E-Companion for Identifying Influential Users by Topic in Unstructured User-generated Content

EC.1 Model Inference Procedure

According to the model specification in Section 3, the joint likelihood of our proposed model becomes:

$$\begin{aligned} & p(w,z,\eta,\phi,\alpha,\beta,\gamma,\delta) \\ &= \prod_{u=1}^{U}\prod_{t=1}^{T}\left\{\prod_{i=1}^{N_{ut}} p(w_{uti}\mid z_{uti},\phi)p(z_{uti}\mid \eta_{ut})\right\}\prod_{k=1}^{K} p(\eta_{utk}\mid \eta_{\cdot t-1k},\alpha_k,\beta_{u\cdot k},\gamma_{tk},\delta_{uk})\times \\ & \quad p(\phi,\eta_{\cdot 0},\alpha,\beta,\gamma,\delta) \end{aligned} \tag{EC.1}$$

where $p(\phi,\eta_{\cdot 0},\alpha,\beta,\gamma,\delta)$ represents the joint prior distribution of the model unknowns, which will be specified in Section EC.2.

We now derive the procedure for model inference. Our approach utilizes the Markov Chain Monte Carlo (MCMC) method to generate samples from the posterior distributions for a predetermined number of K topics. A notable challenge in this process is the lack of a closed-form expression for the posterior distribution, primarily due to the logistic normal prior, as defined in Equation 2. To overcome this, we employ the technique proposed by Polson et al. (2013), which represents the categorical likelihood as a mixture of Gaussians with respect to a Pólya-Gamma distribution. As a result, the likelihood of the topic distribution for corresponding topic assignments is defined as follows:

$$\begin{aligned} p(z_{ut}\mid \eta_{ut}) &\propto \left(\frac{\exp(\eta_{ut1})}{\sum_{k'}\exp(\eta_{utk'})}\right)^{N_{ut1}}\cdots\left(\frac{\exp(\eta_{utK})}{\sum_{k'}\exp(\eta_{utk'})}\right)^{N_{utK}} \\ &\propto \exp\left(\kappa_{utk}\psi_{utk}-\frac{\zeta_{utk}}{2}\psi_{utk}^2\right), \end{aligned} \tag{EC.2}$$

where $\psi_{utk}=\eta_{utk}-\log\sum_{k'\neq K}\exp(\eta_{utk'}), \kappa_{utk}=N_{utk}-\frac{N_{ut}}{2}$, and N_{utk} is the number of elements assigned to topic k in the content generated by user u at time t. ζ_{utk} is an auxiliary variable following the Pólya-Gamma distribution $\zeta_{utk}\sim PG(N_{ut},0)$. As a result, we can derive the posterior distribution in a closed form.

To make inference on the dynamic processes that govern the topic distribution, we apply the forward filtering and backward sampling algorithm (Cater and Kohn 1994). Thus, we can rewrite the model representation of the hierarchical regression structure of the topic distribution in Equation 3 for derivation of the posterior distribution. Let $\eta_{tk}=(\eta_{1tk},\ldots,\eta_{Utk})^\top$ be a stacked form of η_{utk} over all users.

$$\eta_{tk}\sim N(B_k\eta_{t-1k}+C_{tk},I), \tag{EC.3}$$

$$\text{where } [B_k]_{uu'}=\begin{cases}\alpha_k & \text{if } u=u'\\ \beta_{uu'k} & \text{if } u\neq u', u'\in\mathcal{F}_u\\ 0 & \text{otherwise}\end{cases}, \qquad C_{tk}=\gamma_{tk}+\begin{pmatrix}\delta_{1k}\\ \vdots\\ \delta_{Uk}\end{pmatrix}$$

A coefficient matrix B_k is a $U \times U$ square matrix that can be interpreted as a social network weighted by autoregressive coefficients for self-loops (diagonal) and social influences on network edges (non-diagonal). This reformulation allows us to derive the filtering distribution and the smoothing distribution as follows.

$$
\begin{aligned}
p(\eta_{tk} \mid z_{1:t}, \cdots) &\propto N(\mu_{tk}, \Sigma_{tk}) \qquad \text{(EC.4)}\\
\Sigma_{tk} &= \left(S_{tk}^{-1} + diag(\zeta_{tk})\right)^{-1}, \quad S_{tk} = I + B_k \Sigma_{t-1k} B_k^\top \\
\mu_{tk} &= \Sigma_{tk}\left(\kappa_{tk} + \zeta_{tk} \cdot \log \sum_{k' \neq k} \exp(\eta_{tk'}) + S_{tk}^{-1}(B_k \mu_{t-1k} + C_{tk})\right) \\
p(\eta_{tk} \mid z_{1:T}, \cdots) &\propto N(\tilde{\mu}_{tk}, \tilde{\Sigma}_{tk}) \qquad \text{(EC.5)}\\
\tilde{\mu}_{tk} &= \mu_{tk} + G_{tk}\left(\tilde{\mu}_{t+1k} - B_k \mu_{tk} - C_{t+1k}\right), \quad G_{tk} = \Sigma_{tk} B_k^\top \left(I + B_k \Sigma_{tk} B_k^\top\right)^{-1} \\
\tilde{\Sigma}_{tk} &= \tilde{S}_{tk} + G_{tk} \tilde{\Sigma}_{t+1k} G_{tk}^\top, \quad \tilde{S}_{tk} = \Sigma_{tk} - G_{tk}(I + B_k \Sigma_{tk} B_k^\top) G_{tk}
\end{aligned}
$$

More details of this derivation and posterior distributions of the remaining parameters are in EC.2. Furthermore, we conduct numerical experiments on synthetic data to validate the inference procedure regarding the performance of recovering model parameters. These results are presented in EC.3.

EC.2 Details of the Posterior Distributions

In this appendix, we show the detailed derivation process of the filtering distribution and smoothing distribution (the derivation is based on Sarkka 2013), which were omitted in the text. As described in Section EC.1, we adopt the forward filtering and backward sampling algorithm to sample from the posterior distribution of the topic distribution. Note that in the following, we omit the notation of parameters without our focus in this section for simplicity. First, let the filtering distribution at time $t-1$ be $p(\eta_{t-1k} \mid z_{1:t-1}) = N(\mu_{t-1k}, \Sigma_{t-1k})$, and then the joint distribution of η_{t-1k} and η_{tk} given data up to $t-1$ is defined as follows.

$$
\begin{aligned}
p(\eta_{t-1k}, \eta_{tk} \mid z_{1:t-1}) &= p(\eta_{tk} \mid \eta_{t-1k}) p(\eta_{t-1k} \mid z_{1:t-1}) \\
&= N(\eta_{tk}; B_k \eta_{t-1k} + C_{tk}, I) N(\eta_{t-1k}; \mu_{t-1k}, \Sigma_{t-1k}) \\
&= N(m_1, S_1), \qquad \text{(EC.6)}\\
\text{where } m_1 = \begin{pmatrix} \mu_{t-1k} \\ B_k \mu_{t-1k} + C_{tk} \end{pmatrix}, &\qquad S_1 = \begin{pmatrix} \Sigma_{t-1k} & \Sigma_{t-1k} B_k^\top \\ B_k \Sigma_{t-1k} & I + B_k \Sigma_{t-1k} B_k^\top \end{pmatrix}
\end{aligned}
$$

The last line is obtained by using Lemma A.1 of Sarkka (2013). By marginalizing the joint distribution with respect to η_{t-1k}, we obtain the following conditional distribution.

$$
\begin{aligned}
p(\eta_{tk} \mid z_{1:t-1}) &= N(m_2, S_2) \qquad \text{(EC.7)}\\
\text{where } m_2 = B_k \mu_{t-1k} + C_{tk}, &\qquad S_2 = I + B_k \Sigma_{t-1k} B_k^\top
\end{aligned}
$$

The likelihood of η_{tk} with respect to z_t can be obtained by using (EC.2).

$$
\begin{aligned}
p(z_t \mid \eta_{tk}) &\propto \exp\left(\psi_{tk}^\top \kappa_{tk} - \frac{1}{2}\psi_{tk}^\top diag(\zeta_{tk})\psi_{tk}\right) \qquad \text{(EC.8)}\\
\text{where } \psi_{tk} &= \{\psi_{1tk}, \ldots, \psi_{Utk}\}^\top = \eta_{tk} - \log \sum_{k' \neq k} \exp(\eta_{tk'})\\
\kappa_{tk} &= \{\kappa_{1tk}, \ldots, \kappa_{Utk}\}^\top, \qquad \zeta_{tk} = \{\zeta_{1tk}, \ldots, \zeta_{Utk}\}^\top
\end{aligned}
$$

Therefore, the posterior distribution of η_{tk} when observing the data up to t (i.e., filtering distribution) is given as follows.

$$
\begin{aligned}
p(\eta_{tk} \mid z_t, z_{1:t-1}) &\propto p(z_t \mid \eta_{tk}) p(\eta_{tk} \mid z_{1:t-1})\\
&\propto \exp\left(\psi_{tk}^\top \kappa_{tk} - \frac{1}{2}\psi_{tk}^\top diag(\zeta_{tk})\psi_{tk}\right) N(m_2, S_s)\\
&= N(\mu_{tk}, \Sigma_{tk}) \qquad \text{(EC.9)}\\
\text{where } \Sigma_{tk} &= \left(S_2^{-1} + diag(\zeta_{tk})\right)^{-1},\\
\mu_{tk} &= \Sigma_{tk}\left(\kappa_{tk} + \zeta_{tk} \cdot \log \sum_{k' \neq k} \exp(\eta_{tk'}) + S_2^{-1} m_2\right)
\end{aligned}
$$

Next, as with the above, the joint distribution of η_{tk} and η_{t+1k} given $z_{1:t}$ is as follows.

$$
\begin{aligned}
p(\eta_{tk}, \eta_{t+1k} \mid z_{1:t}) &= N(\tilde{m}_1, \tilde{S}_1) \qquad \text{(EC.10)}\\
\text{where } \tilde{m}_1 &= \begin{pmatrix} \mu_{tk} \\ B_k \mu_{tk} + C_{t+1k} \end{pmatrix}, \qquad \tilde{S}_1 = \begin{pmatrix} \Sigma_{tk} & \Sigma_{tk} B_k^\top \\ B_k \Sigma_{tk} & I + B_k \Sigma_{tk} B_k^\top \end{pmatrix}
\end{aligned}
$$

Since the joint distribution is Gaussian, the conditional distribution is easily obtained as follows.

$$
\begin{aligned}
p(\eta_{tk} \mid \eta_{t+1k}, z_{1:t}) &= N(\tilde{m}_2, \tilde{S}_2) \qquad \text{(EC.11)}\\
\text{where } \tilde{m}_2 &= \mu_{tk} + G_{tk}(\eta_{t+1k} - B_k \mu_{tk} - C_{t+1k})\\
G_{tk} &= B_k \Sigma_{tk} (I + B_k \Sigma_{tk} B_k^\top)^{-1}\\
\tilde{S}_2 &= \Sigma_{tk} - G_{tk}(I + B_k \Sigma_{tk} B_k^\top) G_{tk}^\top
\end{aligned}
$$

Let the smoothing distribution at $t+1$ be $p(\eta_{t+1k} \mid z_{1:t}) = N(\tilde{\mu}_{t+1k}, \tilde{\Sigma}_{t+1k})$, and since $p(\eta_{tk} \mid \eta_{t+1k}, z_{1:T}) = p(\eta_{tk} \mid \eta_{t+1k}, z_{1:t})$ from the model specification, we can obtain the joint distribution when observing the whole data as follows.

$$
\begin{aligned}
p(\eta_{tk}, \eta_{t+1k} \mid z_{1:T}) &= p(\eta_{tk} \mid \eta_{t+1k}, z_{1:t}) p(\eta_{t+1k} \mid z_{1:T})\\
&= N(\tilde{m}_2, \tilde{S}_2) N(\tilde{\mu}_{t+1k}, \tilde{\Sigma}_{t+1k})\\
&= N(\tilde{m}_3, \tilde{S}_3) \qquad \text{(EC.12)}\\
\text{where } \tilde{m}_3 &= \begin{pmatrix} \tilde{\mu}_{t+1k} \\ \mu_{tk} + G_{tk}(\tilde{\mu}_{t+1k} - B_k \mu_{tk} - C_{t+1k}) \end{pmatrix}\\
\tilde{S}_2 &= \begin{pmatrix} \tilde{\Sigma}_{t+1k} & \tilde{\Sigma}_{t+1k} G_{tk}^\top \\ G_{tk} \tilde{\Sigma}_{t+1k} & \tilde{S}_2 + G_{tk} \tilde{\Sigma}_{t+1k} G_{tk}^\top \end{pmatrix}
\end{aligned}
$$

Therefore, we can obtain the posterior distribution of η_{tk} when observing the whole data (i.e., smoothing distribution) by marginalizing the joint distribution with respect to η_{t+1k}.

$$p(\eta_{tk} \mid z_{1:T}) = N(\tilde{\mu}_{tk}, \tilde{\Sigma}_{tk}) \tag{EC.13}$$
$$\text{where } \tilde{\mu}_{tk} = \mu_{tk} + G_{tk}(\tilde{\mu}_{t+1k} - B_k\mu_{tk} - C_{t+1k}), \qquad \tilde{\Sigma}_{tk} = \tilde{S}_2 + G_{tk}\tilde{\Sigma}_{t+1k}G_{tk}^{\top}$$

In the MCMC iteration, we calculate the filtering distribution forwards, and then let us regard $\tilde{\mu}_{Tk} = \mu_{Tk}, \tilde{\Sigma}_{Tk} = \Sigma_{Tk}$ to sample from the smoothing distribution backwards.

If η_{ut} and z_{ut} are given, ζ_{utk} can be also sampled from the following Pólya-Gamma distribution (Polson et al. 2013).

$$p(\zeta_{utk} \mid \eta_{ut}, z_{ut}) \propto PG\left(N_{ut}, \eta_{utk} - \log \sum_{k' \neq k} \exp(\eta_{utk'})\right) \tag{EC.14}$$

Since we can easily derive the posterior distributions of the remaining parameters as with the conventional Bayesian estimation of the normal linear regression model (Rossi et al. 2005) and topic models (Griffiths and Steyvers 2004), only the obtained distributions are displayed in the following.

$$p(z_{uti} = k \mid \cdots) \propto \exp(\eta_{utk}) \times \frac{N_{kv\setminus uti} + \phi_0}{N_{k\setminus uti} + \phi_0 \cdot V}, \qquad \text{where} \quad \phi_k \sim Dirichlet(\phi_0) \tag{EC.15}$$

$$p(\alpha_k \mid \cdots) \propto N(\mu, \sigma^2), \qquad \text{where} \quad \sigma^2 = \left(\sum_{u=1}^{U}\sum_{t=1}^{T} \eta_{ut-1k}^2 + \frac{1}{\sigma_{\alpha 0}^2}\right)^{-1} \tag{EC.16}$$
$$\mu = \sigma^2 \left(\sum_{u=1}^{U}\sum_{t=1}^{T} \eta_{ut-1k}\left(\eta_{utk} - \sum_{f \in \mathcal{F}_u} \beta_{ufk} \cdot \eta_{ft-1k} - \gamma_{tk} - \delta_{uk}\right)\right)$$

$$p(\beta_{ufk} \mid \pi_{ufk} = 1, \ldots) \propto N(\mu, \sigma^2), \qquad \text{where} \quad \sigma^2 = \left(\sum_{t=1}^{T} \eta_{ft-1k}^2 + \frac{1}{\sigma_{\beta}^2}\right)^{-1} \tag{EC.17}$$
$$\mu = \sigma^2 \left(\sum_{t=1}^{T} \eta_{ft-1k}\left(\eta_{utk} - \alpha_k \cdot \eta_{ut-1k} - \sum_{f' \in \mathcal{F}_u} \beta_{uf'k} \cdot \eta_{f't-1k} - \gamma_{tk} - \delta_{uk}\right) + \frac{x_{uf}^{\top}\rho_k}{\sigma_{\beta}^2}\right)$$

$$p(\beta_{ufk} \mid \pi_{ufk} = 0, \ldots) \propto N(\mu, \sigma^2), \qquad \text{where} \quad \sigma^2 = \left(\sum_{t=1}^{T} \eta_{ft-1k}^2 + \frac{1}{\sigma_{\beta}^2 \cdot \omega_{ufk}}\right)^{-1} \tag{EC.18}$$
$$\mu = \sigma^2 \left(\sum_{t=1}^{T} \eta_{ft-1k}\left(\eta_{utk} - \alpha_k \cdot \eta_{ut-1k} - \sum_{f' \in \mathcal{F}_u} \beta_{uf'k} \cdot \eta_{f't-1k} - \gamma_{tk} - \delta_{uk}\right)\right)$$

$$p(\gamma_{tk} \mid \cdots) \propto N(\mu, \sigma^2), \qquad \text{where} \quad \sigma^2 = \left(U + \frac{1}{\sigma_{\gamma 0}^2}\right)^{-1} \tag{EC.19}$$
$$\mu = \sigma^2 \left(\sum_{u=1}^{U} \eta_{utk} - \alpha_k \cdot \eta_{ut-1k} - \sum_{f \in \mathcal{F}_u} \beta_{ufk} \cdot \eta_{ft-1k} - \delta_{uk}\right)$$

$$p(\delta_{uk} \mid \cdots) \propto N(\mu, \sigma^2), \qquad \text{where} \quad \sigma^2 = \left(T + \frac{1}{\sigma_{\delta 0}^2}\right)^{-1} \tag{EC.20}$$
$$\mu = \sigma^2 \left(\sum_{t=1}^{T} \eta_{utk} - \alpha_k \cdot \eta_{ut-1k} - \sum_{f \in \mathcal{F}_u} \beta_{ufk} \cdot \eta_{ft-1k} - \gamma_{tk}\right)$$

EC.3 Performance of Parameter Recovery

In this section, we conduct parameter recovery experiments using synthetic data to validate the performance of the proposed model and the estimation procedure. As introduced in Section 3, since the proposed model defines social influence as the lagged correlation between users' latent topic distributions, we should demonstrate the reliability of the estimated social influence on imaginary variables through a numerical experiment.

To evaluate the performance of parameter recovery, now we suppose several scenarios. The number of users (U) and the number of times (T) are set to be 100 or 200, and the number of topics (K) is set to be 5, 10, or 20. Another scenario is the sparsity proportion of social influence, specifically, in the case of $s\%$ sparsity, only randomly chosen $s\%$ of all edges in the generated network are given non-zero value of β, while the remaining $1-s\%$ of edges do not have any influence. Following the setting of each scenario, a random network and the values of parameters in Equation 3 are initialized, and then the topic distribution is set according to the hierarchical structure. The element distribution is also randomly set with $\phi_k \sim Dirichlet(\phi_{0k}),\ \phi_{0k} = \{\phi_{0k1}, \ldots, \phi_{0kV}\}^\top$, where ϕ_{0kv} corresponding 50 unique objects for each topic is ten times the others, specifically, in the case of $k = 1$, $\phi_{0k1} = \phi_{0k50} = 10$, while $\phi_{0k51} = \phi_{0kV} = 1$. Thus, the size of vocabulary is $V = 50 \times K$. Given the generated topic distribution and element distribution, the data w is generated according to the generative process in Section 3.1. Using the generated data, the model is estimated by MCMC described in the previous section.

Figure EC.1 and EC.2 show the values of root mean square error (RMSE) and correlation coefficient between the true values and the estimated values for each scenario and parameter. Although the parameters of the hierarchical regression model in Equation 3 are not recovered well, the topic distribution and the element distribution, which are the parameters of interest, are correctly estimated with about 0.8 of the correlation coefficient in most scenarios. For each scenario, the accuracy of the estimation tends to be worse as the number of data (U, T) and the number of topics (K), that is, the number of parameters to be estimated, increase. Since the parameters of interest in this study are topic distribution, element distribution, and social influence as can be seen from the discussion in Section 4.3, improving the estimation accuracy of the remaining parameters (α, γ, δ) is out of scope.

Next, to validate the prior distribution for social influence in the proposed model, several models with different prior settings are estimated for each scenario. In the field of Bayesian statistics, in addition to the Bayesian lasso prior assumed in the proposed model, the horseshoe prior (Carvalho et al. 2010) and the Dirichlet-Laplace prior (Bhattacharya et al. 2015) have been used as shrinkage priors. The definitions of the horseshoe prior and the Dirichlet-Laplace prior are

$$\beta_{ufk} \mid \omega_{ufk}, \tau_k \sim N(0, \omega_{ufk}^2 \cdot \tau_k^2), \qquad \omega_{ufk}^2 \sim C^+(0,1), \qquad \tau_k^2 \sim C^+(0,1) \tag{EC.21}$$

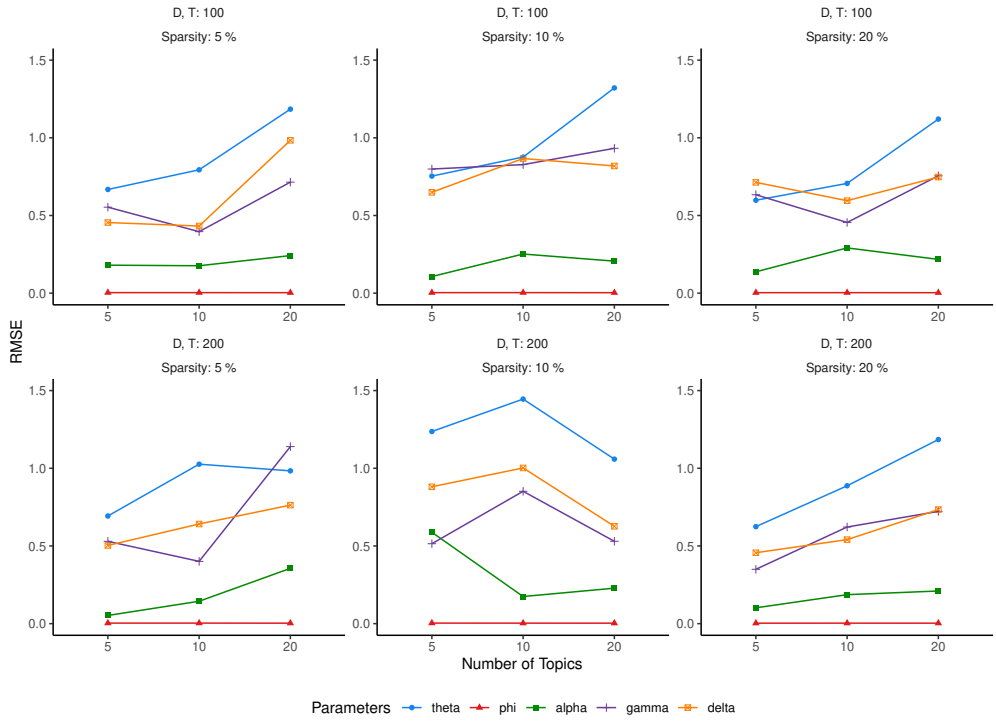

Figure EC.1 RMSE

and

$$\beta_{ufk} \mid \xi_{ufk}, \omega_{ufk}, \tau_k \sim N(0, \xi_{ufk} \cdot \omega_{ufk}^2 \cdot \tau_k^2) \tag{EC.22}$$
$$\xi_{ufk} \sim Exp\left(\frac{1}{2}\right), \qquad \omega_k \sim Dirichlet\left(\frac{1}{\sum_{u=1}^{U} |\mathcal{F}_u|}\right), \qquad \tau_k \sim Gamma\left(1, \frac{1}{2}\right),$$

respectively. In this simulation, we compare the performance of the Bayesian lasso prior with three different prior distributions, including these plus weakly informative prior ($\beta_{ufk} \sim N(0, 10^2)$), which does not assume sparsity. Figure EC.3 shows the values of RMSE between true and estimates. All models can accurately recover the true values, among which Bayesian lasso is superior to others in most of the scenarios. Moreover, even when the scale of the model (U, T, K) is increased, the RMSEs of social influence do not get so worse as the other parameters. Figure EC.4 shows the F-measure which is calculated by regarding users as influential when the estimated β is 0.5 or higher in absolute value. The F-measures are high in all scenarios, among which Bayesian lasso outperforms the others, and it indicates that the proposed model can provide reliable estimates of social influences among users.

EC.4 Simulation Experiments for Large Network Data

In this section, we discuss the computational cost of applying the proposed model to larger-scale networks, rather than the medium-scale network dataset used in the empirical analysis. Figure EC.5 illustrates the computation time required per iteration for estimating the proposed model as the number of users. As

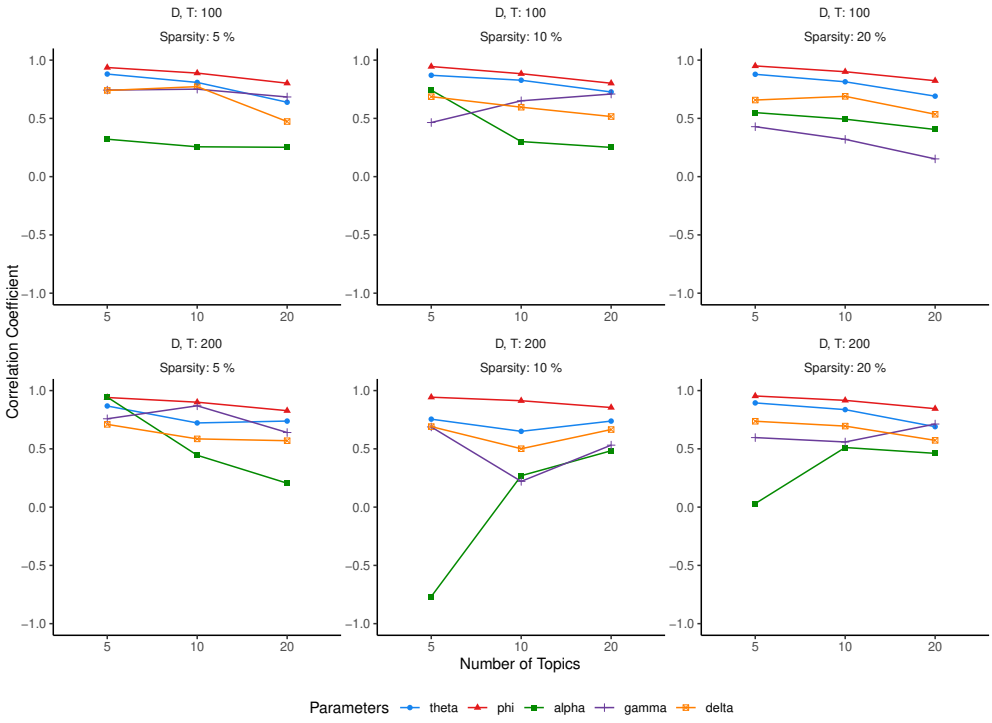

Figure EC.2 Correlation coefficient

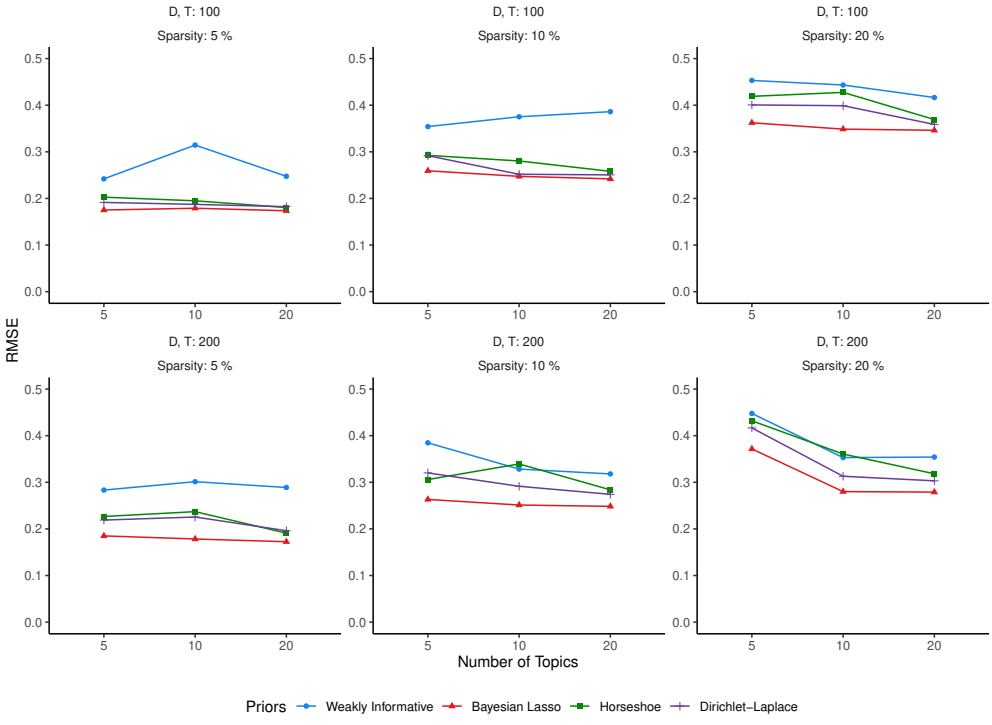

Figure EC.3 RMSE

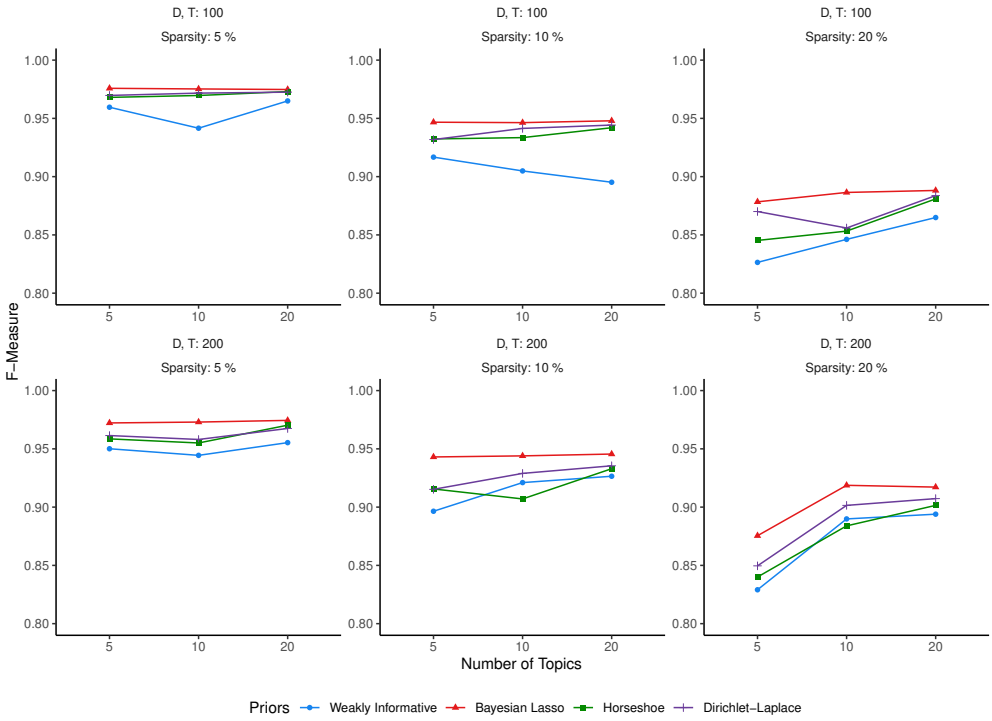

Figure EC.4 F-measure

shown, the computation time increases exponentially with the number of users. This exponential growth is likely due to the bottleneck caused by the matrix multiplication and inversion in Equation EC.10 and EC.14, which could pose a significant challenge when applying the proposed model to large-scale networks. Although this study has not identified a solution to this high computational cost, recent advancements in machine learning, particularly in low-rank matrix approximations, offer promising avenues for reducing computational demands. By leveraging these techniques, it may be possible to estimate user-specific topic distributions with realistic computational costs by approximating the large-scale social influence network, while still accurately estimating the social influence between user pairs. While addressing this challenge is beyond the scope of the current study and remains a topic for future research, it is an essential issue given the high demand in practice for effectively designing SMCs within large-scale networks involving thousands or even millions of users.

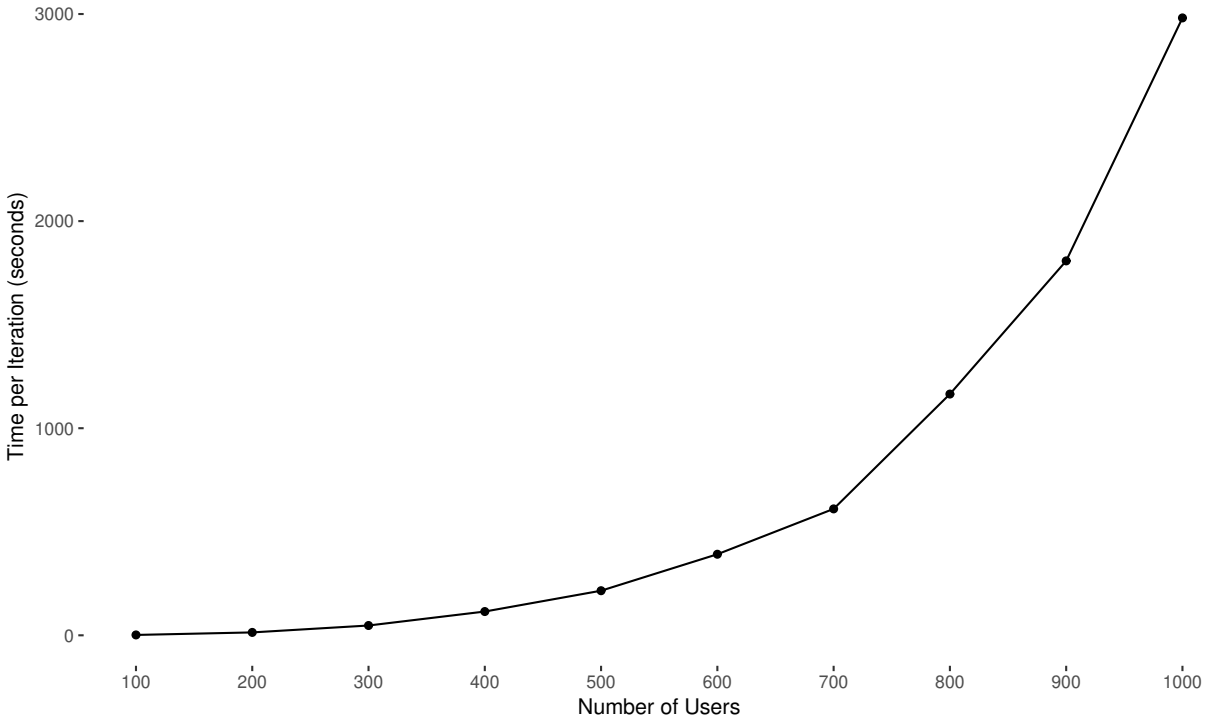

Figure EC.5 Computation time of the proposed model