The Annals of Applied Statistics
2020, Vol. 14, No. 3, 1304–1325
https://doi.org/10.1214/20-AOAS1346

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

BY HOHYUN JUNG[1,*], JAE-GIL LEE[2], NAMGIL LEE[3] AND SUNG-HO KIM[1,†]

[1]*Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology,* [*]*hhjung@kaist.ac.kr;* [†]*sung-ho.kim@kaist.edu*

[2]*Graduate School of Knowledge Service Engineering, Korea Advanced Institute of Science and Technology, jaegil@kaist.ac.kr*

[3]*Department of Information Statistics, Kangwon National University, namgil.lee@kangwon.ac.kr*

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,[1] StackExchange,[2] Naver Kin[3] and Quora.[4] 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 information such as profile and rank on the community and some badges are exposed together.

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 (Kondor et al. (2014), Merton (1968), Perc (2014), van de Rijt et al. (2014), Jung et al. (2018)).

Received August 2019; revised February 2020.

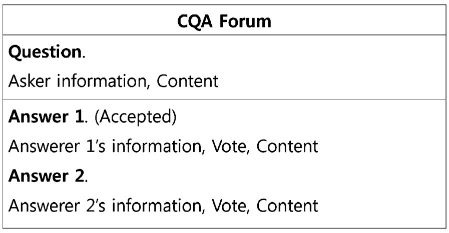
Key words and phrases. Topic analysis, EM algorithm, community question answering, rich get richer, user behavior.

[1] https://answers.yahoo.com

[2] https://stackexchange.com

[3] https://kin.naver.com

[4] https://www.quora.com



CQA Forum

Question.
Asker information, Content

Answer 1. (Accepted)
Answerer 1's information, Vote, Content
Answer 2.
Answerer 2's information, Vote, Content

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

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.

For the study of global expertise analysis, Aslay et al. (2013) employed graph-based methods through community expertise networks. Jurczyk and Agichtein (2007) proposed Hyperlink-Induced Topic Search (HITS) algorithm to estimate the expertise level of users. Zhang, Ackerman and Adamic (2007) proposed ExpertiseRank which is a variation of Google's PageRank. Liu, Song and Lin (2011) suggested competition-based expertise networks for the estimation of the expertise level.

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 models. 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 (Ma et al. (2015), Yang and Manandhar (2014), Yang et al. (2013), 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 (Patra (2017), Srba and Bielikova (2016), 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, for example, 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 $u = 1, 2, \dots, U$, and the answers uploaded by user u are expressed in time order as $a = 1, 2, \dots, A_u$. 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 $d = 1, 2, \dots, D$. 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 said answer's vote. We further assume that each answer covers only one topic, and each topic k has a distribution of words, denoted by $\phi_k = (\phi_{k,1}, \dots, \phi_{k,D})$, 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 $w_{u,a,l}$, $l = 1, 2, \dots, L_{u,a}$. 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 $\psi_u = (\psi_{u,1}, \dots, \psi_{u,K})$, where $\sum_k \psi_{u,k} = 1$. The expertise level of users on each topic is expressed as $x_u = (x_{u,1}, \dots, x_{u,K})$, and $x_{u,k}$ indicates the expertise level of user u on topic k.

We summarize the notations in Table 1.

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 $u = 1, 2, \dots, U$:
  - Draw user-topic distribution $\psi_u = (\psi_{u,1}, \dots, \psi_{u,K}) \sim \text{Dirichlet}(\alpha)$.

TABLE 1
Notations and their descriptions

Notation	Description
U	total number of users
A_u	total number of user u's answers
$L_{u,a}$	total number of words in user u's answer a
K	total number of topics
D	total number of unique words
ψ_u	topic distribution of user u
ϕ_k	word distribution of topic k
$x_{u,k}$	expertise level of user u on topic k
α	hyperparameter of prior on user-topic distributions
γ	hyperparameter of prior on topic-word distributions
$t_{u,a}$	posting time of user u's answer a
$z_{u,a}$	topic of user u's answer a
$v_{u,a}$	vote of user u's answer a

  - For topic $k = 1, 2, \ldots, K$:
    - $*$ Draw user-topic-expertise level $x_{u,k} \sim N(0, 1)$.
- For topic $k = 1, 2, \ldots, K$:
  - Draw topic-word distribution $\phi_k = (\phi_{k,1}, \ldots, \phi_{k,D}) \sim \text{Dirichlet}(\gamma)$.
- For user $u = 1, 2, \ldots, U$:
  - For answer $a = 1, 2, \ldots, A_u$:
    - $*$ Draw topic $z_{u,a} \sim \text{Multinomial}(n = 1, \psi_u)$.
    - $*$ For $l = 1, 2, \ldots, L_{u,a}$:
      - $+$ Draw word

$$w_{u,a,l} \sim \text{Multinomial}(n = 1, \phi_{z_{u,a}}). \tag{2.1}$$

    - $*$ Draw vote

$$v_{u,a} \sim \text{NegativeBinomial}(m = m_{u,a}, \xi), \tag{2.2}$$

where the probability mass function is defined by

$$P(v_{u,a} = q) = \frac{\Gamma(\xi + q)}{q!\Gamma(\xi)}\left(\frac{\xi}{\xi + m}\right)^{\xi}\left(\frac{m}{\xi + m}\right)^{q}, \quad q = 0, 1, \ldots$$

which is parametrized by the mean parameter m and the shape parameter ξ with mean m and variance $m + \frac{1}{\xi}m^2$.

The priors of user-topic and topic-word distributions are Dirichlet distributions, and their hyperparameters are denoted as α and γ, 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

$$m_{u,a} = \exp(\beta_0 + \beta_1 x_{u,z_{u,a}} + \beta_2 y_{u,t_{u,a}}), \tag{2.3}$$

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, β_0, β_1, and β_2 are called the intercept, expertise coefficient and popularity coefficient, respectively. We simplify the model parameters as $\theta = (\beta_0, \beta_1, \beta_2, \xi)$.

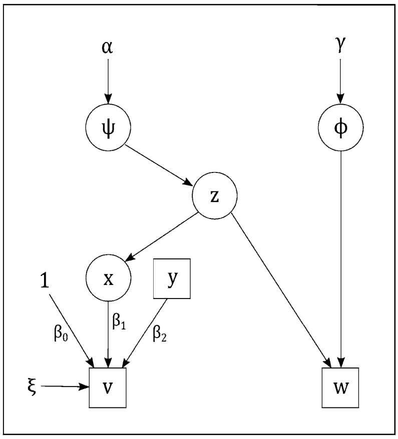


FIG. 2. *The plate notation of the PTEM. Rectangles and circles are used for observed and latent variables, respectively.*

2.3. *Popularity measure.* The popularity is related to the votes through equation (2.3). It is required to define popularity measurement corresponding to the ath 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 $y'_{u,t_{u,a}}$, $u = 1, 2, \dots, U$, $a = 1, 2, \dots, A_u$ is usually much larger than that of expertise levels $x_{u,k}$, $k = 1, 2, \dots, K$ 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

$$y_{u,t} = \frac{y'_{u,t}}{t - T_0}, \tag{2.4}$$

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 θ, the topics of answers $z_{u,a}$, $u = 1, 2, \dots, U$, $a = 1, 2, \dots, A_u$ and the topical expertise of users, $x_{u,k}$, $u = 1, 2, \dots, U$, $k = 1, 2, \dots, K$. The hyperparameters $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_K)$ and $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_D)$ are assumed given. The parameter θ and latent variables, that is, 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, that is, $b, b+1, \dots, c$. For example, $z_{u,1:A_u} = (z_{u,1}, \dots, z_{u,A_u})$ and $z_{1:U,1:A_u} = (z_{1,1}, \dots, z_{1,A_1}, \dots, z_{U,1}, \dots, z_{U,A_U})$.

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

$$\begin{aligned} & p(v_{u,a}|x_{u,z_{u,a}}, z_{u,a}, y_{u,t_{u,a}}, \theta) \\ & \quad = \frac{\Gamma(\xi + v_{u,a})}{(v_{u,a})!\Gamma(\xi)} \xi^{\xi} \big[\exp(\beta_0 + \beta_1 x_{u,z_{u,a}} + \beta_2 y_{u,t_{u,a}})\big]^{v_{u,a}} \\ & \qquad \times \big[\xi + \exp(\beta_0 + \beta_1 x_{u,z_{u,a}} + \beta_2 y_{u,t_{u,a}})\big]^{-(\xi + v_{u,a})}. \end{aligned} \tag{3.1}$$

Let $\mathcal{A}_{u,k}$ be the set of topic k answers for user u, that is, $\mathcal{A}_{u,k} = \{a \in \{1, 2, \dots, A_u\} : z_{u,a} = k\}$. Let $v_{u,k} = \{v_{u,a} : a \in \mathcal{A}_{u,k}\}$ and $y_{u,k} = \{y_{u,t_{u,a}} : a \in \mathcal{A}_{u,k}\}$ 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

$$p(v_{u,k}|x_{u,k}, z_{u,1:A_u}, y_{u,k}, \theta) = \prod_{a \in \mathcal{A}_{u,k}} p(v_{u,a}|x_{u,z_{u,a}}, z_{u,a}, y_{u,t_{u,a}}, \theta). \tag{3.2}$$

The probability distribution function of all the votes, which covers every topic of all the users, are given by

$$p(v|x, z, y, \theta) = \prod_{u=1}^{U} \prod_{k=1}^{K} p(v_{u,k}|x_{u,k}, z_{u,1:A_u}, y_{u,k}, \theta), \tag{3.3}$$

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*

$$\begin{aligned} l(\theta|w, v, x, z, \psi, \phi, y, \alpha, \gamma) &= \ln p(v|x, z, y, \theta) + (\textit{constant on } \theta) \\ &= \sum_{u=1}^{U} \sum_{k=1}^{K} \ln p(v_{u,k}|x_{u,k}, z_{u,1:A_u}, y_{u,k}, \theta) + (\textit{constant on } \theta), \end{aligned}$$

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. □

Before applying the EM algorithm, we define a function Q as the conditional expectation

$$\begin{aligned} Q(\theta|\hat{\theta}^{(s)}) &= E\big[l(\theta|w, v, X, Z, \Psi, \Phi, y, \alpha, \gamma)|w, v, y, \alpha, \gamma, \hat{\theta}^{(s)}\big] \\ &= E\big[\ln p(v|X, Z, y, \theta)|w, v, y, \alpha, \gamma, \hat{\theta}^{(s)}\big] + (\text{constant on } \theta), \end{aligned}$$

where the expectation is taken on the latent variables x, z, ψ and ϕ. Considering equation (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

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

It follows from equation (3.3) that

$$Q(\theta|\hat{\theta}^{(s)}) = \sum_{u=1}^{U} \sum_{k=1}^{K} Q_{u,k}(\theta|\hat{\theta}^{(s)}) + (\text{constant on } \theta). \tag{3.4}$$

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

$$z_{u,a}|z_{-(u,a)}, x, w, v, y, \alpha, \gamma, \theta, \quad u = 1, 2, \ldots, U, a = 1, 2, \ldots, A_u, \tag{3.5}$$

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 ψ and ϕ 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*

$$\begin{aligned} &p(z_{u,a} = k|z_{-(u,a)}, x, w, v, y, \alpha, \gamma, \theta) \\ &\propto \frac{n_{u,k,-(u,a)} + \alpha_k}{\sum_{k'=1}^{K}(n_{u,k',-(u,a)} + \alpha_{k'})} \cdot \frac{\prod_{d=1}^{D}\prod_{b=1}^{n_{k,d,(u,a)}}(n_{k,d,-(u,a)} + \gamma_d + b - 1)}{\prod_{c=1}^{n_{k,1:D,(u,a)}}(\sum_{d=1}^{D}(n_{k,d,-(u,a)} + \gamma_d) + c - 1)} \\ &\quad \times p(v_{u,a}|x_{u,k}, z_{u,a} = k, y_{u,t_{u,a}}, \theta), \end{aligned} \tag{3.6}$$

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.

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 $x_{u,k}$,

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

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 θ is given, we estimate the latent variables z and x through the collapsed Gibbs sampling algorithm. The detailed procedure is presented in Algorithm 1.

Algorithm 1: Collapsed Gibbs Sampling

```
input : $w, v, y, \alpha, \gamma, \theta$
Initialize $z^{(0)}_{1:U,1:A_u}$ and $x^{(0)}_{1:U,1:K}$
for $g = 1, 2, \dots, G$ do
    for $u = 1, 2, \dots, U$ do
        for $a = 1, 2, \dots, A_u$ do
            Draw $z^{(g)}_{u,a}$ according to (3.6) using expertise levels $x_{u,1:K} = x^{(g-1)}_{u,1:K}$ and
            topics
            $z_{-(u,a)} = \{z_{u',a'} : (u', a') \neq (u, a), u = 1, 2, \dots, U, a = 1, 2, \dots, A_u\}$,
            which are assigned by $z_{u',a'} = z^{(g)}_{u',a'}$ if $u' < u$ or if $u' = u$ and $a' < a$, and
            $z_{u',a'} = z^{(g-1)}_{u',a'}$ otherwise.
        end
        for $k = 1, 2, \dots, K$ do
            Draw $x^{(g)}_{u,k}$ according to (3.8) using topics $z_{u,1:A_u} = z^{(g)}_{u,1:A_u}$.
        end
    end
end
output: $z^{(g)}_{1:U,1:A_u}, x^{(g)}_{1:U,1:K}, g = 1, 2, \dots, G$.
```

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

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

where samples $z^{(g)}_{1:U,1:A_u}, x^{(g)}_{1:U,1:K}, g = 1, 2, \dots, G$, are obtained by Algorithm 1 with parameter $\hat{\theta}^{(s)}$ as input. Using equation (3.4), the function Q can be approximated by

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

and the parameter θ 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 Gth 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 θ that maximize $\hat{Q}(\theta|\hat{\theta}^{(s)})$ in the M-step, and we use the sequential quadratic progamming.

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 $\hat{Q}(\theta|\hat{\theta}^{(s)})$. 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, for example, the computation of (3.10).

Algorithm 2: EM Algorithm

```
input : $w, v, y, \alpha, \gamma, \hat{\theta}^{(0)}$
Initialize: $s = 0$, *converged* $=$ *False*
while *not converged* do
    (E-step) Run Algorithm 1 with $\hat{\theta}^{(s)}$, we get Gibbs samples $z_{1:U,1:A_u}^{(g)}, x_{1:U,1:K}^{(g)}$, $g = 1, 2, \ldots, G$
    (E-step) Find $\hat{Q}(\theta|\hat{\theta}^{(s)})$ in equation (3.10)
    (M-step) Find $\hat{\theta}^{(s+1)} = \mathrm{argmax}_\theta \hat{Q}(\theta|\hat{\theta}^{(s)})$
    if *convergence criteria is satisfied* then
        *converged* $\leftarrow$ *True*
        $\hat{\theta} \leftarrow \hat{\theta}^{(s+1)}$
    end
    $s \leftarrow s + 1$
end
Run Algorithm 1 with converged parameter estimate $\hat{\theta}$; we get the final Gibbs samples
$z_{1:U,1:A_u}^{(g)}, x_{1:U,1:K}^{(g)}, g = 1, 2, \ldots, G$
output: $\hat{\theta}, z_{1:U,1:A_u}^{(g)}, x_{1:U,1:K}^{(g)}, g = 1, 2, \ldots, G$
```

4. Inference.

Let $\hat{\theta}$ be the estimated parameter by Algorithm 2. Moreover, let $z_{1:U,1:A_u}^{(g)}$, $x_{1:U,1:K}^{(g)}$, $g = 1, 2, \ldots, G$ be, respectively, the Gibbs samples of topic assignments and topical expertise levels obtained by Algorithm 2.

The standard errors of the parameter θ can be obtained by an information matrix. By Louis' method (Louis (1982)), the estimated observed information matrix is given by

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

where

$$D(g) = \sum_{u=1}^{U} \sum_{k=1}^{K} \{\nabla \ln p(v_{u,k}|x_{u,k}^{(g)}, z_{u,1:A_u}^{(g)}, y_{u,k}, \theta)\}_{\theta=\hat{\theta}}$$

and $\nabla = (\partial/\partial\beta_0, \partial/\partial\beta_1, \partial/\partial\beta_2, \partial/\partial\xi)'$. The asymptotic covariance matrix of θ 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

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

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, that is,

$$\hat{z}_{u,a} = \underset{k}{\operatorname{argmax}}\, p(z_{u,a}=k). \tag{4.1}$$

The topical expertise of user u is estimated by

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

It is informative to provide the user's topics of interest. Let us denote $n_{u,k}^{(g)} = |\{a \in \{1,2,\dots,A_u\} : z_{u,a}^{(g)} = k\}|$ and $n_{k,d} = |\{(u,a,l), u \in \{1,\dots,U\}, a \in \{1,\dots,A_u\}, l \in \{1,\dots,L_{u,a}\} : w_{u,a,l} = d, z_{u,a}^{(g)} = k\}|$ 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

$$\hat{\psi}_{u,k} = \frac{1}{G-G_0} \sum_{g=G_0+1}^{G} \hat{\psi}_{u,k}^{(g)}, \tag{4.3}$$

where

$$\hat{\psi}_{u,k}^{(g)} = \frac{n_{u,k}^{(g)} + \alpha_k}{\sum_{k'=1}^{K} (n_{u,k'}^{(g)} + \alpha_{k'})}. \tag{4.4}$$

Next, the estimated probability of word d covered by topic k, that is, topic-word distribution is given by

$$\hat{\phi}_{k,d} = \frac{1}{G-G_0} \sum_{g=G_0+1}^{G} \hat{\phi}_{k,d}^{(g)}, \tag{4.5}$$

where

$$\hat{\phi}_{k,d}^{(g)} = \frac{n_{k,d}^{(g)} + \gamma_d}{\sum_{d'=1}^{D} (n_{k,d'}^{(g)} + \gamma_{d'})}. \tag{4.6}$$

Equations (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, $x_{u,all} = \sum_{k=1}^{K} \psi_{u,k} x_{u,k}$. It can be estimated by

$$\hat{x}_{u,all} = \sum_{k=1}^{K} \hat{\psi}_{u,k} \hat{x}_{u,k}. \tag{4.7}$$

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 $L_{u,a} = 20$, $u = 1,\dots,100$, $a = 1,\dots,A_u$ 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 $t_{u,a} - t_{u,a-1}$, $a = 1,\dots,A_u$, 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 hyperparameters be $\alpha = (50/K,\dots,50/K)$

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

K	β_0 (intercept)		β_1 (expertise)		β_2 (popularity)		ξ	
	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

and $\gamma = (0.01, \dots, 0.01)$ 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 five answers. The parameter estimation results for SD5 and SD10 are shown in Tables 2 and 3, respectively.

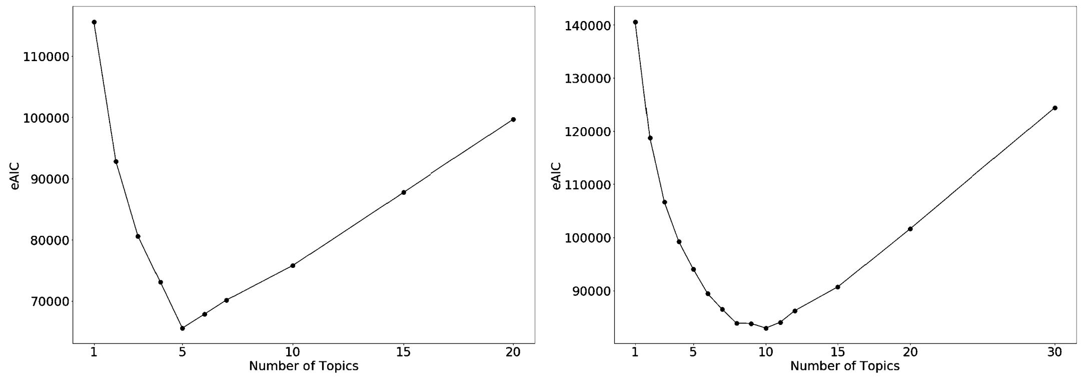
We can see that the mean of the parameter estimate deviates from the true value when applied K is less than the true number of topics. However, for the number of topics that is larger than or equal to the true number of topics, the mean of the parameter estimate is similar to the true parameter.

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

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

K	β_0 (intercept)		β_1 (expertise)		β_2 (popularity)		ξ	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
1	−0.3739	0.0629	0.1851	0.1195	0.1590	0.0833	1.3961	0.2900
2	−0.3953	0.0591	0.2569	0.1194	0.1661	0.0582	1.4982	0.2966
3	−0.4065	0.0587	0.2635	0.1259	0.1823	0.0490	1.5166	0.2912
4	−0.4199	0.0572	0.3232	0.1155	0.1805	0.0681	1.6095	0.2507
5	−0.4254	0.0624	0.3181	0.1423	0.1878	0.0419	1.6678	0.4248
6	−0.4387	0.0562	0.3879	0.0960	0.1777	0.0558	1.8071	0.3959
7	−0.4424	0.0747	0.3557	0.1449	0.1954	0.0407	1.8070	0.5956
8	−0.4454	0.0676	0.3699	0.1123	0.1954	0.0420	1.7999	0.4696
9	−0.4520	0.0617	0.3884	0.1289	0.1932	0.0507	1.8865	0.5320
10 (true)	−0.4601	0.0675	0.4196	0.1117	0.1914	0.0403	1.9601	0.4880
11	−0.4723	0.0668	0.4335	0.1349	0.1948	0.0425	2.1878	0.9324
12	−0.4619	0.0701	0.4207	0.1204	0.1927	0.0392	2.0064	0.5980
15	−0.4781	0.0746	0.4584	0.1477	0.1904	0.0477	2.2976	0.8079
20	−0.4809	0.0813	0.4615	0.1451	0.1885	0.0416	2.2798	0.7117
30	−0.4791	0.0718	0.4682	0.1178	0.1922	0.0432	2.3268	0.8620
50	−0.4746	0.0732	0.4441	0.1119	0.1950	0.0448	2.1066	0.5713



(A) *SD5*

(B) *SD10*

FIG. 3. *The average eAIC values over* 20 *datasets with the true number of topics* $K = 5$ (*SD*5) *and* $K = 10$ (*SD*10).

might dilute the effect of expertise. On the other hand, if the topics were properly refined, the effect of expertise should 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

$$\begin{aligned} eAIC &= -2\ln p(v, w|\hat{x}, \hat{z}, \hat{\psi}, \hat{\phi}, y, \hat{\theta}) + 2\big(|x| + |z| + |\psi| + |\phi| + |\theta|\big) \\ &= -2\ln p(v|\hat{x}, \hat{z}, y, \hat{\theta}) - 2\ln p(w|\hat{z}, \hat{\phi}) + 2\big(|x| + |z| + |\psi| + |\phi| + |\theta|\big), \end{aligned} \tag{5.1}$$

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.

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 five or more answers are considered in the analysis.

We use the hyperparameters $\alpha = (50/K, \dots, 50/K)$ and $\gamma = (0.01, \dots, 0.01)$ 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 data and requirements are given in the Supplementary Material (Jung et al. (2020)).

[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

Field	Users	U	Answers	$\sum_u A_u$	D	T_0
android	11,069	820	34,914	21,610	696	04/06/2009
philosophy	1753	289	9633	7380	1524	03/06/2011

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 equation (5.1). The plots of eAIC values are in Figure 4.

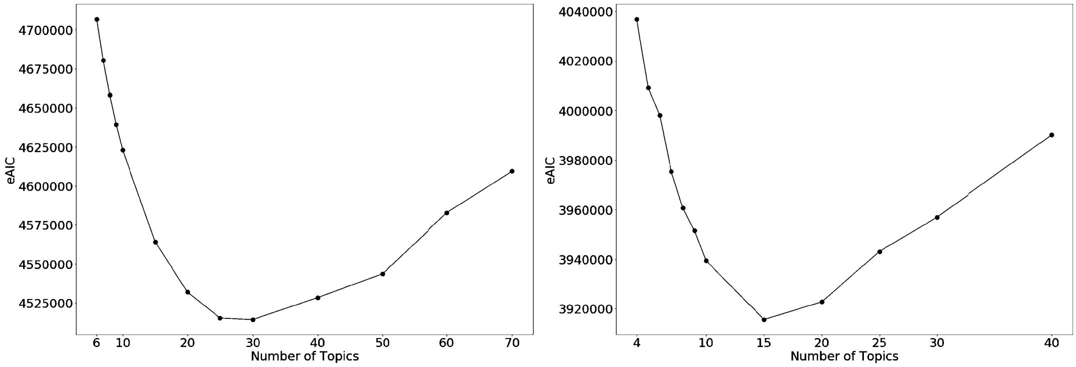
Topics for each field can be described by the estimated topic-word distribution $\hat{\phi}_{k,d}$ in equation (4.5). Top five words, according to $\hat{\phi}_{k,d}$ values, are shown in descending order in Tables 5 and 6 for each field. The topics are titled in a representative manner. Table 5 indicates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

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



(A) *android* (B) *philosophy*

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

TABLE 5
Thirty topics with five top frequency words in the field of android. Representative titles are assigned to topics

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

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 equation (2.3) is estimated by

$$\hat{m}_{u,a} = \exp(\hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{u,\hat{z}_{u,a}} + \hat{\beta}_2 y_{u,t_{u,a}}). \tag{6.1}$$

The estimated mean parameter $\hat{m}_{u,a}$, $u = 1, 2, \dots, U$, $a = 1, 2, \dots, A_u$ 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 $\hat{m} + \frac{1}{\xi}\hat{m}^2$. 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 equation (4.3). We present user-topic distributions of four users 2216, 2702, 5877 and 8056 in

TABLE 6
Fifteen topics with five top frequency words in the field of philosophy. Representative titles are assigned to topics

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

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.

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 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 remaining 2%.

The topical expertise estimates $\hat{x}_{u,k}$ in equation (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

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

Field	β_0 (intercept)		β_1 (expertise)		β_2 (popularity)		ξ	
	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E
android	0.4242	0.0154	0.7262	0.0182	0.0561	0.0212	1.5410	0.0327
philosophy	0.5197	0.0272	0.7106	0.0238	0.2543	0.0508	2.5897	0.1001

[8] https://philosophy.stackexchange.com/questions/14932/what-if-aliens-had-diffrent-morals

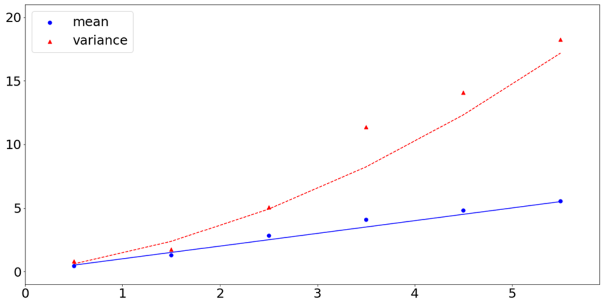


FIG. 5. *Scatter plots of estimated* (*in line*) *and observed* (*as dots*) *means and variances of the votes for the field of philosophy. The estimated means* ($\hat{m}$) *are on the* x*-axis and the* y*-axis is for the observed means and variances of the the votes for each group. The blue solid line 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* $\hat{\xi}$.

topic 6 (see Figure 6) but with a low level of expertise (-0.9194). By employing equation (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

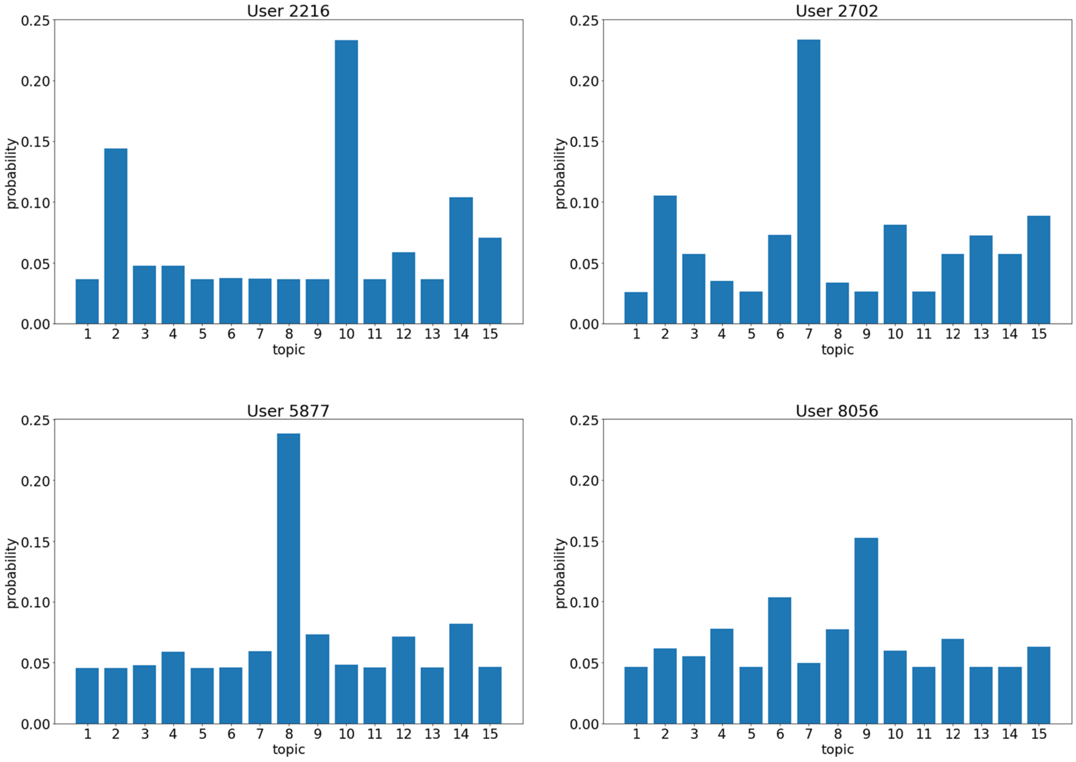


FIG. 6. *User-topic distributions* $\hat{\psi}_{u,k}$ *of four users* 2216, 2702, 5877 *and* 8056 *in the field of philosophy. The topic labels are on the* x*-axis.*

TABLE 8
The topical expertise estimates $\hat{x}_{u,k}$ of user 8056 in the field of philosophy

Topic	1	2	3	4	5
Expertise	−0.0441	−0.3071	−0.1530	−0.8380	0.0493
Topic	6	7	8	9	10
Expertise	−0.9194	−0.2440	0.7352	0.1325	−0.2864
Topic	11	12	13	14	15
Expertise	−0.1257	−0.6249	0.0463	0.1170	−0.5450

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. However, the expertise levels 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

$$\begin{aligned}
&p(w, v, x, z, \psi, \phi | y, \alpha, \gamma, \theta)\\
&\quad = \prod_{k=1}^{K} p(\phi_k|\gamma) \prod_{u=1}^{U} p(\psi_u|\alpha) \prod_{k=1}^{K} p(x_{u,k})\\
&\qquad \times \prod_{a=1}^{A_u} p(z_{u,a}|\psi_u) p(v_{u,a}|x_{u,z_{u,a}}, z_{u,a}, y_{u,t_{u,a}}, \theta) \prod_{l=1}^{L_{u,a}} p(w_{u,a,l}|\phi_{z_{u,a}})\\
&\quad = p(w, z, \psi, \phi|\alpha, \gamma) \cdot p(x) \cdot p(v|x, z, y, \theta),
\end{aligned} \tag{A.1}$$

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

$$p(w, z, \psi, \phi|\alpha, \gamma) = \prod_{k=1}^{K} p(\phi_k|\gamma) \prod_{u=1}^{U} p(\psi_u|\alpha) \prod_{a=1}^{A_u} p(z_{u,a}|\psi_u) \prod_{l=1}^{L_{u,a}} p(w_{u,a,l}|\phi_{z_{u,a}}), \tag{A.2}$$

$$p(x) = \prod_{u=1}^{U} \prod_{k=1}^{K} p(x_{u,k}), \tag{A.3}$$

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

$$\begin{aligned}L(\theta|w,v,x,z,\psi,\phi,y,\alpha,\gamma) &= p(w,v,x,z,\psi,\phi|y,\alpha,\gamma,\theta)\\ &\propto p(v|x,z,y,\theta).\end{aligned}$$

Note that only $p(v|x,z,y,\theta)$ term involves θ.

A.2. Proof of Proposition 2. We can separate ψ and ϕ related terms as

$$\begin{aligned}p(w,z|\alpha,\gamma) &= \int_\psi \int_\phi p(w,z,\psi,\phi|\alpha,\gamma)\,d\phi\,d\psi\\ &= \int_\psi \prod_{u=1}^{U} p(\psi_u|\alpha) \prod_{a=1}^{A_u} p(z_{u,a}|\psi_u)\,d\psi\\ &\quad\times \int_\phi \prod_{k=1}^{K} p(\phi_k|\gamma) \prod_{u=1}^{U}\prod_{a=1}^{A_u}\prod_{l=1}^{L_{u,a}} p(w_{u,a,l}|\phi_{z_{u,a}})\,d\phi.\end{aligned} \tag{A.4}$$

We can further separate the ψ term by users as

$$\int_\psi \prod_{u=1}^{U} p(\psi_u|\alpha) \prod_{a=1}^{A_u} p(z_{u,a}|\psi_u)\,d\psi = \prod_{u=1}^{U} \int_{\psi_u} p(\psi_u|\alpha) \prod_{a=1}^{A_u} p(z_{u,a}|\psi_u)\,d\psi_u. \tag{A.5}$$

Since ψ_u follows a Dirichlet distribution, we have

$$\begin{aligned}&\int_{\psi_u} p(\psi_u|\alpha) \prod_{a=1}^{A_u} p(z_{u,a}|\psi_u)\,d\psi_u\\ &\quad= \int_{\psi_u} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^{K} \psi_{u,k}^{\alpha_k-1} \prod_{a=1}^{A_u} p(z_{u,a}|\psi_u)\,d\psi_u\\ &\quad= \int_{\psi_u} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^{K} \psi_{u,k}^{n_{u,k}+\alpha_k-1}\,d\psi_u\\ &\quad= \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\prod_{k=1}^K \Gamma(n_{u,k}+\alpha_k)}{\Gamma(\sum_{k=1}^K (n_{u,k}+\alpha_k))},\end{aligned} \tag{A.6}$$

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

$$\begin{aligned}&\int_\phi \prod_{k=1}^{K} p(\phi_k|\gamma) \prod_{u=1}^{U}\prod_{a=1}^{A_u}\prod_{l=1}^{L_{u,a}} p(w_{u,a,l}|\phi_{z_{u,a}})\,d\phi\\ &\quad= \prod_{k=1}^{K} \frac{\Gamma(\sum_{d=1}^D \gamma_d)}{\prod_{d=1}^D \Gamma(\gamma_d)} \frac{\prod_{d=1}^D \Gamma(n_{k,d}+\gamma_d)}{\Gamma(\sum_{d=1}^D (n_{k,d}+\gamma_d))},\end{aligned} \tag{A.7}$$

where $n_{k,d} = |\{(u,a,l), u \in \{1,\dots,U\}, a \in \{1,\dots,A_u\}, l \in \{1,\dots,L_{u,a}\} : w_{u,a,l} = d, z_{u,a} = k\}|$ is the number of word d in topic k answers posted by all users. Using equations

(A.5), (A.6) and (A.7), $p(w, z|\alpha, \gamma)$ in equation (A.4), we have

$$\begin{aligned} p(w, z|\alpha, \gamma) = {} & \prod_{u=1}^{U} \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma(n_{u,k} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} (n_{u,k} + \alpha_k))} \\ & \times \prod_{k=1}^{K} \frac{\Gamma(\sum_{d=1}^{D} \gamma_d)}{\prod_{d=1}^{D} \Gamma(\gamma_d)} \frac{\prod_{d=1}^{D} \Gamma(n_{k,d} + \gamma_d)}{\Gamma(\sum_{d=1}^{D} (n_{k,d} + \gamma_d))}. \end{aligned}$$

To obtain the distribution in (3.5), we calculate

$$\begin{aligned} & p(z_{u,a} = k | z_{-(u,a)}, w, \alpha, \gamma) \\ & \quad \propto p(z_{u,a} = k, z_{-(u,a)}, w | \alpha, \gamma) \\ & \quad = \left(\frac{\Gamma(\sum_{k'=1}^{K} \alpha_{k'})}{\prod_{k'=1}^{K} \Gamma(\alpha_{k'})} \right)^{U} \prod_{u' \neq u} \frac{\prod_{k'=1}^{K} \Gamma(n_{u',k'} + \alpha_{k'})}{\Gamma(\sum_{k'=1}^{K} (n_{u',k'} + \alpha_{k'}))} \cdot \frac{\prod_{k'=1}^{K} \Gamma(n_{u,k'} + \alpha_{k'})}{\Gamma(\sum_{k'=1}^{K} (n_{u,k'} + \alpha_{k'}))} \\ & \qquad \times \left(\frac{\Gamma(\sum_{d=1}^{D} \gamma_d)}{\prod_{d=1}^{D} \Gamma(\gamma_d)} \right)^{K} \prod_{k'=1}^{K} \frac{\prod_{d=1}^{D} \Gamma(n_{k',d} + \gamma_d)}{\Gamma(\sum_{d=1}^{D} (n_{k',d} + \gamma_d))} \\ & \quad \propto \frac{\prod_{k'=1}^{K} \Gamma(n_{u,k'} + \alpha_{k'})}{\Gamma(\sum_{k'=1}^{K} (n_{u,k'} + \alpha_{k'}))} \cdot \prod_{k'=1}^{K} \frac{\prod_{d=1}^{D} \Gamma(n_{k',d} + \gamma_d)}{\Gamma(\sum_{d=1}^{D} (n_{k',d} + \gamma_d))}. \end{aligned} \tag{A.8}$$

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

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

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

$$\begin{aligned} \Gamma\left(\sum_{k'=1}^{K} (n_{u,k'} + \alpha_{k'}) \right) & = \Gamma\left(\sum_{k'=1}^{K} (n_{u,k',-(u,a)} + \alpha_{k'}) + 1 \right) \\ & = \Gamma\left(\sum_{k'=1}^{K} (n_{u,k',-(u,a)} + \alpha_{k'}) \right) \cdot \left(\sum_{k'=1}^{K} (n_{u,k',-(u,a)} + \alpha_{k'}) \right), \end{aligned} \tag{A.10}$$

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

If $z_{u,a} = k$, then we have

$$n_{k,d} = n_{k,d,-(u,a)} + n_{k,d,(u,a)}, \qquad n_{k',d} = n_{k',d,-(u,a)} \quad \text{if } k' \neq k. \tag{A.12}$$

Using equation (A.12), we have

$$\begin{aligned} & \prod_{k'=1}^{K} \Gamma\left(\sum_{d=1}^{D} (n_{k',d} + \gamma_d) \right) \\ & \quad = \prod_{k'=1}^{K} \Gamma\left(\sum_{d=1}^{D} (n_{k',d,-(u,a)} + \gamma_d) \right) \cdot \prod_{c=1}^{n_{k,1:D,(u,a)}} \left(\sum_{d=1}^{D} (n_{k,d,-(u,a)} + \gamma_d) + c - 1 \right). \end{aligned} \tag{A.13}$$

Also, we have

$$\begin{aligned}\prod_{k'=1}^{K}\prod_{d=1}^{D}\Gamma(n_{k',d}+\gamma_d)\\ = \prod_{k'=1}^{K}\prod_{d=1}^{D}\Gamma(n_{k',d,-(u,a)}+\gamma_d)\cdot\prod_{b=1}^{n_{k,d,(u,a)}}(n_{k,d,-(u,a)}+\gamma_d+b-1).\end{aligned}\tag{A.14}$$

Using equations (A.10), (A.11), (A.13) and (A.14), we can simplify equation (A.8) by eliminating the terms that are not related to user u's answer a as

$$\begin{aligned}&p(z_{u,a}=k|z_{-(u,a)},w,\alpha,\gamma)\\ &\quad\propto \frac{n_{u,k,-(u,a)}+\alpha_k}{\sum_{k'=1}^{K}(n_{u,k',-(u,a)}+\alpha_{k'})}\cdot\frac{\prod_{d=1}^{D}\prod_{b=1}^{n_{k,d,(u,a)}}(n_{k,d,-(u,a)}+\gamma_d+b-1)}{\prod_{c=1}^{n_{k,1:D,(u,a)}}(\sum_{d=1}^{D}(n_{k,d,-(u,a)}+\gamma_d)+c-1)}.\end{aligned}\tag{A.15}$$

To find the conditional distributions in (3.5), we write

$$\begin{aligned}p(z_{u,a}|z_{-(u,a)},x,w,v,y,\alpha,\gamma,\theta)&\propto\int_{\psi}\int_{\phi}p(z,x,w,v,\psi,\phi|y,\alpha,\gamma,\theta)\,d\phi\,d\psi\\ &=\int_{\psi}\int_{\phi}p(z,w,\psi,\phi|\alpha,\gamma)\,d\phi\,d\psi\cdot p(x)\cdot p(v|x,z,y,\theta)\\ &\propto p(z,w|\alpha,\gamma)p(v|x,z,y,\theta).\end{aligned}$$

Since topic $z_{u,a}$ can have discrete values of $k=1,2,\dots,K$, the probability can be expressed, by using the relation (A.15) and the vote of user u's answer a, as

$$\begin{aligned}&p(z_{u,a}=k|z_{-(u,a)},x,w,v,y,\alpha,\gamma,\theta)\\ &\quad\propto p(z_{u,a}=k,z_{-(u,a)},w|\alpha,\gamma)\cdot p(v|x,z_{u,a}=k,z_{-(u,a)},y,\theta),\end{aligned}$$

and we have (3.6).

A.3. Proof of Proposition 3. Using the Bayes formula and equations (A.1), (A.2) and (A.3), we have

$$\begin{aligned}&p(x_{u,k}|x_{-(u,k)},w,v,z,y,\alpha,\gamma,\theta)\\ &\quad\propto p(w,v,x,z,\psi,\phi|y,\alpha,\gamma,\theta)\\ &\quad= p(w,z,\psi,\phi|\alpha,\gamma)\cdot p(x)\cdot p(v|x,z,y,\theta)\\ &\quad\propto p(x)\cdot p(v|x,z,y,\theta)\\ &\quad=\prod_{k=1}^{K}p(\phi_k|\gamma)\prod_{u=1}^{U}p(\psi_u|\alpha)\prod_{a=1}^{A_u}p(z_{u,a}|\psi_u)\prod_{l=1}^{L_{u,a}}p(w_{u,a,l}|\phi_{z_{u,a}})\cdot\prod_{u=1}^{U}\prod_{k=1}^{K}p(x_{u,k})\\ &\quad\propto p(v_{u,k}|x_{u,k},z_{u,1:A_u},y_{u,k},\theta)\cdot p(x_{u,k}).\end{aligned}$$

Acknowledgments. The second author was supported by the KAIA grant (No. 19NSIP-B081011-06).

The fourth author was supported by the NRF grants (No. 2016R1D1A1B03936155, 2020R1A2C1A01008767) of the Republic of Korea.

SUPPLEMENTARY MATERIAL

Supplement A to "PTEM: A popularity-based topical expertise model for community question answering." (DOI: 10.1214/20-AOAS1346SUPPA; .zip). We provide the StackExchange data dump. Supporting Python codes are also provided.

Supplement B to "PTEM: A popularity-based topical expertise model for community question answering." (DOI: 10.1214/20-AOAS1346SUPPB; .pdf). We provide a description of the StackExchange data and an instruction for running Python codes.

REFERENCES

ASLAY, Ç., O'HARE, N., AIELLO, L. M. and JAIMES, A. (2013). Competition-based networks for expert finding. In *Proceedings of the* 36*th International ACM SIGIR Conference on Research and Development in Information Retrieval* 1033–1036. ACM.

BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

BOUGUESSA, M., DUMOULIN, B. and WANG, S. (2008). Identifying authoritative actors in question-answering forums: The case of yahoo! answers. In *Proceedings of the* 14*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 866–874. ACM.

CAI, Y. and CHAKRAVARTHY, S. (2013). Expertise ranking of users in QA community. In *International Conference on Database Systems for Advanced Applications* 25–40. Springer.

CAO, X., CONG, G., CUI, B. and JENSEN, C. S. (2010). A generalized framework of exploring category information for question retrieval in community question answer archives. In *Proceedings of the* 19*th International Conference on World Wide Web* 201–210. ACM.

GILKS, W. R. and WILD, P. (1992). Adaptive rejection sampling for Gibbs sampling. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **41** 337–348.

GRIFFITHS, T. L. and STEYVERS, M. (2004). Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **101** 5228–5235.

HOFMANN, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* 289–296. Morgan Kaufmann Publishers Inc.

JUNG, H., LEE, J.-G., LEE, N. and KIM, S.-H. (2018). Comparison of fitness and popularity: Fitness-popularity dynamic network model. *J. Stat. Mech. Theory Exp.* **12** 123403, 15. MR3898969 https://doi.org/10.1088/1742-5468/aaeb40

JUNG, H., LEE, J.-G., LEE, N. and KIM, S.-H. (2020). Supplement to "PTEM: A popularity-based topical expertise model for community question answering." https://doi.org/10.1214/20-AOAS1346SUPPA, https://doi.org/10.1214/20-AOAS1346SUPPB

JURCZYK, P. and AGICHTEIN, E. (2007). Discovering authorities in question answer communities by using link analysis. In *Proceedings of the* 16*th ACM Conference on Information and Knowledge Management* 919–922. ACM.

KONDOR, D., PÓSFAI, M., CSABAI, I. and VATTAY, G. (2014). Do the rich get richer? An empirical analysis of the Bitcoin transaction network. *PLoS ONE* **9** e86197. https://doi.org/10.1371/journal.pone.0086197

LIU, J., SONG, Y.-I. and LIN, C.-Y. (2011). Competition-based user expertise score estimation. In *Proceedings of the* 34*th International ACM SIGIR Conference on Research and Development in Information Retrieval* 425–434. ACM.

LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44** 226–233. MR0676213

MA, Z., SUN, A., YUAN, Q. and CONG, G. (2015). A tri-role topic model for domain-specific question answering. In *Proceedings of the* 29*th AAAI Conference on Artificial Intelligence*.

MERTON, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science* **159** 56–63.

MOVSHOVITZ-ATTIAS, D., MOVSHOVITZ-ATTIAS, Y., STEENKISTE, P. and FALOUTSOS, C. (2013). Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. In *Proceedings of the* 2013 *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* 886–893. ACM.

PAL, A., FARZAN, R., KONSTAN, J. A. and KRAUT, R. E. (2011). Early detection of potential experts in question answering communities. In *International Conference on User Modeling, Adaptation, and Personalization* 231–242. Springer.

PAPADIMITRIOU, C. H., RAGHAVAN, P., TAMAKI, H. and VEMPALA, S. (2000). Latent semantic indexing: A probabilistic analysis. *J. Comput. System Sci.* **61** 217–235. MR1802556 https://doi.org/10.1006/jcss.2000.1711

PATRA, B. (2017). A survey of community question answering. arXiv preprint arXiv:1705.04009.

PAUL, S. A., HONG, L. and CHI, E. H. (2012). Who is authoritative? Understanding reputation mechanisms in quora. arXiv preprint arXiv:1204.3724.

PERC, M. (2014). The Matthew effect in empirical data. *J. R. Soc. Interface* **11** 20140378.

SRBA, I. and BIELIKOVA, M. (2016). A comprehensive survey and classification of approaches for community question answering. *ACM Trans. Web* **10** 18.

TAUSCZIK, Y. R. and PENNEBAKER, J. W. (2011). Predicting the perceived quality of online mathematics contributions from users' reputations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 1885–1888. ACM.

VAN DE RIJT, A., KANG, S. M., RESTIVO, M. and PATIL, A. (2014). Field experiments of success–breeds–success dynamics. *Proc. Natl. Acad. Sci. USA* **111** 6934–6939. https://doi.org/10.1073/pnas.1316836111

VER HOEF, J. M. and BOVENG, P. L. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology* **88** 2766–2772.

WANG, X., HUANG, C., YAO, L., BENATALLAH, B. and DONG, M. (2018). A survey on expert recommendation in community question answering. *J. Comput. Sci. Tech.* **33** 625–653.

XU, F., JI, Z. and WANG, B. (2012). Dual role model for question recommendation in community question answering. In *Proceedings of the* 35*th International ACM SIGIR Conference on Research and Development in Information Retrieval* 771–780. ACM.

YANG, B. and MANANDHAR, S. (2014). Exploring user expertise and descriptive ability in community question answering. In *Proceedings of the* 2014 *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* 320–327. IEEE Press.

YANG, L., QIU, M., GOTTIPATI, S., ZHU, F., JIANG, J., SUN, H. and CHEN, Z. (2013). Cqarank: Jointly model topics and expertise in community question answering. In *Proceedings of the* 22*nd ACM International Conference on Information and Knowledge Management* 99–108. ACM.

ZHANG, J., ACKERMAN, M. S. and ADAMIC, L. (2007). Expertise networks in online communities: Structure and algorithms. In *Proceedings of the* 16*th International Conference on World Wide Web* 221–230. ACM.

ZHOU, G., ZHAO, J., HE, T. and WU, W. (2014). An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. *Knowl.-Based Syst.* **66** 136–145.