Supplementary Material for Deep latent Dirichlet allocation with topic-layer-adaptive stochastic gradient Riemannian MCMC

Yulai Cong, Bo Chen, Hongwei Liu, and Mingyuan Zhou

A. Naive derivation of the Fisher information matrix of the Poisson gamma belief network

For simplicity, we take for example a two-layer Poisson gamma belief network (PGBN), expressed as

$$\begin{aligned}\boldsymbol{\theta}_j^{(2)} &\sim \mathrm{Gam}\left(\boldsymbol{r}, 1/c_j^{(3)}\right),\\ \boldsymbol{x}_j^{(1)} \sim \mathrm{Pois}\left(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}\right), \boldsymbol{\theta}_j^{(1)} &\sim \mathrm{Gam}\left(\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_j^{(2)}, \tfrac{p_j^{(2)}}{1-p_j^{(2)}}\right),\end{aligned} \tag{20}$$

and focus on a specific element $\boldsymbol{\Phi}_{vk}^{(2)}$ only.

With the definition in (8), it is straight to show that the $\boldsymbol{\Phi}^{(2)}$-relevant part in $\ln p\left(\boldsymbol{\Omega}\,|\boldsymbol{z}\right)$ is

$$\sum\nolimits_{vj}\left[\boldsymbol{\Phi}_{v:}^{(2)}\boldsymbol{\theta}_{:j}^{(2)}\ln\left(c_j^{(2)}\boldsymbol{\theta}_{vj}^{(1)}\right)-\ln\Gamma\left(\boldsymbol{\Phi}_{v:}^{(2)}\boldsymbol{\theta}_{:j}^{(2)}\right)\right]. \tag{21}$$

Accordingly, for $\boldsymbol{\Phi}_{vk}^{(2)}$, we have

$$\mathbb{E}\left[-\frac{\partial^2}{\partial[\boldsymbol{\Phi}_{vk}^{(2)}]^2}\ln p\left(\boldsymbol{\Omega}\,|\boldsymbol{z}\right)\right]=\mathbb{E}\left[\sum_j\psi'\left(\boldsymbol{\Phi}_{v:}^{(2)}\boldsymbol{\theta}_{:j}^{(2)}\right)\left[\boldsymbol{\theta}_{:j}^{(2)}\right]^2\right], \tag{22}$$

where $\psi'(\cdot)$ is the trigamma function. This expectation involving the trigamma function is difficult to calculate.

B. Derivation of the $\Gamma(\cdot)$ functions in Section 3.2

With $\mathbf{D}\left(\boldsymbol{z}\right)=\mathbf{G}(\boldsymbol{z})^{-1}$, $\mathbf{Q}\left(\boldsymbol{z}\right)=\mathbf{0}$, and the block-diagonal Fisher information matrix (FIM) $\mathbf{G}(\boldsymbol{z})$ in (9), it is straight to show that $\frac{\partial}{\partial\boldsymbol{\varphi}_k}\left[\mathbf{D}\left(\boldsymbol{z}\right)+\mathbf{Q}\left(\boldsymbol{z}\right)\right]$ is non-zero only in the $\boldsymbol{\varphi}_k$-related block $\mathbf{I}\left(\boldsymbol{\varphi}_k\right)$ in (10). Therefore, we focus on this block and have

$$\Gamma_v\left(\boldsymbol{\varphi}_k\right)=\sum\nolimits_u\frac{\partial}{\partial\varphi_{uk}}\left[\mathbf{I}_{vu}^{-1}\left(\boldsymbol{\varphi}_k\right)\right], \tag{23}$$

where $\mathbf{I}^{-1}\left(\boldsymbol{\varphi}_k\right)=M_k^{-1}\left[\mathrm{diag}(\boldsymbol{\varphi})-\boldsymbol{\varphi}\boldsymbol{\varphi}^T\right]$. Accordingly, we have

$$\begin{aligned}\Gamma_v\left(\boldsymbol{\varphi}_k\right)&=M_k^{-1}\sum\nolimits_u\frac{\partial}{\partial\varphi_{uk}}\left[\delta_{u=v}\varphi_{uk}-\varphi_{vk}\varphi_{uk}\right]\\&=M_k^{-1}(1-V\varphi_{vk}).\end{aligned} \tag{24}$$

Since $\mathbf{G}(\boldsymbol{z})$ is block-diagonal with its $\boldsymbol{r}$-relevant block being $\mathbf{I}\left(\boldsymbol{r}\right)=M^{(L+1)}\mathrm{diag}\left(1/\boldsymbol{r}\right)$, according to (3), it is straightforward to show that

$$\begin{aligned}\Gamma_k\left(\boldsymbol{r}\right)&=\sum\nolimits_u\frac{\partial}{\partial r_u}\left[\mathbf{I}_{ku}^{-1}\left(\boldsymbol{r}\right)\right],\\&=\sum\nolimits_u\frac{\partial}{\partial r_u}\left[\delta_{u=k}\frac{r_u}{M^{(L+1)}}\right],\\&=1/M^{(L+1)}.\end{aligned} \tag{25}$$

C. Proof of Lemma 3.1

Note that the counts in $x_{vj}^{(l)}\sim\mathrm{Pois}\big(q_j^{(l)}\sum_{k=1}^{K_l}\phi_{vk}^{(l)}\theta_{kj}^{(l)}\big)$ can be augmented as

$$\begin{aligned}x_{vj}^{(l)}&=\sum\nolimits_{k=1}^{K_l}x_{vkj}^{(l)},\\x_{vkj}^{(l)}&\sim\mathrm{Pois}\big(q_j^{(l)}\phi_{vk}^{(l)}\theta_{kj}^{(l)}\big),\end{aligned} \tag{26}$$

which, according to Lemma 4.1 of Zhou et al. (2012), can be equivalently expressed as

$$\begin{aligned}\left(x_{vkj}^{(l)}\right)_v&\sim\mathrm{Mult}\left(m_{kj}^{(l)(l+1)},\boldsymbol{\phi}_k^{(l)}\right),\\m_{kj}^{(l)(l+1)}&\sim\mathrm{Pois}\left(q_j^{(l)}\theta_{kj}^{(l)}\right),\end{aligned} \tag{27}$$

where $m_{kj}^{(l)(l+1)}:=\sum_{v=1}^{K_{l-1}}x_{vkj}^{(l)}$. Marginalizing out $\theta_{vj}^{(l)}\sim\mathrm{Gam}\left(\sum_{k=1}^{K_{l+1}}\boldsymbol{\phi}_{vk}^{(l+1)}\boldsymbol{\theta}_{kj}^{(l+1)},1/c_j^{(l+1)}\right)$ from (27) leads to

$$m_{vj}^{(l)(l+1)}\sim\mathrm{NB}\left(\sum\nolimits_{k=1}^{K_{l+1}}\boldsymbol{\phi}_{vk}^{(l+1)}\boldsymbol{\theta}_{kj}^{(l+1)},\,p_j^{(l+1)}\right), \tag{28}$$

which can be augmented as

$$\begin{aligned}m_{vj}^{(l)(l+1)}&\sim\mathrm{SumLog}(x_{vj}^{(l+1)},p_j^{(l+1)}),\\x_{vj}^{(l+1)}&\sim\mathrm{Pois}\left(q_j^{(l+1)}\sum\nolimits_{k=1}^{K_{l+1}}\boldsymbol{\phi}_{vk}^{(l+1)}\boldsymbol{\theta}_{kj}^{(l+1)}\right).\end{aligned} \tag{29}$$

When $l=L$, we have

$$m_{kj}^{(L)(L+1)}\sim\mathrm{NB}(r_k,p_j^{(L+1)}), \tag{30}$$

marginalizing the gamma process $G\sim\mathrm{GaP}(G_0,1/c_0)$ from which leads to a gamma-negative binomial process random count matrix, as expressed in the first two lines of (6).

D. Corollary D.1

Corollary D.1. *The gamma-negative binomial process PFA can be equivalently expressed as*

$$\begin{aligned}\ell_{k\cdot}\sim\mathrm{Log}\left(\frac{q_\cdot}{c_0+q_\cdot}\right),\ K&\sim\mathrm{Pois}\left(\gamma_0\ln\frac{c_0+q_\cdot}{c_0}\right),\\\mathcal{L}&=\sum\nolimits_{k=1}^{K}\ell_{k\cdot}\delta_{\boldsymbol{\phi}_k},\\\left(\ell_{kj}\right)_j&\sim\mathrm{Mult}\left[\ell_{k\cdot},\left(q_j\right)_j/q_\cdot\right],\\m_{kj}&\sim\mathrm{SumLog}(\ell_{kj},p_j)\\x_{vj}=\sum\nolimits_{k=1}^{K}x_{vkj},\ \left(x_{vkj}\right)_v&\sim\mathrm{Mult}\left(m_{kj},\boldsymbol{\phi}_k\right).\end{aligned} \tag{31}$$

E. Visualizations of the extracted topics and networks

In the following, we provide some example results, obtained using DLDA where $[K_1, K_2, K_3] = [128, 64, 32]$ and $\eta^{(l)} = 1/K_l$ for the lth layer, on extracting multilayer representations/topics from both the RCV1 and Wiki corpora. Clearly interpretable results, which are similar to those reported in Zhou et al. (2016a) and hence omitted here for brevity, are also extracted from the 20Newsgroups corpus.

E.1. RCV1

Following the visualization techniques in Zhou et al. (2016a), we plot 54 example topics of layer one in Figure 4, the top 30 topics of layer two in Figure 5, and the top 30 topics of layer three in Figure 6. Figure 4 clearly shows that the topics of layer one are very specific. For example, topics 41, 71 and 62 in the first row are about "Germany," "Polish," and "France," respectively; topics 53 and 54 in the second row are about "airline" and "European union," respectively; and topics 85 and 36 in the third row are about "ship & island" and "comput & techn," respectively. By contrast, the topics of layers two and three, shown in Figures 5 and 6, respectively, are increasingly more general. Such topics can be better interpreted via the following informative tree structured visualizations. Note that a tree defined in this paper allows a child node of a layer to be connected to more than one parent node of the adjacent higher layer.

Shown in Figure 7 is a [10, 3, 1] tree rooted at node 4 of the top layer on "bonds, rates, & credit markets." Apparently, the topics become more and more specific when moving from top to bottom following the branches. For example, the root node splits into three nodes from layers three to two, which focus differently on "treasury bill," "dollar rate," and "bond, credit, & debt," respectively. When moving from layers two to one, all three topics in layer two splits into multiple ones that is clearly more specific. For example, topics 1, 17, and 87 are about "months," "loan & credit," and "bond & pay," respectively.

Shown in Figure 8 is another analogous tree rooted at node 24 of layer three. It is clear that, as the nodes of this tree, topics 55, 38, 34, and 30 of layer two are mainly about "Germany," "France," "airline," and "labor union," respectively. Moreover, these four topics at layer two are all connected to topic 8 of layer one, which is very specific on "office meetings."

To understand the relationships and distinctions between different trees, we construct subnetworks as shown in Figures 9-10. Figure 9 clearly shows that all three trees, rooted at nodes 16, 10, and 17 of layer three, respectively, are highly related to topic 3 of layer two on "low & expect". However, the two trees rooted at node 10 and 17, respectively, both have their own specificities. For example, topic 52 of layer two on "wall street," is unique to node 10 of layer three, and topic 35 of layer two on "India" is unique to node 17 of layer three. Similar phenomena can also be observed from another subnetwork on "China," shown in Figure 10, where both nodes of the top layer are connected to topic 19 of layer two on "corp & techn," topic 36 on "China," and topic 12 on "profit & expect." Though related to each other, the tree rooted at node 18 of the top layer is also strongly connected to topic 31 on "project" and topic 34 on "airline," whereas the other one focuses differently on topic 49 on "car & Korea" and topic 44 on "growth rate".

E.2. Wiki

What follows are analogous figures illustrating some interpretable topics extracted from Wiki.

Figures 11-13 show the top example topics at layers one, two, and three, respectively. It is obvious that topics of layer one are specific, such as topic 31 on "university & research," topic 72 on "news, magazine, & times," topic 83 on "military & army," topic 74 on "police, crime & prison," topic 75 on "birds & species," topic 36 on "British & England," and so on. By contrast, when going to higher layers, topics become more general, as shown in Figures 12 and 13. To better illustrate topics of higher layers, we explicitly show their hierarchical structures via the following trees and subnetworks.

Figure 14 shows a tree rooted at node 1 of layer three on "music & song," whose topics at layer two are about "song & band" and "music, piano, & theatre," respectively. Figure 15 demonstrates another tree consisting of topic 9 of layer two on "London & British," topic 50 on "church & Catholic," and topic 25 on "king & prince," which is mainly about "United Kingdom." Given in Figure 16 is another tree on "art & museum," where the left side is about "art" while the right is on "history & building." These trees are all clearly interpretable.

In the subnetwork shown in Figure 17, all three trees are related to topic 9 of layer two on "London, British, & Sir." But they focus differently on topic 16 of layer two on "Irish Americans," topic 24 on "life, birth, education, career, family, & death," and topic 25 on "king & prince," respectively. Similar phenomena can also be observed in Figure 18. Both trees are related to topic 52 of layer two on "ship" and topic 49 on "air," but the left one is about various means of transportation and communication while the right one is about various components of "war." Figure 19 shows another subnetwork on "team & race," where three trees, all include topic 6 of layer two, focus differently on "goals, clubs, & league," "world cup," and "rank, first, second, & third," respectively.

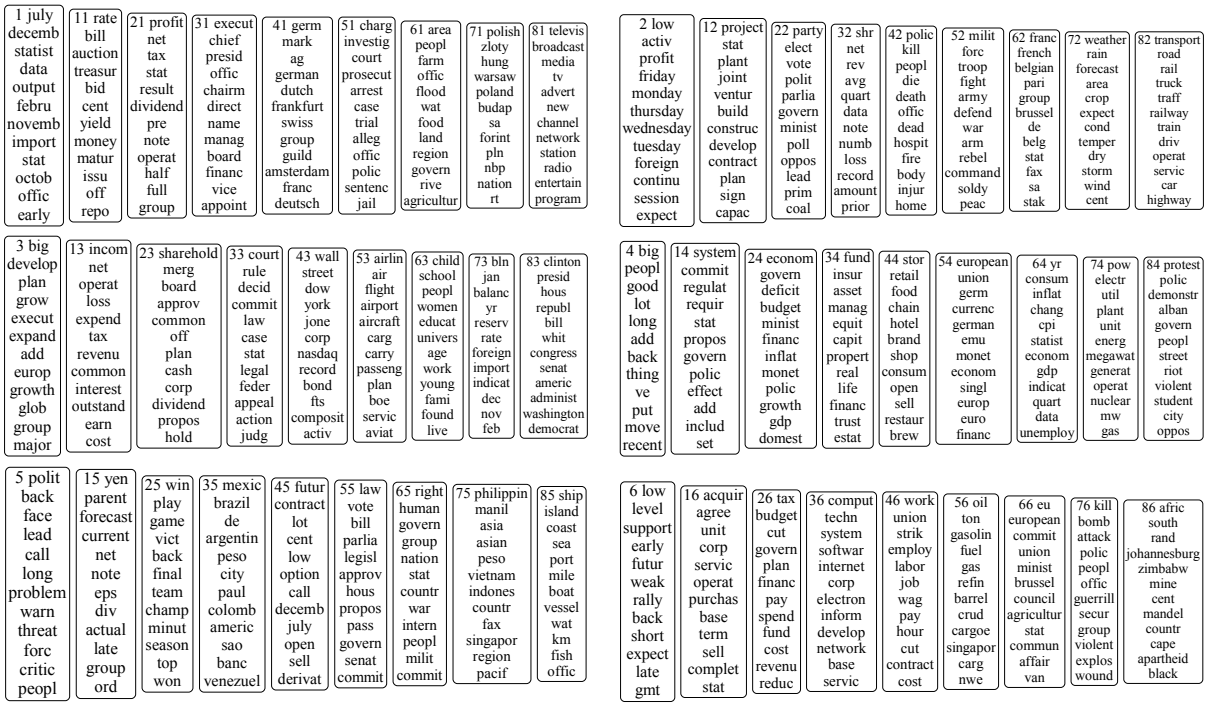

Figure 4. Example topics of layer one of DLDA trained with TLASGR MCMC on RCV1.

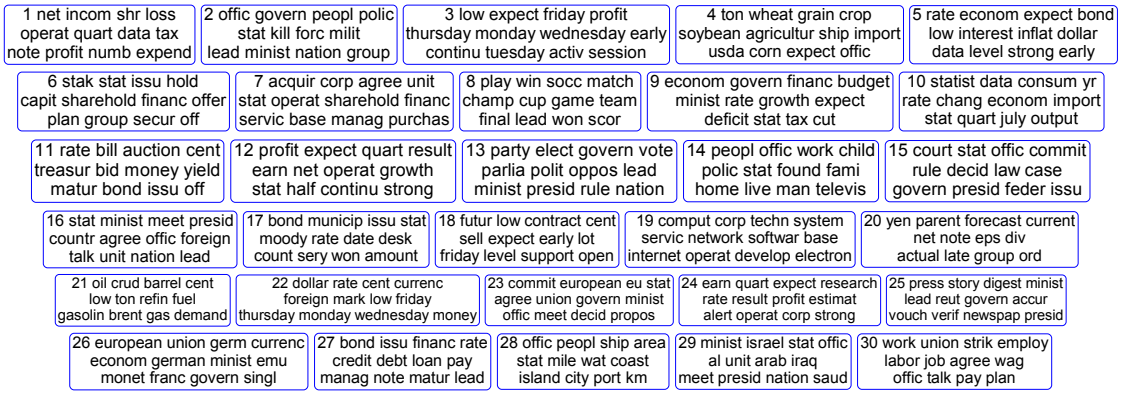

Figure 5. The top 30 topics of layer two of DLDA trained with TLASGR MCMC on RCV1.

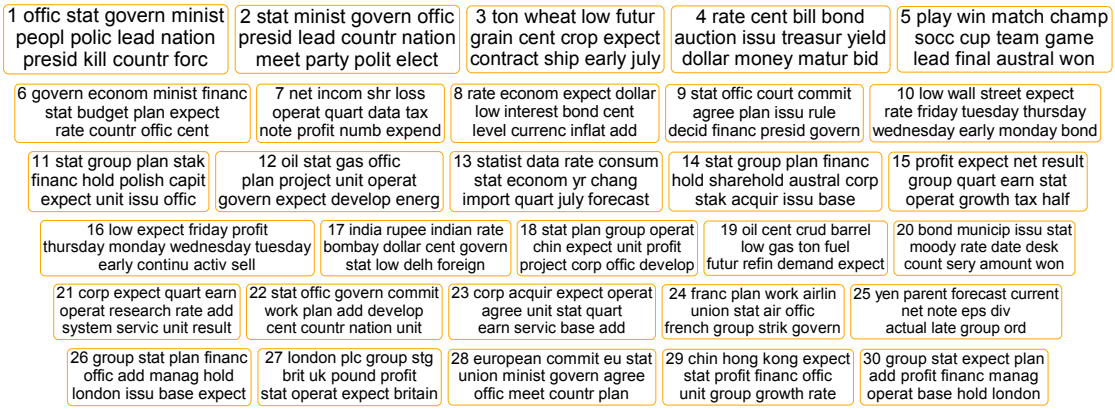

Figure 6. The top 30 topics of layer three of DLDA trained with TLASGR MCMC on RCV1.

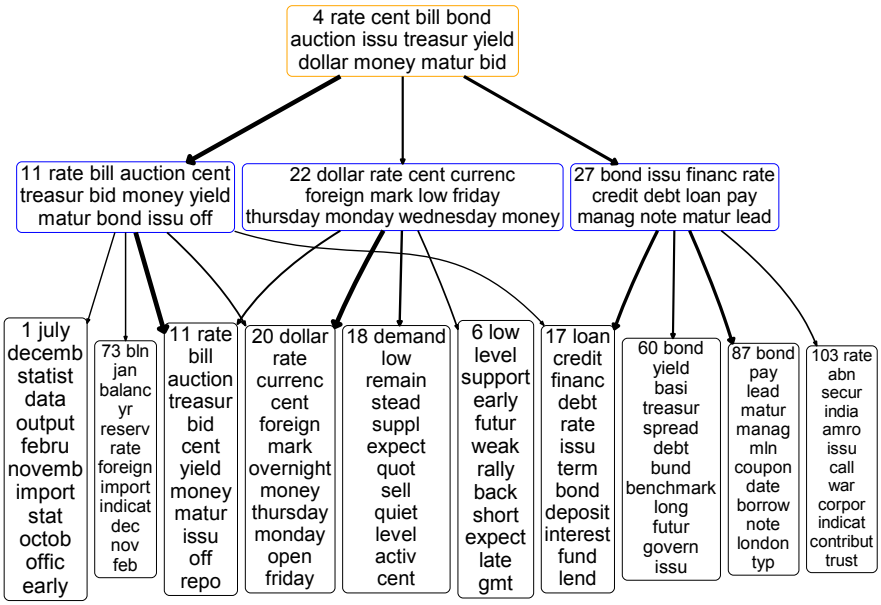

Figure 7. A $[10, 3, 1]$ tree on "bonds, rates, & credit markets" that includes all the lower-layer nodes (directly or indirectly) linked with non-negligible weights to node 4 of the top layer, taken from the full $[128, 64, 32]$ DLDA network trained with TLASGR MCMC on the 794,414 training documents of the RCV1 corpus, with $\eta^{(l)} = 1/K_l$ for the lth layer. A line from node k at layer l to node k' at layer $l-1$ indicates that $\mathbf{\Phi}^{(l)}(k', k) > 5/K_{l-1}$, with the width of the line proportional to $\sqrt{\mathbf{\Phi}^{(l)}(k', k)}$. For each node, the rank (in terms of popularity) at the corresponding layer and the top 12 words of the corresponding topic are displayed inside the text box, where the text font size monotonically decreases as the popularity of the node decreases, and the outside border of the text box is colored as orange, blue, or black if the node is at layer three, two, or one, respectively.

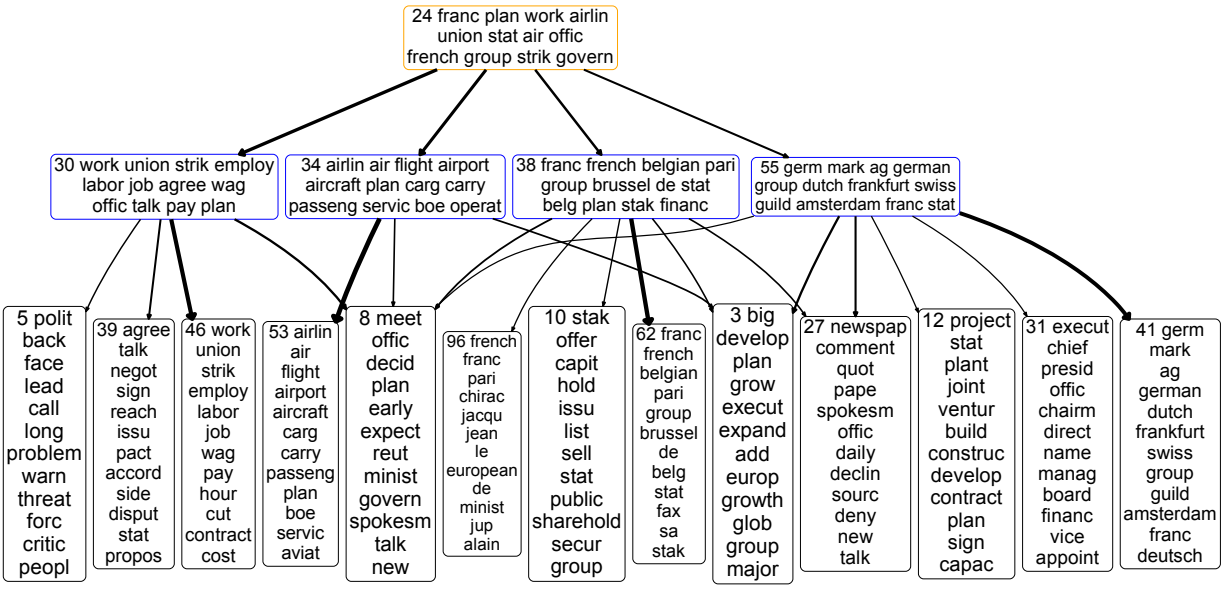

Figure 8. Analogous plots to Figure 7 for a tree rooted at node 24 on "France, Germany, & airline" from RCV1.

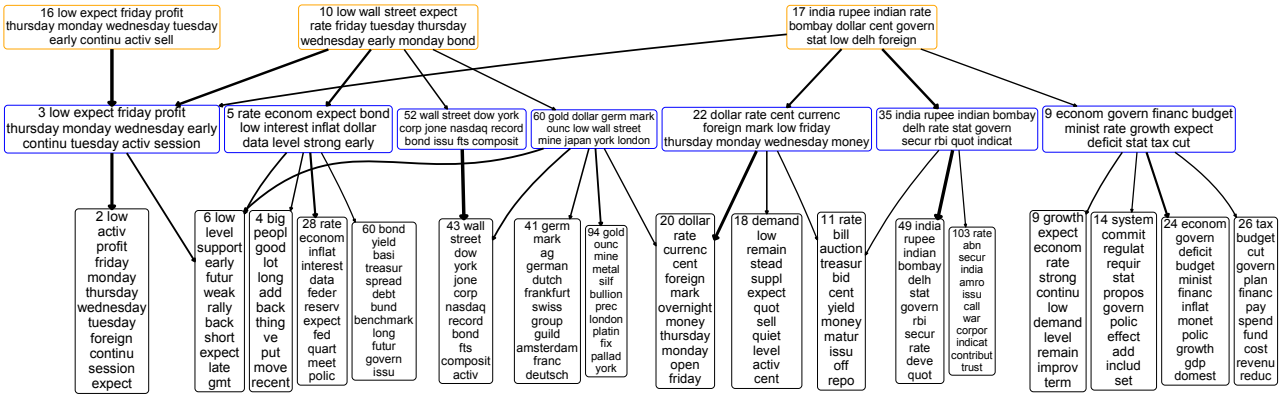

Figure 9. Analogous plots to Figure 7 for a subnetwork related to "low & expect" from RCV1, consisting of three trees rooted at nodes 16, 10 and 17, respectively, of layer three.

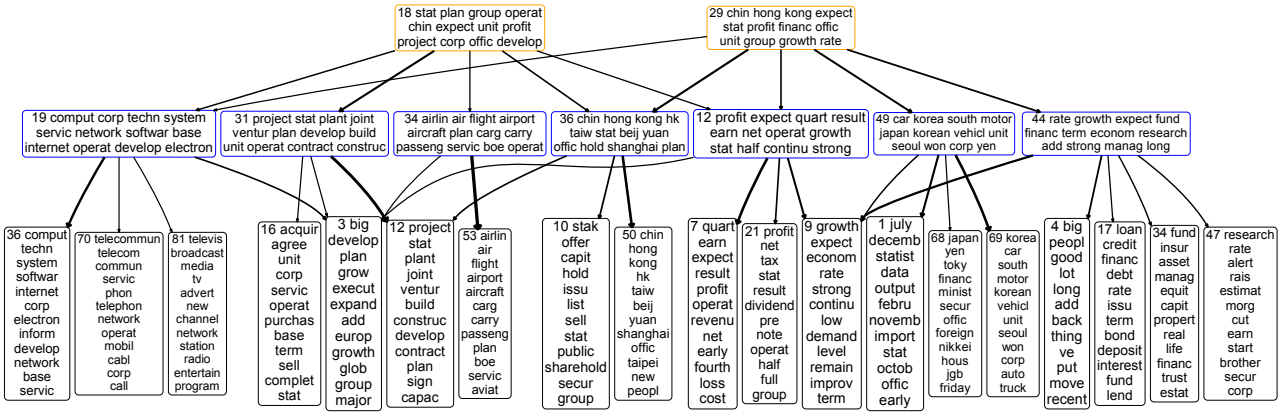

Figure 10. Analogous plots to Figure 7 for a subnetwork on "China" from RCV1, consisting of two trees rooted at nodes 18 and 29, respectively, of layer three.

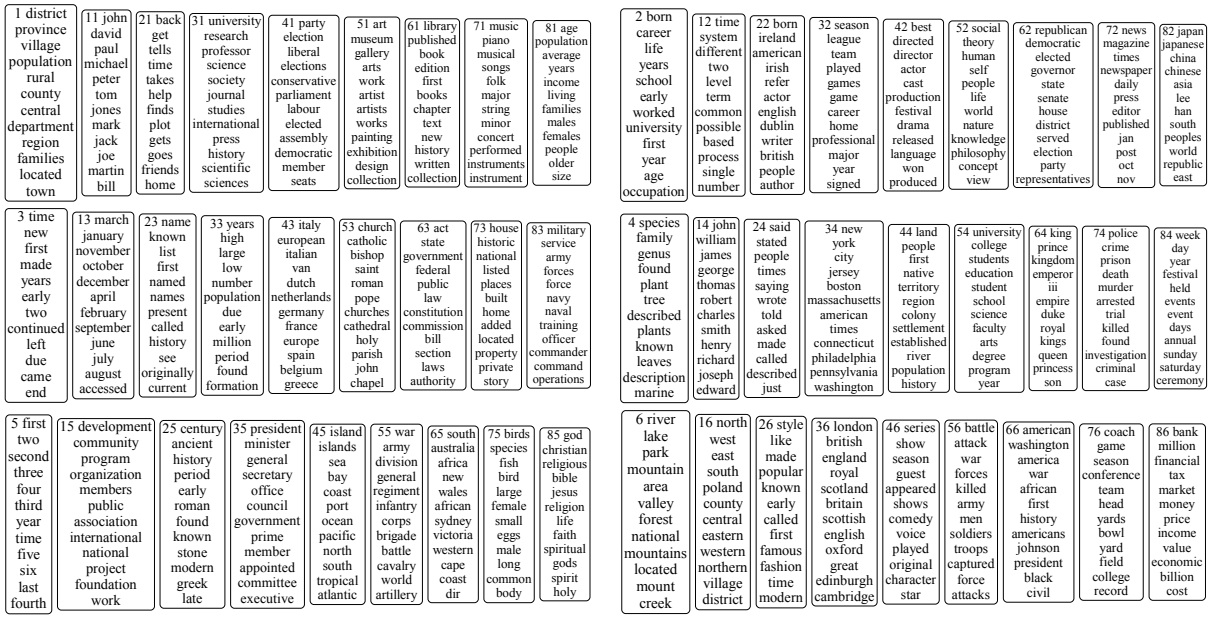

Figure 11. Example topics of layer one of DLDA trained with TLASGR MCMC on Wiki.

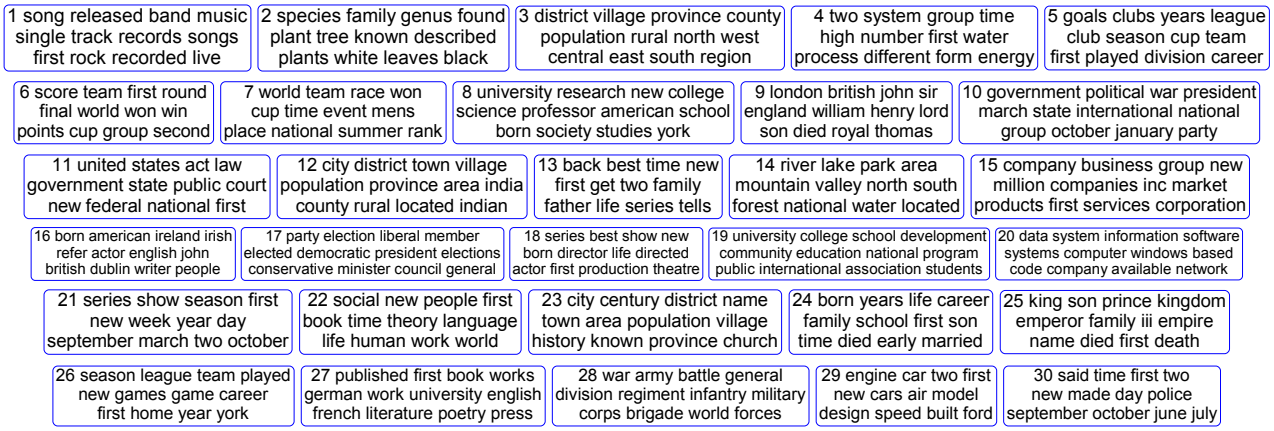

Figure 12. The top 30 topics of layer two of DLDA trained with TLASGR MCMC on Wiki.

Figure 13. The top 30 topics of layer three of DLDA trained with TLASGR MCMC on Wiki.

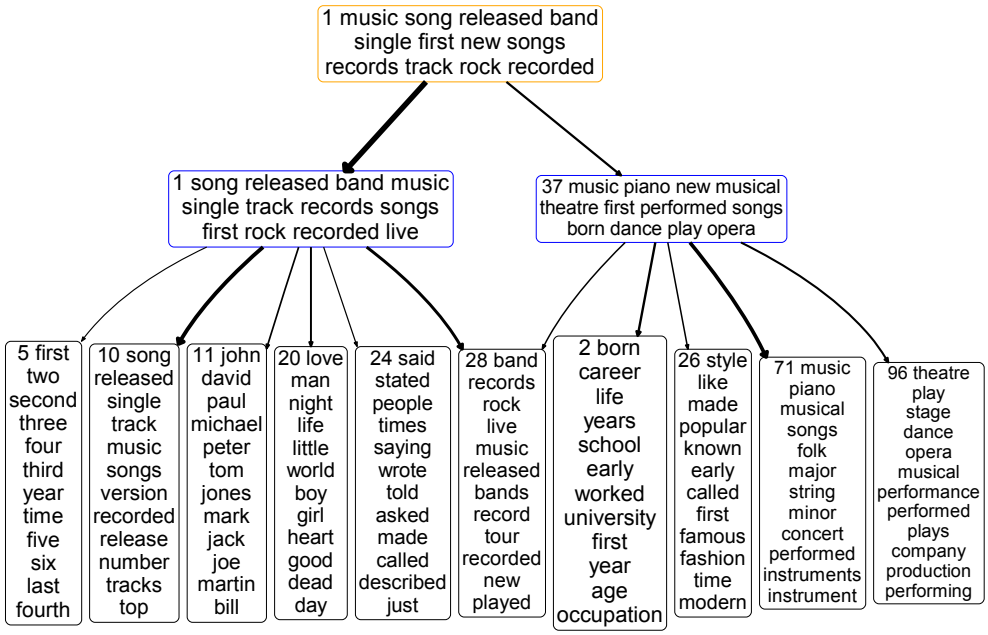

Figure 14. Analogous plots to Figure 7 for a tree rooted at node 1 on "music & song" from Wiki.

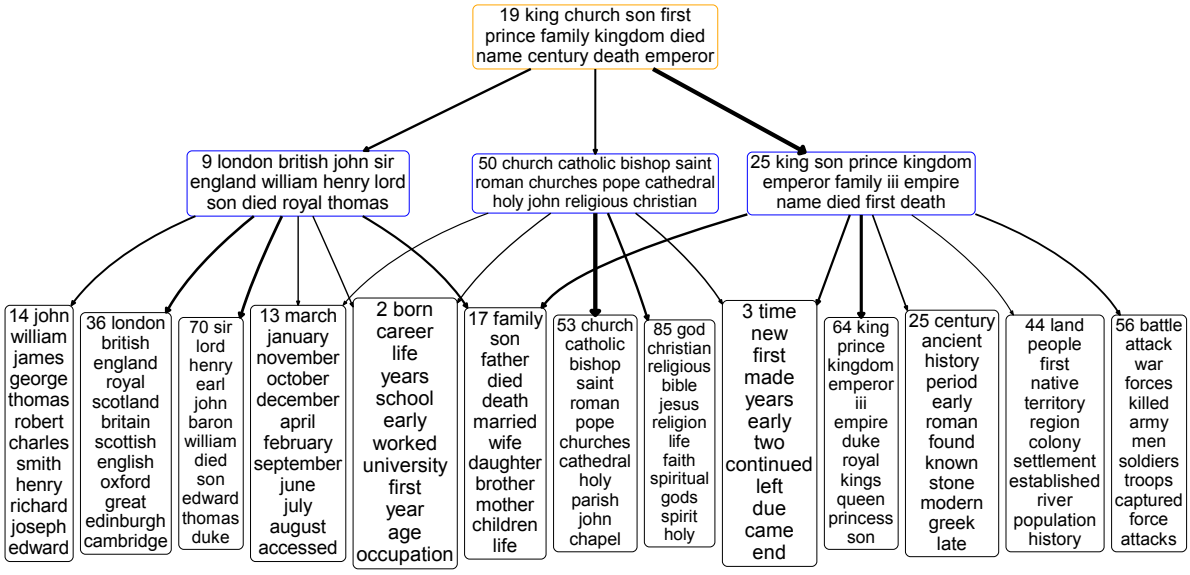

Figure 15. Analogous plots to Figure 7 for a tree rooted at node 19 on "United Kingdom" from Wiki.

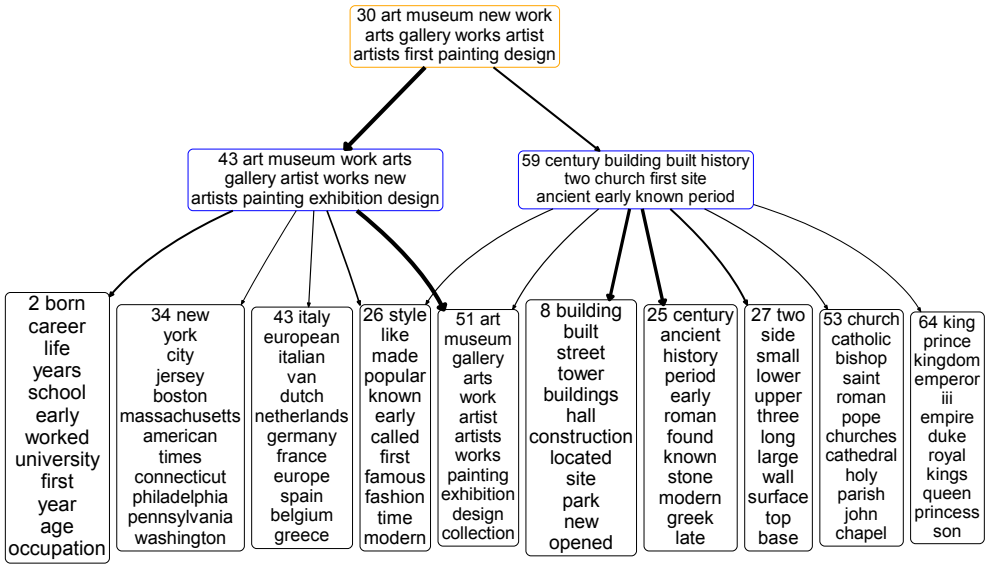

Figure 16. Analogous plots to Figure 7 for a tree rooted at node 30 on "art & museum" from Wiki.

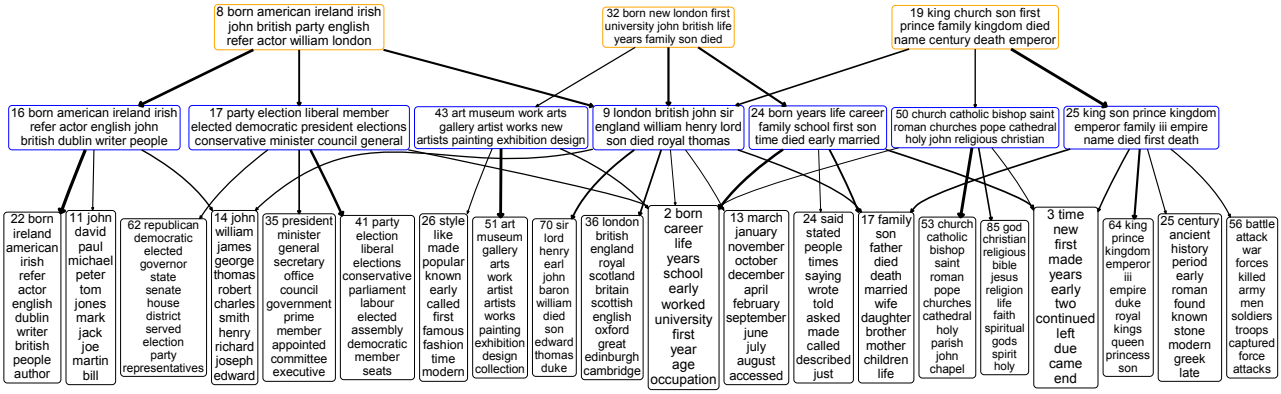

Figure 17. Analogous plots to Figure 7 for a subnetwork on "British" from Wiki, consisting of three trees rooted at nodes 8, 32, and 19, respectively, of layer three.

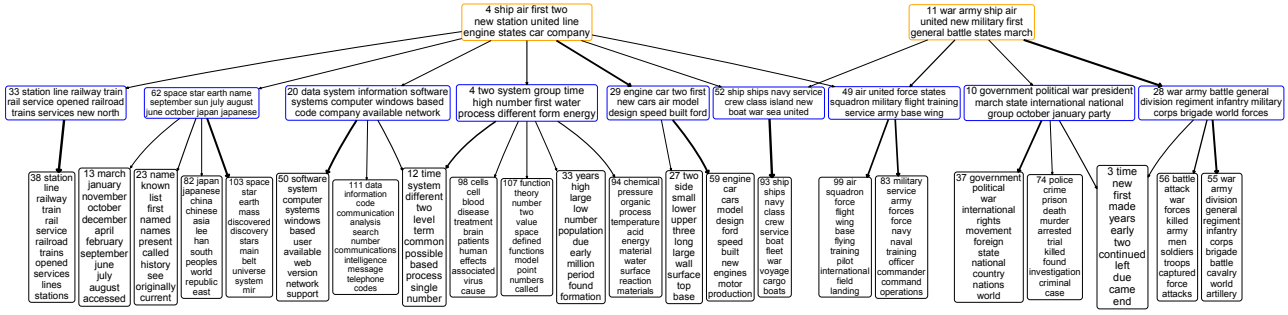

Figure 18. Analogous plots to Figure 7 for a subnetwork on "ship & air" from Wiki, consisting of two trees rooted at nodes 4 and 11, respectively, of layer three.

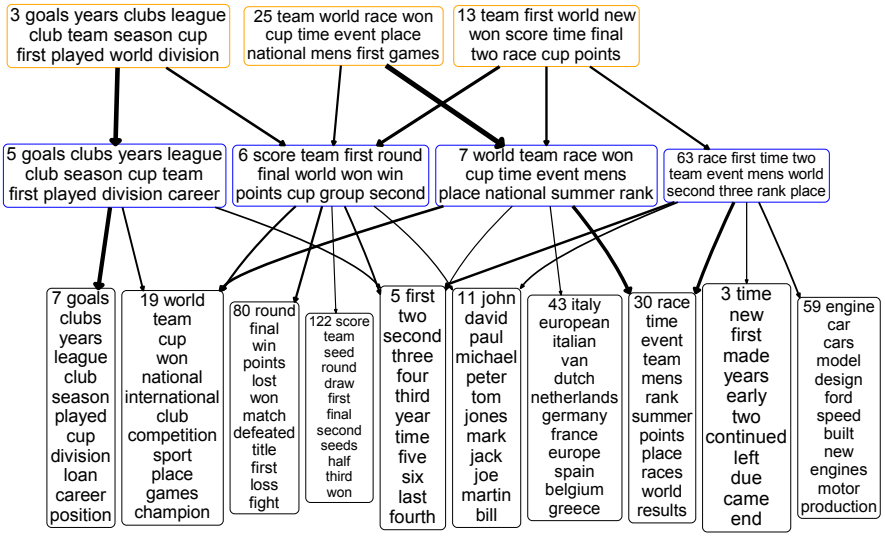

Figure 19. Analogous plots to Figure 7 for a subnetwork on "team & race" from Wiki, consisting of three trees rooted at nodes 3, 25, and 13, respectively, of layer three.