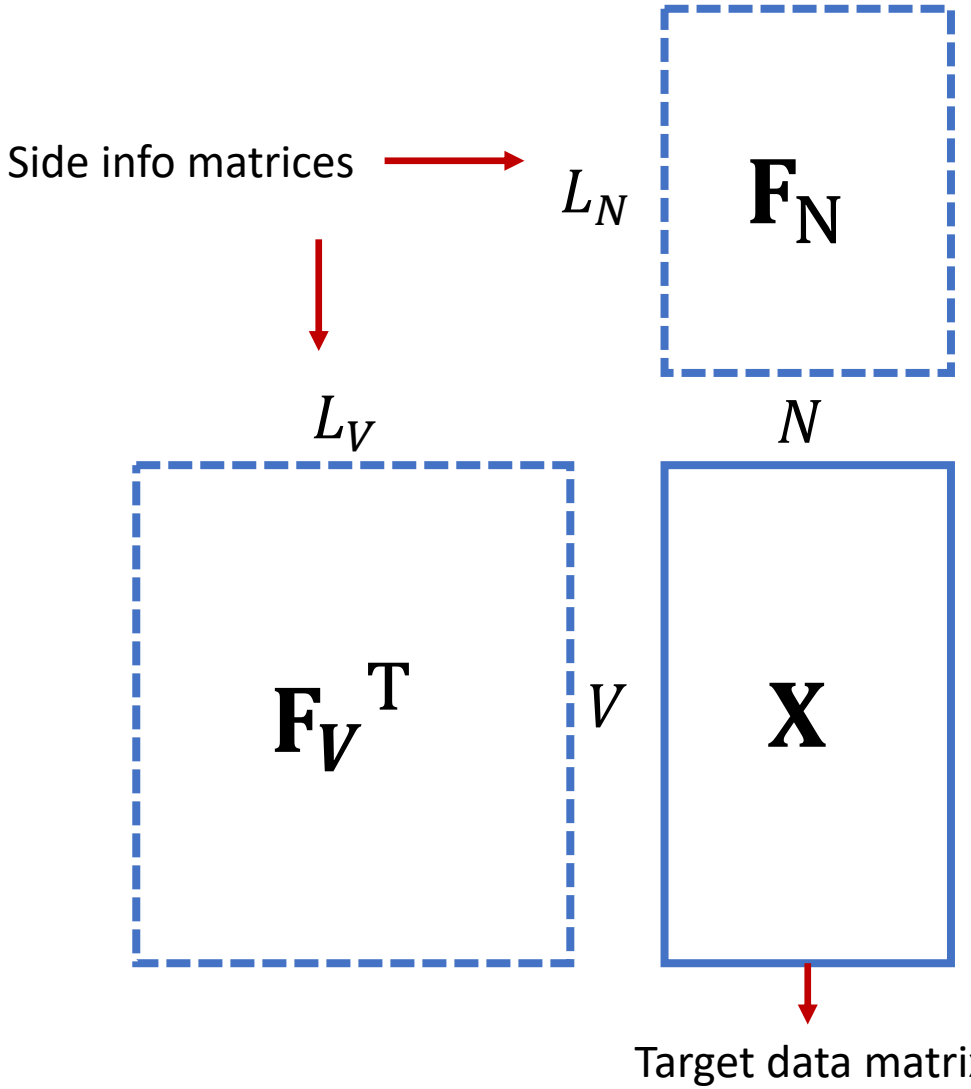


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Our recent work in Poisson matrix factorisation with side info

Side information

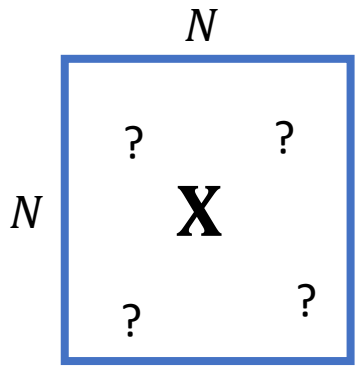


- User-movie interactions
  - User side info: age, gender, education, …
  - Movie side info: genres, year, actors, directors, …
  - Why a user watched a movie
- Paper citation network
  - Paper side info: venue, authors, key words, …
  - Why a paper cited another
- Documents
  - Document side info: labels, authors, dates, …
  - Why a document has those topics
  - Word side info: word correlations, word embeddings, …
  - Why a topic is described by those words

Complementary data

Relational network analysis

Adjacent matrix of a relational network with N nodes



- Square -> Only one set of nodes
- Binary -> No weight on edge
- Sparse -> Many nodes, but few links
- A/Symmetric -> Directed/Undirected Relation

- Friendship of users in social networks
- Co-authorship of authors
- Citation networks of papers
- Medicine interactions
- …

Predict missing links

Edge Partition Model (EPM)

(Zhou, AISTATS, 2015)

$$x_{ji} = 1_{y_{ji}>0}$$

$$y_{ji} = \sum_{k_1,k_2}^{K_1,K_2} y_{k_1 j k_2 i}$$

$$y_{k_1 j k_2 i} \sim \text{Poisson}(\phi_{k_1 j} \lambda_{k_1,k_2} \phi_{k_2 i})$$

$$\lambda_{k_1,k_2} \sim \begin{cases} \text{Gamma}(\epsilon r_k, 1/a_0), & \text{if } k_1 = k_2 = k, \\ \text{Gamma}(r_{k_1} r_{k_2}, 1/a_0), & \text{otherwise.} \end{cases}$$

$$r_k \sim \text{Gamma}(r_0/K, 1/c_0)$$

$$\phi_{ki} \sim \text{Gamma}(\alpha_i, 1/c_i)$$

$$\alpha_i \sim \text{Gamma}(e_0, 1/f_0)$$

$\mathbf{X}_{N \times N} \sim \text{BerPois}\left(\mathbf{\Phi}^{\mathrm{T}}_{N \times K} \times \mathbf{\Lambda}_{K \times K} \times \mathbf{\Phi}_{K \times N}\right)$

Embeddings of nodes

Stochastic Block Model (SBM)

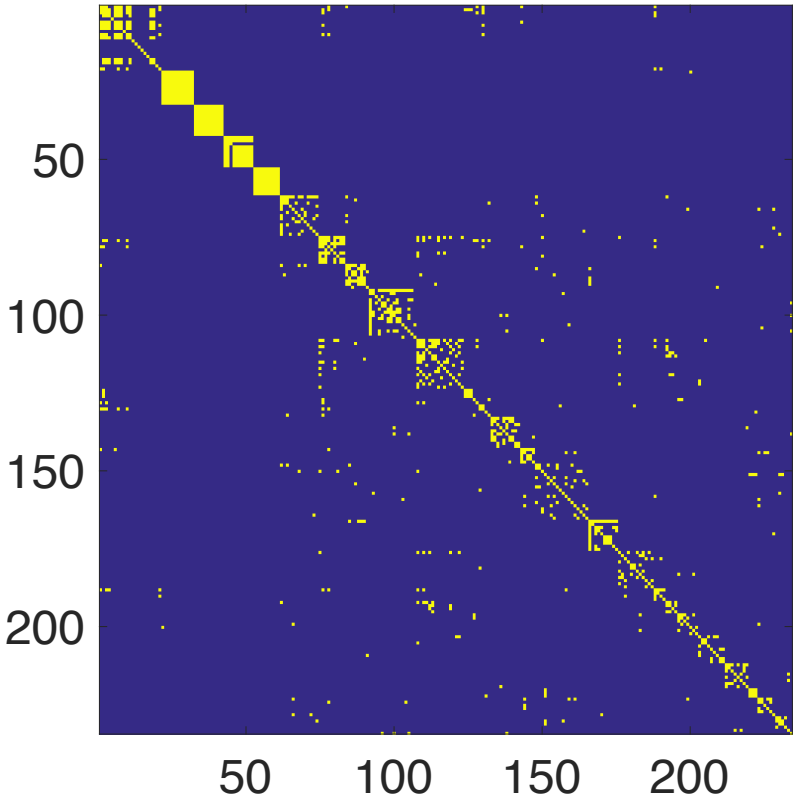
Correlations of embeddings

Block matrix

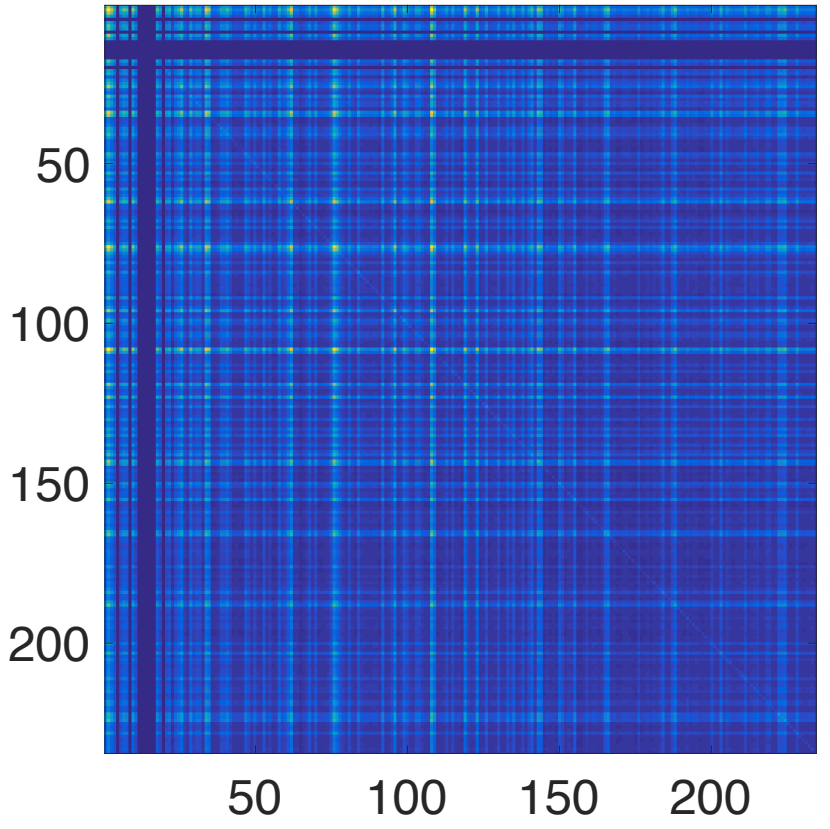
Binary data matrix

Edge Partition Model (EPM)

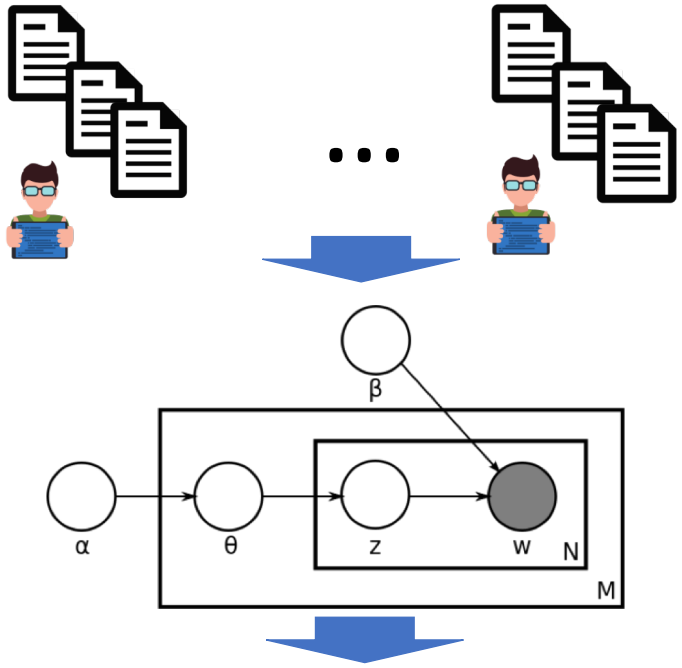
Co-authorship network of
234 authors in NIPS conferences 1988-2003



EPM's reconstruction with
80% of the node pairs



Can node attribute (side info) help?



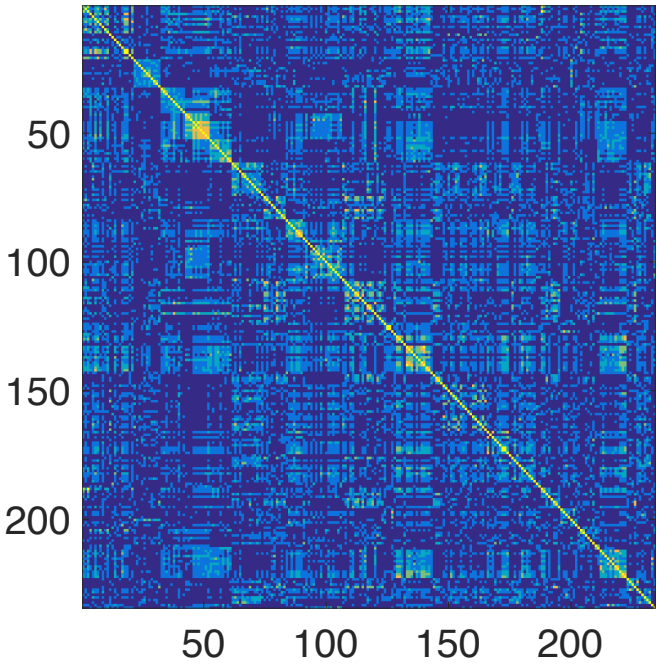
LDA with 100 topics

Topic distribution: 100 dimensional real-valued attributes for each author

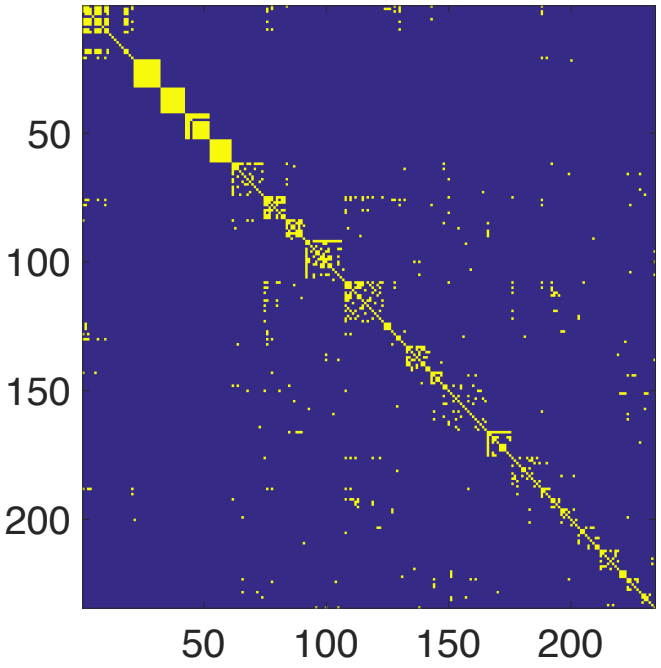
Top 5 topics with the largest weights

100 dimensional binary attributes for each author

Cosine distance of the topic distributions of authors

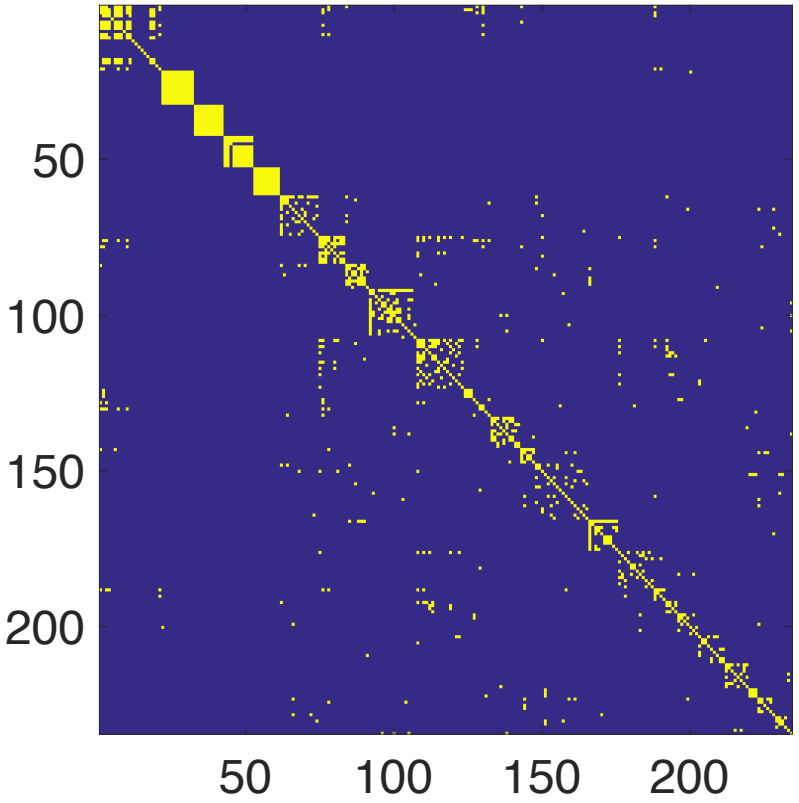


Co-authorship network

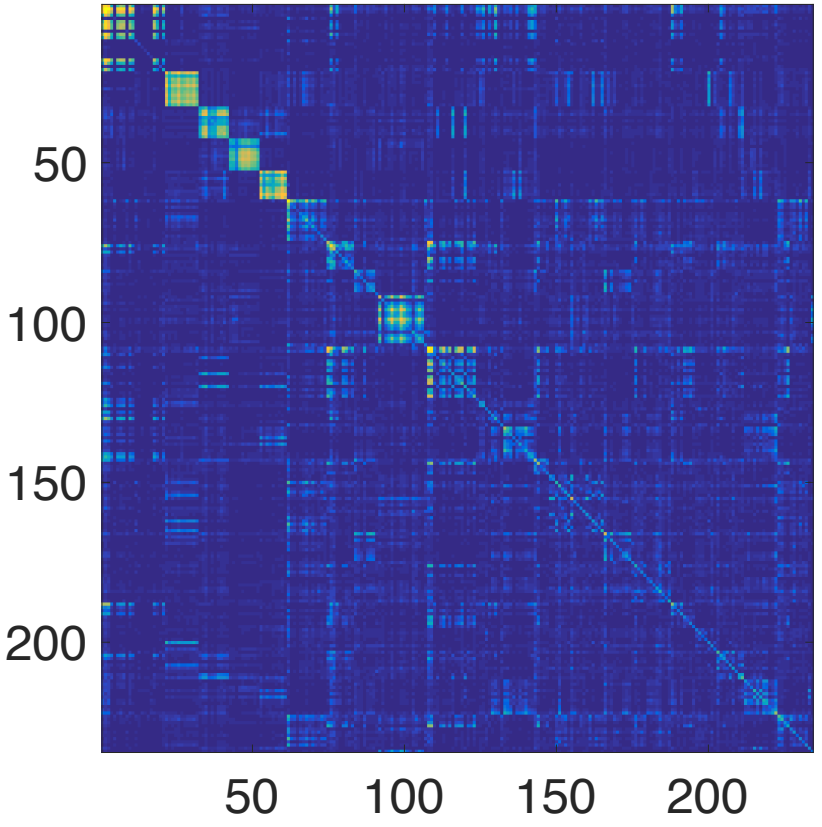


Node attribute helps!

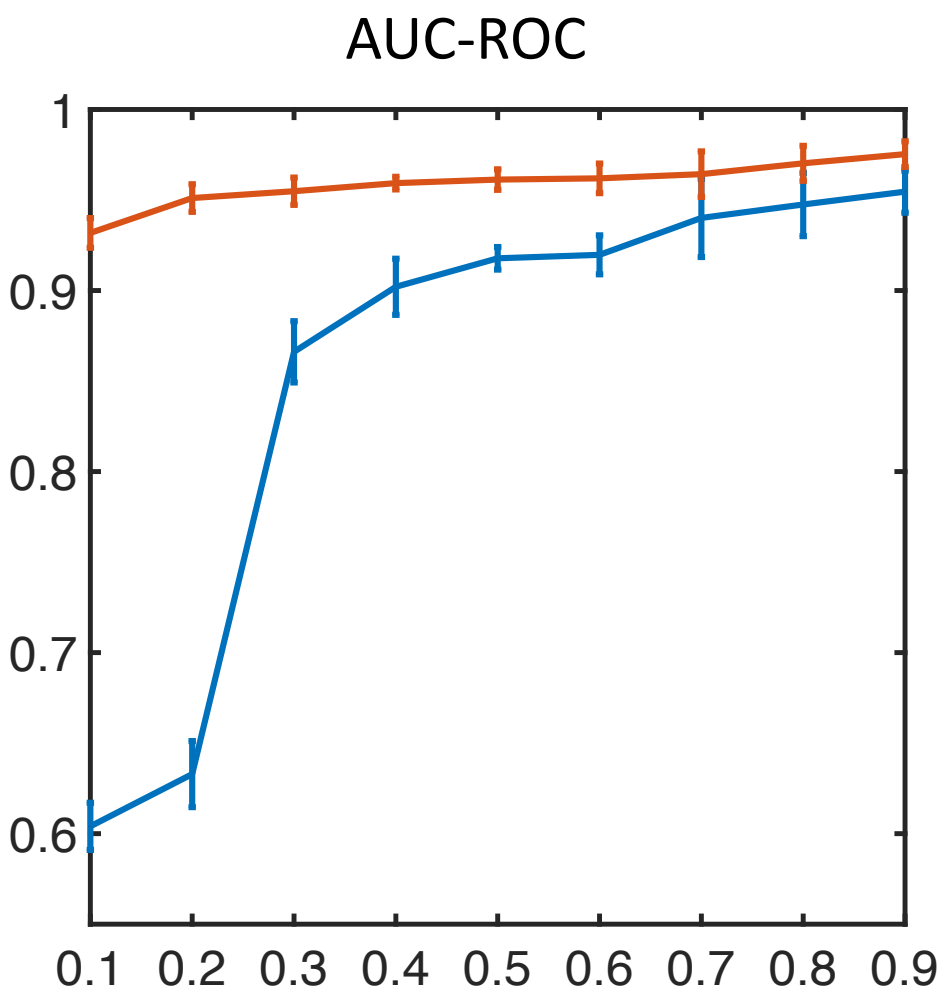
The co-authorship network of
234 authors in NIPS 1988-2003 conferences

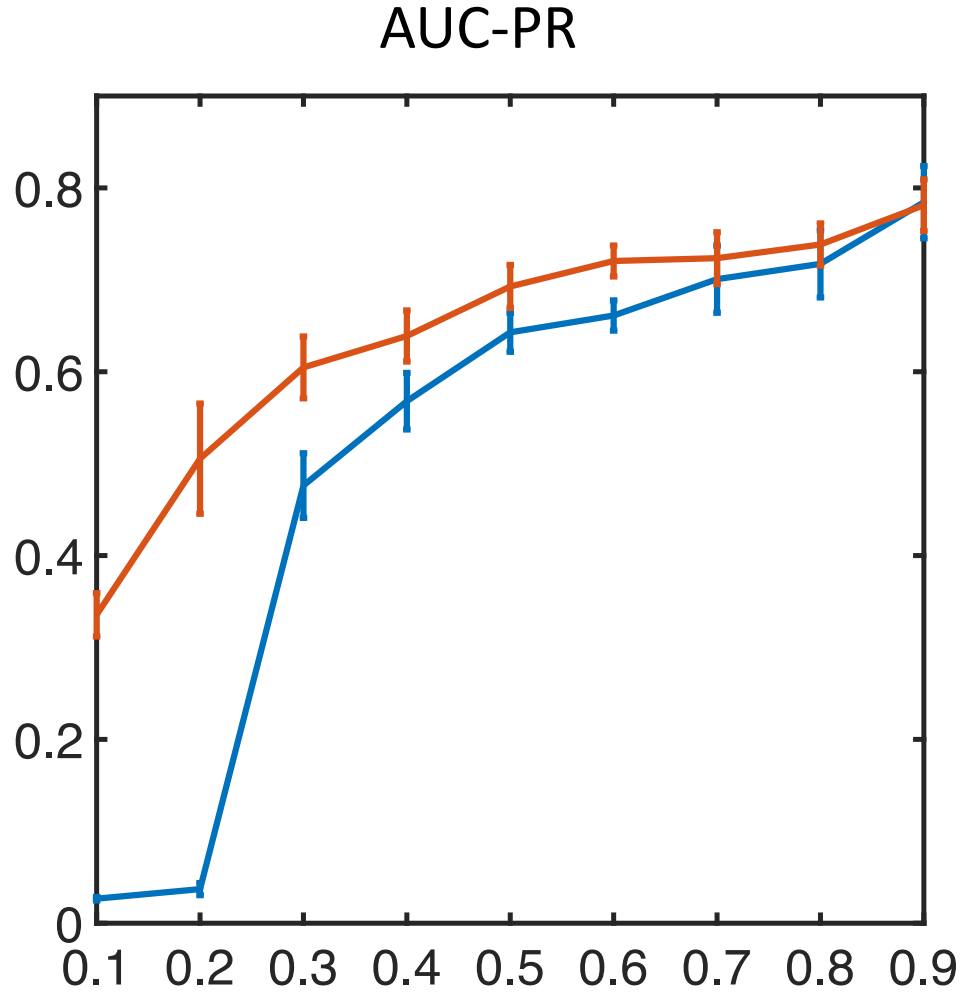


Our model's reconstruction with
20% of the node pairs



Node attribute helps!

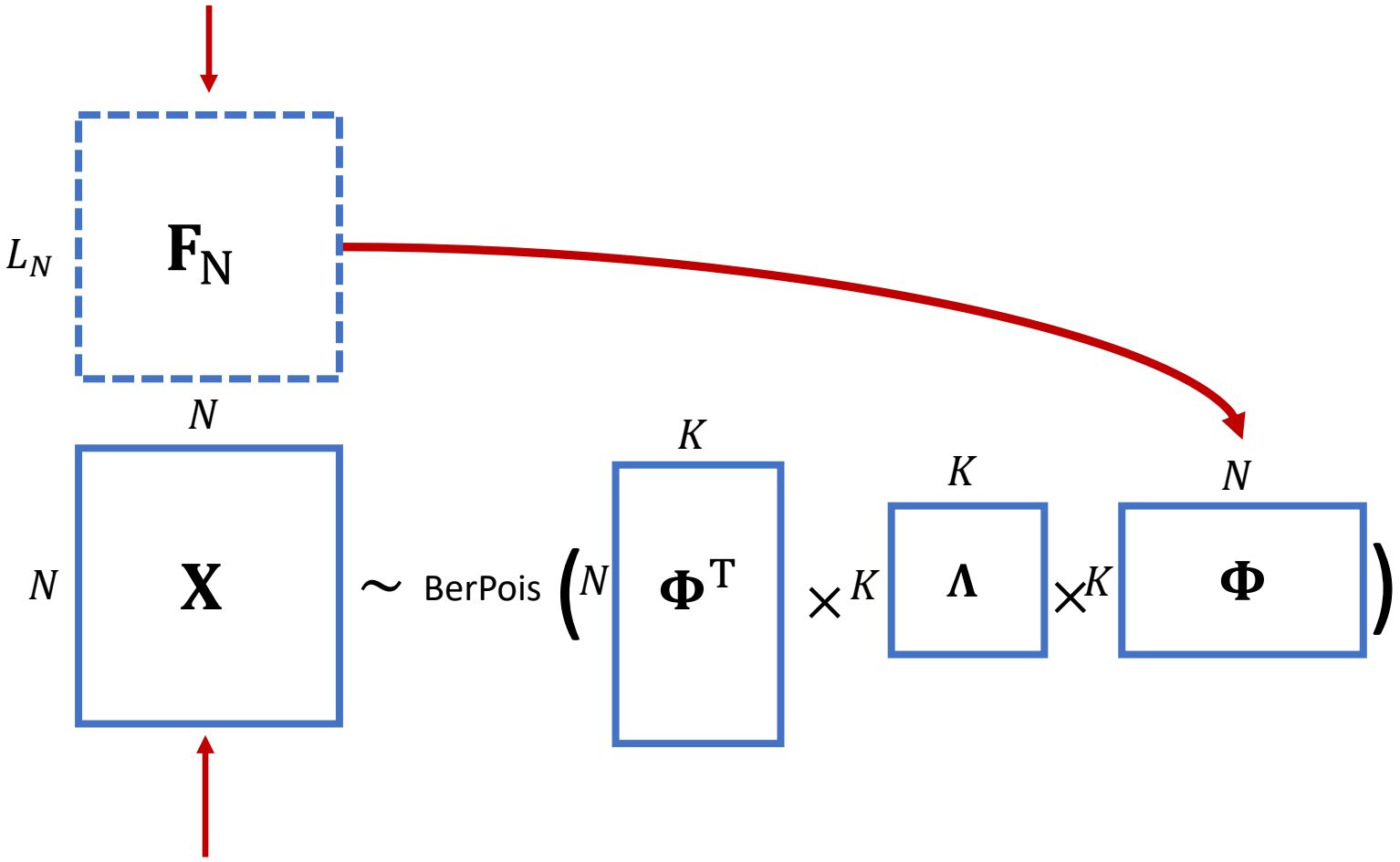




Node Attribute Relational Model (NARM)

(Zhao et.al, ICML 2017)

Binary note attribute matrix



Binary data matrix

EPM

$$\phi_{ki} \sim \mathrm{Gamma}(\alpha_i, 1/c_i)$$
$$\alpha_i \sim \mathrm{Gamma}(e_0, 1/f_0)$$

Ours

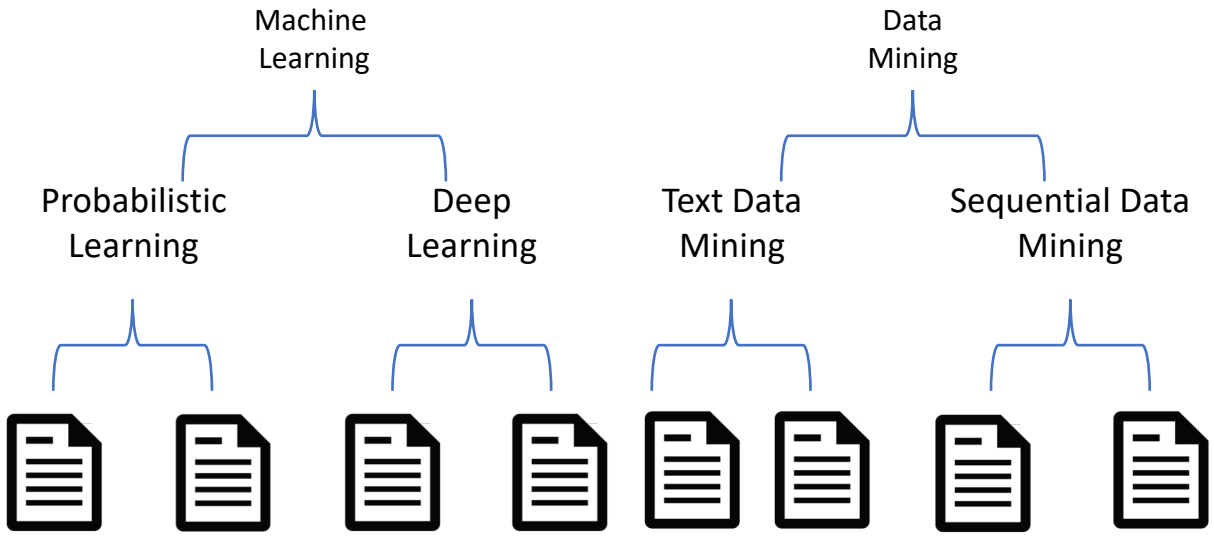
$$\phi_{ki} \sim \mathrm{Gamma}(g_{ki}, 1/c_i)$$
$$g_{ki} = b_k \prod_l^L (h_{kl})^{f_{li}}$$
$$h_{kl}, b_k \sim \mathrm{Gamma}(\mu_0, \mu_0)$$

Regression from binary features to positive latent variables

More on NARM

Modelling

- Simple but effective
- Flexible
  - Gamma-Poisson
  - Dirichlet-multinomial
  - Multi-labels
  - Hierarchical side information



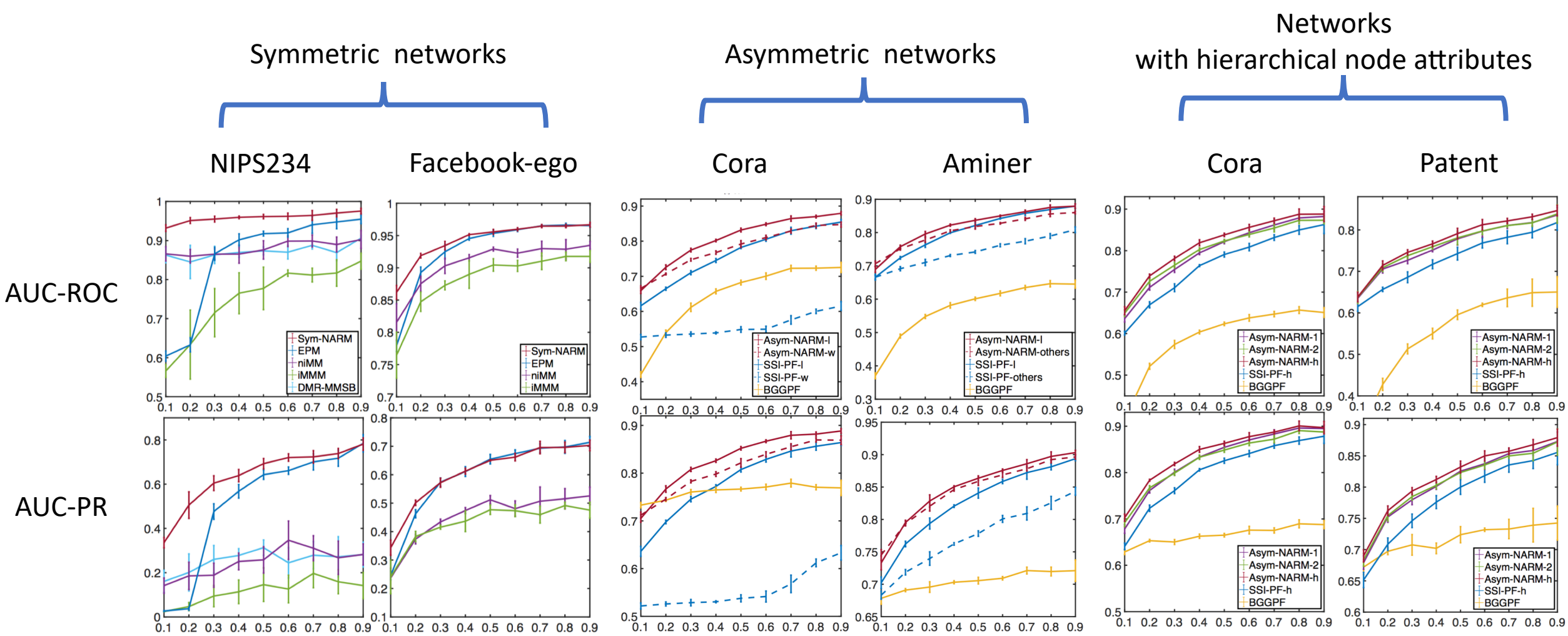
Inference

- Fully local conjugacy with auxiliary variables
- Efficient Gibbs sampling
- Computation only on non-zeros of $\mathbf{X}$ (network) and $\mathbf{F}_N$ (side info)

Our model structure

Bernoulli-Poisson link

More results



Efficiency

Computational Complexity

Model	Complexity
Models with the block matrix	
*NMDR (Kim et al., 2012)	$\mathcal{O}(N^2K + NKL)$
*niMM (Fan et al., 2016)	$\mathcal{O}(N^2K^2 + NKL)$
EPM (Zhou, 2015)	$\mathcal{O}(NK^2D)$
***Sym-NARM**	$\mathcal{O}(NK^2D + D'KL)$
Models without the block matrix	
BGGPF (Zhou et al., 2012)	$\mathcal{O}(NKD)$
*SSI-PF (Hu et al., 2016a)	$\mathcal{O}(NKDL)$
***Asym-NARM**	$\mathcal{O}(NKD + D'KL)$

Sparsity of network

Sparsity of node attributes

Running time per iteration in seconds

Attr	Non-zeros & attr size	**Asym-NARM**	SSI-PF	niMM	DMR-MMSB $K = 50$
Label	2660 2555*10	**0.26**	0.48	134.11	89.12
AT	12775 2555*100	**0.29**	0.87	135.22	126.44
Authors	5647 2555*2597	**0.33**	2.99	136.41	-
All	31273 2555*3058	**0.51**	5.21	136.14	-

Take-away messages of NARM

- Simple, effective, efficient component to incorporate side information
- Flexible to work with many models
- Takes advantage of the sparsity of side information
- Runs faster with better results than previous

- Only work with binary side information

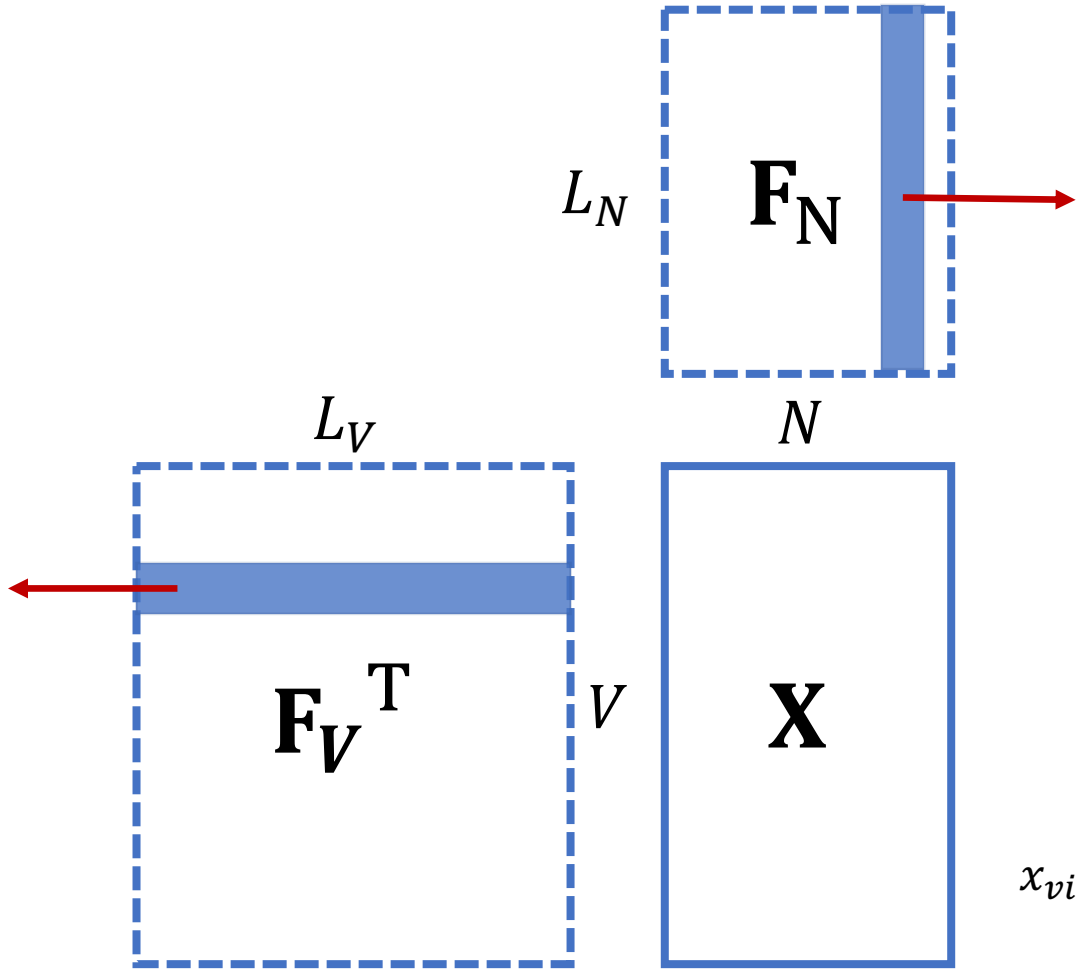
Next …

PMF ⟷ LDA

Networks ⟷ Documents

Topic modelling with side information

Side info for documents



- Author
- Category
- Year
- Venue

- Semantic correlation with other words
- POS tags
- Word embeddings

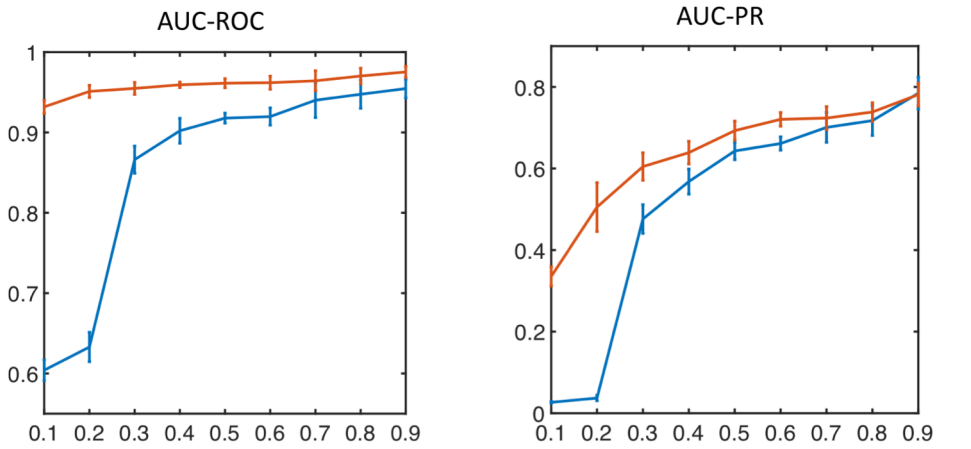
N: the number of documents

V: the vocabulary size

x_{vi}: how many words of v in document i

Short text topic modelling

Node attribute helps!



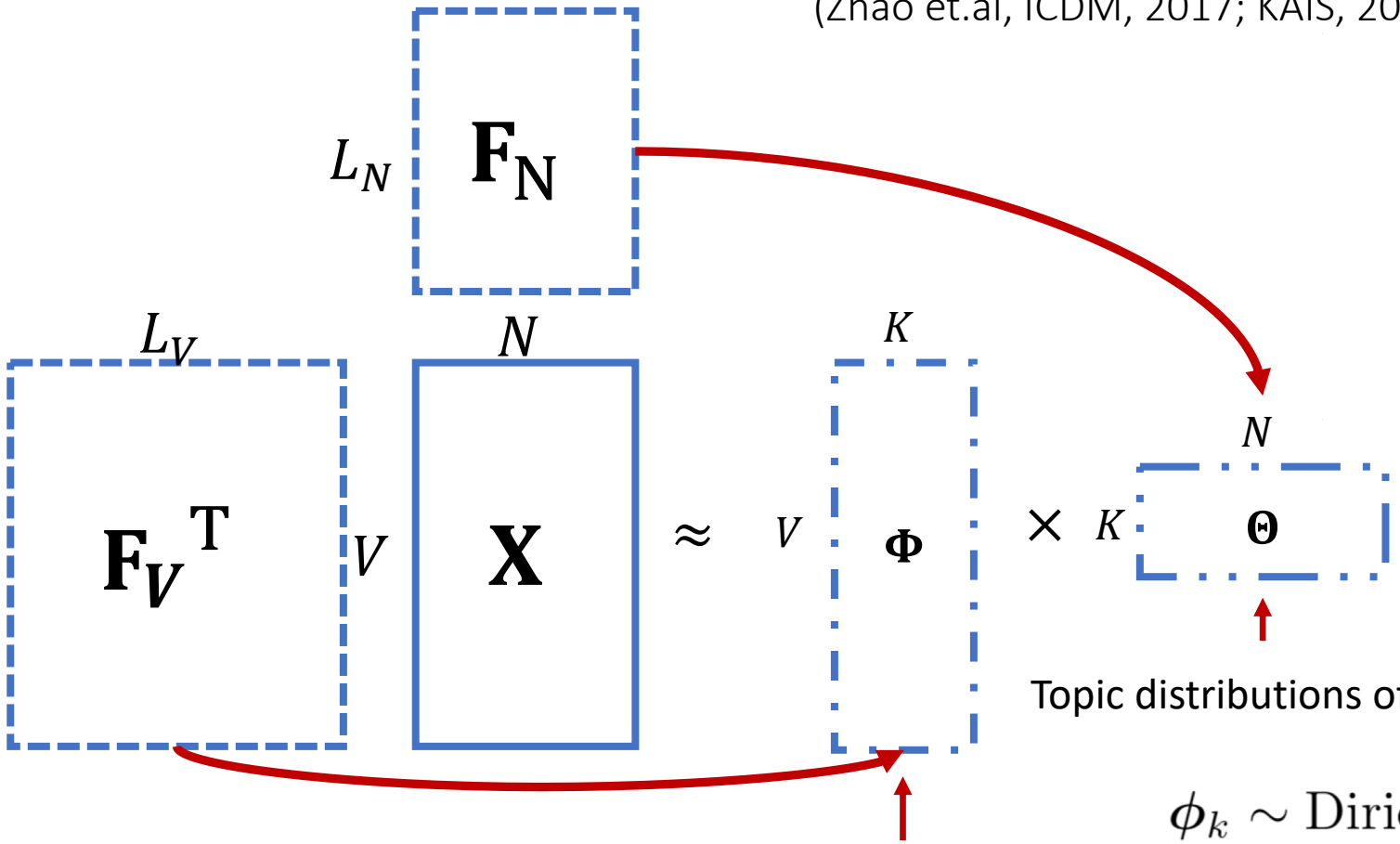
21

Short texts

- Tweets
- News headlines
- Abstracts

MetaLDA

(Zhao et.al, ICDM, 2017; KAIS, 2018)



$$\boldsymbol{\theta}_i \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_i)$$

$$\alpha_{ki} = b_k \prod_{l}^{L_N} (h_{kl})^{f_{li}^N}$$

$$h_{kl}, b_k \sim \mathrm{Gamma}(\mu_0, \mu_0)$$

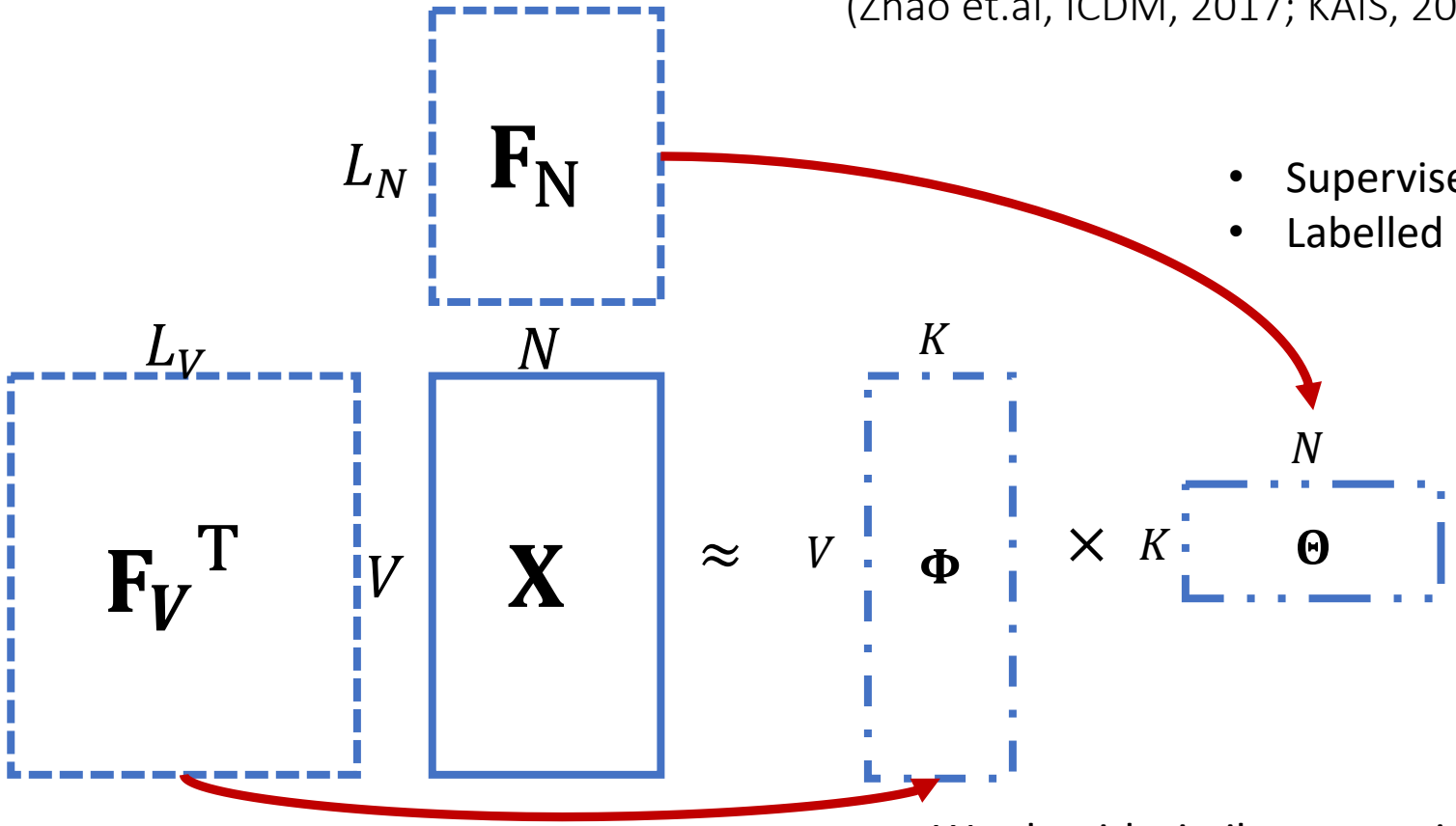
$$\boldsymbol{\phi}_k \sim \mathrm{Dirichlet}(\boldsymbol{\beta}_k)$$

$$\beta_{vk} = b'_k \prod_{l}^{L_V} (h'_{kl})^{f_{lv}^V}$$

$$h'_{kl}, b'_k \sim \mathrm{Gamma}(\nu_0, \nu_0)$$

MetaLDA

(Zhao et.al, ICDM, 2017; KAIS, 2018)



- Supervised Topic Model (Blei, NIPS, 2008)
- Labelled LDA (Ramage, ACL, 2009; KDD 2011)

- Latent feature LDA (Nguyen et al, TACL, 2015)
- Generalised Pó lya-urn Dirichlet multinomial mixture (Li et al, SIGIR, 2016)

Words with similar semantic features -> Similar probabilities of showing up in topics

- If a topic “prefers” a certain word
- We expect that it will also prefer other words with similar features to that word
- Dog -> Puppy

MetaLDA as a supervised topic model

Label	Topic number	Top 5 words
Business	72	exchange stock estate currency trading
	93	trade capital export venture import
	94	jobs marketing job stress advertising
	49	bank financial banking finance insurance
	28	business management services resources solutions
Computers	20	intel device digital apple chip
	66	internet bandwidth speed connection test
	35	computer software engineering architecture graphics
	48	linux operating system unix library
	86	memory computer virtual cache security
Culture & Arts & Entertainment	47	art arts museum painting surrealism
	45	guitar piano jazz orchestra instruments
	7	religion ancient culture roman christian
	41	album tom beatles band julia
	22	culture american chinese history japanese
Education & Science	68	journal journals international conference research
	19	theoretical models model reasoning framework
	81	thesis dissertation technical empirical edu
	15	physics quantum theory mechanics mathematics
	37	research discovery scientific science scientists

Topic number	Top 5 words	Labels
46	programming web java server code	Computers Education&Science Engineering
54	diet calorie nutrition health energy	Health Engineering Business
20	intel device digital apple chip	Computers Culture&Arts&Entertainment Business
17	movie fiction documentary film soundtrack	Culture&Arts&Entertainment Education&Science Sports

Perplexity

How well a model predicts the missing words in the test documents, the lower the better

12,237 docs
vocabulary size 10,052
8 categories

32,597 docs
vocabulary size 13,370
7 categories

Dataset	Web Snippet				TagMyNews			
#Topics	50	100	150	200	50	100	150	200
LDA	961	878	869	888	1969	1873	1881	1916
DMR	845	683	607	562	1750	1506	1391	1323
WF-LDA	894	839	827	842	1853	1766	1830	1854
LF-LDA	1164	1039	1019	992	2415	2393	2371	2374
MetaLDA	**774**	**627**	**572**	**534**	**1657**	**1415**	**1304**	**1235**

No side info ←
Doc side info ←
Word side info
Both ←

Topic coherence

How coherent the top words in a topic, the higher the better

Normalised Pointwise Mutual Information

	All 100 topics			Top 20 topics		
	WS	TMN	AN	WS	TMN	AN
LDA	-0.0030±0.0047	0.0319±0.0032	-0.0636±0.0033	0.1025±0.0067	0.137±0.0043	-0.0010±0.0052
PTM	-0.0029±0.0048	0.0355±0.0016	-0.0640±0.0037	0.1033±0.0081	0.1527±0.0052	0.0004±0.0037
DMR	0.0091±0.0046	0.0396±0.0044	-0.0457±0.0024	0.1296±0.0085	0.1472±0.1507	0.0276±0.0101
LF-LDA	0.0130±0.0052	0.0397±0.0026	-0.0523±0.0023	0.1230±0.0153	0.1456±0.0087	0.0272±0.0042
WF-LDA	0.0091±0.0046	0.0390±0.0051	-0.0457±0.0024	0.1296±0.0085	0.1507±0.0055	0.0276±0.0101
GPU-DMM	-0.0934±0.0106	-0.0970±0.0034	-0.0769±0.0012	0.0836±0.0105	0.0968±0.0076	-0.0613±0.0020
MetaLDA	**0.0311**±0.0038	**0.0451**±0.0034	**-0.0326**±0.0019	**0.1511**±0.0093	**0.1584**±0.0072	**0.0590**±0.0065

Efficiency

Conjugacy

Sparsity of
side information

On the Reuters dataset with 100 topics:

- 11,367 docs
- Vocabulary size: 8,817
- 120 labels

MetaLDA with both doc labels and 50 dimensional GloVe word embeddings

- 2× faster than DMR (2008) with doc labels only
- 10× faster than LF-LDA (2015) with word embeddings only
- 3× faster than WF-LDA (2010) with word embeddings only
- 10× faster than PTM (2016), a short text topic model

Runs in parallel

MetaLDA runs in parallel with multi-threads on splits of documents

On the NYT dataset with 500 topics:

- 52,521 docs
- Vocabulary size: 21,421
- 545 labels

- 16 seconds per iteration with 10 threads on a low-speed Xeon desktop

Take-away messages of MetaLDA

- MetaLDA: a topic model that efficiently incorporates side information
- Better perplexity, topic quality, and running speed
- Large improvement for short texts

- Only work with binary side information
- Especially annoying for word embeddings

Deal with real-valued word embeddings (WETM)

(Zhao et.al, ICML, 2018)

MetaLDA

$$\boldsymbol{\phi}_k \sim \text{Dirichlet}(\boldsymbol{\beta}_k)$$

$$\beta_{vk} = b'_k \prod_l^{L_V} (h'_{kl})^{f_{lv}^V}$$

→ Binary word embeddings

$$h'_{kl}, b'_k \sim \text{Gamma}(\nu_0, \nu_0)$$

The new model

$$\boldsymbol{\phi}_k \sim \text{Dirichlet}(\boldsymbol{\beta}_k)$$

$$\beta_{vk} \sim \text{Gamma}(\beta_0, e^{\boldsymbol{f}_v^\top \boldsymbol{w}_k})$$

→ Real-valued word embeddings

WETM

Regression from real-valued features to positive normalised latent variables

Word distribution of topic k

$$\phi_k \sim \text{Dirichlet}(\beta_k)$$

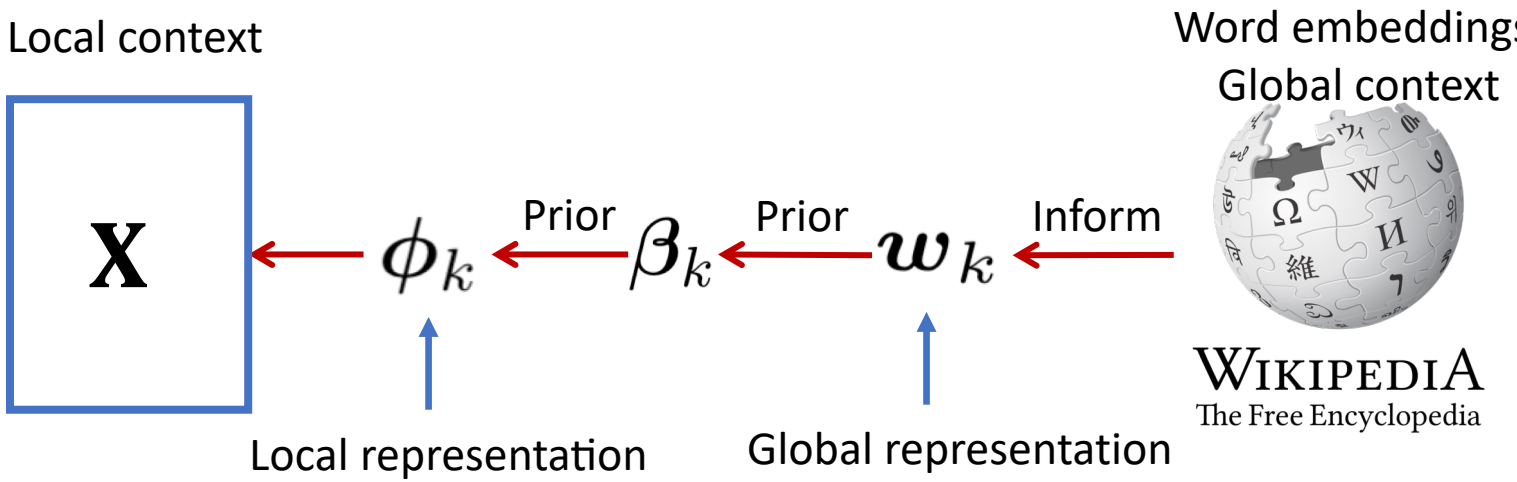
$$\beta_{vk} \sim \text{Gamma}(\beta_0, e^{\boldsymbol{f}_v^\top \boldsymbol{w}_k})$$

Prior of word distribution of topic k

Topic embeddings of k

- Same dimensions with word embeddings
- A real-valued vector drawn from Gaussian

Word embeddings of v

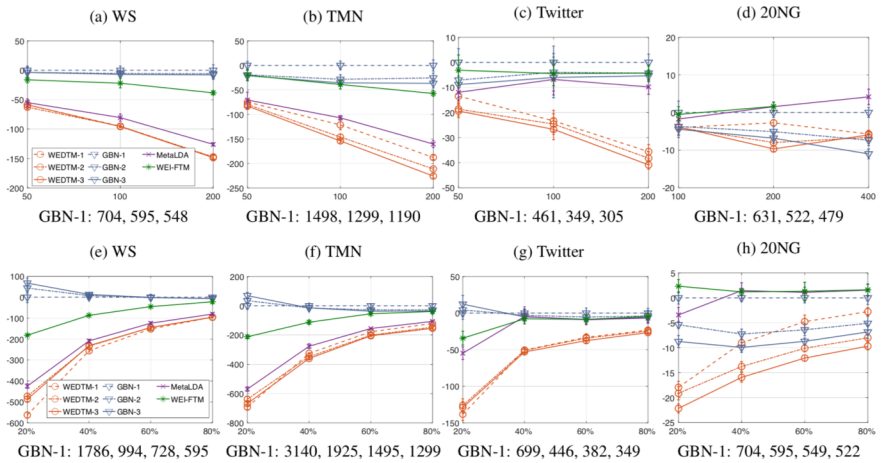


Representations of topics

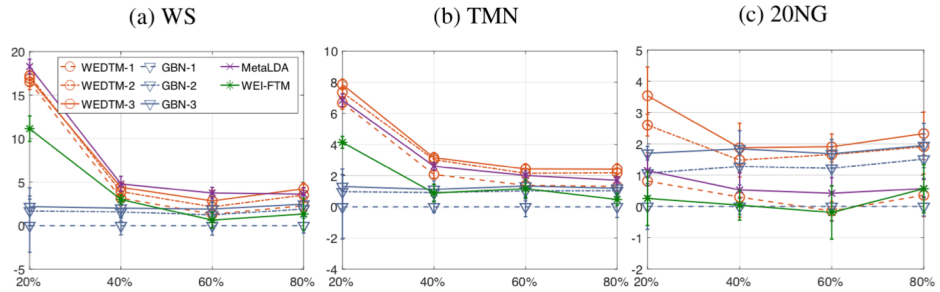
Representation	Top 10 words	NPMI
Local	cancer lung tobacco information health smoking treatment gov research symptoms	0.050
Global	cancer breast diabetes pulmonary cancers patients asthma cardiovascular cholesterol obesity	0.050
Local	art awards oscar academy gallery museum surrealism sculpture picasso arts	0.076
Global	paintings awards award art museum gallery sculpture painting picasso portrait	0.087
Local	security network wireless access networks spam spyware networking national computer	0.061
Global	wireless internet networks devices phone broadband users network wi-fi providers	0.143

WETM works well

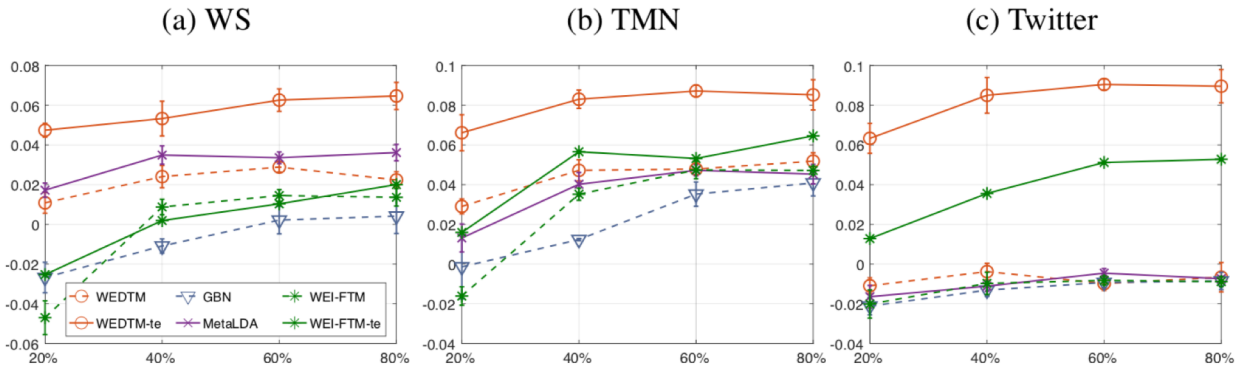
Perplexity



Document classification accuracy



Topic coherence



Topics are not semantically indivisible …

Topics can mix the words which co-occur locally in the target corpus but are less semantically related in general

Intra-topic structure

1	journal science biology research journals international cell psychology scientific bioinformatics
2	fitness piano guitar swimming violin weightlifting lessons training swim weight

- Journal
- Biology

- Fitness
- Music

Local context

X

Word embeddings:
Global context

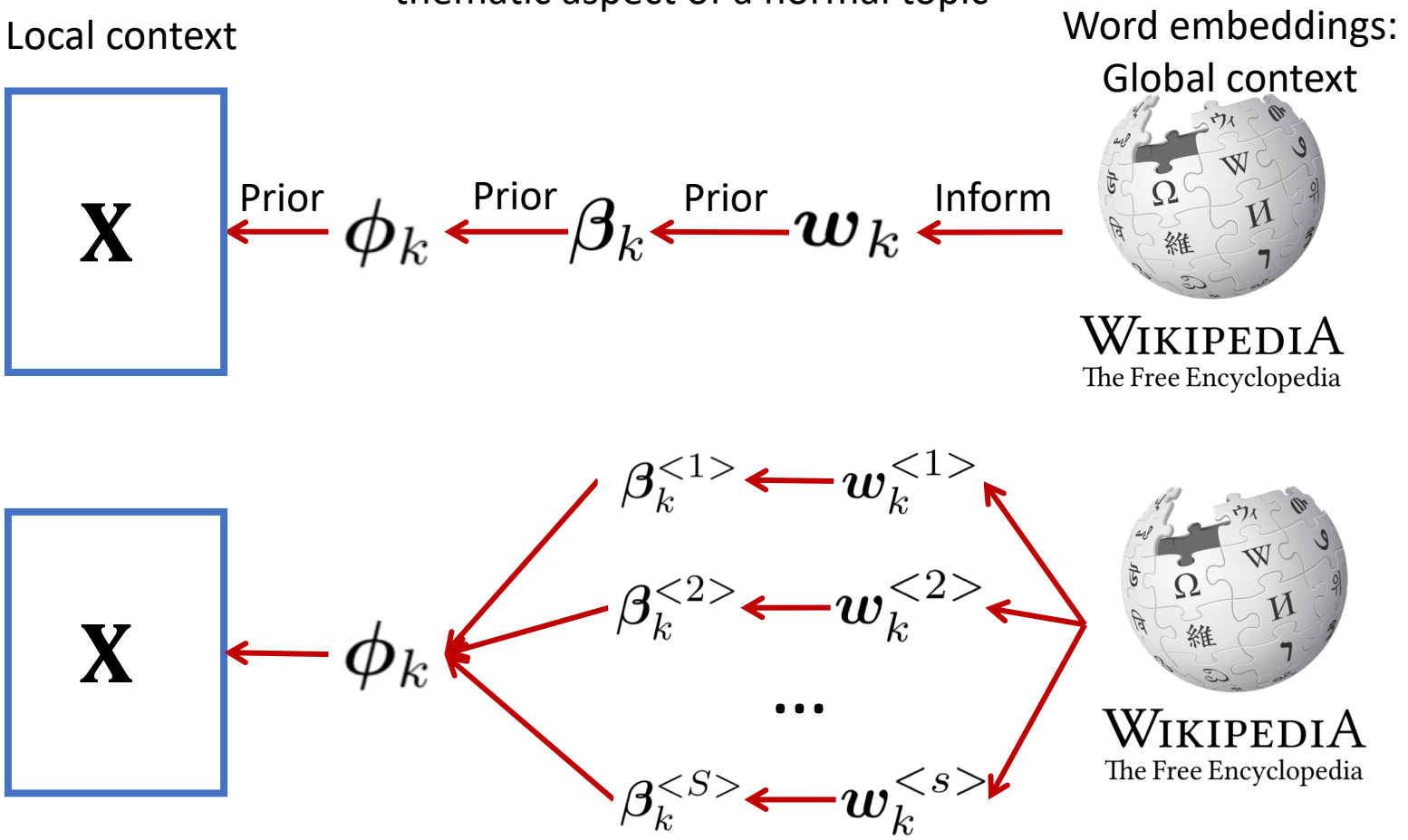
WIKIPEDIA
The Free Encyclopedia

Discover intra-topic structure with word embeddings

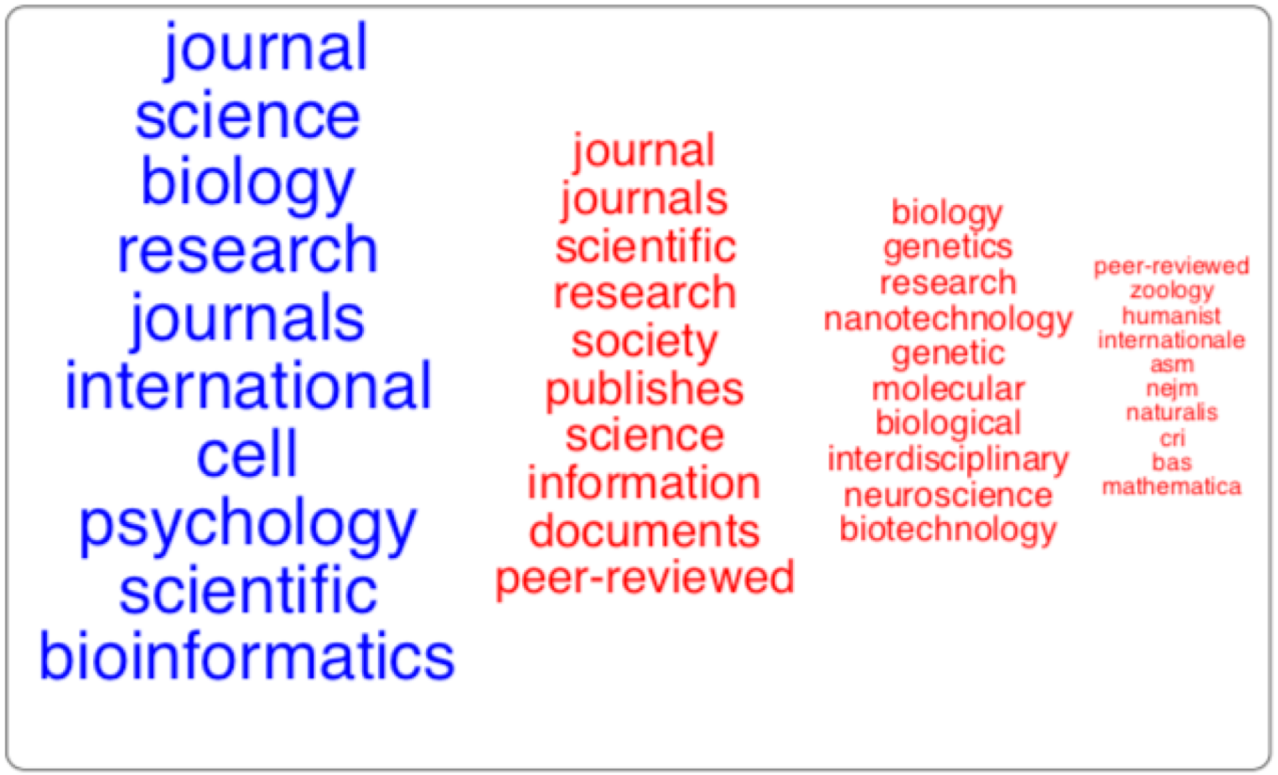
(Zhao et.al, ICML, 2018)

Sub-topics

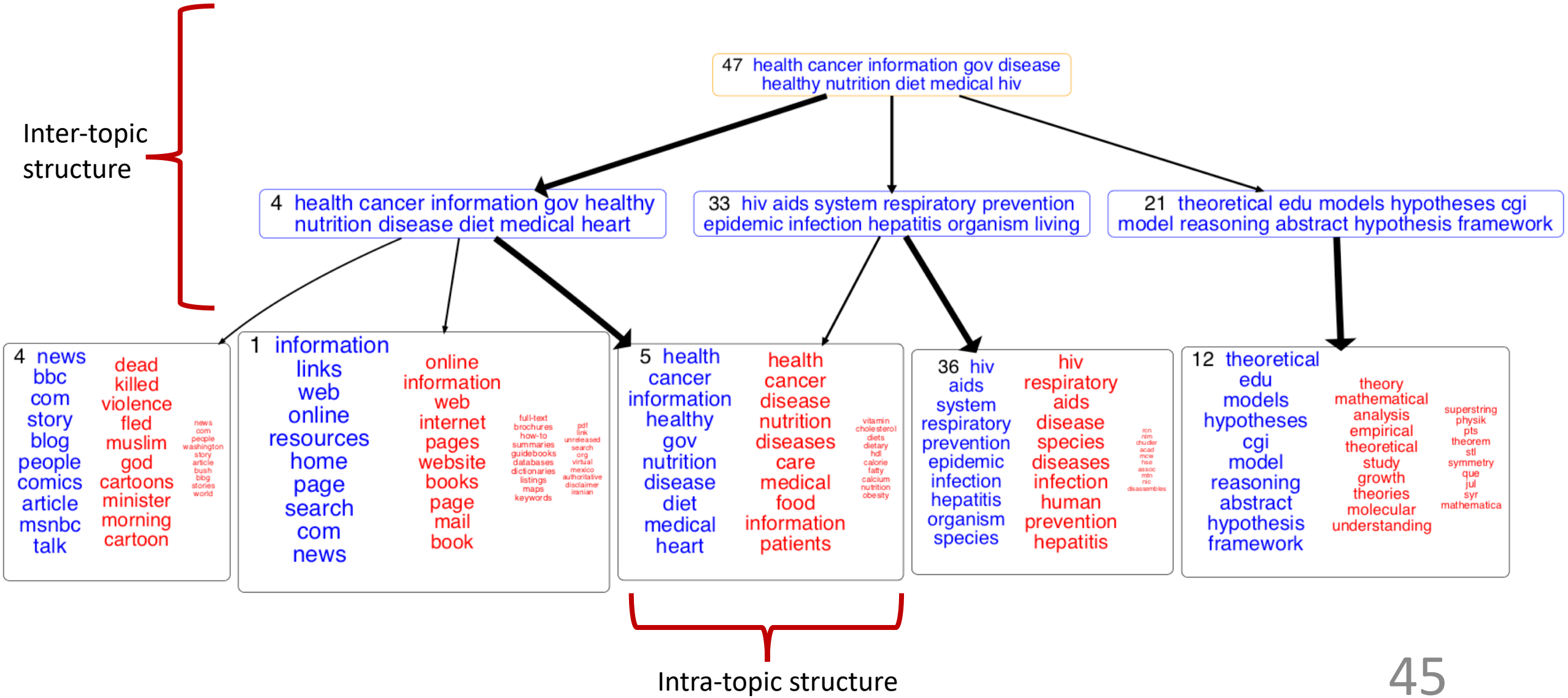
For each normal topic, there is a set of sub-topics, each of which is informed by word embeddings and captures a fine-grained thematic aspect of a normal topic



Discover intra-topic structure with word embeddings



A full picture of topic structure learning with word embeddings in WETM



Take-away messages of WETM

- Work with real-valued word embeddings
- More flexible to discover inter and intra topic structures

- Runs not as fast as MetaLDA

Summary of our recent work

Poisson matrix factorisation with side information

- Node Attributes Relational Model (NARM)
  - Binary node attributes
- MetaLDA
  - Binary document and word side information

- WETM
  - Real-valued word embeddings
  - Topic structure

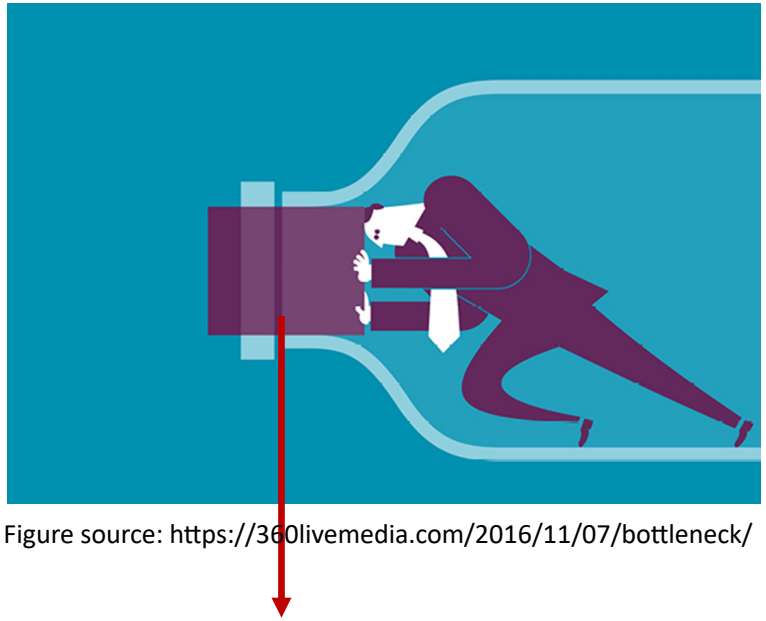


Figure source: https://360livemedia.com/2016/11/07/bottleneck/

Scalability of Gibbs sampling

Code available at https://ethanhezhao.github.io

Future work

Future work

Deep generative models

- Variational auto-encoder (VAE)
  - Topic models with VAE (Miao et al, ICML, 2017)
  - Matrix factorisations with VAE (Liang et al, WWW, 2017)

Scalable inference

- Batch training with stochastic gradient descent
  - Stochastic variational inference
  - Stochastic gradient MCMC

+

TensorFlow

Questions?

Figure source: http://milewalk.com/mwblog/4-worst-job-interview-questions/