JNET: Learning User Representations via Joint Network Embedding and Topic Embedding

Lin Gong, Lu Lin, Weihao Song, Hongning Wang
Department of Computer Science, University of Virginia
{lg5bt,ll5fy,ws5dw,hw5x}@virginia.edu

ABSTRACT

User representation learning is vital to capture diverse user preferences, while it is also challenging as user intents are latent and scattered among complex and different modalities of user-generated data, thus, not directly measurable. Inspired by the concept of user schema in social psychology, we take a new perspective to perform user representation learning by constructing a shared latent space to capture the dependency among different modalities of user-generated data. Both users and topics are embedded to the same space to encode users' social connections and text content, to facilitate joint modeling of different modalities, via a probabilistic generative framework. We evaluated the proposed solution on large collections of Yelp reviews and StackOverflow discussion posts, with their associated network structures. The proposed model outperformed several state-of-the-art topic modeling based user models with better predictive power in unseen documents, and state-of-the-art network embedding based user models with improved link prediction quality in unseen nodes. The learnt user representations are also proved to be useful in content recommendation, e.g., expert finding in StackOverflow.

CCS CONCEPTS

• **Information systems** → **Social networks**; **Document topic models**; • **Mathematics of computing** → *Probabilistic inference problems*;

KEYWORDS

Network embedding; topic modeling; social networks; representation learning

ACM Reference Format:
Lin Gong, Lu Lin, Weihao Song, Hongning Wang. 2020. JNET: Learning User Representations via Joint Network Embedding and Topic Embedding. In *The Thirteenth ACM International Conference on Web Search and Data Mining (WSDM '20), February 3–7, 2020, Houston, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3336191.3371770

1 INTRODUCTION

Inferring user intent from recorded user behavior data has been studied extensively for user modeling [11, 22, 31, 39, 40]. Essentially, user modeling builds up conceptual representations of users, which help automated systems to better capture users' needs and enhance user experience in such systems [9, 17]. The rapid development of social media enables users to participate in online activities and create vast amount of observational data, such as social interactions [15, 16] and opinionated text content [8, 12, 27], which in turn provides informative signs about user intents and enables more accurate user representation learning. Extensive efforts have proved the value of user representation learning in various real-world applications, such as latent factor models for collaborative filtering [18, 29], topic models for content modeling [23, 38], network embedding models for social link prediction [5, 20], and many more [31, 42].

User representation learning is challenging, and it can never be a straightforward application of existing statistical learning algorithms on user-generated data. First, user-generated data is noisy, incomplete, highly unstructured, and tied with social interactions [34], which imposes serious challenges in modeling such data. For example, in an environment where users are connected, e.g., social network, user-generated data is potentially related, which directly breaks the popularly imposed independent and identically distributed assumptions in most learning solutions [10, 20, 32]. Second, users often participate in various online activities simultaneously, which creates instrumental contextual signals across different modalities. Although oftentimes scattered and sparse, such multimodal observations reflect users' underlying intents as a whole and call for a holistic modeling approach [19]. Ad-hoc data-driven solutions inevitably isolate the dependency and hence fail to create a comprehensive representation of users. For example, users' social interactions [5, 28] and their generated text data [4, 23, 38] have been extensively studied for user representation learning, but they are largely modeled in isolation. Third, consequently, a unified user representation learning solution is preferred to serve different applications, by taking advantage of data-rich applications to help those data-poor applications.

Even among a few attempts for joint modeling of different types of user-generated data [12, 43], *explicit modeling of dependency* among multiple behavior modalities is still missing. For example, Yang et al. [43] incorporated user-generated text content into network representation learning via joint matrix factorization. In their solution, content modeling is only used as a regularization for network modeling; and thus the learnt model is not in a position to predict unseen text content. Gong and Wang [12] paired the task of sentiment classification with that of social network modeling, and represented each user as a mixture over the instances of these paired tasks. Though text and network are jointly considered, they are only correlated by sharing the same mixing component, without explicitly modeling of the mutual influence between them.

WSDM '20, February 3–7, 2020, Houston, TX, USA

ACM ISBN 978-1-4503-6822-3/20/02…$15.00
https://doi.org/10.1145/3336191.3371770

In social psychology and cognitive science, the concept of *user schema* defines the knowledge structure a person holds which organizes categories of information and the relationships among such categories [37]. Putting it into the scenario of user modeling, we naturally interpret the knowledge structure as user representation described by the collection of associated data, such as the set of textual reviews and behavioral logs associated with individual users. The interrelation existing among multiple types of data further motivates us to perform user modeling in a joint manner while the concept of distributed representation learning [2], i.e., embedding, provides us one possible solution. By constructing a shared latent space, we can embed different modalities of user-generated data in the same low-dimensional space, where the structural dependency among them can be realized by the proximity among different embeddings. The space should be constructed in such a way: 1) the properties of each modality of user-generated data is preserved; 2) the closeness among different modalities of user-generated data can be characterized by the similarity measured in the latent space. For example, connected users in a social network should be closer to each other in this latent space; and by mapping other types of user behavior data into this same space, e.g., text data or behavioral logs, users should be surrounded by their own generated data.

To realize this new perspective of user representation learning, we exploit two most widely available and representative forms of user-generated data, i.e., text content and social interactions. We develop a probabilistic generative model to integrate user modeling with content and network embedding. Due to the unstructured nature of text, we appeal to statistical topic models to model user-generated text content [4, 38], with a goal to capture the underlying semantics. We define a topic as a probability distribution over a fixed vocabulary [4]. We embed both users and topics to the same low-dimensional space to capture of their mutual dependency. On one hand, a user's affinity to a topic is characterized by his/her proximity to the topic's embedding in this latent space, which is utilized to generate each text document of the user. On the other hand, the affinity between users is directly modeled by the proximity between users' embeddings, which are utilized to generate the corresponding social network connections. In this latent space, the two modalities of user-generated data are correlated explicitly, indicated by the user's topical preferences. The user representation is obtained by posterior inference of those embedding vectors over a set of training data, via variational Bayesian. To reflect the nature of our proposed user representation learning method, we name the solution **J**oint **N**etwork **E**mbedding and **T**opic Embedding, or JNET for short.

Extensive empirical evaluations are performed on two large collections of user-generated text documents from Yelp and StackOverflow, together with their network structures. Compared with a set of state-of-the-art user representation learning solutions, clear advantages of JNET are observed: the model's predictive power in content modeling is enhanced on users with rich social connections, and similar improvement is observed in its prediction in network modeling on users with rich text data. The use of learnt user representation generalizes beyond content modeling and social network modeling: it accurately suggests technical discussion threads for users to participate in StackOverflow, e.g., expert recommendation.

2 RELATED WORK

In order to learn effective user representations, a lot efforts have been devoted to modeling diverse modalities of user-generated data: 1) in an isolated manner, i.e., focusing on one particular modality of user-generated data such as text reviews; 2) in a joint manner, i.e., utilizing multiple types of user data. Our proposed solution falls into the second category as it learns user representations from both network structure and text content by explicitly capturing the dependency between the two modalities in the latent topic space.

When performing user representation learning in an isolated way, much attention has been paid on exploring user-user interactions to learn users' distributed representations, which are essential for better understanding users' interactive preferences in social network analysis. Inspired from word embedding techniques [25], random walk models are exploited to generate random paths over a network to learn dense, continuous and low-dimensional representations of users [13, 28, 35]. Matrix factorization technique is also commonly used to learn user embeddings [26, 41], as learning a low-rank space for an adjacency matrix representing the network naturally fits the need of learning low-rank user/node embeddings. For instance, Tang and Liu [36] factorize an input network's modularity matrix and use discriminative training to extract representative dimensions for learning user representation.

In parallel, the user-generated text data is utilized to understand users' emphasis on specific entities or aspects. Topic models [4, 14] serve as a building block for statistical modeling of text data. Typical solutions model individual users as a bag of topics [30], which govern the generation of associated text documents. Wang and Blei [38] combine topic modeling with collaborative filtering to estimate topical user representations with additional observations from user-item ratings. Wang et al. [39] use topic modeling to estimate users' detailed aspect-level preferences from their opinionated review content. Lin et al. [21] learn users' personalized topical compositions to differentiate user's subjectivity from item's intrinsic property in the review documents. McAuley and Leskovec [23] uncover the implicit preferences of each user as well as the properties of each product by mapping users and items into a shared topic space. Some recent works use deep neural networks to obtain user embedding from their generated text data [7, 33].

Although most previous works studied social networks and user-generated text content in isolation, little attention has been paid in combining the two sources for better user modeling. Earlier work [24] regularizes a statistical topic model with a harmonic regularizer defined on the network structure. Yang et al. [43] incorporate text features of users into network representation learning via joint matrix factorization. Gong and Wang [12] pair tasks of opinionated content modeling and network structure modeling in a group-wise fashion, and model each user as a mixture over the tasks. Though both text and network are utilized for user modeling in the aforementioned works, explicit modeling of dependence among different modalities is still missing. Archarya et al. [1] explore the dependency among documents and network but on a per-community basis instead of a per-user basis. Our work proposes a holistic view to model users' social preferences and topical interests jointly, thus to provide a more general understanding of user intents from multiple perspectives.

3 JOINT NETWORK EMBEDDING AND TOPIC EMBEDDING

To interrelate different modalities of user-generated data for user representation learning, we propose to perform joint network embedding and topic embedding. In this section, we first provide the details of our probabilistic generative model, JNET, which imposes a complete generative process over user-generated social interactions and text data in each individual user. Then we describe our variational Bayesian based Expectation Maximization algorithm, which retrieves the learnt user representation from a given corpus.

3.1 Model Specification

We denote a collection of U users as $\mathcal{U} = \{u_1, u_2, ...u_U\}$, in which each user u_i is associated with a set of documents $\mathcal{D}_i = \{x_{i,d}\}_{d=1}^{D_i}$. Each document is represented as a bag of words $x_d = \{w_1, w_2, .., w_N\}$, where each w_n is chosen from a vocabulary of size V. Each user is also associated with a set of social connections denoted as $\mathcal{E}_i = \{e_{ij}\}_{j\neq i}^{U}$, where $e_{ij} = 1$ indicates user u_i and u_j are connected in the network; otherwise, $e_{ij} = 0$.

We represent each user as a real-valued continuous vector $u_i \in \mathbb{R}^M$ in a low-dimensional space. And we seek to impose a joint distribution over the observations in each user's associated text documents and social interactions, so as to capture the underlying structural dependency between these two types of data. Based on our assumption that both types of users-generated data are governed by the same underlying user intent, we explicitly model the joint distribution as $p(\mathcal{D}_i, \mathcal{E}_i) = \int p(\mathcal{D}_i, \mathcal{E}_i, u_i)\mathrm{d}u_i$, which can be further decomposed into $p(\mathcal{D}_i, \mathcal{E}_i, u_i) = p(\mathcal{D}_i|\mathcal{E}_i, u_i)p(\mathcal{E}_i|u_i)p(u_i)$. We assume given the user representation u_i, the generation of text documents in $\mathcal{D}_i$ is independent from the generation of social interactions in $\mathcal{E}_i$, i.e., $p(\mathcal{D}_i|\mathcal{E}_i, u_i) = p(\mathcal{D}_i|u_i)$. As a result, the modeling of joint probability over a user's observational data with his/her latent representation can be decomposed into three related modeling tasks: 1) $p(\mathcal{D}_i|u_i)$ for content modeling, 2) $p(\mathcal{E}_i|u_i)$ for social connection modeling, and 3) $p(u_i)$ for user embedding modeling.

We appeal to topic models [4, 14] due to their effectiveness shown in existing empirical studies for content modeling. The concept of user schema inspires us to embed both users and topics to the same latent space in order to capture the dependency between them. By projecting a user's embedding vector to topic embedding vectors, we can easily measure affinity between a user and a topic, and thus capture users' topical preferences. It also allows us to capture the topical variance in documents from the same user and establish a valid predictive distribution of his/her documents.

Formally, we assume there are in total K topics underlying the corpus with each represented as an embedding vector $\phi_k \in \mathbb{R}^M$ in the same latent space; denote $\Phi \in \mathbb{R}^{K\times M}$ as the matrix of topic embeddings, which facilitate our representation of each user's affinity towards different topics: $\Phi \cdot u_i$, which reflects user u_i's topical preferences, and serves as the prior of topic distribution in each text document from him/her. Specifically, denote the document-level topic vector as $\theta_{id} \in \mathbb{R}^K$, we have $\theta_{id} \sim \mathcal{N}(\Phi \cdot u_i, \tau^{-1}I)$, where τ characterizes the uncertainty when user u_i is choosing topics from his/her global topic preferences for each single document. By projecting the document-level topic vector θ_{id} into a probability simplex, we obtain the topic distribution for document $x_{i,d}$:

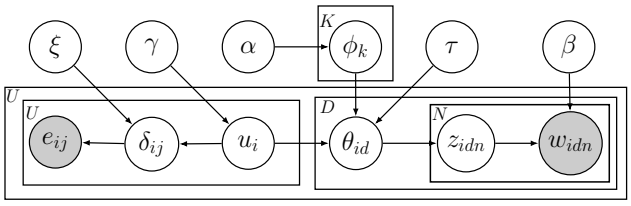

Figure 1: Graphical model representation of JNET. The upper plate indexed by K denotes the learnt topic embeddings. The outer plate indexed by U denotes distinct users in the collection. The inner plates indexed by U and D denote each user's social connections and text documents respectively. The inner plate indexed by N denotes the word content in one text document.

$\pi_{idk} = \mathrm{softmax}(\theta_{idk}) = \exp(\theta_{idk})/\sum_{l=1}^{K}\exp(\theta_{idl})$, from which we sample a topic indicator $z_{idn} \in \{1, ..., K\}$ for each word w_{idn} in $x_{i,d}$ by $z_{idn} \sim \mathrm{Multi}(\pi_{idk})$. As in conventional topic models, each topic k is also associated with a multinomial distribution β_k over a fixed vocabulary, and each word w_{idn} is then drawn from the respective word distribution indicated by corresponding topic assignment, i.e., $w_{idn} \sim p(w|\beta_{z_{idn}})$. Putting all pieces together, the task of content modeling for each user can be summarized as $p(\mathcal{D}_i|u_i) = \prod_{d=1}^{D_i} p(\theta_{id}|u_i, \Phi, \tau)\prod_{n=1}^{N} p(z_{idn}|\theta_{id})p(w_{idn}|z_{idn}, \beta)$.

The key in modeling social connections is to understand the closeness among users. As we represent users with a real-valued continuous vector, this can be easily measured by the vector inner product in the learnt low-dimensional space. Define the underlying affinity between a pair of users u_i and u_j as δ_{ij}, we assume $\mathbb{E}[\delta_{ij}] = u_i^\mathsf{T}u_j$. To capture uncertainty of the affinity between different pairs of users, we further assume δ_{ij} is drawn from a Gaussian distribution centered at the measured closeness, $\delta_{ij} \sim \mathcal{N}(u_i^\mathsf{T}u_j, \xi^2)$, where ξ characterizes the concentration of this distribution. The observed social connection e_{ij} between user u_i and u_j is then assumed as a realization of this underlying user affinity: $e_{ij} \sim \mathrm{Bernoulli}(\mathrm{logistic}(\delta_{ij}))$ where $\mathrm{logistic}(\delta_{ij}) = 1/(1 + \exp(-\delta_{ij}))$. As a result, the task of social connection modeling can be achieved by $p(\mathcal{E}_i|u_i) = \prod_{j\neq i}^{U} p(e_{ij}|\delta_{ij})p(\delta_{ij}|u_i, u_j)$.

We do not have any specific constraint on the form of latent user embedding vectors $\{u_i\}_{i=1}^{U}$ and topic embedding $\{\phi_k\}_{k=1}^{K}$, as long as they are in a M-dimensional space. For simplicity, we assume they are drawn from isotropic Gaussian distributions respectively, i.e., $u_i \sim \mathcal{N}(\mathbf{0}, \gamma^{-1}\boldsymbol{I})$, where γ measures the concentration of different users' embedding vectors, and $\phi_k \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\boldsymbol{I})$. Other types of prior distribution can also be introduced, if one has more knowledge about the user and topic embeddings, such as sparsity or a particular geometric shape. But it is generally preferred to have conjugate priors, so as to simplify later posterior inference steps.

Putting these components together, the generative process of our solution can be described as follows:

- For each topic ϕ_k:
  - Draw its topic compact representation $\phi_k \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\boldsymbol{I})$
- For each user u_i:
  - Draw its user compact representation $u_i \sim \mathcal{N}(\mathbf{0}, \gamma^{-1}\boldsymbol{I})$
  - For every other user u_j:
    - Draw affinity δ_{ij} between u_i and u_j, $\delta_{ij} \sim \mathcal{N}(u_i^\mathsf{T}u_j, \xi^2)$

    - Draw interaction e_{ij} between u_i and u_j, $e_{ij} \sim$ Bernoulli(logistic(δ_{ij}))
- For each document of user u_i:
  - Draw the user-document topic preference vector $\theta_{id} \sim \mathcal{N}(\Phi \cdot u_i, \tau^{-1}\boldsymbol{I})$
  - For each word w_{idn}:
    - Draw topic assignment $z_{idn} \sim$ Multi(softmax(θ_{id}))
    - Draw word $w_{idn} \sim$ Multi($\beta_{z_{idn}}$)

We make two explicit assumptions here: 1) the dimensionality M of the compact representation of topics and users is predefined and fixed; 2) the word distributions under topics are parameterized by a $K \times V$ matrix β where $\beta_{kv} = p(w^v|z^k)$ over a fixed vocabulary of size V. The generative model captures the interrelation between multiple modalities of user-generated data for user representation learning. In essence, we are performing a **J**oint **N**etwork **E**mbedding and **T**opic Embedding, thus, we name the resulting model as JNET in short.

3.2 Variational Bayesian Inference

The compact user representations can be obtained via posterior inference over the latent variables on a given set of data. However, posterior inference is not analytically tractable in JNET due to the coupling among latent variables, i.e., user-user affinity δ, user embedding u, topic embedding Φ, document-level topic proportion θ and word-level topic assignment z. We appeal to a mean-field variational method to approximate the posterior distributions, and further utilize Taylor expansion [3] to address the difficulty introduced by non-conjugate logistic-normal priors.

We begin by postulating a factorized distribution: $q(\Phi, U, \Delta, \Theta, Z) = \prod_{k=1}^{K} q(\phi_k) \prod_{i=1}^{U} q(u_i) \Big[\prod_{j=1, j\neq i}^{U} q(\delta_{ij}) \prod_{d=1}^{D} q(\theta_{id}) \prod_{n=1}^{N} q(z_{idn}) \Big]$, where the factors have the following parametric forms:

$$q(\phi_k) = \mathcal{N}(\phi_k|\mu^{(\phi_k)}, \Sigma^{(\phi_k)}), q(u_i) = \mathcal{N}(u_i|\mu^{(u_i)}, \Sigma^{(u_i)}),$$
$$q(\delta_{ij}) = \mathcal{N}(\delta_{ij}|\mu^{(\delta_{ij})}, {\sigma^{(\delta_{ij})}}^2), q(\theta_{id}) = \mathcal{N}(\theta_{id}|\mu^{(\theta_{id})}, \Sigma^{(\theta_{id})}),$$
$$q(z_{idn}) = \text{Mult}(z_{idn}|\eta_{idn})$$

Because the topic proportion vector θ_{id} is inferred in each document, it is not necessary to estimate a full covariance matrix for it [3]. Hence, in its variational distribution, we only estimate the diagonal variance parameters.

Variational algorithms aim to minimize the KL divergence from the approximated posterior distribution q to the true posterior distribution p. It is equivalent to tightening the evidence lower bound (ELBO) by Jensen's inequality [4]:

$$\log p(\boldsymbol{w}, \boldsymbol{e}|\alpha, \beta, \gamma, \tau) \tag{1}$$
$$\geq \mathbb{E}_q[\log p(U, \Theta, Z, \Phi, \Delta, \boldsymbol{w}, \boldsymbol{e}|\alpha, \beta, \gamma, \tau)] - \mathbb{E}_q[\log q(U, \Theta, Z, \Phi, \Delta)]$$

where the expectation is taken with respect to the factorized variaitonal distribution of the latent variables $q(\Phi, U, \Delta, \Theta, Z)$. Let $\mathcal{L}(q)$ denote the right-hand side of Eq (1), the first step of maximizing this lower bound is to derive the analytic form of posterior expectations required in $\mathcal{L}(q)$. Thanks to the conjugate priors introduced on $\{u_i\}_{i=1}^{U}$ and $\Phi = \{\phi_k\}_{k=1}^{K}$, the expectations related to these latent variables have closed form solutions, while due to non-conjugate logistic-normal priors, we use Taylor expansions to approximate the expectations related to θ_{id}, δ_{ij}. Next we describe the detailed inference procedure for each latent variable.

• **Estimate topic embedding.** For each topic k, we relate the terms associated with $q(\phi_k|\mu^{(\phi_k)}, \Sigma^{(\phi_k)})$ in Eq (1) and take maximization w.r.t. $\mu^{(\phi_k)}$ and $\Sigma^{(\phi_k)}$. Closed form estimations of $\mu^{(\phi_k)}, \Sigma^{(\phi_k)}$ exist,

$$\begin{aligned} \mu^{(\phi_k)} &= \tau\Sigma^{(\phi_k)} \sum_{i=1}^{U} \sum_{d=1}^{D_i} \mu_k^{(\theta_{id})} \mu^{(u_i)} \\ \Sigma^{(\phi_k)} &= \Big[\alpha\boldsymbol{I} + \tau \sum_{i=1}^{U} \sum_{d=1}^{D_i} (\Sigma^{(u_i)} + \mu^{(u_i)}\mu^{(u_i)\mathsf{T}})\Big]^{-1} \end{aligned} \tag{2}$$

The estimation of $\Sigma^{(\phi_k)}$ is not related to a specific topic k, because we impose an isotropic Gaussian prior for all $\{\phi_k\}_{k=1}^{K}$ in JNET. It suggests that the correlations between different topic embedding dimensions are homogeneous across topics. Interestingly, we can notice that the posterior covariance $\Sigma^{(\phi_k)}$ of topic embeddings is closely related to user embeddings, which indicates direct dependency from network structure to text content.

• **Estimate user embedding.** For each user i, we relate the terms associated with $q(u_i|\mu^{(u_i)}, \Sigma^{(u_i)})$ in Eq (1) and maximize it with respect to $\mu^{(u_i)}, \Sigma^{(u_i)}$. Closed form estimations can also be achieved for these two parameters as follows:

$$\begin{aligned} \mu^{(u_i)} =& \Sigma^{(u_i)} \big(\tau \sum_{d=1}^{D_i} \sum_{k=1}^{K} \mu_k^{(\theta_{id})} \mu^{(\phi_k)} + \sum_{j\neq i}^{U} \xi^{-2} \mu^{(\delta_{ij})} \mu^{(u_j)}\big) \\ \Sigma^{(u_i)} =& \gamma\boldsymbol{I} + \tau D_i \sum_{k=1}^{K} (\Sigma^{(\phi_k)} + \mu^{(\phi_k)}\mu^{(\phi_k)\mathsf{T}}) \\ &+ \sum_{j\neq i}^{U} \xi^{-2} (\Sigma^{(u_j)} + \mu^{(u_j)}\mu^{(u_j)\mathsf{T}}) \end{aligned} \tag{3}$$

The effect of joint content modeling and network modeling for user representation learning is clearly depicted in this posterior estimation of user embedding vectors. The updates of $\mu^{(u_i)}$ and $\Sigma^{(u_i)}$ come from two types of influence: the text content and social interactions of the current user. For example, the posterior mode estimation of user embedding vector u_i is a weighted average over the topic vectors that this user has used in his/her past text documents and the user vectors from his/her friends. And the weights measure his/her affinity to those topics and users in each specific observation. The updates exactly reflect the formation of "user schema" in social psychology from two perspectives: both modalities of user-generated data shape user embeddings, while the structural dependency between them is reflected in this unified user representation.

• **Estimate per-document topic proportion vector.** Similar procedures as above can be taken to estimate $\mu^{(\theta_{id})}$ and $\Sigma^{(\theta_{id})}$. Due to the lack of conjugate prior for logistic Normal distributions, we apply Taylor expansion and introduce an additional free variational parameter ζ in each document. Because there is no closed form solution for the resulting optimization problem, we use gradient ascent to optimize $\mu^{(\theta_{id})}$ and $\Sigma^{(\theta_{id})}$ with the following gradients,

$$\begin{aligned} \partial L/\partial \mu_k^{(\theta_{id})} =& -\tau\mu_k^{(\theta_{id})} + \tau{\mu^{(\phi_k)}}^{\mathsf{T}} \mu^{(u_i)} \\ &+ \sum_{n=1}^{N} \big[\eta_{idnk} - \zeta^{-1} \exp(\mu_k^{(\theta_{id})} + \Sigma_{kk}^{(\theta_{id})}/2)\big] \\ \partial L/\partial \Sigma_{kk}^{(\theta_{id})} =& -\tau - N\exp(\mu_k^{(\theta_{id})} + \Sigma_{kk}^{(\theta_{id})}/2)/\zeta + 1/\Sigma_{kk}^{(\theta_{id})} \end{aligned} \tag{4}$$

where $\zeta = \sum_{k=1}^{K} \exp(\mu_k^{(\theta_{id})} + \Sigma_{kk}^{(\theta_{id})}/2)$. Since only the diagonal elements in $\Sigma_{id}^{(\theta)}$ are statistically meaningful (i.e., variance), we simply set its off-diagonal elements to zero in gradient update. The gradient

function suggests that the document-level topic proportion vector should align with the corresponding compact user representation and topic representation. Although no closed form estimations of $\mu^{(\theta_{id})}$ and $\Sigma^{(\theta_{id})}$ exist, the expected property of $\mu^{(\theta_{id})}$ is clearly reflected: the proportion of each topic in document $x_{i,d}$ should align with this user's preference on this topic (i.e., affinity in the embedding space) and the topic assignment in document content. And the variance is introduced by the uncertainty of per-word topic choice and the intrinsic uncertainty of a user's affinity with a topic.

• **Estimate user affinity.** Similar approach can be applied here to estimate $\mu^{(\delta_{ij})}$ and $\sigma^{(\delta_{ij})^2}$ which govern the latent user affinity. Again, gradient ascent is utilized to optimize $\mu^{(\delta_{ij})}$ and $\Sigma^{(\theta_{id})}$,

$$\partial L/\partial \mu^{(\delta_{ij})} = e_{ij} - \varepsilon^{-1} \exp(\mu^{(\delta_{ij})} + \sigma^{(\delta_{ij})^2}/2) - \xi^{-2}(\mu^{(\delta_{ij})} - \mu^{(u_i)^\top} \mu^{(u_j)})$$

$$\partial L/\partial \sigma^{(\delta_{ij})} = -\varepsilon^{-1}\sigma^{(\delta_{ij})} \exp(\mu^{(\delta_{ij})} + \sigma^{(\delta_{ij})^2}/2) - \xi^{-2}\sigma^{(\delta_{ij})} + 1/\sigma^{(\delta_{ij})}$$

The gradient functions suggest that the latent affinity between a pair of users is closely related with their observed connectivity and their closeness in the embedding space.

• **Estimate word topic assignment.** The topic assignment z_{idn} for each word w_{idn} in document $x_{i,d}$ can be estimated by,

$$\eta_{idnk} \propto \exp\{\mu_k^{(\theta_{id})} + \sum_{v=1}^{V} w_{idnv} \log \beta_{kv}\}$$

We execute the above variational inference procedures in an alternative fashion until the lower bound $\mathcal{L}(q)$ defined in Eq (1) converges. The variational inference algorithm postulates strong independence structures between the variational parameters, allowing straightforward **parallel computing**. Since the variational parameters can be grouped by documents: $\mu^{(\theta_{id})}$, $\Sigma^{(\theta_{id})}$ and η, by topics: $\mu^{(\phi_k)}$ and $\Sigma^{(\phi_k)}$, and by users: $\mu^{(u_i)}$, $\Sigma^{(u_i)}$, $\mu^{(\delta_{ij})}$ and $\sigma^{(\delta_{ij})^2}$, we perform alternative update in parallel to improve computational efficiency: for example, we fix topic-level parameters and user-level parameters, and distribute the documents across different machines to estimate their own $\mu^{(\theta_{id})}$, $\Sigma^{(\theta_{id})}$ and η in parallel for large collections of user-generated data.

3.3 Parameter Estimation

When performing the variational inference described above, we have assumed the knowledge of model parameters $\alpha, \gamma, \tau, \xi$ and β. Based on the inferred posterior distribution of latent variables in JNET, these model parameters can be readily estimated by the Expectation-Maximization (EM) algorithm. The most important model parameters are priors for user embedding γ and topic embedding α, and word-topic distribution β. As ξ and τ serve as the variance for user affinity δ_{ij} and document topic proportion vector θ_{id}, and we have large amount of observations in text documents and social connections across all users, our model is less sensitive to their settings. Therefore, we estimate ξ and τ less frequently than α, γ and β.

By taking the gradient of $\mathcal{L}(q)$ in Eq (1) with respect to α, and set the resulting gradient to 0, we get the closed form estimation of α as follows:

$$\alpha = \frac{KM}{\sum_{k=1}^{K}[\sum_{m=1}^{M} \Sigma_{mm}^{(\phi_k)} + \mu^{(\phi_k)^\top} \mu^{(\phi_k)}]},$$

Similarly, the closed form estimation of γ can be easily derived as,

$$\gamma = \frac{UM}{\sum_{i=1}^{U}[\sum_{m=1}^{M} \Sigma_{mm}^{(u_i)} + \mu^{(u_i)^\top} \mu^{(u_i)}]}.$$

And the closed form estimation for word-topic distribution β can be achieved by,

$$\beta_{kv} \propto \sum_{i=1}^{U} \sum_{d=1}^{D_i} \sum_{n=1}^{N} w_{idnv} \eta_{idnv},$$

where w_{idnv} indicates the nth word in u_i's dth document is the vth term in the vocabulary. The estimation for ξ and τ is omitted for space limit, but they can be easily derived based on Eq (1).

The resulting EM algorithm consists of E-step and M-step. In E-step, the variational parameters are inferred based the procedures described in Section 3.2; and in M-step, the model parameters are estimated based on collected sufficient statistics from E-step. These two steps are repeated until the lower bound $\mathcal{L}(q)$ converges over all training data.

Inferring the latent variables with each user and each topic are computationally cheap. Specifically, by Eq (2), updating the variables for each topic imposes a complexity of $O(KM^2|D|)$, where K is the total number of topics, M is the latent dimension, $|D|$ is the total number of documents. By Eq (3), updating the variables for each user imposes a complexity of $O(M^2U^2)$ where U is the total number of users. Estimating the latent variables for the per-document topic proportion imposes a complexity of $O(|D|K(\bar{N} + M))$ by Eq (4), where $\bar{N}$ is the average document length. And updating variables for each pair of user affinity takes constant time while there are U^2 affinity variables. With the consideration of the total number of users and topics, the overall complexity for the proposed algorithm is $O(KM^2|D| + M^2U^2)$.

4 EXPERIMENTS

We evaluated the proposed model on large collections of Yelp reviews and StackOverflow forum discussions, together with their user network structures. Qualitative analysis demonstrates the descriptive power of JNET through direct mapping of user and topic embeddings into a 2-D space. The explicit modeling of dependency among user-generated data confirms the effectiveness of JNET, as indicated by the model's predictive power in recovering missing links and modeling unseen documents. The learnt user representation also enables accurate content recommendation to users.

4.1 Experiment Settings

Datasets. We employed two large publicly available user-generated text datasets together with the associated user networks: 1) **Yelp**, collected from Yelp dataset challenge [1], consists of 187,737 Yelp restaurant reviews generated by 10,830 users. The Yelp dataset provides user friendship imported from their Facebook friend connections. Among the whole set of users, 10,194 of them have friends with an average of 10.65 friends per user. 2) **StackOverflow**, collected from Stackoverflow.com [2], consists of 244,360 forum discussion posts generated by 10,808 users. While there is no explicit network structure in StackOverflow dataset, we utilized the "*reply-to*" information in the discussion threads to build a user network,

[1] Yelp dataset challenge. http://www.yelp.com/dataset_challenge
[2] StackOverflow. http://stackoverflow.com

Figure 2: Visualization of user embedding and learnt topics in 2-D space of Yelp (left) and StackOverflow (right).

because this relation suggests implicit social connections among users based on their expertise and technical topic interest. We ended up with 10,041 connected users, with an average of 5.55 connections per user. We selected 5,000 unigram and bigram text features based on Document Frequency (DF) in both datasets. We randomly split the data for 5-fold cross validation in all the reported experiments.

Baselines. We compared the proposed JNET model against a rich set of user representation learning methods, including topic modeling based solutions, the network embedding methods, and models performing joint modeling of text and network. 1) **Latent Dirichlet Allocation (LDA)** [4] generates the topic distribution in documents across different users, and the user presentation is constructed by averaging the posterior topic proportion of documents associated with a user. 2) **Relational Topic Model (RTM)** [6] explicitly models the connection between two documents and we constructed a user-level network by concatenating all documents of one user in this baseline. 3) **Hidden Factors and Hidden Topics (HFT)** [23] combines latent rating dimensions of users with latent review topics for user modeling. Users' "*upvote*" toward a question is utilized as a proxy of rating in StackOverflow. 4) **Collaborative Topic Regression (CTR)** [38] combines collaborative filtering with topic modeling to explain the observed text content and ratings. 5) **DeepWalk (DW)** [28] takes truncated random walks as input to learn social representations of vertices in the network. 6) **Text-Associated DeepWalk (TADW)** [43] further incorporates text content of vertices into network representation learning under the framework of joint matrix factorization.

Parameter Settings. We set the latent dimensions of user and topic embeddings to 10 in both JNET and baselines as larger dimension gives limited performance improvement but slows down all models considerably. As we tuned the topic size from 10 to 100, we found the learnt topics are most representative and meaningful at around 40 topics. Hence, we set topic number to 40 in the reported experiments. The maximum number of iteration in our EM algorithm is set to 100. Both the source codes and data are available online [3].

4.2 The Learnt User Representations

We first study the quality of the learnt user representations from JNET. The learnt user embeddings are mapped to a 2-D space using the t-SNE algorithm and is visualized in Figure 2. For illustration purpose, we simply assign each user to the topic that he/she is closest to, i.e., $\arg\max_k(\phi_k \cdot u_i)$ and we mark users sharing the same interested topic with the same color. We also plot the most representative words of each topic learnt from JNET (i.e., $\arg\max_w p(w|\beta_z)$), with the same color of the corresponding set of users.

As we can find from the visualization of StackOverflow, users of similar interests are clearly clustered in the 2-D space, which indicates the descriptive power of our learnt user vectors. Meanwhile, we can easily identify the theme of each learnt topic, such as C++ (in light green circle), SQL (in dark purple circle) and java (in light blue circle). It is also interesting to find correlations among the users and topics by looking into their distances. The users in dark green are mainly interested in website development, thus are far away from the users who are interested in C++ (in light green). The users in orange care more about the network communication and they are overlapped with other clusters of users focusing on SQL (in dark purple) and C++ (in light green) as network communication is an important component among different programming languages. Similar observations can also be found on Yelp dataset.

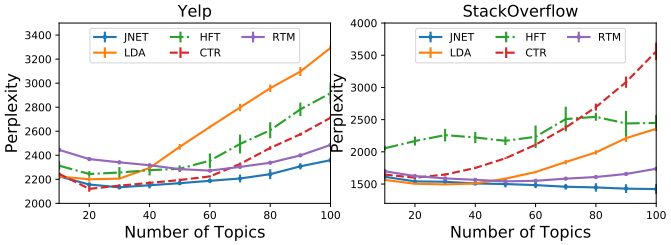

Figure 3: Perplexity comparison on Yelp and StackOverflow.

4.3 Document Modeling

In order to verify the predictive power of the proposed model, we first evaluated the generalization quality of JNET on the document modeling task. We compared all the topic model based solutions by their *perplexity* on a held-out test set. Formally, the perplexity for a set of held-out documents is calculated as follows [4]:

$$perplexity(D_{test}) = \exp\Big(-\frac{\sum_{d\in D_{test}} \log p(\boldsymbol{w}_d)}{\sum_{d\in D_{test}} |d|}\Big)$$

[3] JNET. https://github.com/Linda-sunshine/JNET.

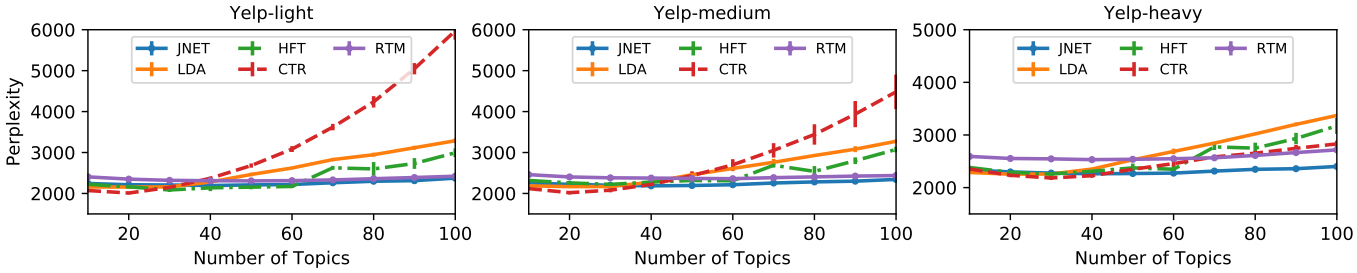

Figure 4: Comparison of perplexity in cold-start users on Yelp.

where $p(\boldsymbol{w}_d)$ is the likelihood of each held-out document given by a trained model. A lower perplexity indicates better generalization quality of a model.

Figure 3 reports the mean and variance of the perplexity for each model with 5-fold cross validation over different topic sizes. JNET achieved the best predictive power on the hold-out dataset, especially when an appropriate topic size is assigned. RTM achieved comparative performance as it utilizes the connectivity information among users, but it is limited by not being able to capture the variance within each user's different documents. The other baselines do not explicitly model network data, i.e., LDA, HFT and CTR, and therefore suffer in their performance.

A good joint modeling of network structure and text content should complement each other to facilitate a more effective user representation learning. Hence, we expect a good model to learn reasonable representations on users lacking text information, a.k.a., cold-start users, by utilizing network structure. We randomly selected 200 users and held out all their text content for testing. Regarding to the number of social connections each testing user has in training data, we further consider three different sets of users, and name them as light, medium and heavy users, to give a finer analysis with respect to the degree of connectivity in cold-start setting. The threshold for categorizing different sets of users is based on the statistics of each dataset; and each group contains 200 users. In particular, we selected 5 and 20 as the connectivity threshold for Yelp, 5 and 15 as the threshold for StackOverflow respectively. That is, in Yelp, light users have fewer than 5 friends, medium users have more than 5 friends while fewer than 20 friends and heavy users have more than 20 friends. We compared JNET against four baselines, i.e., LDA, HFT, RTM, CTR for evaluation purpose. We reported the perplexity on the held-out test documents regarding to the three sets of users, in Figure 4.

As we can observe in Figure 4, JNET performed consistently better on the testing documents for the three different sets of unseen users on Yelp dataset, which indicates the advantage of utilizing network information in addressing cold-start content prediction issue. The benefit of network is further verified across different sets of users as heavily connected users can achieve better performance improvement compared with text only user representation model, i.e., LDA. Similar conclusion is obtained for StackOverflow dataset, while we neglect it due to the space limit.

4.4 Link Prediction

The predictive power of JNET is not only reflected in unseen documents, but also in missing links. In the task of link prediction, the key component is to infer the similarity between users. We split the observed social connections into 5 folds. Each time, we held out one fold of edges for testing and utilized the rest for model training, together with users' text content. In order to construct a valid set of ranking candidates for each testing user, we randomly injected irrelevant users (non-friends) for evaluation purpose. And the number of irrelevant users is proportional to the number of connections a testing user has, i.e., $t \times$ number of social connections. We rank users based on the cosine similarity between their embedding vectors. Normalized discounted cumulative gain (NDCG) and mean average precision (MAP) are used to measure the quality of ranking. We started with the ratio between irrelevant users and relevant users being $t = 2$ and increased the ratio to $t = 8$ to make the task more challenging to further verify the effectiveness of the learnt user representations.

To compare the prediction performance, we tested five baselines, i.e., LDA, HFT, RTM, DW and TADW. We reported the NDCG and MAP for the two datasets in Figure 5. It is clear JNET achieved encouraging performance on both datasets, which indicates effective user representations are learnt to recover network structure. In Yelp dataset, network-only solutions, i.e., DW, and text-only solutions, i.e., LDA and HFT, cannot take the full advantage of both modalities of user-generated data to capture user intents, while RTM achieved descent performance due to the integration of content and network modeling. Since the way of constructing network in StackOverflow is more content oriented, the performance of link prediction on StackOverflow prefers the text based solutions, which explains the comparable performance of LDA. Though TADW utilizes both modalities for user modeling, it fails to capture the dependency between them, leading to the poor performance on this task.

In practice, link prediction for unseen users is especially useful. For example, friend recommendation for new users in a system: they have very few or no friends, while they may associate with rich text content. This is also known as "cold-start" link prediction. Network-only solutions will suffer from the lack of information in such users. However, a joint model can overcome this limitation by utilizing user-generated text content to learn representative user vectors, thus to provide helpful link prediction results.

In order to study the models' predictive power in the cold-start setting, we randomly sampled three sets of users, regarding to the number of documents each user has, and name them as light, medium and heavy users accordingly. Each set of users consists of 200 users, and we selected 10 and 50 as the threshold for Yelp, 15 and 50 as the threshold for StackOverflow respectively. For example, in StackOverflow, light users have fewer than 15 posts, medium

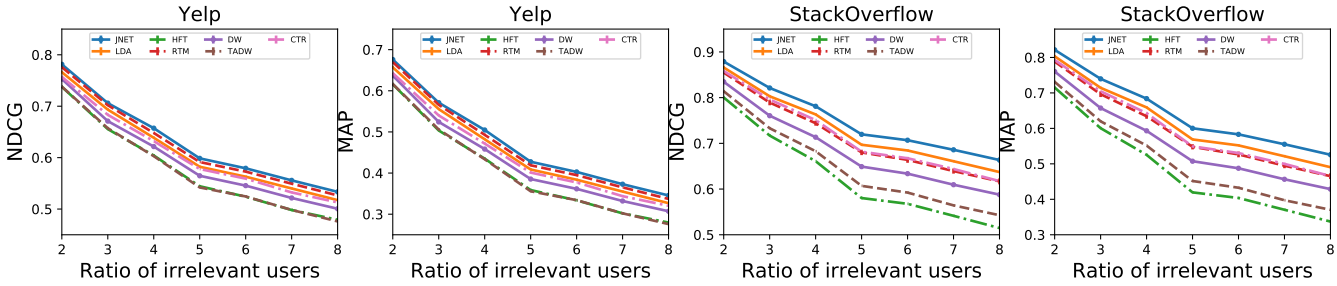

Figure 5: The performance comparison of link suggestion on Yelp and StackOverflow.

Table 1: The performance comparison of link prediction for cold-start users on StackOvewrflow.

Models	Light			Medium			Heavy		
	Ratio=2 NDCG/MAP	Ratio=4 NDCG/MAP	Ratio=6 NDCG/MAP	Ratio=2 NDCG/MAP	Ratio=4 NDCG/MAP	Ratio=6 NDCG/MAP	Ratio=2 NDCG/MAP	Ratio=4 NDCG/MAP	Ratio=6 NDCG/MAP
LDA	0.786/0.648	0.664/0.477	0.632/0.431	0.774/0.597	0.677/0.451	0.612/0.364	0.818/0.581	0.745/0.443	0.697/0.366
HFT	0.666/0.493	0.543/0.333	0.483/0.259	0.671/0.461	0.562/0.313	0.492/0.226	0.682/0.389	0.591/0.250	0.532/0.179
RTM	0.777/0.642	0.688/0.514	0.627/0.433	0.801/0.638	0.709/0.495	0.654/0.419	0.837/0.624	0.760/0.481	0.711/0.399
TADW	0.695/0.525	0.583/0.373	0.515/0.291	0.696/0.481	0.591/0.336	0.532/0.263	0.739/0.448	0.639/0.298	0.587/0.229
JNET	**0.794/0.664**	**0.697/0.534**	**0.643/0.453**	**0.812/0.649**	**0.724/0.511**	**0.663/0.425**	**0.842/0.626**	**0.763/0.483**	**0.713/0.399**

users have more than 15 but fewer than 50 posts, and heavy users have more than 50 posts. We compared JNET against four baselines, i.e., LDA, HFT, RTM and TADW for evaluation purpose. Because DW cannot learn representations for users without any network information, it is excluded in this experiment. We also randomly injected irrelevant users as introduced before for evaluation and we varied the ratio to change the difficulty of the task. We reported the NDCG and MAP performance on the three sets of users in Stackoverflow dataset with three different ratios, i.e., 2, 4 and 6, in Table 1, respectively.

JNET achieved consistently favorable performance in cold-start users, as accurate proximity between users is properly identified with its user representations learnt from text data. Comparing across user groups, better performance is achieved for users with more text documents. Similar results were obtained on Yelp dataset as well, but omitted due to space limit.

4.5 Expert Recommendation

In the sampled StackOverflow dataset, the average number of answers for questions is as low as 1.14, which indicates the difficulty for getting an expert to answer the question. If the system can suggest the right user to answer the posted questions, e.g., push the question to the selected user, more questions would be answered more quickly and accurately. We conjecture the learnt topic distribution of each question in StackOverflow, together with the identified user representation, can facilitate the task of expert recommendation for question answering. The task can be further decomposed into two components: whether the question falls into a user's skill set; and whether the user who asked the question shares similar interests with the potential candidate experts. With the learnt topic embeddings Φ and each user's embedding u_i, each user's interest over topics can be characterized as a mapping from the topic embeddings to the user's embedding, i.e., $\Phi \cdot u_i$. Together with the learnt topic distribution of each question, we can estimate the proximity between a question and a user's expertise to score the alignment

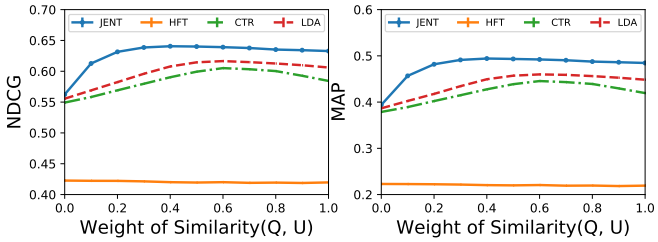

Figure 6: Expert recommendation on StackOverflow.

between them. In the meanwhile, the closeness between users can be simply measured by the distance of their corresponding embedding vectors. As a result, the task can be formalized as finding the user that achieves the highest relatedness with the given question, where we define the relatedness as follows:

$$\text{score} = \alpha \cdot \text{cosine}(u_i \cdot \Phi, \theta_{id}) + (1 - \alpha) \cdot \text{cosine}(u_i, u_j) \quad (5)$$

Due to the limited number of answers for each question in our dataset, we selected 1,816 questions with more than 2 answers for the experiment. Besides the users that answered the given question, we also incorporated irrelevant users for each question for evaluation purpose. And the number of irrelevant users is 10 times of the number of answers. We compared against the learnt topic distributions of questions and user representations from LDA, HFT and CTR as we cannot get the topic distribution of each question from the other baselines. As we tune the weight between the two components in Eq (5), we plot the corresponding NDCG and MAP in Figure 6.

JNET achieved very promising performance in this recommendation task, as it explicitly models a user's expertise and a given question in the topic space. The estimated similarities between user-user and user-content accurately align the question to the right user. The baseline models can only capture the similarity between questions and users based on their topical similarity, which

is insufficient in this task. Interestingly, as we gradually increased the weight of question-content similarity from 0 to 1, JNET's performance peaked, which indicates the relative importance between user-user and user-content similarities for this specific problem.

5 CONCLUSION AND FUTURE WORK

In the paper, we studied user representation learning by explicitly modeling the structural dependency among different modalities of user-generated data. We proposed a complete generative model named JNET to integrate user representation learning with content modeling and social network modeling. The learnt user representations are interpretable and predictive, indicated by the performance improvement in many important tasks such as link prediction and expert recommendation.

Several areas are left open for our future explorations. The current model focuses on the first-order proximity among users in network modeling, while higher-order proximity can be explored to better capture the network connectivity. Also, temporal information of the text content and connections are not considered in the current model. By properly utilizing the temporal information, we would be able to learn the dynamics of user representations, together with the evolution of topics and social network.

6 ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments. This paper is based upon work supported by the National Science Foundation under grant IIS-1553568, IIS-1718216.

REFERENCES

[1] Ayan Acharya, Dean Teffer, Jette Henderson, Marcus Tyler, Mingyuan Zhou, and Joydeep Ghosh. 2015. Gamma process Poisson factorization for joint modeling of network and documents. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 283–299.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[3] David Blei and John Lafferty. 2006. Correlated topic models. *Advances in neural information processing systems* 18 (2006), 147.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[5] Simon Bourigault, Cedric Lagnier, Sylvain Lamprier, Ludovic Denoyer, and Patrick Gallinari. 2014. Learning social network embeddings for predicting information diffusion. In *Proceedings of the 7th ACM WSDM*. ACM, 393–402.

[6] Jonathan Chang and David Blei. 2009. Relational topic models for document networks. In *Artificial Intelligence and Statistics*. 81–88.

[7] Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016. Neural sentiment classification with user and product attention. In *Proceedings of the 2016 EMNLP*. 1650–1659.

[8] Hongbo Deng, Jiawei Han, Hao Li, Heng Ji, Hongning Wang, and Yue Lu. 2014. Exploring and inferring user–user pseudo-friendship for sentiment analysis with heterogeneous networks. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 7, 4 (2014), 308–321.

[9] Gerhard Fischer. 2001. User modeling in human–computer interaction. *User modeling and user-adapted interaction* 11, 1-2 (2001), 65–86.

[10] Lin Gong, Mohammad Al Boni, and Hongning Wang. 2016. Modeling social norms evolution for personalized sentiment classiofication. In *Proceedings of the 54th ACL*, Vol. 1. 855–865.

[11] Lin Gong, Benjamin Haines, and Hongning Wang. 2017. Clustered Model Adaption for Personalized Sentiment Analysis. In *Proceedings of the 26th WWW*. 937–946.

[12] Lin Gong and Hongning Wang. 2018. When Sentiment Analysis Meets Social Network: A Holistic User Behavior Modeling in Opinionated Data. In *Proceedings of the 24th ACM SIGKDD*. ACM, 1455–1464.

[13] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD*. ACM, 855–864.

[14] Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 289–296.

[15] Emily M Jin, Michelle Girvan, and Mark EJ Newman. 2001. Structure of growing social networks. *Physical review E* 64, 4 (2001), 046132.

[16] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD*. ACM, 137–146.

[17] Alfred Kobsa. 2001. Generic user modeling systems. *User modeling and user-adapted interaction* 11, 1-2 (2001), 49–63.

[18] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 8 (2009), 30–37.

[19] Prescott Lecky. 1945. Self-consistency; a theory of personality. (1945).

[20] David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology* 58, 7 (2007), 1019–1031.

[21] Lu Lin, Lin Gong, and Hongning Wang. 2019. Learning Personalized Topical Compositions with Item Response Theory. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 609–617.

[22] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces*. ACM, 31–40.

[23] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 165–172.

[24] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. 2008. Topic modeling with network regularization. In *Proceedings of the 17th WWW*. ACM, 101–110.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[26] Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, and Wenwu Zhu. 2016. Asymmetric transitivity preserving graph embedding. In *Proceedings of the 22nd ACM SIGKDD*. ACM, 1105–1114.

[27] Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining.. In *LREc*, Vol. 10.

[28] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD*. ACM, 701–710.

[29] Steffen Rendle. 2010. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 995–1000.

[30] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on UAI*. AUAI Press, 487–494.

[31] Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Implicit user modeling for personalized search. In *Proceedings of the 14th ACM CIKM*. ACM, 824–831.

[32] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD*. ACM, 1397–1405.

[33] Duyu Tang, Bing Qin, and Ting Liu. 2015. Learning semantic representations of users and products for document level sentiment classification. In *Proceedings of the 53rd ACL*, Vol. 1. 1014–1023.

[34] Jiliang Tang, Yi Chang, and Huan Liu. 2014. Mining social media with social theories: A survey. *ACM SIGKDD Explorations Newsletter* 15, 2 (2014), 20–29.

[35] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th WWW*. 1067–1077.

[36] Lei Tang and Huan Liu. 2009. Relational learning via latent social dimensions. In *Proceedings of the 15th ACM SIGKDD*. ACM, 817–826.

[37] Michelle Rae Tuckey and Neil Brewer. 2003. The influence of schemas, stimulus ambiguity, and interview schedule on eyewitness memory over time. *Journal of Experimental Psychology: Applied* 9, 2 (2003), 101.

[38] Chong Wang and David M Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD*. ACM, 448–456.

[39] Hongning Wang, Yue Lu, and ChengXiang Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD*. ACM, 618–626.

[40] Hongning Wang, ChengXiang Zhai, Feng Liang, Anlei Dong, and Yi Chang. 2014. User modeling in search logs via a nonparametric bayesian approach. In *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 203–212.

[41] Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. 2017. Community Preserving Network Embedding.. In *AAAI*. 203–209.

[42] Ryen W White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. 2013. Enhancing personalized search by mining and modeling task behavior. In *Proceedings of the 22nd WWW*. ACM, 1411–1420.

[43] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y Chang. 2015. Network representation learning with rich text information.. In *IJCAI*. 2111–2117.

A VARIATIONAL INFERENCE

In this section, we provide the detailed derivation of the likelihood lower bound in Eq (1).

Recall that we begin by postulating a factorized distribution:

$$q(\Phi, U, \Delta, \Theta, Z) = \prod_{k=1}^{K} q(\phi_k) \prod_{i=1}^{U} q(u_i) \left[\prod_{j=1, j\neq i}^{U} q(\delta_{ij}) \prod_{d=1}^{D} q(\theta_{id}) \prod_{n=1}^{N} q(z_{idn}) \right]$$

where the factors have the following parametric forms:

$$q(\phi_k) = \mathcal{N}(\phi_k | \mu^{(\phi_k)}, \Sigma^{(\phi_k)}), q(u_i) = \mathcal{N}(u_i | \mu^{(u_i)}, \Sigma^{(u_i)}),$$

$$q(\delta_{ij}) = \mathcal{N}(\delta_{ij} | \mu^{(\delta_{ij})}, {\sigma^{(\delta_{ij})}}^2), q(\theta_{id}) = \mathcal{N}(\theta_{id} | \mu^{(\theta_{id})}, \Sigma^{(\theta_{id})}),$$

$$q(z_{idn}) = \text{Mult}(z_{idn} | \eta_{idn})$$

The log likelihood of observed user behaviors, e.g., posted texts and connected social relations, is bounded by a lower bound using Jensen's inequality:

$$\begin{aligned}
&\log p(\boldsymbol{w}, \boldsymbol{e} | \alpha, \beta, \gamma, \tau) \\
&\geq \mathbb{E}_q[\log p(U, \Theta, Z, \Phi, \Delta, \boldsymbol{w}, \boldsymbol{e} | \alpha, \beta, \gamma, \tau)] - \mathbb{E}_q[\log q(U, \Theta, Z, \Phi, \Delta)] \\
&= \mathbb{E}_q[\log p(\Phi|\alpha)] + \mathbb{E}_q[\log p(U|\gamma)] + \mathbb{E}_q[\log p(\Delta|U)] + \mathbb{E}_q[\log p(\boldsymbol{e}|\Delta)] \\
&+ \mathbb{E}_q[\log p(\Theta|U, \Phi, \tau)] + \mathbb{E}_q[\log p(Z|\Theta)] + \mathbb{E}_q[\log p(\boldsymbol{w}|Z, \beta)] \\
&- \mathbb{E}_q[\log q(\Phi, U, \Delta, \Theta, Z)]
\end{aligned}$$

Thanks to the conjugate priors introduced on U and Φ, the expectations related to these latent variables are straightforward. However, the calculations of $\mathbb{E}_q[\log p(\boldsymbol{e}|\Delta)]$ and $\mathbb{E}_q[\log p(Z|\Theta)]$ are difficult due to no conjugate prior for logistic Normal distribution. We will first provide details of these two nontrivial expectations, and then list the other terms for reproducibility.

• **Compute** $\mathbb{E}_q[\log p(\boldsymbol{e}|\Delta)]$**.** The nonconjugacy of logistic normal leads to difficulty in computing the expected probability of edge assignment between u_i and u_j:

$$\mathbb{E}_q[\log p(e_{ij}|\delta_{ij})] = \mathbb{E}_q[e_{ij}\delta_{ij}] - \mathbb{E}_q[\log(1 + \exp(\delta_{ij}))]$$

We utilize the inequality of logarithm $\log x \leq x - 1$, and set $x = \varepsilon^{-1}(1 + \exp(\delta_{ij}))$, to approximate the second term:

$$\log(1 + \exp(\delta_{ij})) \leq \varepsilon^{-1}(1 + \exp(\delta_{ij})) - 1 + \log \epsilon$$

Thus, the corresponding expectation is as follows,

$$\mathbb{E}_q[\log(1 + \exp(\delta_{ij}))] \leq \varepsilon^{-1}\mathbb{E}_q[1 + \exp(\delta_{ij})] - 1 + \log \varepsilon$$

where the expectation is mean of log normal:

$$\mathbb{E}_q[1 + \exp(\delta_{ij})] = 1 + \exp(\mu^{(\delta_{ij})} + \frac{1}{2}{\sigma^{(\delta_{ij})}}^2)$$

Put them together, we get the expectation as follows:

$$\begin{aligned}
&\mathbb{E}_q[\log p(e_{ij}|\delta_{ij})] \\
&\geq e_{ij}\mu^{(\delta_{ij})} - \varepsilon^{-1}(1 + \exp(\mu^{(\delta_{ij})} + \frac{1}{2}{\sigma^{(\delta_{ij})}}^2)) + 1 - \log \varepsilon
\end{aligned}$$

where a new variational parameter ε is introduced, and we set $\varepsilon = 1 + exp(\delta_{ij})$ to approach the equality.

• **Compute** $\mathbb{E}_q[\log p(Z|\Theta)]$**.** The nonconjugacy of logistic normal also exists in computing the expectation of topic assignment for each word of u_i's d-th document:

$$\mathbb{E}_q[\log p(z_{idn}|\theta_{id})] = \mathbb{E}_q[\theta_{id}^{\mathsf{T}} z_{idn}] - \mathbb{E}_q[\log(\sum_{k=1}^{K} \exp(\theta_{idk}))]$$

We again utilize the equality of logarithm $\log x \leq x - 1$, and set $x = \zeta^{-1}\sum_{k=1}^{K} \exp(\theta_{idk})$ to compute the second term:

$$\log \sum_{k=1}^{K} \exp(\theta_{idk}) \leq \zeta^{-1} \sum_{k=1}^{K} \exp(\theta_{idk}) - 1 + \log \zeta$$

Thus the second term is calculated as:

$$\mathbb{E}_q[\log(\sum_{k=1}^{K} \exp(\theta_{idk}))] \leq \zeta^{-1}(\sum_{k=1}^{K} \mathbb{E}_q[\exp(\theta_{idk})]) - 1 + \log \zeta$$

where the expectation is mean of log normal distribution:

$$\mathbb{E}_q[\exp(\theta_{idk})] = \exp(\mu_k^{(\theta_{id})} + \frac{1}{2}\Sigma_{kk}^{(\theta_{id})})$$

Putting them together, we get the expectation as follows:

$$\begin{aligned}
&\mathbb{E}_q[\log p(z_{idn}|\theta_{id})] \\
&\geq {\mu^{(\theta_{id})}}^{\mathsf{T}} \eta_{idn} - \zeta^{-1} \sum_{k=1}^{K} \exp(\mu_k^{(\theta_{id})} + \frac{1}{2}\Sigma_{kk}^{(\theta_{id})}) + 1 - \log \zeta
\end{aligned}$$

where another variational parameter ζ is introduced, and we set $\zeta = \sum_{k=1}^{K} \exp(\theta_{idk})$ to approach the equality.

• **Compute** $\mathbb{E}_q[\log p(\Phi|\alpha)]$**.** Topic embedding follows Gaussian distributions for p and q, and the corresponding expectation is:

$$\mathbb{E}_q[\log p(\phi_k|\alpha)] \propto \frac{M}{2}\log \alpha - \frac{\alpha}{2}[\sum_{m=1}^{M} \Sigma_{mm}^{(\phi_k)} + {\mu^{(\phi_k)}}^{\mathsf{T}} \mu^{(\phi_k)}]$$

• **Compute** $\mathbb{E}_q[\log p(U|\gamma)]$**.** User embedding also follows Gaussian distributions, thus:

$$\mathbb{E}_q[\log p(u_i|\gamma)] \propto \frac{M}{2}\log \gamma - \frac{\gamma}{2}[\sum_{m=1}^{M} \Sigma_{mm}^{(u_i)} + {\mu^{(u_i)}}^{\mathsf{T}} \mu^{(u_i)}]$$

• **Compute** $\mathbb{E}_q[\log p(\Delta|U)]$**.** The affinity δ_{ij} between a pair of users u_i and u_j follows Gaussian Distributions. Thus, the corresponding expectation can be written as follows:

$$\begin{aligned}
&\mathbb{E}_q[\log p(\delta_{ij}|u_i, u_j)] \\
&\propto -\log \xi - \frac{1}{2\xi^2}\mathbb{E}_q[(\delta_{ij} - u_i^{\mathsf{T}} u_j)^2] \\
&= -\log \xi - \frac{1}{2\xi^2}\mathbb{E}_q[\delta_{ij}^2] + \frac{1}{\xi^2}\mathbb{E}_q[u_i^{\mathsf{T}} u_j]\mathbb{E}_q[\delta_{ij}] - \frac{1}{2\xi^2}\mathbb{E}_q[(u_i^{\mathsf{T}} u_j)^2]
\end{aligned}$$

where the expectations of Gaussian distributions for δ and u can be directly written, thus we get:

$$\begin{aligned}
&\mathbb{E}_q[\log p(\delta_{ij}|u_i, u_j)] \\
&\propto -\log \xi - \frac{1}{2\xi^2}({\mu^{(\delta_{ij})}}^2 + {\sigma^{(\delta_{ij})}}^2) + \frac{\mu^{(\delta_{ij})}}{\xi^2}{\mu^{(u_i)}}^{\mathsf{T}} \mu^{(u_j)} \\
&- \frac{1}{2\xi^2}(\Sigma^{(u_i)} + \mu^{(u_i)}{\mu^{(u_i)}}^{\mathsf{T}})^{\mathsf{T}}(\Sigma^{(u_j)} + \mu^{(u_j)}{\mu^{(u_j)}}^{\mathsf{T}})
\end{aligned}$$

• **Compute** $\mathbb{E}_q[\log p(\Theta|U, \Phi, \tau)]$**.** The topic proportion of each user's document follows Gaussian distribution. The corresponding expectation can be written as:

$$\begin{aligned}
&\mathbb{E}_q[\log p(\theta_{id}|u_i, \Phi, \tau)] \\
&\propto \frac{K}{2}\log \tau - \frac{\tau}{2}\{\mathbb{E}_q[\theta_{id}^{\mathsf{T}}\theta_{id}] - \mathbb{E}_q[\theta_{id}^{\mathsf{T}}\Phi u_i] - \mathbb{E}_q[u_i^{\mathsf{T}}\Phi^{\mathsf{T}}\theta_{id}] + \mathbb{E}_q[u_i^{\mathsf{T}}\Phi^{\mathsf{T}}\Phi u_i]\}
\end{aligned}$$

where Φ and u also follows Gaussian and the calculation is straightforward, thus we get:

$$\mathbb{E}_q[\log p(\theta_{id}|u_i, \Phi, \tau)]$$
$$\propto \frac{K}{2}\log\tau - \frac{\tau}{2}[\sum_{k=1}^{K}\Sigma_{kk}^{(\theta_{id})} + \mu^{(\theta_{id})^\top}\mu^{(\theta_{id})}] + \tau\sum_{k=1}^{K}\mu_k^{(\theta_{id})}\mu^{(\phi_k)^\top}\mu^{(u_i)}$$
$$-\frac{\tau}{2}\sum_{k=1}^{K}(\Sigma^{(u_i)} + \mu^{(u_i)}\mu^{(u_i)^\top})^\top(\Sigma^{(\phi_k)} + \mu^{(\phi_k)}\mu^{(\phi_k)^\top})$$

• **Compute** $\mathbb{E}_q[\log p(\boldsymbol{w}|Z, \beta)]$**.** The expectation of word assignment is given by:

$$\begin{aligned}\mathbb{E}_q[\log p(w_{idn}|z_{idn}, \beta)] &= \sum_{k=1}^{K}\sum_{v=1}^{V}\mathbb{E}_q[z_{idn}^k w_{idn}^v \log\beta_{kv}] \\ &= \sum_{k=1}^{K}\sum_{v=1}^{V} w_{idn}^v \eta_{id,k}\log\beta_{kv}\end{aligned}$$

B PARAMETER ESTIMATION

By taking the gradient of $\mathcal{L}(q)$ in Eq (1) with respect to the variance of user affinity ξ, and set it to 0, we get the closed form estimation as follows:

$$\xi^2 = [U(U-1)]^{-1}\sum_{i=1}^{U}\sum_{j\neq i}^{U}\Big[\mu^{(\delta_{ij})^2} + \sigma^{(\delta_{ij})^2} - 2\mu^{(\delta_{ij})}\mu^{(u_i)^\top}\mu^{(u_j)}$$
$$+ (\Sigma^{(u_i)} + \mu^{(u_i)}\mu^{(u_i)^\top})^\top(\Sigma^{(u_j)} + \mu^{(u_j)}\mu^{(u_j)^\top})\Big]$$

Similarly, the closed form estimation for the variance of document topic proportion τ is given by:

$$\tau^{-1} = (KD)^{-1}\sum_{i=1}^{U}\sum_{d=1}^{D_i}\Big[(\sum_{k=1}^{K}\Sigma_{kk}^{(\theta_{id})} + \mu^{(\theta_{id})^\top}\mu^{(\theta_{id})})$$
$$-\sum_{k=1}^{K}(2\mu_k^{(\theta_{id})}\mu^{(\phi_k)^\top}\mu^{(u_i)} - (\Sigma^{(u_i)} + \mu^{(u_i)}\mu^{(u_i)^\top})^\top(\Sigma^{(\phi_k)} + \mu^{(\phi_k)}\mu^{(\phi_k)^\top}))]$$