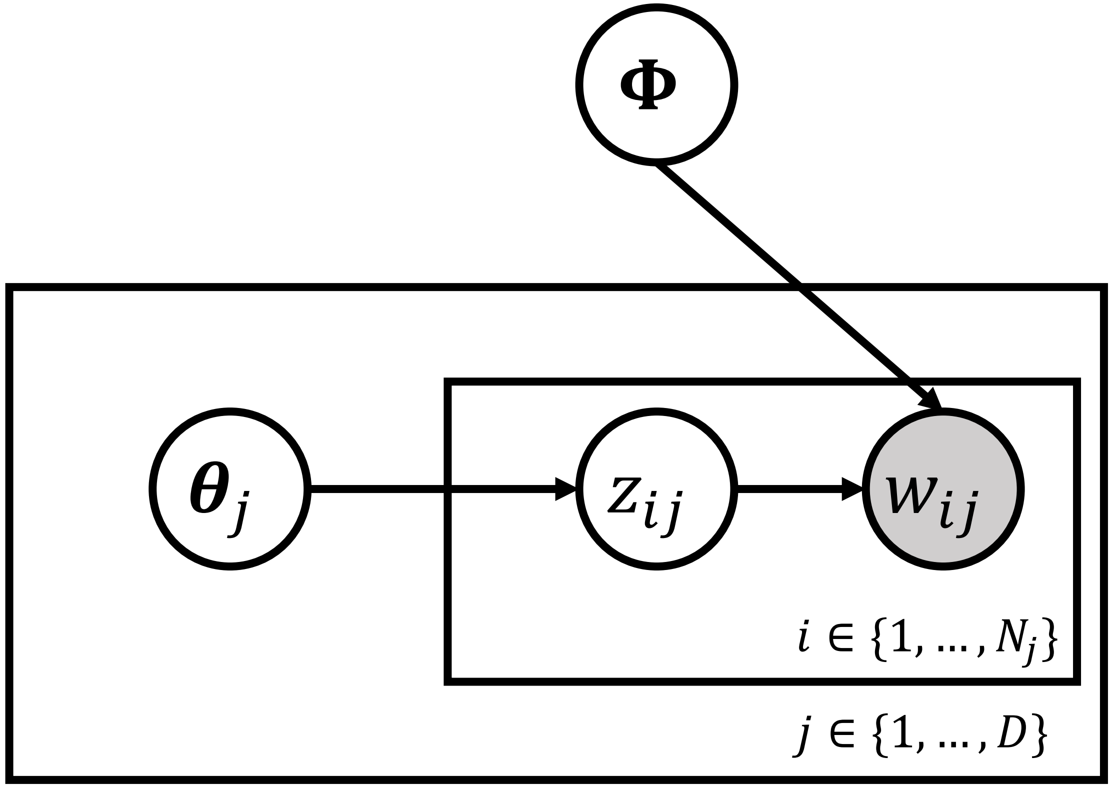
Topic Structure Learning for Interpretable Text Understanding

He (Ethan) Zhao
PhD Student
Monash University

Topic modelling

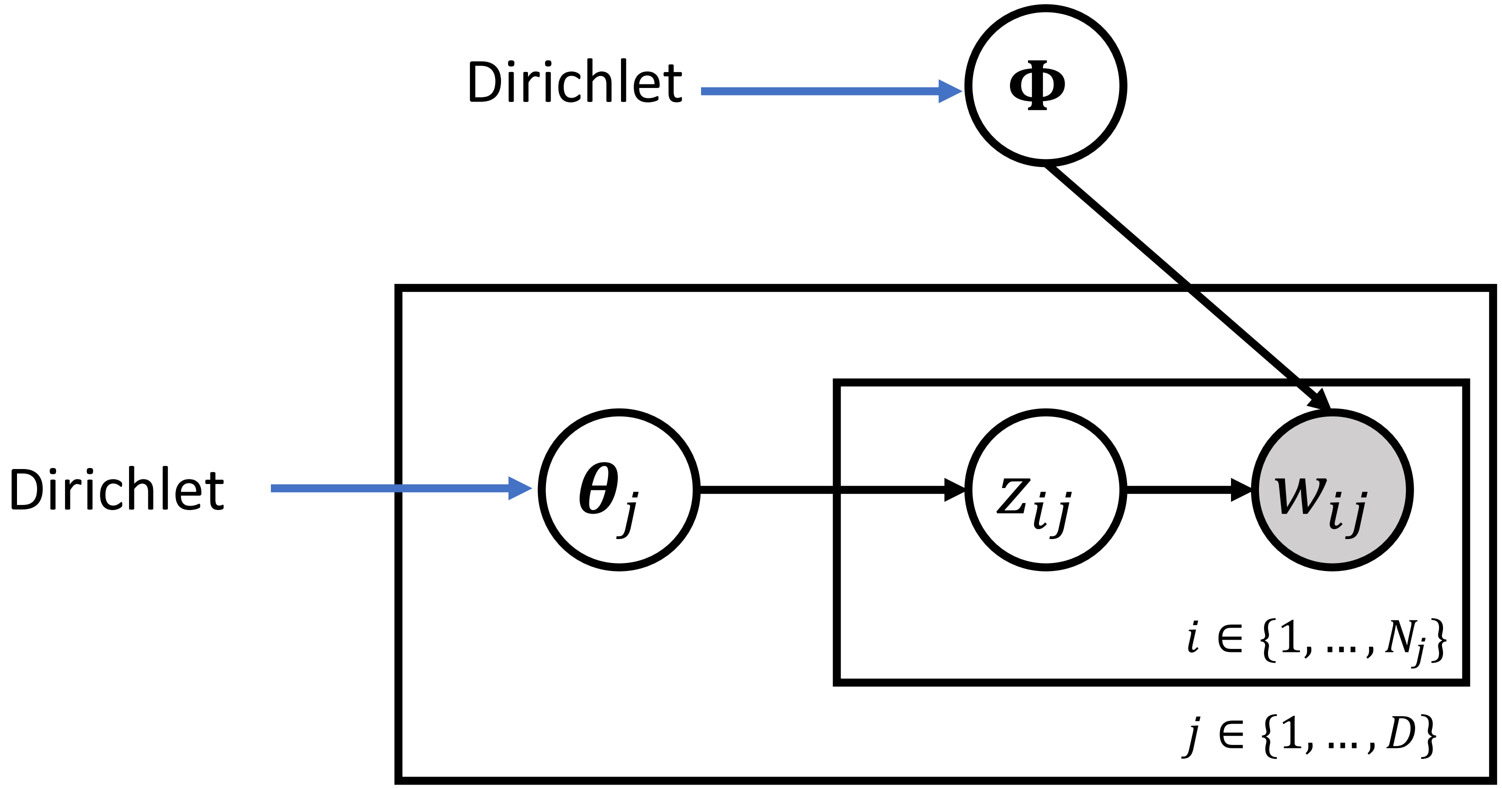


Prior on θ and ϕ:

- LDA: Dirichlet on θ and ϕ
- HDPLDA: HDP on θ
- Correlated topic model: logistic normal on θ
- Focused topic model: IBP on θ
- …

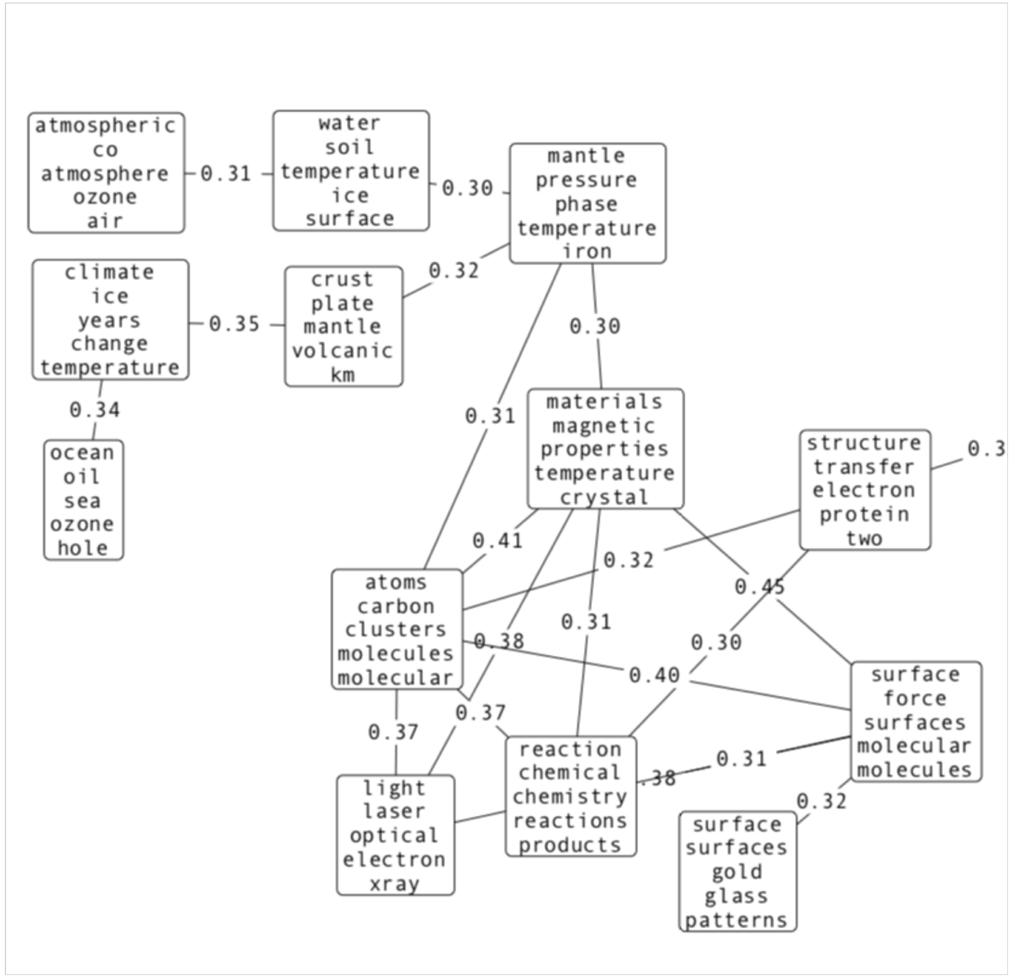
Latent Dirichlet Allocation

Topics are assumed to be independent

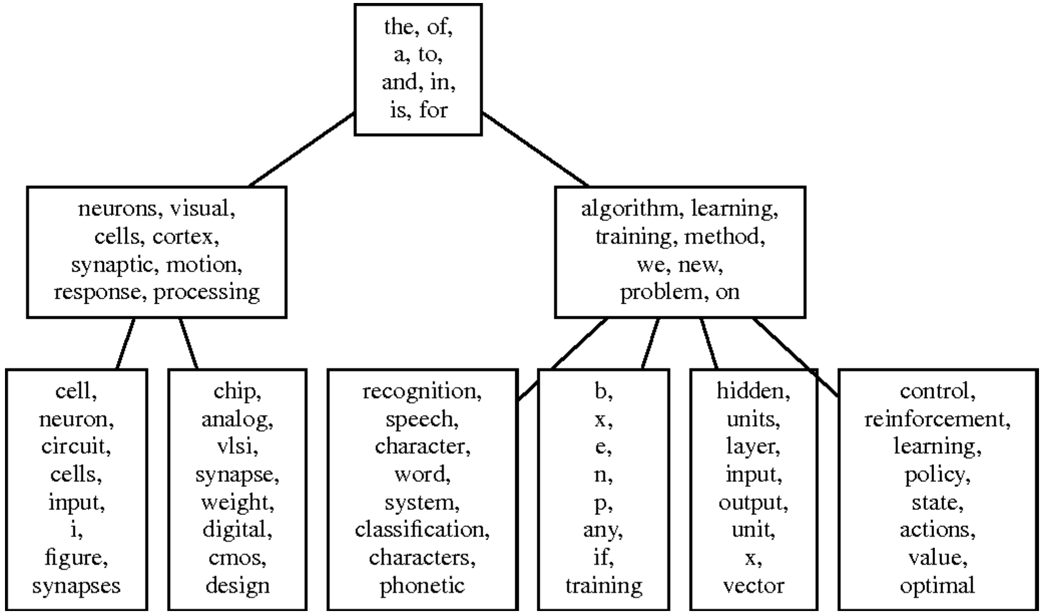


Topic structure learning

Topics are naturally semantically dependent



Correlated Topic Model (CTM, 2006, NIPS)

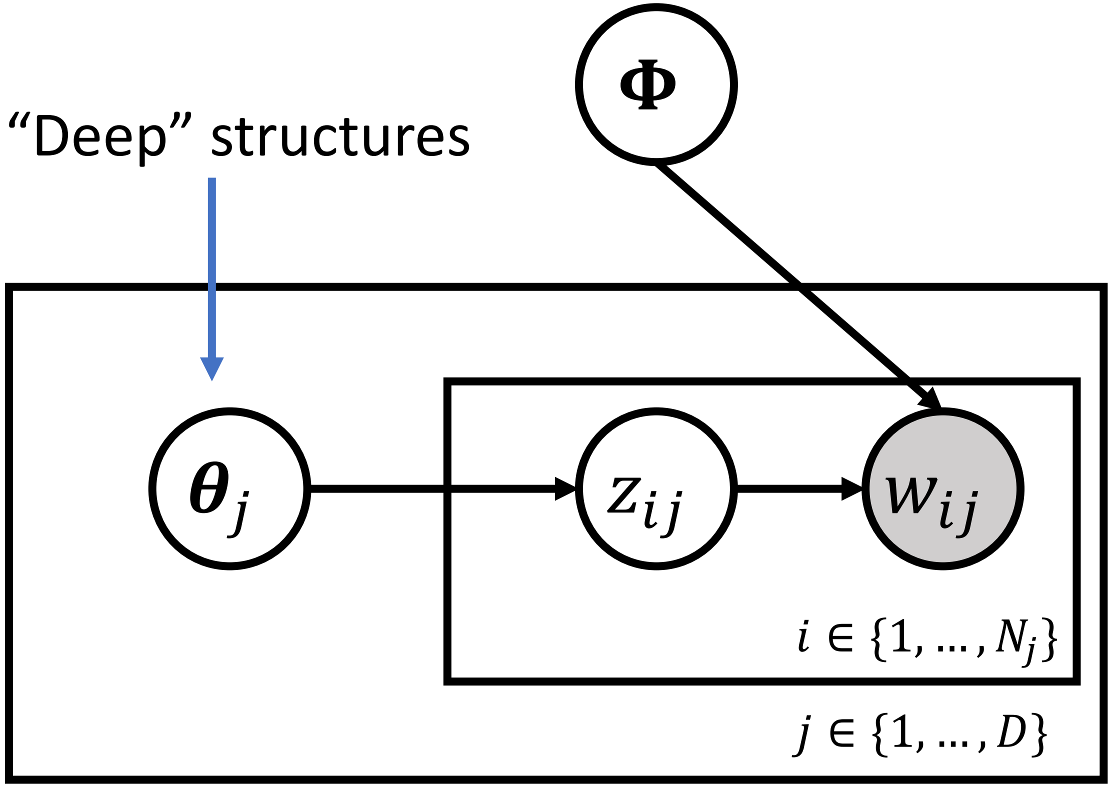


Nested Chinese Restaurant Process (nCRP, JACM, 2010)

Part One

Recent "deep" models that learn topic structures

- Deep Poisson Factor Analysis (ICML, 2015)
- Deep Poisson Factor Modeling (NIPS, 2015)
- Gamma Belief Networks (GBN, JMLR, 2016)
- …



Gamma Belief Networks

$$\boldsymbol{\theta}_j^{(T)} \sim \text{Gam}\left(\boldsymbol{r}, 1/c_j^{(T+1)}\right),$$
$$\cdots$$
$$\boldsymbol{\theta}_j^{(t)} \sim \text{Gam}\left(\boldsymbol{\Phi}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)}, 1/c_j^{(t+1)}\right),$$
$$\cdots$$
$$\underbrace{\boldsymbol{x}_j^{(1)}}_{V\times 1} \sim \text{Pois}\left(\underbrace{\boldsymbol{\Phi}^{(1)}}_{V\times K_1}\underbrace{\boldsymbol{\theta}_j^{(1)}}_{K_1\times 1}\right), \quad \boldsymbol{\theta}_j^{(1)} \sim \text{Gam}\left(\underbrace{\boldsymbol{\Phi}^{(2)}}_{K_1\times K_2}\underbrace{\boldsymbol{\theta}_j^{(2)}}_{K_2\times 1}, p_j^{(2)}/(1-p_j^{(2)})\right).$$

V×1, Bag-of-words of document j

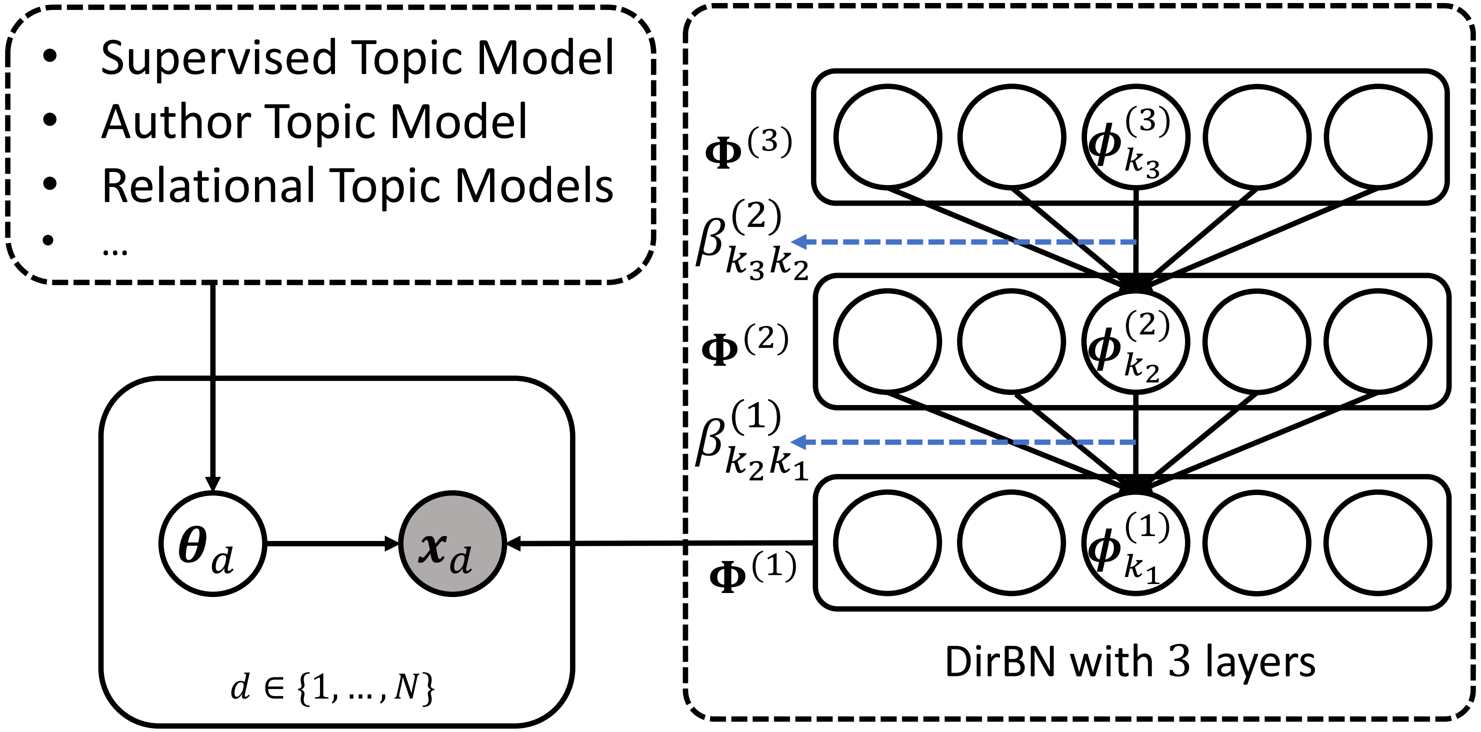
GBN

- It factorises the local variable θ hierarchically
- It abstracts the semantic information in a document layer-wisely
- $\boldsymbol{\Phi}^{(2)}$ is the correlations of the first and second-layer topics

😭

- The higher-layer topics are not directly interpretable
- It is complex to combine with many other topic models on θ
- If a document is short, e.g. tweets and news headlines, it is hard to learn good topics

Dirichlet Belief Networks (DirBN, NIPS, 2018)

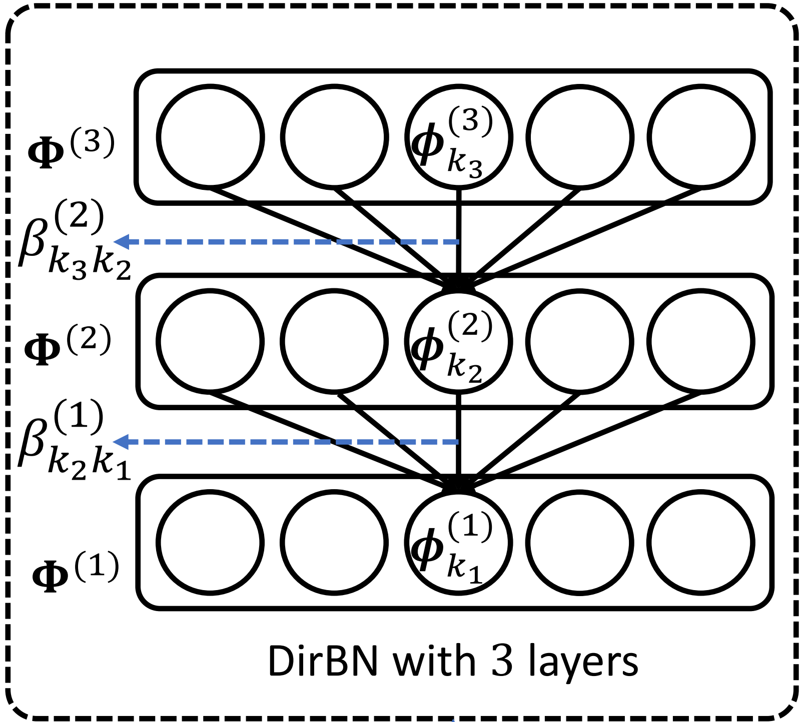


DirBN

$$
\begin{aligned}
\boldsymbol{\phi}_{k_T}^{(T)} &\sim \mathrm{Dir}_V(\eta), \\
&\cdots \\
\boldsymbol{\phi}_{k_t}^{(t)} &\sim \mathrm{Dir}_V(\boldsymbol{\psi}_{k_t}^{(t)}), \boldsymbol{\psi}_{k_t}^{(t)} = \sum_{k_{t+1}}^{K_{t+1}} \boldsymbol{\phi}_{k_{t+1}}^{(t+1)} \beta_{k_{t+1}k_t}^{(t)}, \beta_{k_{t+1}k_t}^{(t)} \sim \mathrm{Ga}(\gamma_{k_{t+1}}^{(t)}, 1/c^{(t)}), \\
&\cdots \\
\boldsymbol{\phi}_{k_1}^{(1)} &\sim \mathrm{Dir}_V(\boldsymbol{\psi}_{k_1}^{(1)}),\ \boldsymbol{\psi}_{k_1}^{(1)} = \sum_{k_2}^{K_2} \boldsymbol{\phi}_{k_2}^{(2)} \beta_{k_2k_1}^{(1)}, \beta_{k_2k_1}^{(1)} \sim \mathrm{Ga}(\gamma_{k_2}^{(1)}, 1/c^{(1)}).
\end{aligned}
$$

Inference/Learning of DirBN

1. Propagating the latent counts from the bottom up



$$\boldsymbol{x}^{(1)}_{k_1} = [x^{(1)}_{1k_1}, \cdots, x^{(1)}_{Vk_1}]$$

2. Updating the latent variables from the top down

Propagating the latent counts from the bottom up

$$\boldsymbol{\phi}_{k_1}^{(1)} \sim \mathrm{Dir}_V(\boldsymbol{\psi}_{k_1}^{(1)}),\ \boldsymbol{\psi}_{k_1}^{(1)} = \sum_{k_2}^{K_2} \boldsymbol{\phi}_{k_2}^{(2)} \beta_{k_2 k_1}^{(1)},\ \beta_{k_2 k_1}^{(1)} \sim \mathrm{Ga}(\gamma_{k_2}^{(1)}, 1/c^{(1)}).$$

$$\boldsymbol{x}_{k_1}^{(1)} = [x_{1k_1}^{(1)}, \cdots, x_{Vk_1}^{(1)}]$$

$$\mathcal{L}\left(\boldsymbol{\psi}_{k_1}^{(1)}\right) \propto \frac{\Gamma(\psi_{\cdot k_1}^{(1)})}{\Gamma(\psi_{\cdot k_1}^{(1)} + x_{\cdot k_1}^{(1)})} \prod_{v}^{V} \frac{\Gamma(\psi_{vk_1}^{(1)} + x_{vk_1}^{(1)})}{\Gamma(\psi_{vk_1}^{(1)})}.$$

$$q_{k_1}^{(1)} \sim \mathrm{Beta}(\psi_{\cdot k_1}^{(1)}, x_{\cdot k_1}^{(1)})$$

$$y_{vk_1}^{(1)} \sim \mathrm{CRT}\left(x_{vk_1}^{(1)}, \psi_{vk_1}^{(1)}\right)$$

$$\mathcal{L}\left(\boldsymbol{\psi}_{k_1}^{(1)}, q_{k_1}^{(1)}, y_{vk_1}^{(1)}\right) \propto \prod_{v}^{V} \left(q_{k_1}^{(1)}\right)^{\psi_{vk_1}^{(1)}} \left(\psi_{vk_1}^{(1)}\right)^{y_{vk_1}^{(1)}}$$

$$\psi_{vk_1}^{(1)} = \sum_{k_2}^{K_2} \phi_{vk_2}^{(2)} \beta_{k_2 k_1}^{(1)}$$

$$\left(z_{v1k_1}^{(1)}, \cdots, z_{vK_2k_1}^{(1)}\right) \sim \mathrm{Mult}\left(y_{vk_1}^{(1)}, \frac{\phi_{v1}^{(2)} \beta_{1k_1}^{(1)}}{\psi_{vk_1}^{(1)}}, \cdots, \frac{\phi_{vK_2}^{(2)} \beta_{K_2k_1}^{(1)}}{\psi_{vk_1}^{(1)}}\right)$$

$$x_{vk_2}^{(2)} = \sum_{k_1}^{K_1} z_{vk_2k_1}^{(1)}$$

$$\boldsymbol{x}_{k_2}^{(2)} = [x_{1k_2}^{(2)}, \cdots, x_{Vk_2}^{(2)}]$$

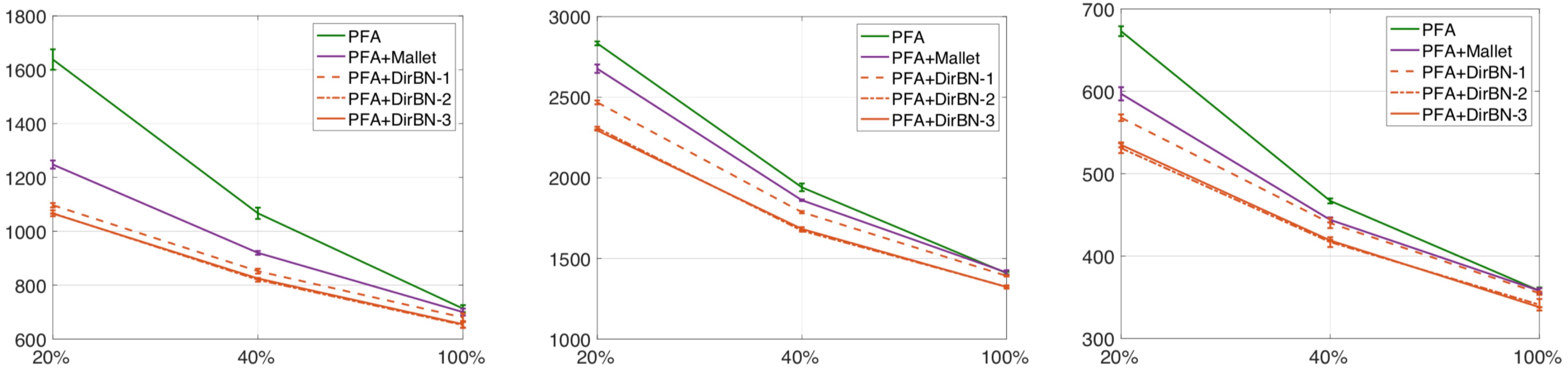
Updating the latent variables from the top down

$$\boldsymbol{\phi}_{k_t}^{(t)} \sim \mathrm{Dir}(\boldsymbol{\psi}_{k_t}^{(t)} + \boldsymbol{x}_{k_t}^{(t)})$$

$$\beta_{k_{t+1}k_t}^{(t)} \sim \mathrm{Ga}\left(\gamma_{k_{t+1}}^{(t)} + z_{\cdot k_{t+1}k_t}^{(t)}, 1.0\right) / \left(c^{(t)} - \log q_{k_t}^{(t)}\right)$$

Numerical results

Perplexity

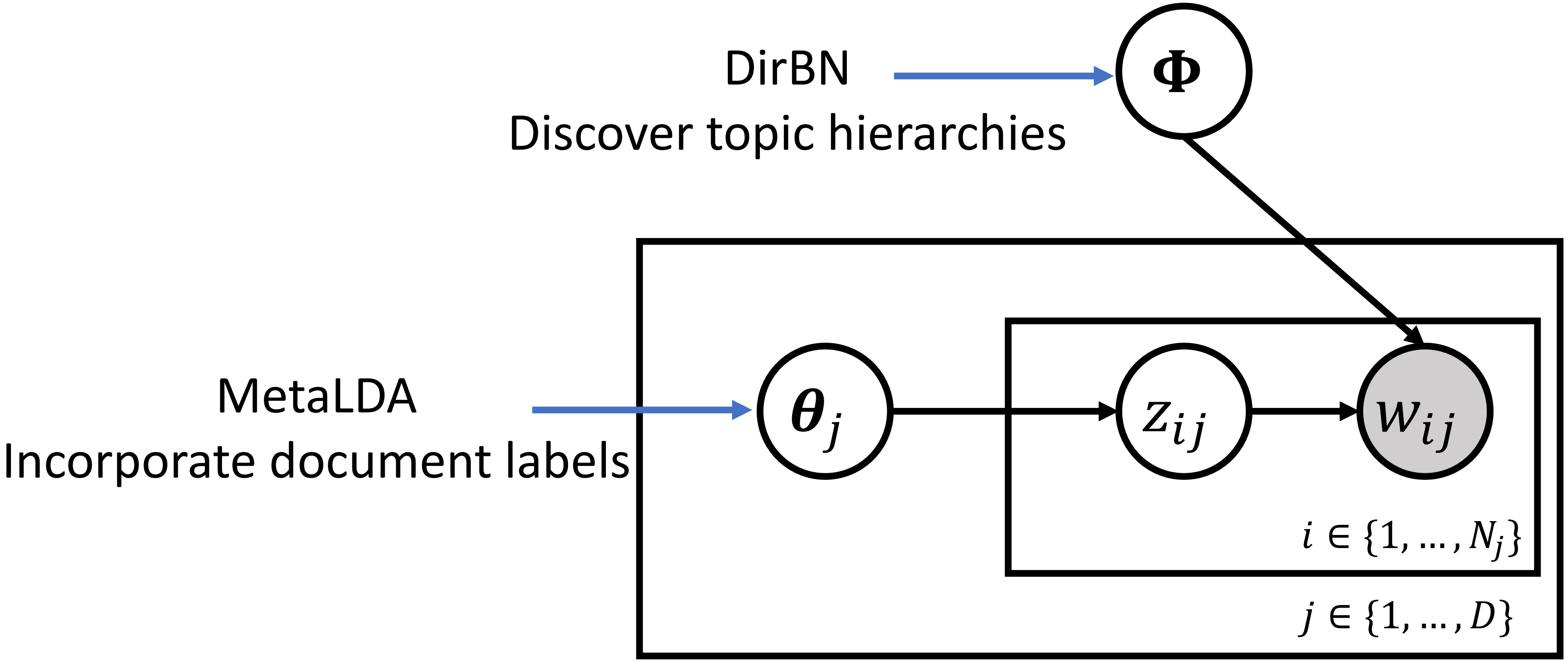


(a) (b) (c)

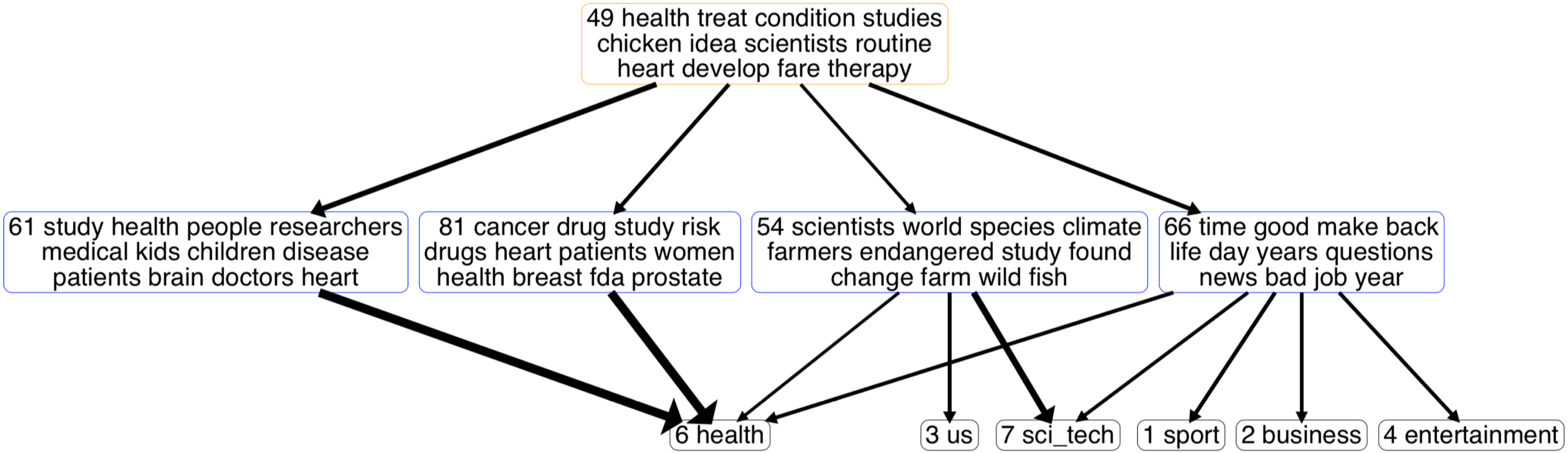
Topic coherence

	WS			TMN			Twitter		
Training words	20%	40%	100%	20%	40%	100%	20%	40%	100%
PFA	-0.070±0.010	0.008±0.002	0.062±0.011	-0.059±0.008	0.064±0.009	0.103±0.006	-0.003±0.003	0.031±0.003	0.046±0.002
PFA+Mallet	0.008±0.004	0.049±0.005	0.063±0.003	0.035±0.006	0.083±0.005	0.108±0.005	0.022±0.003	0.037±0.002	0.045±0.003
PFA+DirBN-1	0.013±0.003	0.052±0.004	0.060±0.006	0.031±0.003	0.080±0.001	0.108±0.008	0.019±0.004	0.037±0.004	0.049±0.007
PFA+DirBN-3	**0.021**±0.005	0.059±0.002	0.068±0.004	0.046±0.003	0.090±0.003	0.111±0.004	0.024±0.001	0.038±0.002	0.049±0.002
GBN	-0.072±0.013	0.007±0.005	0.069±0.009	-0.065±0.008	0.063±0.006	0.106±0.004	-0.005±0.005	0.032±0.002	0.047±0.00
GBN+DirBN-1	0.015±0.005	0.057±0.002	0.069±0.005	0.032±0.002	0.086±0.002	0.112±0.007	0.021±0.004	0.040±0.005	0.050±0.005
GBN+DirBN-3	0.018±0.006	**0.061**±0.004	**0.075**±0.002	0.048±0.003	**0.094**±0.004	**0.113**±0.004	**0.025**±0.003	**0.040**±0.002	**0.051**±0.003

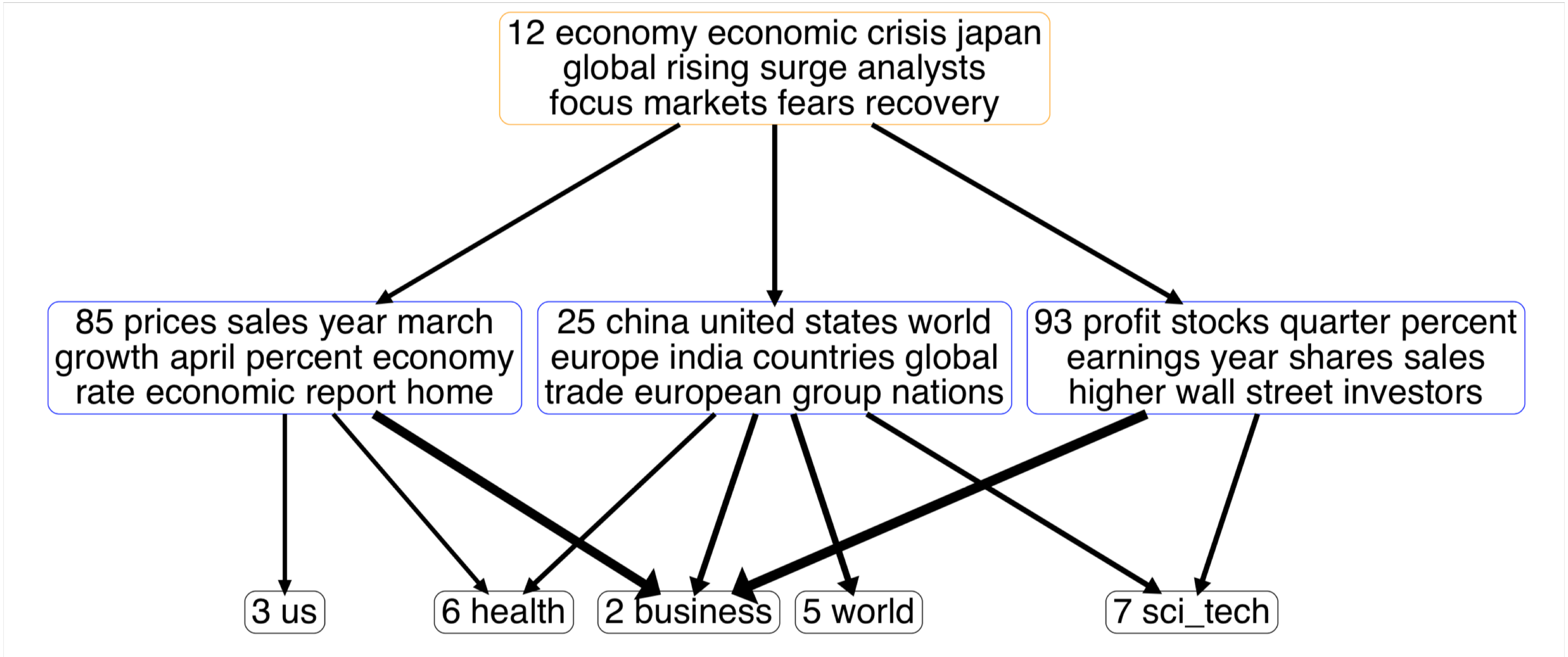
MetaLDA+DirBN



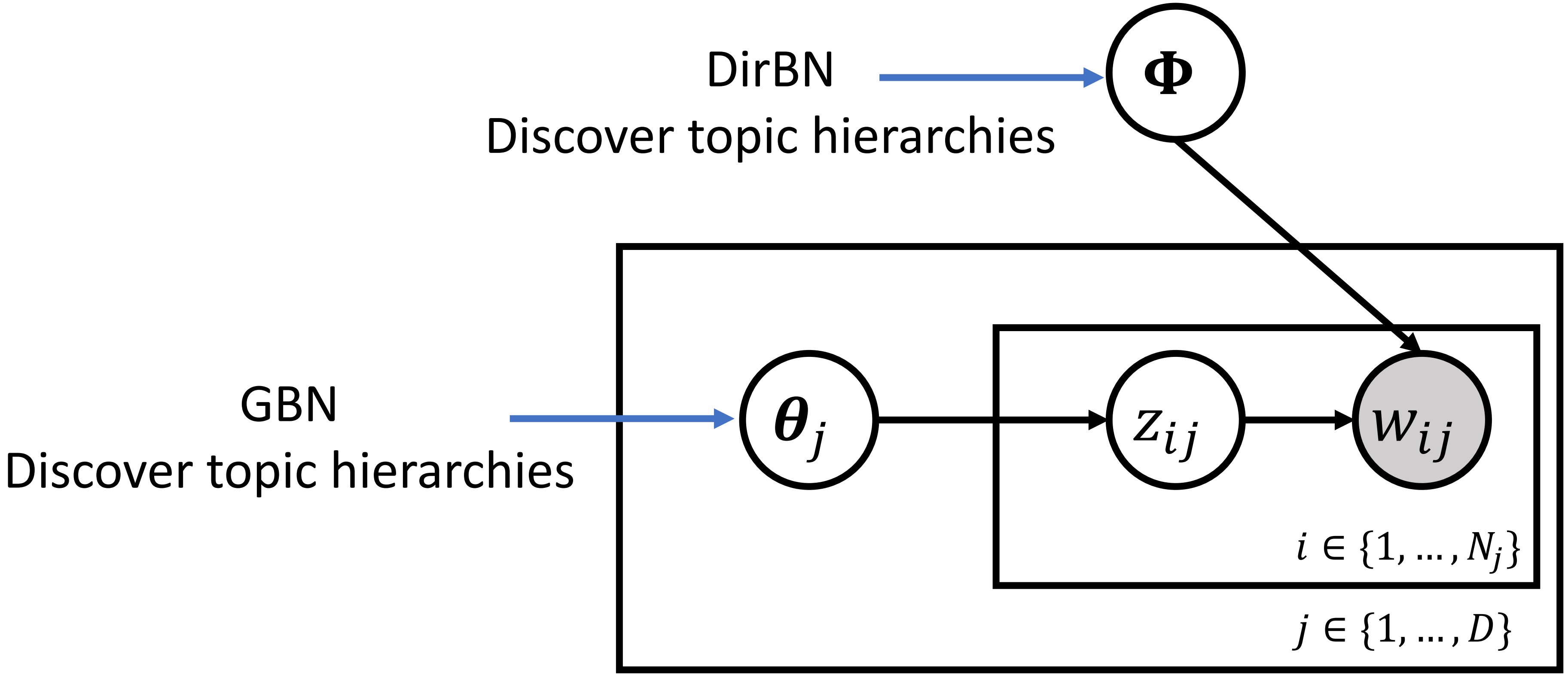
MetaLDA+DirBN



MetaLDA+DirBN



GBN+DirBN


DirBN
Discover topic hierarchies
$\mathbf{\Phi}$
GBN
Discover topic hierarchies
$\boldsymbol{\theta}_j$
z_{ij}
w_{ij}
$i \in \{1, \ldots, N_j\}$
$j \in \{1, \ldots, D\}$

GBN+DirBN

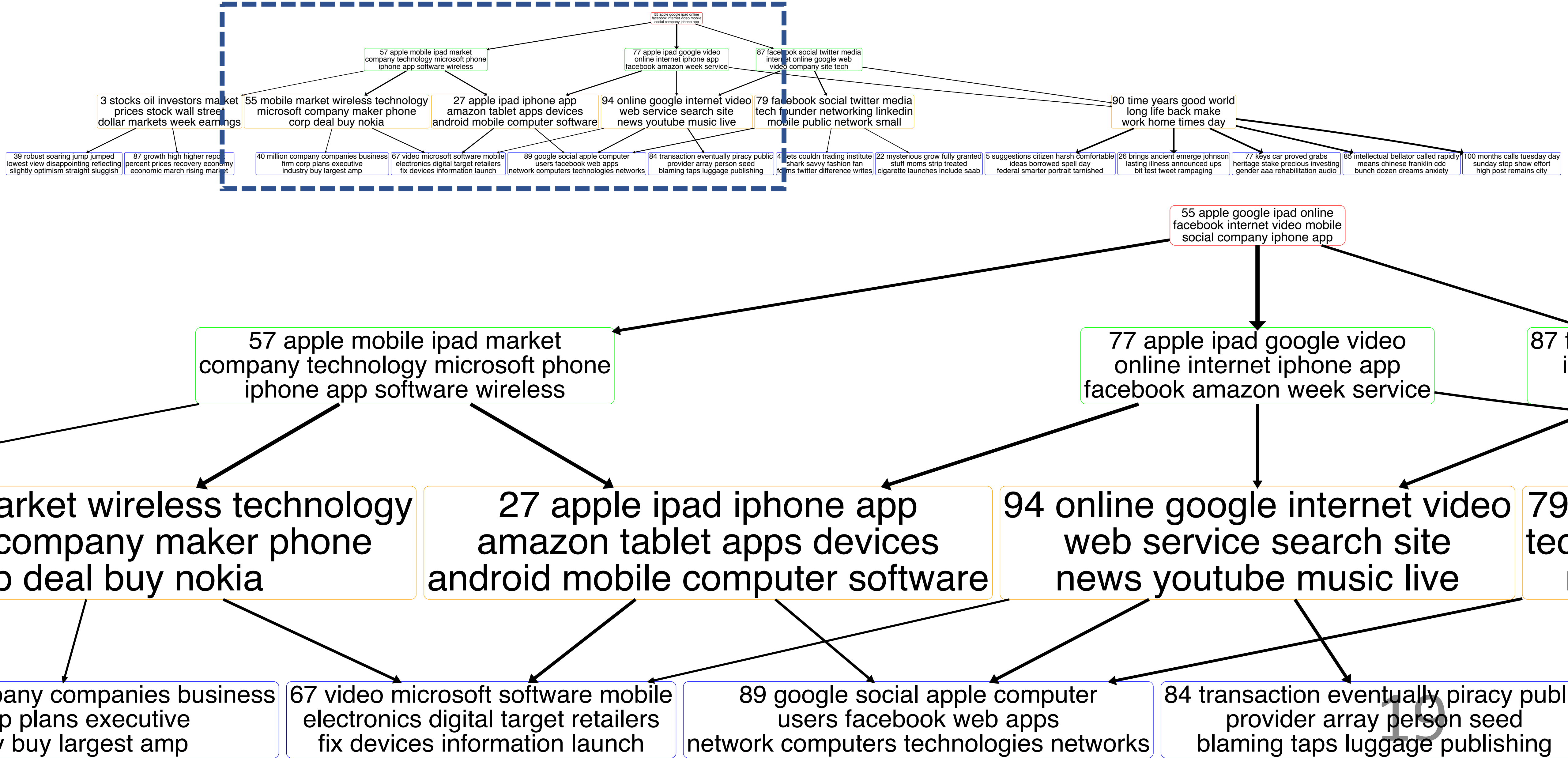
First-layer topic

Top second-layer topics of DirBN

Top second-layer topics of GBN

police arrested man charged woman authorities death year found accused			
0.13	case charges accused trial courtattorney investigation judge allegations criminal	0.38	police arrested man charged year accused found charges woman death
0.13	police official killing attack deaddeath army security man family	0.19	police prison man china years arrested charges charged year chinese
0.11	woman men drug suicide girl sexual death found human york	0.15	china police chinese bomb fire people blast city artist officials
heat miami james lebron game nba finals celtics bulls wade			
0.43	season team game play run night star series fans career	0.97	heat miami james game nba finals lebron bulls mavericks dallas
0.15	nba playoffs court brink seeds defeated berth seed opponent semifinals	0.00	trial rajaratnam insider trading fund hedge raj anthony galleon case
0.10	win victory beat lead winning top fourth loss straight beating	0.00	music album lady gaga justin star pop band rock tour
facebook google internet social twitter online web media site search			
0.18	phone plan video technology mobile devices computer tech ceo content	0.22	facebook social internet google online twitter chief executive media web
0.14	company million buy billion corp industry sales companies consumers products	0.19	court lawsuit case facebook judge social federal internet google online
0.12	government report country nation pressure official state move released public	0.18	facebook social internet google online twitter world web media site

GBN+DirBN

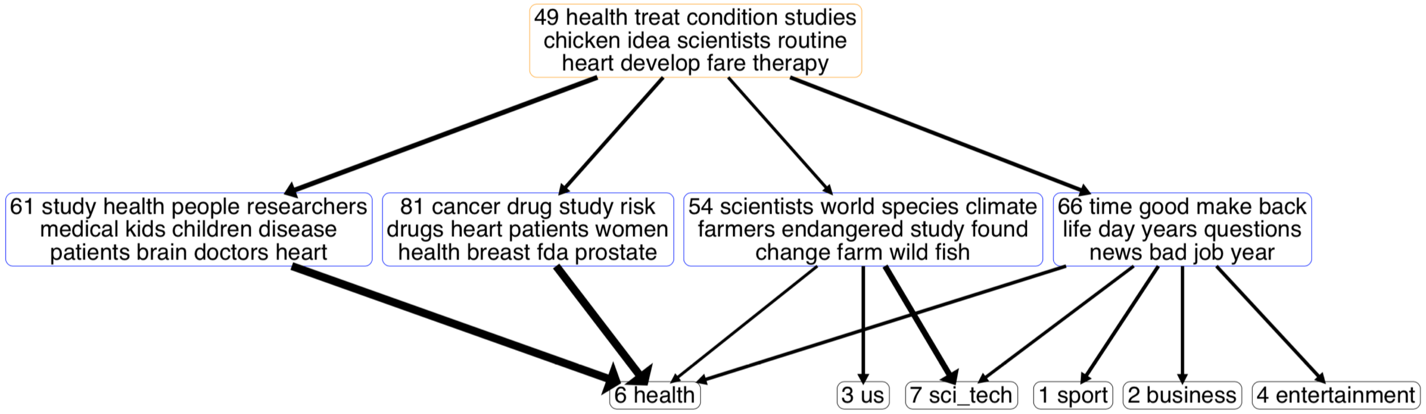


55 apple google ipad online facebook internet video mobile social company iphone app
57 apple mobile ipad market company technology microsoft phone iphone app software wireless
77 apple ipad google video online internet iphone app facebook amazon week service
87 facebook social twitter media internet online google web video company site tech
3 stocks oil investors market prices stock wall street dollar markets week earnings
55 mobile market wireless technology microsoft company maker phone corp deal buy nokia
27 apple ipad iphone app amazon tablet apps devices android mobile computer software
94 online google internet video web service search site news youtube music live
79 facebook social twitter media tech founder networking linkedin mobile public network small
90 time years good world long life back make work home times day
39 robust soaring jump jumped lowest view disappointing reflecting slightly optimism straight sluggish
87 growth high higher report percent prices recovery economy economic march rising market
40 million company companies business firm corp plans executive industry buy largest amp
67 video microsoft software mobile electronics digital target retailers fix devices information launch
89 google social apple computer users facebook web apps network computers technologies networks
84 transaction eventually piracy public provider array person seed blaming taps luggage publishing
22 mysterious grow fully granted stuff moms strip treated cigarette launches include saab
5 suggestions citizen harsh comfortable ideas borrowed spell day federal smarter portrait tarnished
26 brings ancient emerge johnson lasting illness announced ups bit test tweet rampaging
77 keys car proved grabs heritage stake precious investing gender aaa rehabilitation audio
85 intellectual bellator called rapidly means chinese franklin cdc bunch dozen dreams anxiety
100 months calls tuesday day sunday stop show effort high post remains city

Part Two

Previously …

MetaLDA+DirBN



16

Inter-topic structure

An individual topic may not be semantically indivisible

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

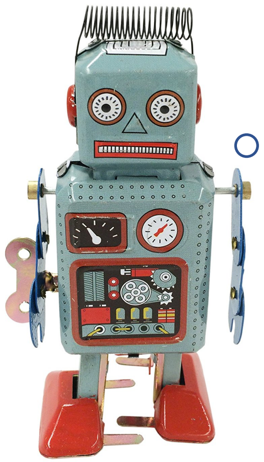
Less interpretable topics

fitness piano guitar swimming violin weightlifting lessons training swim weight

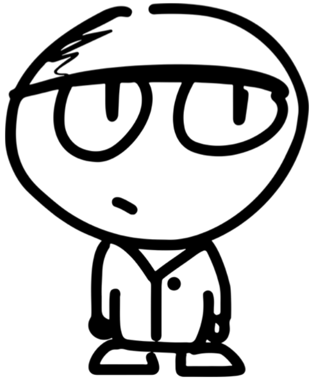
Fitness **Music**

journal science biology research journals international cell psychology scientific bioinformatics

Journal **Biology**

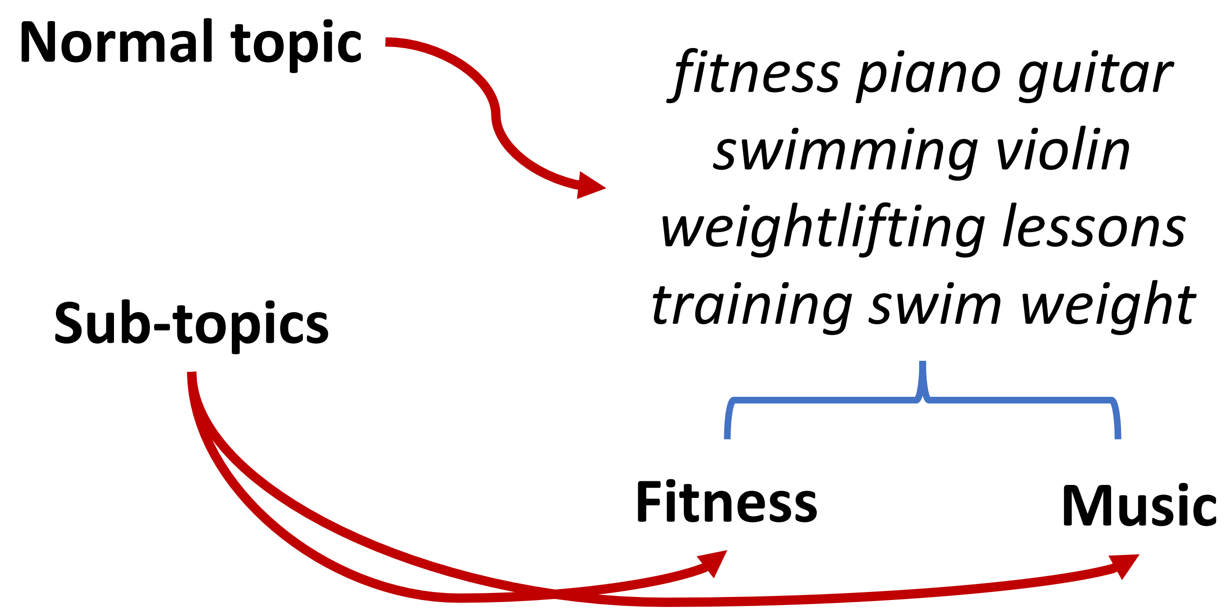


Word cooccurrences in the target corpus



Global word semantics

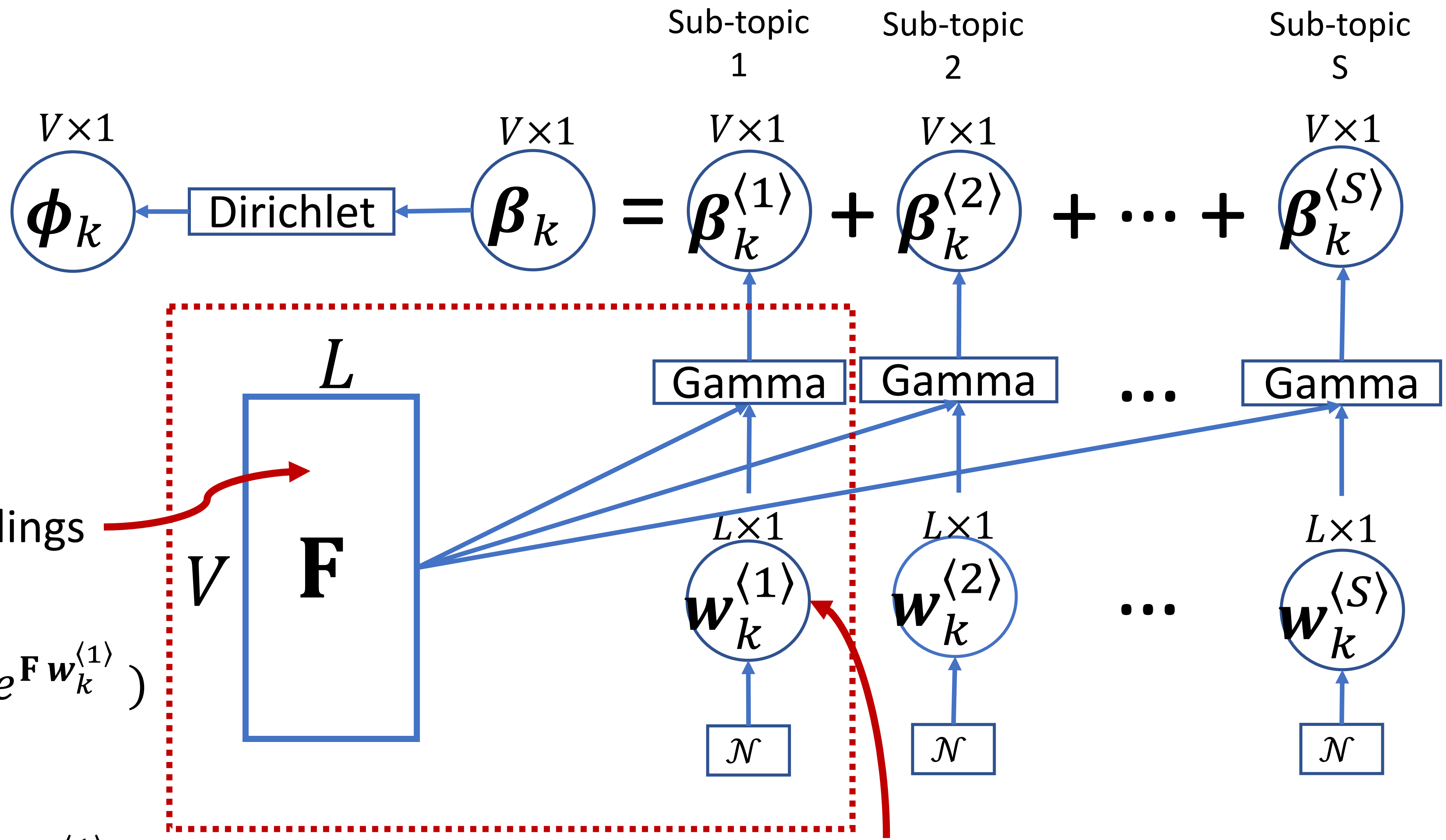
Fine-grained thematic aspects of individual topic



Fine-grained thematic aspects inside a normal topic → **Intra-topic structure learning**

Word embeddings pre-trained on a huge corpus → **Global word semantics**

Intra-topic structure with word embeddings (ICML, 2018)



Word embeddings

$$\boldsymbol{\beta}_k^{\langle 1 \rangle} \sim \mathrm{Gamma}(\alpha_k^{\langle 1 \rangle}, e^{\mathbf{F}\,\boldsymbol{w}_k^{\langle 1 \rangle}})$$

Sub-topic 1: $\boldsymbol{\beta}_k^{\langle 1 \rangle}$ and $\boldsymbol{w}_k^{\langle 1 \rangle}$

- Same dimensions with word embeddings
- Embeddings of sub-topic 1

Inference/Learning

$$\mathcal{L}\left(\beta_{vk_1}\right) \propto \frac{\Gamma(\beta_{\cdot k_1})}{\Gamma(\beta_{k_1\cdot} + x^{(1)}_{\cdot\cdot k_1})} \prod_v^V \frac{\Gamma(\beta_{vk_1} + x^{(1)}_{v\cdot k_1})}{\Gamma(\beta_{vk_1})}$$

$$q_{k_1} \sim \mathrm{Beta}(\beta_{\cdot k_1}, x^{(1)}_{\cdot\cdot k_1})$$

$$h_{vk_1} \sim \mathrm{CRT}\left(x^{(1)}_{v\cdot k_1}, \beta_{vk_1}\right)$$

$$\mathcal{L}\left(\beta_{vk_1}, q_{k_1}, h_{vk_1}\right) \propto (q_{k_1})^{\beta_{vk_1}} \left(\beta_{vk_1}\right)^{h_{vk_1}}$$

$$\beta_{vk_1} = \sum_s^S \beta^{<s>}_{vk_1}$$

$$(h^{<1>}_{vk_1}, \cdots, h^{<S>}_{vk_1}) \sim \mathrm{Mult}\left(h_{vk_1}, \frac{\beta^{<1>}_{vk_1}}{\beta_{vk_1}}, \cdots, \frac{\beta^{<S>}_{vk_1}}{\beta_{vk_1}}\right)$$

$$\beta^{<s>}_{vk_1} \sim \frac{\mathrm{Gam}(\alpha^{<s>}_{k_1} + h^{<s>}_{vk_1}, 1)}{e^{-\pi^{<s>}_{vk_1}} + \log\frac{1}{q_{k_1}}},$$

$$\pi^{<s>}_{vk_1} := \boldsymbol{f}_v^\top \boldsymbol{w}^{<s>}_{k_1}$$

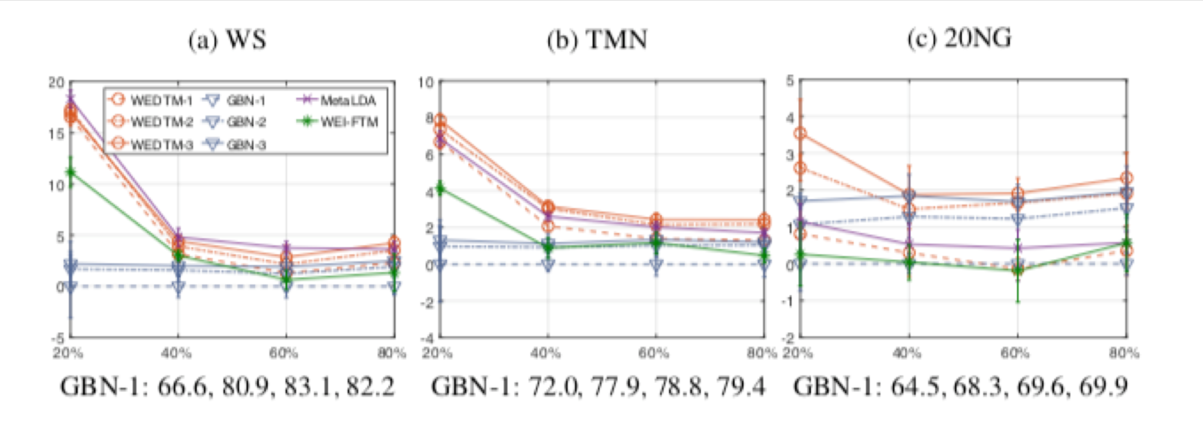
$$\mathcal{L}\left(\pi^{<s>}_{vk_1}\right) \propto \frac{\left(e^{\pi^{<s>}_{vk_1} + \log\log\frac{1}{q_{k_1}}}\right)^{h^{<s>}_{vk_1}}}{\left(1 + e^{\pi^{<s>}_{vk_1} + \log\log\frac{1}{q_{k_1}}}\right)^{\alpha^{<s>}_{k_1} + h^{<s>}_{vk_1}}}$$

$$\omega^{<s>}_{vk_1} \sim \mathrm{PG}\left(h^{<s>}_{vk_1} + \alpha^{<s>}_{k_1}, \pi^{<s>}_{vk_1} + \log\log\frac{1}{q_{k_1}}\right)$$

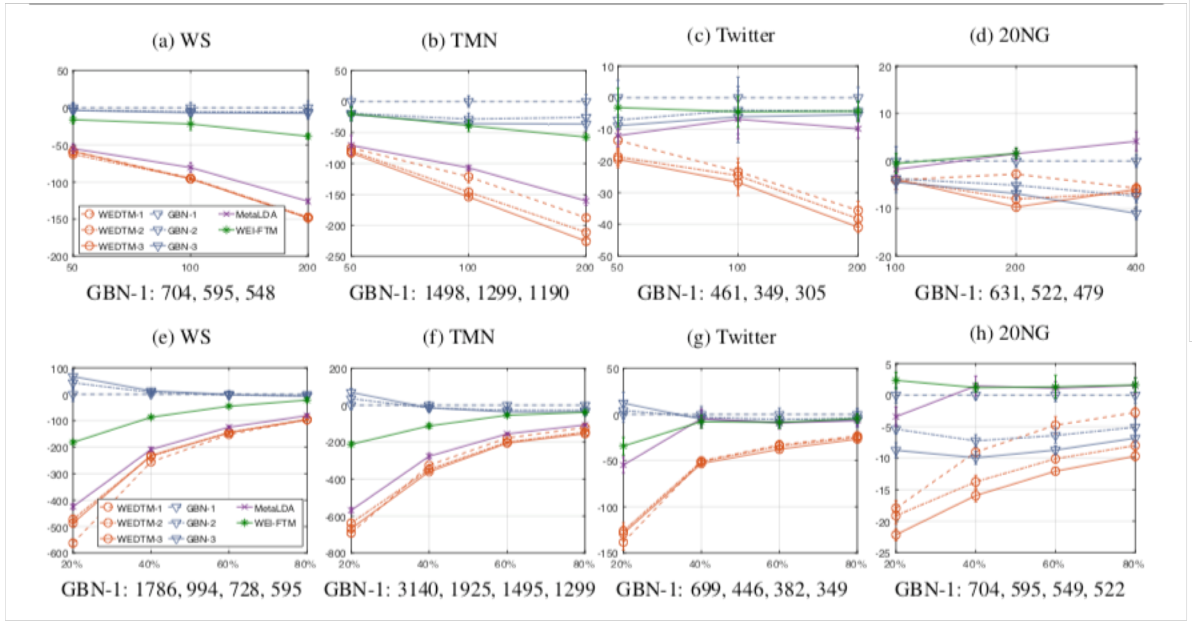
$$\mathcal{L}\left(\pi^{<s>}_{vk_1}, \omega^{<s>}_{vk_1}\right) \propto e^{\frac{h^{<s>}_{vk_1} - \alpha^{<s>}_{k_1}}{2}\pi^{<s>}_{vk_1}} e^{-\frac{\omega^{<s>}_{vk_1}}{2}\pi^{<s>\,2}_{vk_1}}.$$

Numerical results

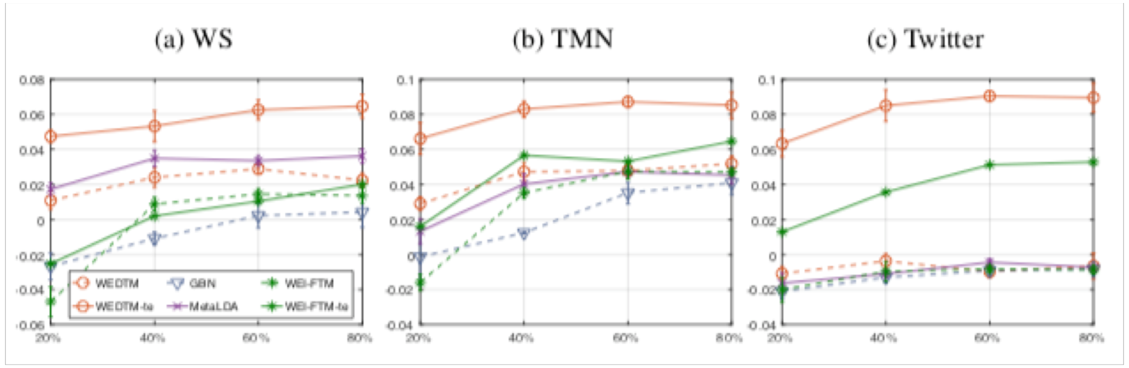
Document classification



Perplexity



Topic coherence



Sub-topics discovered in our model

Font-size: weights of the sub-topics in a normal topic

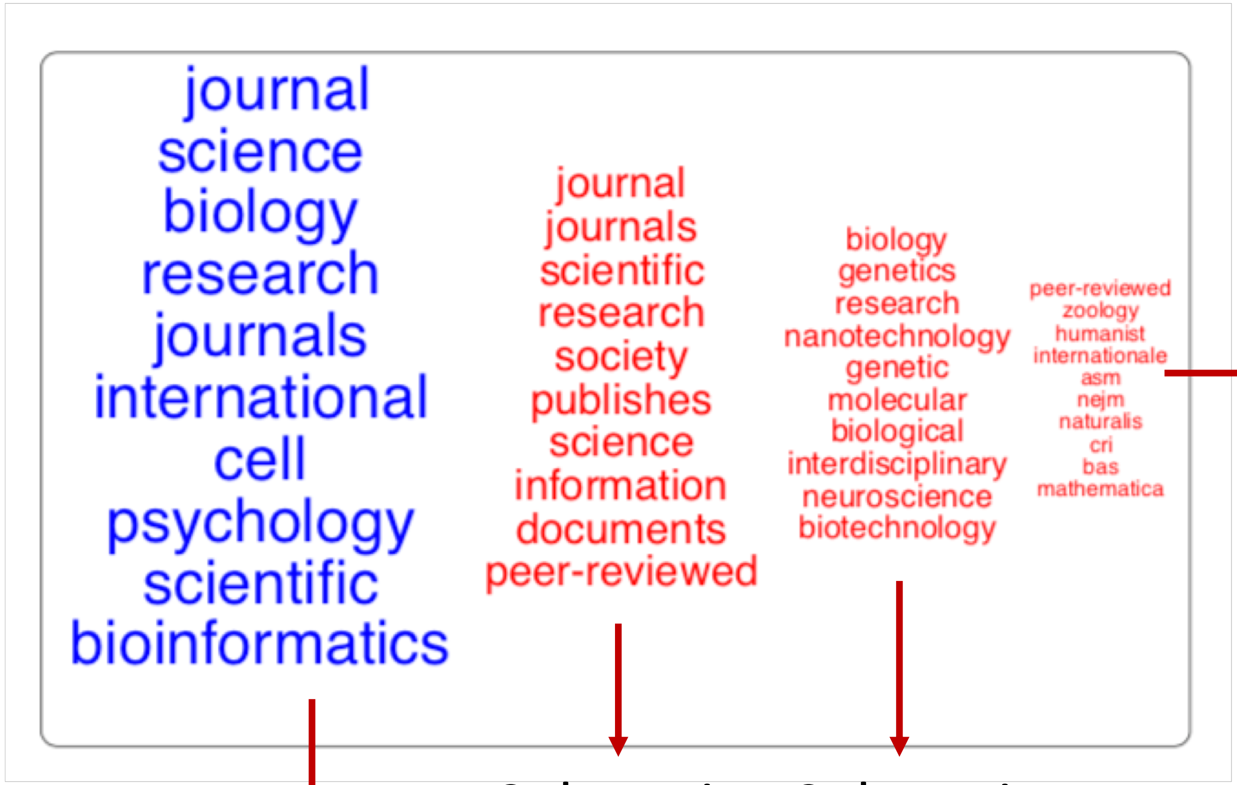
Clean-up sub-topic

Normal topic “Fitness” + “Music”

Sub-topic “Music”

Sub-topic “Fitness”

Sub-topics discovered in our model



Clean-up sub-topic

Sub-topic "Journal"

Sub-topic "Biology"

Normal topic "Journal" + "Biology"

Summary

- Part one: a flexible module discovering three-structured topic hierarchies that is compatible with many other advanced topic models
- Part two: a model that discovers the fine-grained semantic structures inside individual topic, with the help of globally trained word embeddings
- Better modelling performance on perplexity, topic quality, and downstream applications like document classification, plus fantastic interpretability