PTEM: A POPULARITY-BASED TOPICAL EXPERTISE MODEL FOR COMMUNITY QUESTION ANSWERING

By Hohyun Jung†, Jae-Gil Lee*,‡, Namgil Lee§ and Sung-Ho Kim†‡

Korea Advanced Institute of Science and Technology†‡ and Kangwon National University§

Community Question Answering (CQA) websites are widely used in sharing knowledge, where users can ask questions, reply answers, and evaluate answers. So far, the evaluation of answers has been explained by the contents of answers through the investigation of users’ topics of interest and expertise levels. In this paper, we focus on modeling the user’s evaluation behavior in that users can see the answerer’s profile as well as the answer content before evaluating the quality of the answer. We propose a model called Popularity-based Topical Expertise Model (PTEM), a generative model to analyze the rich-get-richer phenomenon that popular user’s answers are more recommended. We can simultaneously estimate the topical expertise of each user and the strength of the rich-get-richer effect through the EM algorithm combined with collapsed Gibbs sampling. Experiments are performed on the StackExchange data, and the results demonstrate a rich-get-richer phenomenon in the community. We further discuss the superiority and usefulness of the proposed model through analysis in the discipline of philosophy.

1. Introduction. People upload, obtain, and share a variety of information over the world wide web. Among them, we can find many websites based on the Community Question Answering (CQA) platform such as Yahoo! Answers, StackExchange, Naver Kin, and Quora. It has the advantage of being able to ask questions directly and obtain the solution through the answers posted by community users. The typical CQA is structured as shown in Figure 1. When a question is registered, interested users post their answers. The asker can select the most helpful answer. Users can also recommend answer by an up-vote. For each question and answer, the posted user’s

*Supported by the KAIA grant (No. 19NSIP-B081011-06)
†Supported by the NRF grant (No. 2016R1D1A1B03936155) of the Republic of Korea

Keywords and phrases: Topic analysis, EM algorithm, Community question answering, Rich-get-richer, User behavior

1https://answers.yahoo.com
2https://stackexchange.com
3https://kin.naver.com
4https://www.quora.com
information such as profile and rank on the community and some badges are exposed together.

<table>
<thead>
<tr>
<th>CQA Forum</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Question.</strong></td>
</tr>
<tr>
<td>Asker information, Content</td>
</tr>
<tr>
<td><strong>Answer 1.</strong> (Accepted)</td>
</tr>
<tr>
<td>Answerer 1’s information, Vote, Content</td>
</tr>
<tr>
<td><strong>Answer 2.</strong></td>
</tr>
<tr>
<td>Answerer 2’s information, Vote, Content</td>
</tr>
</tbody>
</table>

**Fig. 1. The structure of CQA forum.**

In this paper, we aim to investigate factors that affect the answer evaluation of users. In many CQA platforms, we can easily find information on the answerer since the community provides the activity history, community levels or ranks, reputation score, and badges of users. We focus on this user behavior: community users judge the reliability of an answer in consideration of the reputation of the answerer as well as the content of the answer. Users would be biased that an answer might be reliable if it is by a popular user regardless of the quality of the answer. This leads to a rich-get-richer phenomenon which is also called a popularity effect. The rich-get-richer phenomenon is widely seen in many areas of social sciences [Merton (1968), Kondor et al. (2014), Van de Rijt et al. (2014), Perc (2014), Jung et al. (2018)].

1.1. **Related Work.** The CQA community has been studied in terms of the qualities of answers and user expertise, see Srba and Bielikova (2016) for a comprehensive review of literature. Among them, user expertise analysis is of great interest, and there exists two principal approaches: global and topic-specific analyses.


Concerning the topic-specific expertise analysis, the initial topic model was proposed by Papadimitriou et al. (2000), and Hofmann (1999) proposed
the probabilistic latent semantic analysis (PLSA) which was extended toward Latent Dirichlet Allocation (LDA) by Blei, Ng and Jordan (2003). The LDA has been widely used in finding the structure of documents, classification, and so on. It assumes that the documents are composed of a bag of words and that words come from a specific topic.

Topic models have been used extensively on the CQA since users have their topics of interest. Especially, it is challenging to recommend users who would be able to give the best answer for a given question. An expert answerer should have an interest in the topic of the question, and also the expertise level should be high on that topic. The topic models can provide topic-specific expertise levels of users.

Cao et al. (2010) proposed the LDA topic model based on the similarity measure, and Cai and Chakravarthy (2013) proposed the ExpertRank framework to estimate the expertise level using a graph structure as well as topic-specific information. Zhou et al. (2014) considered link structure and topical similarity between askers and answerers by mixing graph-based PageRank and LDA semantic model. Besides, several models have been proposed to predict the expertise level of users [Bouguessa, Dumoulin and Wang (2008), Pal et al. (2011), Movshovitz-Attias et al. (2013)].

Investigations have been made on the impact of the reputation or profile on evaluations of questions and answers. Tausczik and Pennebaker (2011) argued that user reputations play a decisive role in determining question quality through MathOverflow data. They assumed that the question quality is well represented by the vote of a question. Paul, Hong and Chi (2012) interviewed Quora users and found that users judge other users based on their past contributions. The user expertise analysis was not performed in these studies.

1.2. Our contributions. The popularity or reputation of answers has not yet been considered in literature in the evaluation of answers, and contents of answers are solely considered via topical expertise models [Yang et al. (2013), Yang and Manandhar (2014), Ma et al. (2015), Xu, Ji and Wang (2012)]. In this paper, we propose a popularity-based topical expertise model (PTEM) in an effort to explain the popularity effect as well as the topics of the community and the expertise levels of users. To the best of our knowledge, the PTEM is the first model to analyze the rich-get-richer phenomenon concerning user expertise in the CQA community [Srba and Bielikova (2016), Patra (2017), Wang et al. (2018)]. We assume that the vote (the number of recommendations) is influenced by two factors, the expertise and popularity levels of answerers, through a negative binomial model. The proposed model
can analyze how the rich-get-richer phenomenon affects the mechanism of getting recommendations by community users.

We develop an algorithm to simultaneously estimate the strengths of a popularity effect, topics of the community, topics of interests of users, and topic-specific expertise levels of users. We employ an MCMC-EM (Markov Chain Monte Carlo - Expectation Maximization) algorithm based on a collapsed Gibbs sampling. The expertise levels can be estimated in a more flexible manner by allowing continuum of values. We also suggest a model selection method based on the Akaike information criterion.

Finally, we conduct experiments on the StackExchange community. The analysis shows that the rich-get-richer effect is present in the community. If we would estimate the expertise level without considering this rich-get-richer phenomenon, then the estimate could be biased, e.g., the expertise of a popular user could be overestimated. We perform a detailed analysis on topics of the community and expertise levels of users in the field of philosophy considering the rich-get-richer phenomenon.

The remainder of the paper is organized as follows. In Section 2, we define notations and present our model. We propose the estimation procedures and make inferences in Sections 3 and 4, respectively. The model selection criterion of the proposed model is suggested in Section 5 with a supporting simulation study. In Section 6, we analyze StackExchange data by applying the proposed model with interpretations on the result. Section 7 concludes the paper.

2. The Proposed Model.

2.1. Notation and assumption. The askers and answerers in the CQA platform are called users. We denote the users by \( u = 1, 2, \ldots, U \), and the answers uploaded by user \( u \) are expressed in time order as \( a = 1, 2, \ldots, A_u \). Let \( t_{u,a} \) be the time point at which user \( u \)'s answer \( a \) is posted.

The questions and answers are given in text, which can be viewed as a sequence of words. We consider word as the smallest unit of text, and denote distinct words by \( d = 1, 2, \ldots, D \). We can infer the topic of answers based on their contents. We assume that there are \( K \) topics. In this paper, we focus on the content of the answer rather than the question since we aim to find the mechanism of answer’s vote. We further assume that each answer covers only one topic and each topic \( k \) has a distribution of words, denoted by \( \phi_k = (\phi_{k,1}, \ldots, \phi_{k,D}) \), where \( \sum_d \phi_{k,d} = 1 \). Let \( L_{u,a} \) be the number of words in user \( u \)'s answer \( a \), and the list of words is denoted by \( w_{u,a,l}, l = 1, 2, \ldots, L_{u,a} \). Let \( z_{u,a} \) and \( v_{u,a} \) be the topic and vote of user \( u \)'s answer \( a \), respectively.
The user’s topics of interest are represented by the user-topic distribution 
\( \psi_u = (\psi_u,1, \cdots, \psi_u,K) \), where \( \sum_k \psi_{u,k} = 1 \). The expertise level of users on each topic is expressed as 
\( x_u = (x_u,1, \cdots, x_u,K) \) and \( x_{u,k} \) indicates the expertise level of user \( u \) on topic \( k \).

We summarize the notations in Table 1.

### Table 1
Notations and their descriptions.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( U )</td>
<td>total number of users</td>
</tr>
<tr>
<td>( A_u )</td>
<td>total number of user ( u )’s answers</td>
</tr>
<tr>
<td>( L_{u,a} )</td>
<td>total number of words in user ( u )’s answer ( a )</td>
</tr>
<tr>
<td>( K )</td>
<td>total number of topics</td>
</tr>
<tr>
<td>( D )</td>
<td>total number of unique words</td>
</tr>
<tr>
<td>( \psi_u )</td>
<td>topic distribution of user ( u )</td>
</tr>
<tr>
<td>( \phi_k )</td>
<td>word distribution of topic ( k )</td>
</tr>
<tr>
<td>( x_{u,k} )</td>
<td>expertise level of user ( u ) on topic ( k )</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>hyper-parameter of prior on user-topic distributions</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>hyper-parameter of prior on topic-word distributions</td>
</tr>
<tr>
<td>( t_{u,a} )</td>
<td>posting time of user ( u )’s answer ( a )</td>
</tr>
<tr>
<td>( z_{u,a} )</td>
<td>topic of user ( u )’s answer ( a )</td>
</tr>
<tr>
<td>( v_{u,a} )</td>
<td>vote of user ( u )’s answer ( a )</td>
</tr>
</tbody>
</table>

2.2. **The popularity-based topical expertise model.** We present the generative process of the PTEM in CQA platform. The prior distributions and the sampling distribution in a hierarchical Bayesian framework are summarized as follows:

- For user \( u = 1, 2, \cdots, U \):
  - Draw user-topic distribution \( \psi_u = (\psi_u,1, \cdots, \psi_u,K) \sim \text{Dirichlet}(\alpha) \).
  - For topic \( k = 1, 2, \cdots, K \):
    * Draw user-topic-expertise level \( x_{u,k} \sim N(0,1) \).

- For topic \( k = 1, 2, \cdots, K \):
  - Draw topic-word distribution \( \phi_k = (\phi_k,1, \cdots, \phi_k,D) \sim \text{Dirichlet}(\gamma) \).

- For user \( u = 1, 2, \cdots, U \):
  - For answer \( a = 1, 2, \cdots, A_u \):
    * Draw topic \( z_{u,a} \sim \text{Multinomial}(n = 1, \psi_u) \).
    * For \( l = 1, 2, \cdots, L_{u,a} \):
\* Draw word
\hspace{1cm} w_{u,a,l} \sim \text{Multinomial}(n = 1, \phi_{z_{u,a}}).
\hspace{1cm} (2.1)
\* Draw vote
\hspace{1cm} v_{u,a} \sim \text{NegativeBinomial} \left( m = m_{u,a}, \xi \right),
\hspace{1cm} (2.2)
where the probability mass function is defined by
\hspace{1cm} P(v_{u,a} = q) = \frac{\Gamma(\xi + q)}{q!\Gamma(\xi)} \left( \frac{\xi}{\xi + m} \right)^\xi \left( \frac{m}{\xi + m} \right)^q, \hspace{0.5cm} q = 0, 1, \cdots,
\hspace{1cm} \text{which is parametrized by the mean parameter } m \text{ and the shape parameter } \xi \text{ with mean } m \text{ and variance } m + \frac{1}{\xi} m^2.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{plate_notation.png}
\caption{The plate notation of the PTEM. Rectangles and circles are used for observed and latent variables, respectively.}
\end{figure}

The priors of user-topic and topic-word distributions are Dirichlet distributions, and their hyper-parameters are denoted as $\alpha$ and $\gamma$, respectively. The plate notation is given in Figure 2. The vote is assumed to follow a negative binomial distribution. The mean parameter of negative binomial distribution in (2.2) is $m_{u,a}$, and it is affected by the popularity $y_{u,t_{u,a}}$ and topic-specific expertise level $x_{u,z_{u,a}}$ of the answerer $u$, defined by
\hspace{1cm} m_{u,a} = \exp(\beta_0 + \beta_1 x_{u,z_{u,a}} + \beta_2 y_{u,t_{u,a}}),
\hspace{1cm} (2.3)
where $y_{u,t}$ is a scaled popularity of user $u$ at time $t$ and $\beta = (\beta_0, \beta_1, \beta_2)$ are model coefficients. Specifically, $\beta_0$, $\beta_1$, and $\beta_2$ are called the intercept, expertise coefficient, and popularity coefficient, respectively. We simplify the model parameters as $\theta = (\beta_0, \beta_1, \beta_2, \xi)$. 
2.3. Popularity measure. The popularity is related to the votes through Eq. (2.3). It is required to define popularity measurement corresponding to the $a$-th answer of user $u$. For this measure, we define a rule of gaining popularity in CQA:

- An answer gets an upvote: +1 point

It can be modified according to the characteristics of the CQA forum. For the reputation of StackExchange, one earns +10 points for an upvote.

The popularity of user $u$ at time $t$, denoted by $y'_{u,t}$, can be given by

$$y'_{u,t} = \sum_{a:t_{u,a} < t} v_{u,a}.$$ 

The scale of $y'_{u,t}$, $u = 1, 2, \cdots, U$, $a = 1, 2, \cdots, A_u$ is usually much larger than that of expertise levels $x_{u,k}$, $k = 1, 2, \cdots, K$, which follow the standard normal distribution. Moreover, $y'_{u,t}$ increases over time and never decreases. Therefore, we scale the popularity measure with respect to time and let

$$y_{u,t} = \frac{y'_{u,t}}{t - T_0},$$ 

where $T_0$ is the creation time of the CQA forum. This scaling is important in the sense that the popularity is comparable among the users at a specific time $t$ and that we measure the popularity change against time.

3. Algorithm. We present an algorithm for estimating the model parameter $\theta$, the topics of answers $z_{u,a}$, $u = 1, 2, \cdots, U$, $a = 1, 2, \cdots, A_u$, and the topical expertise of users, $x_{u,k}$, $u = 1, 2, \cdots, U$, $k = 1, 2, \cdots, K$. The hyper-parameters $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_K)$ and $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_D)$ are assumed given. The parameter $\theta$ and latent variables, i.e., topics and topical expertise levels, are alternately updated by the MCMC-based EM algorithm. In this process, we use the collapsed Gibbs sampling method to generate latent variables.

In this section, we denote by $b : c$ the sequence of indices from $b$ to $c$, i.e., $b, b + 1, \cdots, c$. For example, $z_{u,1:A_u} = (z_{u,1}, \cdots, z_{u,A_u})$ and $z_{1:U,1:A_u} = (z_{1,1}, \cdots, z_{1,A_1}, \cdots, z_{U,1}, \cdots, z_{U,A_U})$.

3.1. Likelihood functions. The estimation procedure requires the likelihood function. We focus on the vote model in (2.2). The vote $v_{u,a}$ follows a negative binomial distribution, and the probability mass function is given
by
\begin{equation}
\begin{aligned}
P(v_{u,a}|x_{u,z_{u,a}},z_{u,a},y_{u,t_{u,a}},\theta) \\
= \frac{\Gamma(\xi + v_{u,a})}{(v_{u,a})!(\xi)} \left[ \exp(\beta_0 + \beta_1 x_{u,z_{u,a}} + \beta_2 y_{u,t_{u,a}}) \right]^{v_{u,a}} \\
\times \left[ \xi + \exp(\beta_0 + \beta_1 x_{u,z_{u,a}} + \beta_2 y_{u,t_{u,a}}) \right]^{-(\xi + v_{u,a})}.
\end{aligned}
\end{equation}

Let $A_{u,k}$ be the set of topic $k$ answers for user $u$, i.e., $A_{u,k} = \{a \in \{1, 2, \ldots, A_u\} : z_{u,a} = k\}$. Let $v_{u,k} = \{v_{u,a} : a \in A_{u,k}\}$ and $y_{u,k} = \{y_{u,t_{u,a}} : a \in A_{u,k}\}$ be the sets of votes and popularity values of topic $k$ answers, respectively. Using the above notations, the probability distribution function of $v_{u,k}$ is given by
\begin{equation}
P(v_{u,k}|x_{u,k},z_{u,1:A_u},y_{u,k},\theta) = \prod_{a \in A_{u,k}} P(v_{u,a}|x_{u,z_{u,a}},z_{u,a},y_{u,t_{u,a}},\theta),
\end{equation}

The probability distribution function of all the votes, which covers every topic of all the users, are given by
\begin{equation}
P(v|x,z,y,\theta) = \prod_{u=1}^{U} \prod_{k=1}^{K} P(v_{u,k}|x_{u,k},z_{u,1:A_u},y_{u,k},\theta),
\end{equation}
where $v = v_{1:U,1:A_u}$, $x = x_{1:U,1:K}$, $z = z_{1:U,1:A_u}$, and $y = y_{1:U,1:A_u}$.

**Proposition 1.** The complete data log-likelihood function can be written by
\[
l(\theta|w,v,x,z,\psi,\phi,y,\alpha,\gamma) = \ln P(v|x,z,y,\theta) + (\text{constant on } \theta)
\]
\[
= \sum_{u=1}^{U} \sum_{k=1}^{K} \ln P(v_{u,k}|x_{u,k},z_{u,1:A_u},y_{u,k},\theta) + (\text{constant on } \theta),
\]
where $w = w_{1:U,1:A_u,1:L_{u,a}}$, $\psi = \psi_{1:U,1:K}$, and $\phi = \phi_{1:K,1:D}$.

**Proof.** See Appendix A.1. \qed

Before applying the EM algorithm, we define a function $Q$ as the conditional expectation
\[
Q(\theta|\hat{\theta}(s)) = E \left[ l(\theta|w,v,X,Z,\Psi,\Phi,y,\alpha,\gamma)|w,v,y,\alpha,\gamma,\hat{\theta}(s) \right]
\]
\[
= E \left[ \ln P(v|X,Z,y,\theta)|w,v,y,\alpha,\gamma,\hat{\theta}(s) \right] + (\text{constant on } \theta),
\]
where the expectation is taken on the latent variables $x$, $z$, $\psi$ and $\phi$. Considering Eq. (3.2), we can decompose the function $Q$ as a sum of the components which are each indexed by the user and the topic and define a function, $Q_{u,k}$, as

$$Q_{u,k}(\theta|\hat{\theta}(s)) = E \left[ \ln p(v_{u,k}|X_{u,k}, Z_{u:1:A_u}, y_{u,k}, \theta) | w, v, y, \alpha, \gamma, \hat{\theta}(s) \right].$$

It follows from Eq. (3.3) that

$$Q(\theta|\hat{\theta}(s)) = \sum_{u=1}^{U} \sum_{k=1}^{K} Q_{u,k}(\theta|\hat{\theta}(s)) + \text{(constant on } \theta).$$

Unfortunately, $Q$ is not given in a closed form. For this kind of problem, sampling methods are useful, and we use the Gibbs sampling to estimate the conditional distribution $p(x, z | w, v, y, \alpha, \gamma, \hat{\theta}(s))$. In the sampling procedure, we sample $x$ and $z$ alternately until samples are sufficiently gathered. We require the conditional distributions of topics $z$ and topical expertise levels $x$.

First, we need to find the conditional distributions of

$$z_{u,a}| z_{-(u,a)}, x, w, v, y, \alpha, \gamma, \theta, \quad u = 1, 2, \cdots, U, \quad a = 1, 2, \cdots, A_u,$$

where $z_{-(u,a)}$ is topic assignments excluding user $u$’s answer $a$. We use the process of the collapsed Gibbs sampling, which has been widely employed in the estimation method of the LDA [Griffiths and Steyvers (2004)]. The latent variables $\psi$ and $\phi$ can be marginalized out, and hence it is called collapsed Gibbs sampling. The following proposition enables us to sample the topic of each answer.

**Proposition 2.** The conditional distribution in (3.5) is given by

$$p(z_{u,a} = k | z_{-(u,a)}, x, w, v, y, \alpha, \gamma, \theta) \propto \frac{n_{u,k,-(u,a)} + \alpha_k}{\sum_{k'=1}^{K}(n_{u,k',-(u,a)} + \alpha_{k'})} \prod_{d=1}^{D} \prod_{b=1}^{n_{k,d,-(u,a)}} \left( n_{k,d,-(u,a)} + \gamma_d + b - 1 \right)$$

$$\times p(v_{u,a}|x_{u,k}, z_{u,a} = k, y_{u,t_{u,a}}, \theta),$$

where $n_{u,k,-(u,a)}$ is the number of user $u$’s topic $k$ answers excluding user $u$’s specific answer $a$, $n_{k,d,-(u,a)}$ is the number of particular word $d$ in all topic $k$ answers except user $u$’s specific answer $a$, $n_{k,d,(u,a)}$ is the number of a particular word $d$ in user $u$’s answer $a$, and $n_{k,1:D,(u,a)} = \sum_{d=1}^{D} n_{k,d,(u,a)}$ is the total number of words in user $u$’s answer $a$. 

imsart-aoas ver. 2014/10/16 file: ptem_shkim.tex date: February 14, 2020
Proof. See Appendix A.2.

The term $\sum_{k'=1}^{K}(n_{u,k'} - (u,a) + \alpha_{k'})$ in (3.6) is constant over $k$ and thus may be ignored. We leave it there for the analytic point of view that it is closely related to the proportion of topic $k$ over all topics.

Next, we require the conditional distributions of $x_{u,k}$,

$$x_{u,k} | x_{-(u,k)}, w, v, z, y, \alpha, \gamma, \theta, u = 1, 2, \ldots, U, k = 1, 2, \ldots, K. \tag{3.7}$$

Proposition 3. The conditional distribution in (3.7) is given by

$$p(x_{u,k} | x_{-(u,k)}, w, v, z, y, \alpha, \gamma, \theta) \propto p(v_{u,k} | x_{u,k}, z_{u,1:A_u}, y_{u,k}, \theta) \cdot p(x_{u,k}), \tag{3.8}$$

where $p(x_{u,k})$ is the probability density function of the standard normal distribution.

Proof. See Appendix A.3.

From Proposition 3, it is not hard to show that $p(x_{u,k} | x_{-(u,k)}, z, x, w, v, y, \alpha, \gamma, \theta)$ is log-concave on $x_{u,k}$, since the product of log-concave functions is log-concave. Thus, we can efficiently sample $x_{u,k}$ by using the adaptive rejection sampling (ARS) in which a proposal distribution is not required [Gilks and Wild (1992)].

3.2. Algorithm. Assuming that the model parameter $\theta$ is given, we estimate the latent variables $z$ and $x$ through the collapsed Gibbs sampling algorithm. The detailed procedure is presented in Algorithm 1.

The initial samples are influenced by the initialization. Therefore, the first $G_0$ samples are not used in analysis. We set $G = 110$ and $G_0 = 10$ for data analysis. The function $Q_{u,k}$ can be approximated by

$$\hat{Q}_{u,k}(θ | \hat{θ}(s)) = \frac{1}{G - G_0} \sum_{g=G_0+1}^{G} \ln p(v_{u,k} | x_{u,k}^{(g)}, z_{u,1:A_u}^{(g)}, y_{u,k}, \theta), \tag{3.9}$$

where samples $z_{1:U,1:A_u}^{(g)}, x_{1:U,1:K}^{(g)}$, $g = 1, 2, \ldots, G$, are obtained by Algorithm 1 with parameter $\hat{θ}(s)$ as input. Using Eq. (3.4), the function $Q$ can be approximated by

$$\hat{Q}(θ | \hat{θ}(s)) = \sum_{u=1}^{U} \sum_{k=1}^{K} \hat{Q}_{u,k}(θ | \hat{θ}(s)) + (\text{constant on } \theta), \tag{3.10}$$
Algorithm 1: Collapsed Gibbs Sampling

input : \( w, v, y, \alpha, \gamma, \theta \)
1 Initialize \( z^{(0)}_{1:U,1:A_u} \) and \( x^{(0)}_{1:U,1:K} \)
2 for \( g = 1, 2, \ldots, G \) do
3   for \( u = 1, 2, \ldots, U \) do
4     for \( a = 1, 2, \ldots, A_u \) do
5       Draw \( z^{(g)}_{u,a} \) according to (3.6) using expertise levels \( x^{(g-1)}_{u,1:K} \)
6       and topics \( z^{(g-1)}_{-(u,a)} = \{ z^{(g-1)}_{u',a'} : (u',a') \neq (u,a), u = 1, 2, \ldots, U, a = 1, 2, \ldots, A_u \} \),
7       which are assigned by \( z^{(g)}_{u',a'} = z^{(g)}_{u,a} \) if \( u' < u \) or if \( u' = u \) and \( a' < a \),
8       and \( z^{(g-1)}_{u',a'} \) otherwise.
9     end
10    for \( k = 1, 2, \ldots, K \) do
11       Draw \( x^{(g)}_{u,k} \) according to (3.8) using topics \( z^{(g)}_{u,1:A_u} = z^{(g)}_{u,1:A_u} \).
12    end
13  end
14 end
15 output: \( z^{(g)}_{1:U,1:A_u}, x^{(g)}_{1:U,1:K}, g = 1, 2, \ldots, G \).

Algorithm 2: EM Algorithm

input : \( w, v, y, \alpha, \gamma, \hat{\theta}^{(0)} \)
1 Initialize: \( s = 0 \), converged = False
2 while not converged do
3   (E-step) Run Algorithm 1 with \( \hat{\theta}^{(s)} \), we get Gibbs samples \( z^{(g)}_{1:U,1:A_u}, x^{(g)}_{1:U,1:K}, g = 1, 2, \ldots, G \)
4   (E-step) Find \( \tilde{Q}(\theta | \hat{\theta}^{(s)}) \) in Eq. (3.10)
5   (M-step) Find \( \hat{\theta}^{(s+1)} = \arg\max_{\theta} \tilde{Q}(\theta | \hat{\theta}^{(s)}) \)
6   if convergence criteria is satisfied then
7     converged ← True
8   end
9   \( s ← s + 1 \)
10 end
11 Run Algorithm 1 with converged parameter estimate \( \hat{\theta} \), we get the final Gibbs
12 samples \( z^{(g)}_{1:U,1:A_u}, x^{(g)}_{1:U,1:K}, g = 1, 2, \ldots, G \)
13 output: \( \hat{\theta}, z^{(g)}_{1:U,1:A_u}, x^{(g)}_{1:U,1:K}, g = 1, 2, \ldots, G \)

and the parameter \( \theta \) is estimated by the EM algorithm presented in Algorithm 2.

Algorithm 1 is employed in the E-step. We initialize \( z^{(0)} \) by random topics and \( x^{(0)} \) by zero expertise levels \( x^{(0)}_{u,k} = 0 \) for all users \( u \) and topics \( k \), when
s = 0 in the EM iteration. After the first iteration, we use the last \(G\)-th samples \(z^{(G)}\) and \(x^{(G)}\) at the \((s - 1)\)-th iteration for a fast convergence to the target distribution \(p(x, z|w, v, y, \alpha, \gamma, \hat{\theta}^{(s)})\). A gradient descent method can be used to find \(\theta\) that maximize \(Q(\theta|\hat{\theta}^{(s)})\) in the M-step, and we use the sequential quadratic programming.

Remark. The computation of the E-step in Algorithm 2 takes \(O(GKDU)\) time. We use the gradient descent algorithm in the M-step, and its convergence speed is determined by the complexity of the function \(Q(\theta|\hat{\theta}^{(s)})\). The convergence is achieved after a reasonable number of EM iterations (usually less than 50 iterations) in data analysis. Moreover, we can speed up the computation of the summation of log probability distribution functions used in both E- and M-step by parallel computation, e.g., the computation of (3.10).

4. Inference. Let \(\hat{\theta}\) be the estimated parameter by Algorithm 2. Moreover, let \(z^{(g)}_{1:U,1:A_u}, x^{(g)}_{1:U,1:K}, g = 1, 2, \cdots, G\) be respectively the Gibbs samples of topic assignments and topical expertise levels obtained by Algorithm 2.

The standard errors of the parameter \(\theta\) can be obtained by an information matrix. By Louis’ method [Louis (1982)], the estimated observed information matrix is given by

\[
I(\hat{\theta}) \approx -\nabla^2 Q(\hat{\theta}|\hat{\theta}) - \frac{1}{G - G_0} \sum_{g=G_0+1}^G[D(g)] [D(g)]' \\
+ \left[\nabla Q(\hat{\theta}|\hat{\theta})\right] \left[\nabla Q(\hat{\theta}|\hat{\theta})\right]',
\]

where

\[
D(g) = \sum_{u=1}^U \sum_{k=1}^K \left\{\nabla \ln p(v_{u,k}|x^{(g)}_{u,k}, z^{(g)}_{u,k}, y_{u,k}, \theta)\right\}_{\theta=\hat{\theta}},
\]

and \(\nabla = (\partial/\partial \beta_0, \partial/\partial \beta_1, \partial/\partial \beta_2, \partial/\partial \gamma)'\). The asymptotic covariance matrix of \(\theta\) is \([I(\hat{\theta})]^{-1}\), and the standard error of \(\hat{\theta}\) can be estimated by the square root of diagonal elements.

The topic distribution of user \(u\)’s answer \(a\) can be estimated by

\[
\hat{p}(z_{u,a} = k) = \frac{1}{G - G_0} \sum_{g=G_0+1}^G \mathbb{1}(z_{u,a}^{(g)} = k), \quad k = 1, 2, \cdots, K.
\]

In other words,

\[
\hat{p}(z_{u,a}) = \frac{1}{G - G_0} \sum_{g=G_0+1}^G \sum_{k=1}^K \mathbb{1}(z_{u,a}^{(g)} = z_{u,a}).
\]
To provide the most likely topic of a particular answer, we choose a topic $\hat{z}_{u,a}$ such that $p(z_{u,a} = k)$ is the largest, i.e.,

$$\hat{z}_{u,a} = \text{argmax}_k p(z_{u,a} = k).$$

The topical expertise of user $u$ is estimated by

$$\hat{x}_{u,k} = \frac{1}{G - G_0} \sum_{g = G_0 + 1}^{G} \hat{\psi}^{(g)}_{u,k}.$$  \hspace{1cm} (4.2)

It is informative to provide the user’s topics of interest. Let us denote $n_{u,k} = |\{a \in \{1, 2, \ldots, A_u\} : z^{(g)}_{u,a} = k\}|$ and $n_{k,d} = |\{(u, a, l), u \in \{1, \ldots, U\}, a \in \{1, \ldots, A_u\}, l \in \{1, \ldots, L_{u,a}\} : w_{u,a,l} = d, z^{(g)}_{u,a} = k\}|$ as the number of user $u$’s topic $k$ answers and the number of word $d$ in topic $k$ answers posted by all users, respectively. Then the topic distribution of user $u$ is given by

$$\hat{\psi}_{u,k} = \frac{1}{G - G_0} \sum_{g = G_0 + 1}^{G} \hat{\psi}^{(g)}_{u,k},$$

where

$$\hat{\psi}^{(g)}_{u,k} = \frac{n^{(g)}_{u,k} + \alpha_k}{\sum_{k'=1}^{K} \left( n^{(g)}_{u,k'} + \alpha_{k'} \right)}.$$  \hspace{1cm} (4.3)

Next, the estimated probability of word $d$ covered by topic $k$, i.e., topic-word distribution is given by

$$\hat{\phi}_{k,d} = \frac{1}{G - G_0} \sum_{g = G_0 + 1}^{G} \hat{\phi}^{(g)}_{k,d},$$

where

$$\hat{\phi}^{(g)}_{k,d} = \frac{n^{(g)}_{k,d} + \gamma_d}{\sum_{d'=1}^{D} \left( n^{(g)}_{k,d'} + \gamma_{d'} \right)}.$$  \hspace{1cm} (4.5)

Eqs. (4.4) and (4.6) are derived from the posterior of Dirichlet distributions. Let $x_{u,all}$ be the overall expertise level of user $u$, defined by the weighted mean of topical expertise levels over topics, $x_{u,all} = \sum_{k=1}^{K} \hat{\psi}_{u,k} x_{u,k}$. It can be estimated by

$$\hat{x}_{u,all} = \sum_{k=1}^{K} \hat{\psi}_{u,k} \hat{x}_{u,k}.$$  \hspace{1cm} (4.7)
5. Monte Carlo Simulations. Monte Carlo simulations are performed to check validity of our estimation algorithm. We build a synthetic network using the generative process in Section 2.2. We set the true parameter values as \( \theta = (-0.50, 0.50, 0.20, 2.00) \). We also set \( U = 100 \) users, \( D = 1000 \) distinct words, and \( L_{u,a} = 20, u = 1, \ldots, 100, a = 1, \ldots, A_u \) words for each answer. Let the starting time be \( T_0 = 0 \). For each user \( u \), answers are generated from time \( t = 2 \) to \( t = 10 \), and the time intervals \( t_{u,a} - t_{u,a-1}, a = 1, \ldots, A_u \), follow the exponential distribution with rate 0.8, which means 10 answers on average. We generated answers from time \( t_{u,0} = 2 \) to avoid the exploding popularity in (2.4), which can be caused by extremely small \( t - T_0 \). Finally, let hyper-parameters be \( \alpha = (50/K, \ldots, 50/K) \) and \( \gamma = (0.01, \ldots, 0.01) \) according to Griffiths and Steyvers (2004). We generate 20 synthetic datasets with the number of topics \( K = 5 \) (SD5) and \( K = 10 \) (SD10), respectively.

Algorithm 2 is applied to the generated datasets with different numbers of topics. We exclude users with less than 5 answers. The parameter estimation results for SD5 and SD10 are shown in Tables 2 and 3, respectively.

| Table 2 | Estimates of \( \hat{\theta} \) of the PTEM with synthetic data SD5 assuming various number of topics \( K \). The mean and standard deviation of the 20 dataset applications are presented. |
|---------|---------------------------------|-----------------|-----------------|-----------------|
|         | \( \beta_0 \) (intercept) | \( \beta_1 \) (expertise) | \( \beta_2 \) (popularity) | \( \xi \) |
|         | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| 1       | -0.4018 | 0.0743 | 0.2552 | 0.1136 | 0.1661 | 0.1194 | 1.3224 | 0.2469 |
| 2       | -0.4367 | 0.0706 | 0.3236 | 0.0867 | 0.1911 | 0.1050 | 1.4143 | 0.2073 |
| 3       | -0.4609 | 0.0674 | 0.3882 | 0.0914 | 0.1937 | 0.0832 | 1.5403 | 0.2263 |
| 4       | -0.4764 | 0.0695 | 0.4379 | 0.0867 | 0.1880 | 0.0880 | 1.6778 | 0.2584 |
| 5 (true)| -0.5089 | 0.0668 | 0.5011 | 0.0736 | 0.1892 | 0.1000 | 1.9825 | 0.5313 |
| 6       | -0.5087 | 0.0647 | 0.5065 | 0.0753 | 0.1853 | 0.0947 | 1.9504 | 0.2934 |
| 7       | -0.5137 | 0.0702 | 0.5075 | 0.0963 | 0.1927 | 0.0875 | 2.0239 | 0.5074 |
| 10      | -0.5127 | 0.0728 | 0.4849 | 0.1215 | 0.1987 | 0.0884 | 1.9293 | 0.4169 |
| 15      | -0.5133 | 0.0730 | 0.5039 | 0.0781 | 0.1924 | 0.0858 | 1.9856 | 0.4186 |
| 20      | -0.5182 | 0.0682 | 0.5294 | 0.0768 | 0.1903 | 0.0854 | 2.0719 | 0.4167 |
| 30      | -0.5140 | 0.0724 | 0.5101 | 0.0739 | 0.1878 | 0.0804 | 2.0169 | 0.4382 |

We can see that the mean of the parameter estimate deviates from the true value when applied \( K \) is less than the true number of topics. However, for the number of topics that is larger than or equal to the true number of topics, the mean of the parameter estimate is similar to the true parameter.

This phenomenon seems to indicate importance of the refinement level of topics. If the number of topics were smaller than necessary, then different topics would be merged, which might dilute the effect of expertise. On the other hand, if the topics were properly refined, the effect of expertise should
Table 3
Estimates of $\hat{\theta}$ of the PTEM with synthetic data SD10 assuming various number of topics $K$. The mean and standard deviation of the 20 dataset applications are presented.

<table>
<thead>
<tr>
<th>K</th>
<th>$\hat{\beta}_0$ (intercept)</th>
<th>$\hat{\beta}_1$ (expertise)</th>
<th>$\hat{\beta}_2$ (popularity)</th>
<th>$\xi$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean S.D.</td>
<td>Mean S.D.</td>
<td>Mean S.D.</td>
<td>Mean S.D.</td>
</tr>
<tr>
<td>1</td>
<td>-0.3739 0.0629</td>
<td>0.1851 0.1195</td>
<td>0.1590 0.0833</td>
<td>1.3961 0.2900</td>
</tr>
<tr>
<td>2</td>
<td>-0.3953 0.0591</td>
<td>0.2569 0.1194</td>
<td>0.1661 0.0582</td>
<td>1.4982 0.2966</td>
</tr>
<tr>
<td>3</td>
<td>-0.4065 0.0587</td>
<td>0.2635 0.1259</td>
<td>0.1823 0.0490</td>
<td>1.5166 0.2912</td>
</tr>
<tr>
<td>4</td>
<td>-0.4199 0.0572</td>
<td>0.3232 0.1155</td>
<td>0.1805 0.0681</td>
<td>1.6095 0.2507</td>
</tr>
<tr>
<td>5</td>
<td>-0.4254 0.0624</td>
<td>0.3181 0.1423</td>
<td>0.1878 0.0419</td>
<td>1.6678 0.4248</td>
</tr>
<tr>
<td>6</td>
<td>-0.4387 0.0562</td>
<td>0.3879 0.0960</td>
<td>0.1932 0.0507</td>
<td>1.8071 0.3959</td>
</tr>
<tr>
<td>7</td>
<td>-0.4424 0.0747</td>
<td>0.3557 0.1449</td>
<td>0.1954 0.0407</td>
<td>1.8070 0.5956</td>
</tr>
<tr>
<td>8</td>
<td>-0.4454 0.0676</td>
<td>0.3699 0.1123</td>
<td>0.1954 0.0420</td>
<td>1.7999 0.4966</td>
</tr>
<tr>
<td>9</td>
<td>-0.4520 0.0617</td>
<td>0.3884 0.1289</td>
<td>0.1932 0.0507</td>
<td>1.8865 0.5320</td>
</tr>
<tr>
<td>10</td>
<td>-0.4601 0.0675</td>
<td>0.4196 0.1117</td>
<td>0.1914 0.0403</td>
<td>1.9601 0.4880</td>
</tr>
<tr>
<td>11</td>
<td>-0.4723 0.0668</td>
<td>0.4335 0.1349</td>
<td>0.1948 0.0425</td>
<td>2.1878 0.9324</td>
</tr>
<tr>
<td>12</td>
<td>-0.4619 0.0701</td>
<td>0.4207 0.1204</td>
<td>0.1927 0.0392</td>
<td>2.0064 0.5980</td>
</tr>
<tr>
<td>15</td>
<td>-0.4781 0.0746</td>
<td>0.4584 0.1477</td>
<td>0.1904 0.0477</td>
<td>2.2976 0.8079</td>
</tr>
<tr>
<td>20</td>
<td>-0.4809 0.0813</td>
<td>0.4615 0.1451</td>
<td>0.1885 0.0416</td>
<td>2.2798 0.7117</td>
</tr>
<tr>
<td>30</td>
<td>-0.4791 0.0718</td>
<td>0.4682 0.1178</td>
<td>0.1922 0.0432</td>
<td>2.3268 0.8620</td>
</tr>
<tr>
<td>50</td>
<td>-0.4746 0.0732</td>
<td>0.4441 0.1119</td>
<td>0.1950 0.0448</td>
<td>2.1066 0.5713</td>
</tr>
</tbody>
</table>

well correspond to the relevant topic.

5.1. Model selection. Before applying the proposed model to real data, we suggest a method for the selection of the number of topics $K$ based on the simulation result. Note that the model involves latent variables. Let an estimated Akaike information criterion (eAIC) be the Akaike information criterion with the estimated parameter and latent variables obtained by Algorithm 2, which is given by

$$eAIC = -2 \ln p(v, w|\hat{x}, \hat{z}, \hat{\psi}, \hat{\phi}, \hat{\theta}) + 2 (|x| + |z| + |\psi| + |\phi| + |\theta|)$$

where $p(w|\hat{z}, \hat{\phi})$ can be found by (2.1). The numbers of estimated latent variables and parameters are given by $|x| = UK$, $|z| = \sum_u A_u$, $|\psi| = U(K - 1)$, $|\phi| = (D - 1)K$, and $|\theta| = 4$.

Figure 3 shows the average eAIC values over 20 datasets against the number of topics $K$. We can see that eAIC values are minimized at the true $K$. We use the eAIC for the selection of $K$ in the rest of the paper.
Fig. 3. The average $e$AIC values over 20 datasets with the true number of topics $K = 5$ (SD5) and $K = 10$ (SD10).

6. Data Analysis. We investigate the StackExchange data dump in android\(^5\) and philosophy\(^6\) fields. For each field, we use data from the community creation time $T_0$ to March 7, 2015. $T_0$ is set as 30 days before the posting time of the first question. We count the time by dates. The users with 5 or more answers are considered in the analysis.

We use the hyperparameters $\alpha = (50/K, \cdots, 50/K)$ and $\gamma = (0.01, \cdots, 0.01)$ as in the Monte Carlo experiments. For each field, words are considered as consecutive English characters including hyphen(-) separated by space and selected with a frequency between 1% and 25% in the content of answers. The minimum frequency is required to exclude specific words that do not fit the field. The maximum frequency is also required to exclude the inessential words (such as do, like, want, and yes) that appear in many answers. We make all English characters lower case and remove the stopwords provided by Natural Language Toolkit (NLTK)\(^7\) version 3.4.4. We also apply the lemmatization technique of NLTK. There is a little number of negative vote, which is taken as zero in our model applications.

The summary statistics for the two fields are shown in Table 4. We apply the PTEM to the two fields of StackExchange data. The number of topics $K = 30$ and $K = 15$ are selected for the android and philosophy fields, respectively, according to the eAIC in Eq. (5.1). The plots of eAIC values are in Figure 4.

Topics for each field can be described by the estimated topic-word distribution $\hat{\phi}_{k,d}$ in Eq. (4.5). Top 5 words according to $\hat{\phi}_{k,d}$ values are shown in descending order in Tables 5 and 6 for each field. The topics are titled

---

\(^{5}\)https://android.stackexchange.com

\(^{6}\)https://philosophy.stackexchange.com

\(^{7}\)http://www.nltk.org/
Table 4

Summary statistics of the StackExchange data. For each field, we present the total number of answered users, the number of investigated users ($U$), the total number of answers, the number of investigated answers ($\sum_u A_u$), the number of words ($D$), and the community creation time $T_0$.

<table>
<thead>
<tr>
<th>Field</th>
<th>Users</th>
<th>$U$</th>
<th>Answers</th>
<th>$\sum_u A_u$</th>
<th>$D$</th>
<th>$T_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>android</td>
<td>11069</td>
<td>820</td>
<td>34914</td>
<td>21610</td>
<td>696</td>
<td>04/06/2009</td>
</tr>
<tr>
<td>philosophy</td>
<td>1753</td>
<td>289</td>
<td>9633</td>
<td>7380</td>
<td>1524</td>
<td>03/06/2011</td>
</tr>
</tbody>
</table>

Fig. 4. The eAIC values with the applied number of topics $K$ for android (left) and philosophy (right) fields.

in a representative manner. Table 5 indicates that android field has a large number of unambiguous topics with small intersections among topics. The topics of the philosophy field are mostly branches of philosophy, as shown in Table 6.

We look into the analysis result in detail below.

6.1. Rich-get-richer phenomenon. The parameter estimates obtained by the PTEM are shown in Table 7. The strength of the impact of expertise levels is found to be similar to each other between android ($\beta_1 = 0.7262$) and philosophy ($\beta_1 = 0.7106$). Positive popularity effects ($\beta_2$) are observed in both fields. In the StackExchange community, popular users with high reputation tend to receive more votes.

The popularity coefficient value is small in the field of android ($\beta_2 = 0.0561$), where there are many questions that users can easily determine if the answer works well. The questions such as “How to install an app?” are frequently posted in the android field. If the asker can install the app by following the instructions of an answer, the answer will likely receive many votes. It could be argued that the reputation of the answerer does not have a significant impact on the number of votes.
Table 5
Thirty topics with 5 top frequency words in the field of android. Representative titles are assigned to topics.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Title</th>
<th>Top 5 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>message</td>
<td>message call sm number google</td>
</tr>
<tr>
<td>2</td>
<td>adb</td>
<td>adb file root command shell</td>
</tr>
<tr>
<td>3</td>
<td>problem</td>
<td>problem issue try would update</td>
</tr>
<tr>
<td>4</td>
<td>rom version</td>
<td>rom version one support work</td>
</tr>
<tr>
<td>5</td>
<td>memory</td>
<td>apps application memory running ram</td>
</tr>
<tr>
<td>6</td>
<td>backup</td>
<td>backup data apps restore reset</td>
</tr>
<tr>
<td>7</td>
<td>sim card</td>
<td>sim password card google account</td>
</tr>
<tr>
<td>8</td>
<td>recovery</td>
<td>recovery fastboot flash boot partition</td>
</tr>
<tr>
<td>9</td>
<td>battery</td>
<td>battery screen power charge time</td>
</tr>
<tr>
<td>10</td>
<td>keyboard</td>
<td>keyboard key language setting input</td>
</tr>
<tr>
<td>11</td>
<td>rom custom</td>
<td>rom custom root update flash</td>
</tr>
<tr>
<td>12</td>
<td>task</td>
<td>tasker profile task volume set</td>
</tr>
<tr>
<td>13</td>
<td>permission</td>
<td>permission application apps root access</td>
</tr>
<tr>
<td>14</td>
<td>network connection</td>
<td>network wifi connection server connect</td>
</tr>
<tr>
<td>15</td>
<td>google play</td>
<td>google apps play market store</td>
</tr>
<tr>
<td>16</td>
<td>file</td>
<td>file folder music medium google</td>
</tr>
<tr>
<td>17</td>
<td>video</td>
<td>video player http play support</td>
</tr>
<tr>
<td>18</td>
<td>usb</td>
<td>usb file driver adb install</td>
</tr>
<tr>
<td>19</td>
<td>gps</td>
<td>data gps wifi location network</td>
</tr>
<tr>
<td>20</td>
<td>setting</td>
<td>setting application google data apps</td>
</tr>
<tr>
<td>21</td>
<td>contact</td>
<td>contact account google sync gmail</td>
</tr>
<tr>
<td>22</td>
<td>brands</td>
<td>official galaxy ltc samsung nexus</td>
</tr>
<tr>
<td>23</td>
<td>network setting</td>
<td>setting network data wifi mobile</td>
</tr>
<tr>
<td>24</td>
<td>package</td>
<td>package list name command install</td>
</tr>
<tr>
<td>25</td>
<td>card</td>
<td>card apps storage internal file</td>
</tr>
<tr>
<td>26</td>
<td>google apps</td>
<td>apps google might also one</td>
</tr>
<tr>
<td>27</td>
<td>usb cable</td>
<td>usb cable work support port</td>
</tr>
<tr>
<td>28</td>
<td>screen</td>
<td>screen setting launcher notification home</td>
</tr>
<tr>
<td>29</td>
<td>button</td>
<td>button power mode volume press</td>
</tr>
<tr>
<td>30</td>
<td>browser</td>
<td>browser google http chrome link</td>
</tr>
</tbody>
</table>

On the other hand, the popularity coefficient value is large in the field of philosophy ($\beta_2 = 0.2543$). In the philosophy field, we can find the questions like “what is evil?,” which requires profound and subjective interpretations. These kinds of questions are open ended. Philosophy community users might judge the quality of answers considering the popularity of answerers. The answers of popular users tend to become more popular in the community, and the posted answers of such people would be more appreciated and get more votes.

6.2. Case study: philosophy field. We investigate the philosophy field in which the rich-get-richer phenomenon is observed more apparently. For user $u$’s answer $a$, the mean parameter of the vote distribution in Eq. (2.3) is estimated by

\[
\hat{m}_{u,a} = \exp(\hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{u,a} + \hat{\beta}_2 y_{u,t_{u,a}}).
\]
Table 6
Fifteen topics with 5 top frequency words in the field of philosophy. Representative titles are assigned to topics.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Title</th>
<th>Top 5 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>animal</td>
<td>right reaction action animal understanding</td>
</tr>
<tr>
<td>2</td>
<td>human</td>
<td>people human person life good</td>
</tr>
<tr>
<td>3</td>
<td>science</td>
<td>philosophy science theory logic mathematics</td>
</tr>
<tr>
<td>4</td>
<td>argument</td>
<td>true argument premise false statement</td>
</tr>
<tr>
<td>5</td>
<td>awareness</td>
<td>must may aware know perceive</td>
</tr>
<tr>
<td>6</td>
<td>time and universe</td>
<td>time universe theory physic law</td>
</tr>
<tr>
<td>7</td>
<td>god-belief</td>
<td>god belief true knowledge know</td>
</tr>
<tr>
<td>8</td>
<td>logic</td>
<td>logic true world sentence truth</td>
</tr>
<tr>
<td>9</td>
<td>set theory</td>
<td>set number theory axiom logic</td>
</tr>
<tr>
<td>10</td>
<td>morality</td>
<td>moral right god good human</td>
</tr>
<tr>
<td>11</td>
<td>knowledge</td>
<td>idea problem knowledge true argument</td>
</tr>
<tr>
<td>12</td>
<td>concept and object</td>
<td>world object existence concept kant</td>
</tr>
<tr>
<td>13</td>
<td>god-existence</td>
<td>god universe argument existence evil</td>
</tr>
<tr>
<td>14</td>
<td>philosopher</td>
<td>philosophy philosopher work book read</td>
</tr>
<tr>
<td>15</td>
<td>consciousness</td>
<td>human consciousness experience brain mind</td>
</tr>
</tbody>
</table>

Table 7
Estimates of $\hat{\theta}$ of the PTEM with the StackExchange data.

<table>
<thead>
<tr>
<th>Field</th>
<th>$\beta_0$ (intercept)</th>
<th>$\beta_1$ (expertise)</th>
<th>$\beta_2$ (popularity)</th>
<th>$\xi$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est. S.E.</td>
<td>Est. S.E.</td>
<td>Est. S.E.</td>
<td>Est. S.E.</td>
</tr>
<tr>
<td>android</td>
<td>0.4242</td>
<td>0.0154</td>
<td>0.7262</td>
<td>0.0182</td>
</tr>
<tr>
<td>philosophy</td>
<td>0.5197</td>
<td>0.0272</td>
<td>0.7106</td>
<td>0.0238</td>
</tr>
</tbody>
</table>

Fig. 5. Scatter plots of estimated (in line) and observed (as dots) means and variances of the votes for the field of philosophy. The estimated means ($\hat{m}$) are on the $x$-axis and the $y$-axis is for the observed means and variances of the the votes for each group. The blue solid line and red dashed line are for the mean and the variance of the negative binomial distribution with the mean parameter $\hat{m}$ and the shape parameter $\hat{\xi}$.

The estimated mean parameter $\hat{m}_{u,a}, u = 1, 2, \cdots, U, a = 1, 2, \cdots, A_u$ are binned into groups with intervals of 1.0, and Figure 5 shows the mean and variance of votes for each group with its theoretical mean $\hat{m}$ and variance $\hat{m} + \frac{1}{\hat{\xi}} \hat{m}^2$. We can see that the variance increases as the mean parameter $\hat{m}$
increases, and the observed mean and variance tend to follow the theoretical mean and variance, implying that it is reasonable to assume the negative binomial distribution for votes [Ver Hoef and Boveng (2007)]. The overall tendency indicates a reasonable validity of the suggested model.

User’s topics of interest can be found in the estimated user-topic distribution $\hat{\psi}_{u,k}$ in Eq. (4.3). We present user-topic distributions of four users 2216, 2702, 5877, and 8056 in Figure 6. User 2216 is interested in topics 2 and 10. User 2702 is highly interested in topic 7. User 5877 is highly interested in topic 8 compared with the other topics. User 8056 is interested in topics 4, 6, 8, and 9 more than others.

![User-topic distributions](image)

**Fig. 6.** User-topic distributions $\hat{\psi}_{u,k}$ of four users 2216, 2702, 5877, and 8056 in the field of philosophy. The topic labels are on the x-axis.

We can predict the topic of the answer through the model estimation process. We take user 8056’s 8th answer as an example. The content of the answer$^8$ (user: 8056, vote: 1, accepted: False) is as follows:

Why do you need to bring aliens into it? Cats, pythons and octopuses all have different morals than we do. Are we morally required to offer them the same legal protections we offer ourselves?

Ah, you might say, that’s not the same question at all because cats, pythons and wolverines are not intelligent in the same sense that we are. But neither are your

$^8$https://philosophy.stackexchange.com/questions/14932/what-if-aliens-had-different-morals
supposed aliens. Our moral sense is quite thoroughly interwoven with the rest of our
cognitive apparatus, so a species with a very different moral sense must have a very
different sort of intelligence than we do.

So I don’t see how your question is any different from “What duty do we have to
octopii?”. That might be a hard question and one worth thinking about, but I think
that bringing in the aliens only serves to obscure it.

As can be seen from the content of the answer, the topic of this answer is
estimated to be topic 2 (human) for 98% of Gibbs iterations, and topic 10
(morality) for the rest 2%.

<table>
<thead>
<tr>
<th>Topic</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expertise</td>
<td>-0.0441</td>
<td>-0.3071</td>
<td>-0.1530</td>
<td>-0.8380</td>
<td>0.0493</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expertise</td>
<td>-0.9194</td>
<td>-0.2440</td>
<td>0.7352</td>
<td>0.1325</td>
<td>-0.2864</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expertise</td>
<td>-0.1257</td>
<td>-0.8249</td>
<td>0.0463</td>
<td>0.1170</td>
<td>-0.5450</td>
</tr>
</tbody>
</table>

The topical expertise estimates \( \hat{x}_{u,k} \) in Eq. (4.2) of user 8056 are shown
in Table 8. This user is interested in topic 8 with a high level of expertise
(0.7352). A moderate level of expertise (0.1325) is estimated in topic 9. We
can also see that the user has a high interest in topic 6 (see Figure 6) but
with a low level of expertise (−0.9194). By employing Eq. (4.7), the user’s
overall expertise is estimated as \( \hat{x}_{u,all} = -0.2155 \), which means that the
user is lower than average in the level of expertise since the topical expertise
levels are assumed to have mean 0. Note that these topical expertise levels
are estimated with the popularity effect considered in the same model.

7. Concluding Remarks. In this paper, we propose a model, PTEM,
by which we can consider both the rich-get-richer phenomenon and topical
expertise levels of users in the CQA. These factors can be estimated
simultaneously by the algorithm developed in this work. We applied
the model to the StackExchange community and found that the rich-get-richer
phenomenon is in effect in the field of philosophy, showing that the answer
written by a popular user tends to get more recommendations. The size of
the rich-get-richer effect is different across the fields, and it could be interpreted as reflecting the nature of the fields.

It is worthwhile to note that the topical expertise levels are estimated
for each user under reasonable assumptions on the model parameters. The
experimental results with real data are good evidential support for using
mathematical models for analyzing CQA data with reasonable interpretations.

In our model, we assume the user’s expertise levels are constant over time. The expertise levels however may change in a relative sense because the users may change in their expertise while posting questions or answers. The model for changing expertise levels are in progress by the authors of this work.

APPENDIX A: PROOF OF PROPOSITIONS

A.1. Proof of Proposition 1. By the generative process of the model, the total probability distribution function is given by

\[
p(w, v, x, z, \psi, \phi | y, \alpha, \gamma, \theta) = \prod_{k=1}^{K} p(\phi_k | \gamma) \prod_{u=1}^{U} p(\psi_u | \alpha) \prod_{k=1}^{K} p(x_{u,k})
\]

(\text{A.1})

\[
\times \prod_{a=1}^{A_u} p(z_{u,a} | \psi_u) p(v_{u,a} | x_{u,z_{u,a}}, z_{u,a}, y_{u,t_{u,a}}, \theta) \prod_{l=1}^{L_{u,a}} p(w_{u,a,l} | \phi_{z_{u,a}})
\]

\[
= p(w, z, \psi, \phi | \alpha, \gamma) \cdot p(x) \cdot p(v | x, z, y, \theta),
\]

where the probability distribution functions in Eq. (A.1) are given by

\[
p(w, z, \psi, \phi | \alpha, \gamma) = \prod_{k=1}^{K} p(\phi_k | \gamma) \prod_{u=1}^{U} p(\psi_u | \alpha) \prod_{a=1}^{A_u} p(z_{u,a} | \psi_u) \prod_{l=1}^{L_{u,a}} p(w_{u,a,l} | \phi_{z_{u,a}}),
\]

\[(\text{A.2})\]

\[
p(x) = \prod_{u=1}^{U} \prod_{k=1}^{K} p(x_{u,k}),
\]

\[(\text{A.3})\]

and \(p(v | x, z, y, \theta)\) is given in Eq. (3.3). Assuming the latent variables are given, the complete data likelihood function for the model parameter \(\theta\) can be written by

\[
L(\theta | w, v, x, z, \psi, \phi, y, \alpha, \gamma) = p(w, v, x, z, \psi, \phi | y, \alpha, \gamma, \theta)
\]

\[
\propto p(v | x, z, y, \theta).
\]

Note that only \(p(v | x, z, y, \theta)\) term involves \(\theta\).
A.2. Proof of Proposition 2. We can separate $\psi$ and $\phi$ related terms as

\begin{align}
\nonumber p(w, z|\alpha, \gamma) &= \int \int_{\psi} p(w, z, \psi, \phi|\alpha, \gamma) d\phi d\psi \\
\nonumber &= \int \prod_{u=1}^{U} p(\psi_{u}|\alpha) \prod_{a=1}^{A_{u}} p(z_{u,a}|\psi_{u}) d\psi \times \int_{\phi}^{K} \prod_{k=1}^{K} p(\phi_{k}|\gamma) \prod_{u=1}^{U} \prod_{a=1}^{A_{u}} \prod_{l=1}^{L_{u,a}} p(w_{u,a,l}|\phi_{z_{u,a}}) d\phi.
\end{align}

We can further separate the $\psi$ term by users as

\begin{align}
\nonumber \int_{\psi}^{U} \prod_{u=1}^{U} p(\psi_{u}|\alpha) \prod_{a=1}^{A_{u}} p(z_{u,a}|\psi_{u}) d\psi &= \int_{\psi}^{U} p(\psi_{u}|\alpha) \prod_{a=1}^{A_{u}} p(z_{u,a}|\psi_{u}) d\psi \\
\nonumber &= \int_{\psi}^{U} \psi_{u}^{\sum_{K}^{K} (n_{u,k} + \alpha_{k}) - 1} d\psi_{u} \\
\nonumber &= \frac{\Gamma \left( \sum_{K}^{K} (n_{u,k} + \alpha_{k}) \right)}{\prod_{K}^{K} \Gamma (\alpha_{k}) \Gamma \left( \sum_{K}^{K} (n_{u,k} + \alpha_{k}) \right)},
\end{align}

where $n_{u,k} = \{|a \in \{1, 2, \ldots, A_{u}\} : z_{u,a} = k\}$ is the number of user $u$‘s topic $k$ answers. Similarly, the second integral in Eq. (A.4) can be expressed as

\begin{align}
\int_{\phi}^{K} \prod_{k=1}^{K} p(\phi_{k}|\gamma) \prod_{u=1}^{U} \prod_{a=1}^{A_{u}} \prod_{l=1}^{L_{u,a}} p(w_{u,a,l}|\phi_{z_{u,a}}) d\phi \\
\nonumber &= \prod_{K}^{K} \frac{\Gamma \left( \sum_{D}^{D} (n_{k,d} + \gamma_{d}) \right)}{\prod_{D}^{D} \Gamma (\gamma_{d}) \Gamma \left( \sum_{D}^{D} (n_{k,d} + \gamma_{d}) \right)},
\end{align}

where $n_{k,d} = \{|(u, a, l), u \in \{1, \ldots, U\}, a \in \{1, \ldots, A_{u}\}, l \in \{1, \ldots, L_{u,a}\} : w_{u,a,l} = d, z_{u,a} = k\}$ is the number of word $d$ in topic $k$ answers posted by all users. Using Eqs. (A.5), (A.6), and (A.7), $p(w, z|\alpha, \gamma)$
in Eq. (A.4), we have

\[ p(w, z | \alpha, \gamma) = \prod_{u=1}^{U} \frac{\Gamma \left( \sum_{k=1}^{K} \alpha_k \right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma(n_{u,k} + \alpha_k)}{\Gamma \left( \sum_{k=1}^{K} (n_{u,k} + \alpha_k) \right)} \]

\[ \times \prod_{k=1}^{K} \frac{\Gamma \left( \sum_{d=1}^{D} \gamma_d \right)}{\prod_{d=1}^{D} \Gamma(\gamma_d)} \frac{\prod_{d=1}^{D} \Gamma(n_{k,d} + \gamma_d)}{\Gamma \left( \sum_{d=1}^{D} (n_{k,d} + \gamma_d) \right)}. \]

To obtain the distribution in (3.5), we calculate

(A.8)

\[ p(z_{u,a} = k | z_{-(u,a)}, w, \alpha, \gamma) \propto p(z_{u,a} = k, z_{-(u,a)}, w | \alpha, \gamma) \]

\[ = \left( \frac{\Gamma \left( \sum_{k'=1}^{K} \alpha_{k'} \right)}{\prod_{k'=1}^{K} \Gamma(\alpha_{k'})} \right)^U \prod_{u' \neq u} \frac{\prod_{k'=1}^{K} \Gamma(n_{u',k'} + \alpha_{k'})}{\Gamma \left( \sum_{k'=1}^{K} (n_{u',k'} + \alpha_{k'}) \right)} \cdot \frac{\prod_{k'=1}^{K} \Gamma(n_{u,k'} + \alpha_{k'})}{\Gamma \left( \sum_{k'=1}^{K} (n_{u,k'} + \alpha_{k'}) \right)} \]

\[ \times \left( \frac{\Gamma \left( \sum_{d=1}^{D} \gamma_d \right)}{\prod_{d=1}^{D} \Gamma(\gamma_d)} \right)^K \prod_{k'=1}^{K} \frac{\Gamma(n_{k',d} + \gamma_d)}{\Gamma \left( \sum_{d=1}^{D} (n_{k',d} + \gamma_d) \right)} \]

\[ \propto \frac{\prod_{k'=1}^{K} \Gamma(n_{u,k'} + \alpha_{k'})}{\Gamma \left( \sum_{k'=1}^{K} (n_{u,k'} + \alpha_{k'}) \right)} \cdot \frac{\prod_{k'=1}^{K} \Gamma(n_{k',d} + \gamma_d)}{\Gamma \left( \sum_{d=1}^{D} (n_{k',d} + \gamma_d) \right)}. \]

For a simpler expression, we exclude terms that are not related to user u’s answer a. If \( z_{u,a} = k \), then we have

(A.9) \quad n_{u,k} = n_{u,k,-(u,a)} + 1, \quad n_{u,k'} = n_{u,k',-(u,a)} \quad \text{if} \quad k' \neq k.

Eq. (A.9) and the property \( \Gamma(n+1) = n! \) of gamma function yield

(A.10) \quad \Gamma \left( \sum_{k'=1}^{K} (n_{u,k'} + \alpha_{k'}) \right) = \Gamma \left( \sum_{k'=1}^{K} (n_{u,k',-(u,a)} + \alpha_{k'}) + 1 \right) \]

\[ = \Gamma \left( \sum_{k'=1}^{K} (n_{u,k',-(u,a)} + \alpha_{k'}) \right) \cdot \left( \sum_{k'=1}^{K} (n_{u,k',-(u,a)} + \alpha_{k'}) \right), \]

(A.11) \quad \prod_{k'=1}^{K} \Gamma(n_{u,k'} + \alpha_{k'}) = \prod_{k'=1}^{K} \Gamma(n_{u,k',-(u,a)} + \alpha_{k'}) \cdot (n_{u,k,-(u,a)} + \alpha_k).

If \( z_{u,a} = k \), then we have

(A.12) \quad n_{k,d} = n_{k,d,-(u,a)} + n_{k,d,(u,a)}, \quad n_{k',d} = n_{k',d,-(u,a)} \quad \text{if} \quad k' \neq k.
Using Eq. \((A.12)\), we have
\[
\prod_{k'=1}^{K} \Gamma \left( \sum_{d=1}^{D} (n_{k',d} + \gamma_d) \right) = \prod_{k'=1}^{K} \Gamma \left( \sum_{d=1}^{D} (n_{k',d,-(u,a)} + \gamma_d) \right) \cdot \prod_{c=1}^{n_{k,1:D,(u,a)}} \left( \sum_{d=1}^{D} (n_{k,d,-(u,a)} + \gamma_d + c - 1) \right).
\]

Also, we have
\[
\prod_{k'=1}^{K} \prod_{d=1}^{D} \Gamma(n_{k',d} + \gamma_d) = \prod_{k'=1}^{K} \prod_{d=1}^{D} \Gamma(n_{k',d,-(u,a)} + \gamma_d) \cdot \prod_{b=1}^{n_{k,1:D,(u,a)}} \left( \sum_{d=1}^{D} (n_{k,d,-(u,a)} + \gamma_d + b - 1) \right).
\]

Using Eqs. \((A.10)\), \((A.11)\), \((A.13)\), and \((A.14)\), we can simplify Eq. \((A.8)\) by eliminating the terms that are not related to user \(u\)'s answer \(a\) as
\[
p(z_{u,a} = k|z_{-(u,a)}, w, \alpha, \gamma) \propto \frac{n_{u,k,-(u,a)} + \alpha_k}{\sum_{k'=1}^{K} (n_{u,k',-(u,a)} + \alpha_{k'})} \cdot \prod_{d=1}^{D} \prod_{b=1}^{n_{k,d,(u,a)}} \left( \sum_{d=1}^{D} (n_{k,d,-(u,a)} + \gamma_d + b - 1) \right).
\]

To find the conditional distributions in \((3.5)\), we write
\[
p(z_{u,a}|z_{-(u,a)}, x, w, v, y, \alpha, \gamma, \theta) \propto \int_{\phi} \int_{\psi} p(z, x, w, v, \psi, \phi|y, \alpha, \gamma, \theta) \, d\phi \, d\psi
\]
\[
= \int_{\phi} \int_{\psi} p(z, w, v, \phi|\alpha, \gamma) \, d\phi \, d\psi \cdot p(x) \cdot p(v|x, z, y, \theta)
\]
\[
\propto p(z, w|\alpha, \gamma)p(v|x, z, y, \theta).
\]

Since topic \(z_{u,a}\) can have discrete values of \(k = 1, 2, \cdots, K\), the probability can be expressed, by using the relation \((A.15)\) and the vote of user \(u\)'s answer \(a\), as
\[
p(z_{u,a} = k|z_{-(u,a)}, x, w, v, y, \alpha, \gamma, \theta) \propto p(z_{u,a} = k, z_{-(u,a)}, w|\alpha, \gamma) \cdot p(v|x, z_{u,a} = k, z_{-(u,a)}, y, \theta),
\]
and we have \((3.6)\).
A.3. Proof of Proposition 3. Using the Bayes formula and Eqs. (A.1), (A.2), and (A.3), we have

\[ p(x_{u,k}|x_{-u,k}, w, v, z, y, \alpha, \gamma, \theta) \propto p(w, v, x, z, \psi | y, \alpha, \gamma, \theta) \]

\[ = p(w, v, \psi | \alpha, \gamma) \cdot p(x | \psi, \alpha, \gamma, \theta) \]

\[ \propto p(x) \cdot p(v | x, z, y, \theta) \]

\[ \propto p(x) \cdot p(v | x, z, y, \theta) \]

\[ = \prod_{k=1}^{K} p(\phi_k | \gamma) \prod_{u=1}^{U} p(\psi_u | \alpha) \prod_{a=1}^{A_u} p(z_{u,a} | \psi_u) \prod_{l=1}^{L_{u,a}} p(w_{u,a,l} | \phi_{z_{u,a}}) \cdot \prod_{u=1}^{U} \prod_{k=1}^{K} p(x_{u,k}) \]

\[ \propto p(v_{u,k}|x_{u,k}, z_{u,1:A_u}, y_{u,k}, \theta) \cdot p(x_{u,k}). \]

SUPPLEMENTARY MATERIAL

Supplement to “PTEM: A Popularity-based Topical Expertise Model for Community Question Answering.”

( doi: COMPLETED BY THE TYPESETTER; ptemsupp.pdf). We provide the StackExchange data dump. Supporting Python codes are also provided.

REFERENCES


Conference on Advances in Social Networks Analysis and Mining 320–327. IEEE Press.

Hohyun Jung
Sung-Ho Kim
Department of Mathematical Sciences,
Korea Advanced Institute of Science and Technology,
South Korea
E-mail: hhjung@kaist.ac.kr
sung-ho.kim@kaist.edu

Jae-Gil Lee
Graduate School of Knowledge Service Engineering,
Korea Advanced Institute of Science and Technology,
South Korea
E-mail: jaegil@kaist.ac.kr

Namgil Lee
Department of Information Statistics,
Kangwon National University,
South Korea
E-mail: namgil.lee@kangwon.ac.kr

imsart-aocs ver. 2014/10/16 file: ptem_shkim.tex date: February 14, 2020