Attributed Social Network Embedding

Iizi Iiao [ID], Xiangnan He [ID], Hanwang Zhang, and Tat-Seng Chua

Abstract—Embedding network data into a low-dimensional vector space has shown promising performance for many real-world applications, such as node classification and entity retrieval. However, most existing methods focused only on leveraging network structure. For social networks, besides the network structure, there also exists rich information about social actors, such as user profiles of friendship networks and textual content of citation networks. These rich attribute information of social actors reveal the homophily effect, exerting huge impacts on the formation of social networks. In this paper, we explore the rich evidence source of attributes in social networks to improve network embedding. We propose a generic Attributed Social Network Embedding framework (*ASNE*), which learns representations for social actors (i.e., nodes) by preserving both the *structural proximity* and *attribute proximity*. While the *structural proximity* captures the global network structure, the *attribute proximity* accounts for the homophily effect. To justify our proposal, we conduct extensive experiments on four real-world social networks. Compared to the state-of-the-art network embedding approaches, *ASNE* can learn more informative representations, achieving substantial gains on the tasks of link prediction and node classification. Specifically, *ASNE* significantly outperforms *node2vec* with an 8.2 percent relative improvement on the link prediction task, and a 12.7 percent gain on the node classification task.

Index Terms—Social network representation, homophily, deep learning

---

1 INTRODUCTION

SOCIAL networks are an important class of networks that span a wide variety of media, ranging from social websites such as Facebook and Twitter, citation networks of academic papers, and telephone caller–callee networks—to name a few. Many applications need to mine useful information from social networks. For instance, content providers need to cluster users into groups for targeted advertising [1], and recommender systems need to estimate the preference of a user on items for personalized recommendation [2]. In order to apply general machine learning techniques on network-structured data, it is essential to learn informative node representations.

Recently, research interest in representation learning has spread from natural language to network data [3]. Many network embedding methods have shown promising performance for various applications [3], [4], [5], [6]. However, existing methods have primarily focused on general class of networks and leveraged the structural information only. For social networks, we point out that there almost always exists rich information about social actors in addition to the link structure. For example, users on social websites may have profiles like age, gender and textual comments. We term all such auxiliary information as *attributes*, which not only refer to user demographics, but also include other information such as the affiliated texts and the possible labels.

Attributes essentially exert huge impacts on the organization of social networks. Many studies have justified their importance, ranging from user demographics [7], to subjective preference like political orientation and personal interests [8]. To illustrate this point, we plot the user–user friendship matrix of a Facebook dataset from three views.[1] Each row or column denotes a user, and a colored point indicates that the corresponding users are friends. Each subfigure is a re-ordering of users according to a certain attribute such as "class year", "major" and "dormitory". For example, Fig. 1a first groups users by the attribute "class year", and then sorts these resulting groups in chronological order. As can be seen, there exist clear block structures in each subfigure, where users of a block are more densely connected. Each block actually points to users of the same attribute; for example, the right bottom block of Fig. 1a corresponds to users who will graduate in the year of 2009. This real-world example lends support to the importance of attribute homophily. By jointly considering the attribute homophily and network structure, we believe more informative node representations can be learned. Moreover, since we utilize the auxiliary attribute information, the link sparsity and cold-start problem [10] can largely be alleviated.

In this paper, we present a neural framework named *ASNE* for learning node representations from social network data. *ASNE* is a generic machine learner working with real-valued feature vectors, where each feature denotes the ID or an attribute of a node. Through this, we can easily incorporate any type and number of attributes. Under our

- *L. Liao is with the NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore 117456. E-mail: liaolizi.llz@gmail.com.*
- *X. He and T.-S. Chua are with the School of Computing, National University of Singapore, Singapore 117456. E-mail: xiangnanhe@gmail.com, dcscts@nus.edu.sg.*
- *H. Zhang is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798. E-mail: hanwangzhang@gmail.com.*

1. This is the Chapel Hill data constructed by [9], which we will detail later in Section 5.1.1.
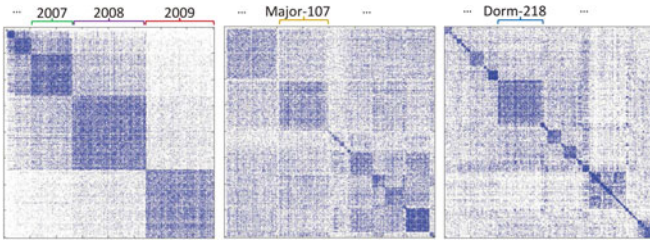
Fig. 1. Attribute homophily largely impacts social network: We group users in each 4018 × 4018 user matrix based on a specific attribute. Clear blocks around the diagonal show the attribute homophily effect.

ASNE framework, each feature is associated with an embedding, and the final embedding for a node is aggregated from its ID embedding (which preserves the *structural proximity*) and attribute embedding (which preserves the *attribute proximity*). To capture the complex interactions between features, we adopt a multi-layer neural network to take advantage of strong representation and generalization ability of deep learning.

In summary, the main contributions are as follows.

- We demonstrate the importance of integrating network structure and attributes for learning more informative node representations for social networks.
- We propose a generic framework *ASNE* to perform network embedding by preserving the *structural proximity* and *attribute proximity* of social networks.
- We conduct extensive experiments on four datasets with two tasks of link prediction and node classification. Empirical results and case studies demonstrate the effectiveness and rationality of *ASNE*.

The rest of the paper is organized as follows. We first discuss the related work in Section 2, followed by providing some preliminaries in Section 3. We then present the *ASNE* framework in Section 4. We show experimental results in Section 5, before concluding the whole paper in Section 6.

2 RELATED WORK

In this section, we briefly summarize studies about attribute homophily. We then discuss network embedding methods that are closely related to our work.

2.1 Attribute Homophily in Social Networks

Social networks belong to a special class of networks, in which the formation of social ties involves not only the self-organizing network process, but also the attribute-based process [11]. The motivation for considering *attribute proximity* in the embedding procedure is rooted in the large impact of attribute homophily, which plays an important role in attribute-based process. Therefore, we provide a brief summarization of homophily studies here as a background. Generally speaking, the "homophily principle"—birds of a feather flock together—is one of the most striking and robust empirical regularities of social life [12], [13], [14]. The hypothesis that people similar to each other tend to become friends dates back to at least the 70s in the last century. In social science, there is a general expectation that individuals develop friendships with others of approximately the same age [15]. In [16] the authors studied the inter-connectedness between homogeneous composition of groups and the emergence of homophily. In [17] the authors

tried to find the role of homophily in online dating choices made by users. They found that online users of the online dating system seek people like them much more often than chance would predict, just as in the offline world. In more recent years, [18] investigated the origins of homophily in a large university community, using network data in which interactions, attributes and affiliations were all recorded over time. Not surprisingly, it has been concluded that besides structural proximity, preferences for attribute similarity also provides an important factor for the social network formation procedure. Thus, to get more informative representations for social networks, we should take attributes information into consideration.

2.2 Network Embedding

There are many ways to "embed" nodes (described by high-dimensional vectors or by pairwise similarities) into a lower-dimensional space. Multidimensional scaling methods [19] preserved distances between nodes such as Euclidean distance, some nonlinear squashing of distances, or shortest graph paths as with IsoMAP [20]. Principal components analysis (PCA) aimed to find a linear projection of the original data which captures as much variance as possible. Other methods attempted to preserve local geometry (e.g., LLE [21]).

Later, latent variable graphical models increased the interpretability of the modeling process. In one line of work, latent class models assume that there are a number of clusters (classes) and that each node belongs to a single cluster [22]. Under these models, the link probability between two nodes depends only on their cluster assignments. To increase the expressiveness, Mixed Membership Stochastic Block Model (MMSB) [23] allowed mixed membership, by associating each object with a distribution over clusters. In another line of work, latent feature models allow each object to possess a vector of features and determine the link probabilities based on interactions among the features. [24] proposed to use a factorial structure to allow each object to be a member of many subclusters as long as these subclusters are in different features. Further, there are studies focusing on jointly modeling information about user behaviors (or attributes) and social interactions. Pairwise Link-LDA [25], [26] combined the Latent Dirichlet Allocation (LDA) and MMSB, but often failed to exploit the synergies between different modalities, as it tends to fragment the latent space. To address this shortcoming, [27], [28] were designed to impose additional constraints on the generative model that would require stronger correlations between links and attributes. However, such models do not always capture the correct community structure of large real-world networks [29]. Most of these works focus on modeling the generative processes, such as drawing latent or observed variables from certain distributions. The discriminative power of generated latent variables might be compromised due to inaccurate or oversimplified assumptions. Furthermore, these methods usually represent each node with a topic distribution [30] where each entry is non-negative and all entries sum to one. Such an inherent constraint limits the representation ability of the resulting node embeddings. Therefore, it has been shown that neural network models such as [30], [31] can achieve superior performance over latent variable graphical models such as NetPLSA [32] and RTM [27].

Recent works focus more on leveraging neural network models to embed an existing network into a low-dimensional

vector space, which achieves better performance than those earlier works. In [3] the authors deployed truncated random walks on networks to generate node sequences. The generated node sequences are treated as sentences in language models and fed to the Skip-gram model to learn the embeddings. In [5] the authors modified the way of generating node sequences by balancing breadth-first sampling and depth-first sampling, and achieved performance improvements. Instead of performing simulated "walks" on the networks, [6] proposed clear objective functions to preserve the *first-order proximity* and *second-order proximity* of nodes while [10] introduced deep models with multiple layers of non-linear functions to capture the highly non-linear network structure. However, all these methods only leverage network structure. In social networks, there exists large amount of attribute information. Purely structure-based methods fail to capture such valuable information, thus may result in less informative embeddings. In addition, these methods get affected easily when the link sparsity problem occurs. Therefore, some recent efforts have explored the possibility of integrating contents to learn informative representations [33]. For example, *TADW* [30] proposed text-associated *DeepWalk* [3] to incorporate text features into the matrix factorization framework. However, only text attributes can be handled. Being with the same problem, *TriDNR* [31] proposed to separately learn embeddings from the structure-based *DeepWalk* model [3] and label-fused *Doc2Vec* attribute model [34], the embeddings learned were linearly combined together in an iterative way. Under such a scheme, the knowledge transferring between the two separate models only goes through a series of weighted sum operations and lacks further convergence constrains. On the contrary, our method models the structure proximity and attribute proximity in an end-to-end neural network that does not have such limitations. In this way, the attribute part only needs to complement the weaknesses of the structure part, rather than a full-size attribute model [35].

There have also been efforts explored semi-supervised learning for network embedding. [36] combined an embedding-based regularizer with a supervised learner to incorporate label information. Instead of imposing regularization, [37] used embeddings to predict the context in graph and leveraged label information to build both transductive and inductive formulations. In our framework, label information can also be incorporated in the same way similar to [37] when available. We leave this extension as future work, as this work focuses on the modeling of attributes for network embedding.

3 DEFINITIONS

Social networks are more than links; in most cases, social actors are associated with rich attributes. We denote a social network as $G = (\mathcal{U}, \mathcal{E}, \mathcal{A})$, where $\mathcal{U} = \{u_1, \ldots, u_M\}$ denotes the social actors, $\mathcal{E} = \{e_{ij}\}$ denotes the links between social actors, and $\mathcal{A} = \{\mathcal{A}_i\}$ denotes the attributes of social actors. Each edge e_{ij} can be associated with a weight s_{ij} denoting the strength of connection between u_i and u_j. Generally, our analysis can apply to any (un) directed, (un)weighted network. While in this paper, we focus on unweighted network, i.e., s_{ij} is 1 for all edges, our method can be easily applied to weighted network through the neighborhood sampling strategy [5].
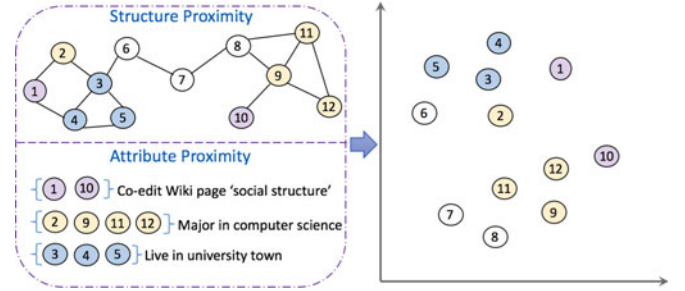

Fig. 2. An illustration of social network embedding. The numbered nodes denote users, and users of the same color share the referred attribute.

The aim of social network embedding is to project the social actors into a low-dimensional vector space (*a.k.a.* embedding space). Since the network structure and attributes offer different sources of information, it is crucial to capture both of them to learn a comprehensive representation of social actors. To illustrate this point, we show an example in Fig. 2. Based on the link structure, a common assumption of network embedding methods [3], [5], [6] is that closely connected users should be close to each other in the embedding space. For example, $(u_1, u_2, u_3, u_4, u_5)$ should be close to each other, and similarly for $(u_8, u_9, u_{11}, u_{12})$. However, we argue that purely capturing structural information is far from enough. Taking the attribute homophily effect into consideration, $(u_2, u_9, u_{11}, u_{12})$ should also be close to each other. This is because they all major in computer science; although u_2 is not directly linked to u_9, u_{11} or u_{12}, we could expect that some computer science articles popular among (u_9, u_{11}, u_{12}) might also be of interest to u_2. To learn more informative representations for social actors, it is essential to capture the attribute information.

In this work, we strive to develop embedding methods that preserve both the *structural proximity* and *attribute proximity* of social network. In what follows, we give the definition of the two notions.

Definition 1 (Structural Proximity). *Denotes the proximity of nodes that is evidenced by links. For u_i and u_j, if there exists a link e_{ij} between them, it indicates the direct proximity; on the other hand, if u_j is within the context of u_i, it indicates the indirect proximity.*

Intuitively, the direct proximity corresponds to the first-order proximity, while the indirect proximity accounts for higher-order proximities [6]. A popular way to generate contexts is by performing random walks in the network [3], i.e., if two nodes appear in a walking sequence, they are treated as in the same context. In our method, we apply the walking procedure proposed by *node2vec* [5], which controls the random walk by balancing the breadth-first sampling (BFS) and depth-first sampling (DFS). In the remaining of the paper, we use the term "neighbors" to denote both the first-order neighbors and the nodes in the same context for simplicity.

Definition 2 (Attribute Proximity). *Denotes the proximity of nodes that is evidenced by attributes. The attribute intersection of $\mathcal{A}_i$ and $\mathcal{A}_j$ indicates the attribute proximity of u_i and u_j.*

By enforcing the constraint of *attribute proximity*, we can model the attribute homophily effect, as social actors with similar attributes will be placed close to each other in the embedding space.

4 PROPOSED METHOD

We first describe how we model the *structural proximity* with a deep neural network architecture. We then elaborate how to model the *attribute proximity* with a similar architecture by casting attributes to a generic feature representation. Our final *ASNE* model integrates the models of structures and attributes by an early fusion on the input layer. Lastly, we discuss the relationships of our *ASNE* model to other relevant models.

4.1 Structure Modeling

Since the focus of this section is on the modeling of network structure, we use only the identity (ID) to represent a node with one-hot representation, in which a node u_i is represented as an M-dimensional sparse vector where only the ith element of the vector is 1. Based on our definition of *structural proximity*, the key to structure modeling is in the estimation of pairwise proximity of nodes. Let f be the function that maps two nodes u_i, u_j to their estimated proximity score. We define the conditional probability of node u_j on u_i using the softmax function as

$$p(u_j|u_i) = \frac{exp(f(u_i, u_j))}{\sum_{j'=1}^{M} exp(f(u_i, u_{j'}))}, \tag{1}$$

which measures the likelihood that node u_j is connected with u_i. Note that when f is set to be the cross-product transformation, the probability is similar to [3], [6].

To account for a node's *structural proximity w.r.t.* all its neighbors, we further define the conditional probability of a node set by assuming conditional independence

$$p(\mathcal{N}_i|u_i) = \prod_{j \in \mathcal{N}_i} p(u_j|u_i), \tag{2}$$

where $\mathcal{N}_i$ denotes the neighbor nodes of u_i. By maximizing this conditional probability over all nodes, we can achieve the goal of preserving the global *structural proximity*. Specifically, we define the likelihood function for the global structure modeling as

$$l = \prod_{i=1}^{M} p(\mathcal{N}_i|u_i) = \prod_{i=1}^{M} \prod_{j \in \mathcal{N}_i} p(u_j|u_i). \tag{3}$$

Having established the target of learning from network data, we now design an embedding model to estimate the pairwise proximity $f(u_i, u_j)$. Most previous efforts have used shallow models for relational modeling, such as matrix factorization [38], [39] and neural networks with one hidden layer [3], [5], [40]. In these formulations, the proximity of two nodes is usually modeled as the inner product of their embedding vectors. However, It is known that simply the inner product of embedding vectors can limit the model's representation ability and incur large ranking loss [41]. To capture the complex non-linearities of real-world networks [10], [42], we propose to adopt a deep architecture to model the pairwise proximity of nodes

$$\begin{aligned} f_{id}(u_i, u_j) \\ = \tilde{\mathbf{u}}_j \cdot \delta_n(\mathbf{W}^{(n)}(\dots \delta_1(\mathbf{W}^{(1)}\mathbf{u}_i + \mathbf{b}^{(1)})\dots) + \mathbf{b}^{(n)}), \end{aligned} \tag{4}$$

where $\mathbf{u}_i$ denotes the embedding vector of node u_i, and n denotes the number of hidden layers to transform an embedding vector to its final representation; $\mathbf{W}^{(n)}$, $\mathbf{b}^{(n)}$ and
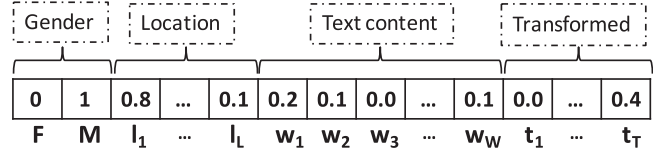
Gender		Location			Text content					Transformed		
0	1	0.8	...	0.1	0.2	0.1	0.0	...	0.1	0.0	...	0.4
F	M	l_1	...	l_L	w_1	w_2	w_3	...	w_W	t_1	...	t_T

Fig. 3. A simple example to show the two kinds of social network attributes information.

δ_n denote the weight matrix, bias vector and activation function of the nth hidden layer, respectively.

It is worth noting that in our model design, each node has two latent vector representations, $\mathbf{u}$ that encodes a node to its embedding and $\tilde{\mathbf{u}}$ that embeds the node as a neighbor. To comprehensively represent a node for downstream applications, practitioners can add/concatenate the two vectors which has empirically shown to have better performance in distributed word representations [43], [44].

4.2 Encoding Attributes

Many real-world social networks contain rich attribute information, which can be heterogeneous and highly diverse. To avoid manual efforts that design specific model components for specific attributes, we convert all attributes to a generic feature vector representation (see Fig. 3 as an example) to facilitate designing a general method for learning from attributes. Regardless of semantics, we can categorize attributes into two types:

- *Discrete attributes*. A prevalent example is categorical variables, such as user demographics like gender and country. We convert a categorical attribute to a set of binary features via one-hot encoding. For example, the gender attribute has two values {*male*, *female*}, so we can express a female user as the vector $\mathbf{v} = \{0, 1\}$ where the second binary feature of value 1 denotes "female".
- *Continuous attributes*. Continuous attributes naturally exist on social networks, e.g., raw features of images and audios. Or they can be artificially generated from the transformation of categorical variables. For example, in document modeling, after obtaining bag-of-words representation of a document, it is common to transform it to a real-valued vector via TF-IDF to reduce noises. Another example is the historical features, such as users' purchases on items and check-ins on locations, which are always normalized to a real-valued vector to reduce the impact of variable length [45].

Suppose there are K feature entries in the attribute feature vector $\mathbf{v}$ as shown in Fig. 3, for each feature entry, we associate it with an low-dimensional embedding vector $\mathbf{e}_k$ which corresponds to the kth column of the weight matrix $\mathbf{W}^{att}$ as shown in Fig. 4. We then aggregate the attribute representation vector $\mathbf{u}'$ for each input social actor by $\mathbf{u}' = \sum_{k=1}^{K} v_k \mathbf{e}_k$.

Similar to structure modeling, we aim to model the *attribute proximity* by adopting a deep model to approximate the complex interactions between attributes and introduce non-linearity, which can be fulfilled by Equation (4) while substituting $\mathbf{u}_i$ with $\mathbf{u}'_i$.

4.3 The ASNE Model

To combine the strength of both structure and attribute modeling, an intuitive way is to concatenate the learned
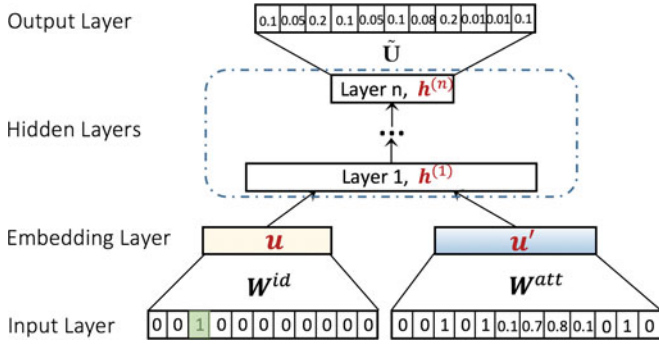
Fig. 4. Attributed social network embedding (*ASNE*).

embeddings from each part by *late fusion* as adopted by [6]. However, the main drawback of late fusion is that individual models are trained separately without knowing each other and results are simply combined after training. On the contrary, *early fusion* allows optimizing all parameters simultaneously. As a result, the attribute modeling can complement the learning of structure modeling, allowing the two parts closely interact with each other. Essentially, the strategy of early fusion is more preferable in recent developments of end-to-end deep learning methods, such as Deep crossing [46] and Neural Factorization Machines [47]. Therefore, we propose a generic attributed social network embedding framework (*ASNE*) as shown in Fig. 4, which integrates the structure and attribute modeling parts by an early fusion on the input layer. In what follows, we elaborate the design of *ASNE* layer by layer.

Embedding Layer. The embedding layer consists of two fully connected components. One component projects the one-hot user ID vector to a dense vector **u** which captures structure information. The other component encodes the generic feature vector and generates a compact vector **u'** which aggregates attributes information.

Hidden Layers. Above the embedding layer, **u** and **u'** are fed into a multi-layer perceptron. The hidden representations for each layer are denoted as $\mathbf{h}^{(0)}, \mathbf{h}^{(1)}, \ldots, \mathbf{h}^{(n)}$, which are defined as follows:

$$
\begin{aligned}
\mathbf{h}^{(0)} &= \begin{bmatrix} \mathbf{u} \\ \lambda \mathbf{u'} \end{bmatrix}, \\
\mathbf{h}^{(k)} &= \delta_k(\mathbf{W}^{(k)} \mathbf{h}^{(k-1)} + \mathbf{b}^{(k)}), \ k = 1, 2, \ldots, n,
\end{aligned}
\tag{5}
$$

where $\lambda \in \mathbb{R}$ adjusts the importance of attributes, δ_k denotes the activation function, n is the number of hidden layers. From the last hidden layer, we obtain an abstractive representation $\mathbf{h}_i^{(n)}$ of the input social actor u_i.

Stacking multiple non-linear layers has been shown to help learning better representations of data [48]. Regarding the architecture design, a common strategy is to use a tower structure, where each successive layer has a smaller number of neurons. The premise is that by using a small number of hidden units for higher layers, they can learn more abstractive features of data [48]. Therefore, as depicted in Fig. 4, we implement the *hidden layers* component following the tower structure with halved layer size for each successive higher layer. Such a design has also been shown to be effective by recent work on recommendation task [41]. Moreover, **u** and **u'** are concatenated with weight adjustments λ before fed into the fully connected layers, which has been shown to help learning higher-order interactions between **u** and **u'** [41], [46].

Output Layer. At last, the output vector of the last hidden layer $\mathbf{h}_i^{(n)}$ is transformed into a probability vector **o**, which contains the predictive link probability of u_i to all the nodes in $\mathcal{U}$

$$
\mathbf{o} = [p(u_1|u_i), p(u_2|u_i), \ldots, p(u_M|u_i)].
\tag{6}
$$

Denoting the abstractive representation of a neighbor u_j as $\tilde{\mathbf{u}}_j$ which corresponds to a row in the weight matrix $\tilde{\mathbf{U}}$ between the last hidden layer and the output layer, the proximity score between u_i and u_j can be defined as below:

$$
f(u_i, u_j) = \tilde{\mathbf{u}}_j \cdot \mathbf{h}_i^{(n)},
\tag{7}
$$

which can be fed into Equation (1) for further obtaining the predictive link probability $p(u_j|u_i)$ in vector **o**

$$
p(u_j|u_i) = \frac{exp(\tilde{\mathbf{u}}_j \cdot \mathbf{h}_i^{(n)})}{\sum_{j'=1}^{M} exp(\tilde{\mathbf{u}}_{j'} \cdot \mathbf{h}_i^{(n)})},
\tag{8}
$$

where all the parameters $\Theta = \{\Theta_h, \mathbf{W}^{id}, \mathbf{W}^{att}, \tilde{\mathbf{U}}\}$ and Θ_h denotes the weight matrices and biases in the *hidden layers* component.

4.3.1 Optimization

To estimate the model parameters of the whole *ASNE* framework, we need to specify an objective function to optimize. As detailed in Equation (3), we aim to maximize the conditional link probability over all nodes. In this way, the whole *ASNE* framework is jointly trained to maximize the likelihood with respect to all the parameters Θ

$$
\begin{aligned}
\Theta^* &= \arg \max_{\Theta} \prod_{i=1}^{M} \prod_{j \in \mathcal{N}_i} p(u_j|u_i) \\
&= \arg \max_{\Theta} \sum_{u_i \in M} \sum_{u_j \in \mathcal{N}_i} log \, p(u_j|u_i)
\end{aligned}
\tag{9}
$$

$$
= \arg \max_{\Theta} \sum_{u_i \in M} \sum_{u_j \in \mathcal{N}_i} log \, \frac{exp(\tilde{\mathbf{u}}_j \cdot \mathbf{h}_i^{(n)})}{\sum_{j' \in M} exp(\tilde{\mathbf{u}}_{j'} \cdot \mathbf{h}_i^{(n)})}.
\tag{10}
$$

Maximizing the softmax scheme in Equation (10) actually has two effects: to enhance the similarity between any u_i and these $u \in \mathcal{N}_i$ as well as to weaken that between any u_i and these $u \notin \mathcal{N}_i$. However, this causes two major problems. The first one lies in the fact that if two social actors are not linked together, it does not necessarily mean they are dissimilar. For example, many users in social websites are not linked, not because they are dissimilar. Most of the times, it is simply because they never had the chance to know each other. Thus forcing dissimilarity between u_i and all the other actors not inside $\mathcal{N}_i$ will be inappropriate. The second problem arises from the calculation of the normalization constant in Equation (10). In order to calculate a single probability, we need to go through all the actors in the whole network, which is computationally inefficient. In order to avoid these problems, we apply negative sampling procedure [40], [49] where only a very small subset of users are sampled from the whole social network.

The main idea is to do approximation in the gradient calculation procedure. When we consider the gradient of the log-probability in Equation (9), the gradient is actually composed of a positive and a negative part as follows:

$$\nabla \, log \, p(u_j|u_i) = \nabla \, f(u_i, u_j) - \sum_{j' \in M} p(u_{j'}|u_i) \nabla \, f(u_i, u_{j'}),$$

where $f(u_i, u_j) = \tilde{\mathbf{u}}_j \cdot \mathbf{h}_i^{(n)}$ as defined in Equation (7). Note that given the actor u_i, the negative part of the gradient is in essence the expected gradient of $\nabla f(u_i, u_{j'})$, denoting as $\mathbb{E}[\nabla f(u_i, u_{j'})]$. The key idea for sampling a subset of social actors is to approximate this expectation, resulting in much lower computational complexity as well as avoiding too strong constraint on those not linked actors.

To optimize the aforementioned framework, we apply the Adaptive Moment Estimation (Adam) [50], which adapts the learning rate for each parameter by performing smaller updates for the frequent parameters and larger updates for the infrequent parameters. The Adam method combines the advantages of two popular optimization methods: the ability of AdaGrad [51] to deal with sparse gradients, and the ability of RMSProp [52] to deal with non-stationary objectives. To address internal covariate shift [53] which slows down the training by requiring careful settings of learning rate and parameter initialization, we adopt batch normalization [53] in our multi-layer *ASNE* framework. In the embedding layer and each hidden layer, we also add dropout component to alleviate overfitting. After proper optimization, we obtain abstractive representation $\mathbf{h}^{(n)}$ and $\tilde{\mathbf{u}}$ for each social actor, similar to [43], [44], we use $\mathbf{h}^{(n)} + \tilde{\mathbf{u}}$ as the final representation for each social actor, which returns us better performance results.

4.4   Connections to Other Models

In this section, we discuss the connection of the proposed *ASNE* framework to other related models. We show that *ASNE* subsumes the state-of-the-art network embedding method *node2vec* [5] and the linear latent factor model *SVD++* [54]. Specially, the two models can be seen as a special case of shallow *ASNE*. We also show that there is close relationship between shallow *ASNE* and *TriDNR*. To facilitate further discussion, we first give the link prediction model of the one-hidden-layer *ASNE* as

$$f(u_i, u_j) = \tilde{\mathbf{u}}_j \cdot \delta_1 \left(\mathbf{W}^{(1)} \begin{bmatrix} \mathbf{u}_i \\ \lambda \mathbf{u}_i' \end{bmatrix} + \mathbf{b}^{(1)} \right). \tag{11}$$

4.4.1   ASNE *versus* node2vec

The *node2vec* applies a shallow neural network model to learning node embeddings. It excels *DeepWalk* by the hybrid BFS and DFS strategies for random walks, which is shared by our *ASNE*. Under the context of *ASNE*, the essence of *node2vec* can be seen as estimating the proximity of two nodes as

$$f_{node2vec}(u_i, u_j) = \tilde{\mathbf{u}}_j \cdot \mathbf{u}_i.$$

By setting λ to 0.0 (i.e., no attribute modeling), δ_1 to an identity function (i.e., no nonlinear transformation), $\mathbf{W}^{(1)}$ to an identity matrix and $\mathbf{b}^{(1)}$ to a zero vector (i.e., no trainable hidden neurons), we can exactly recover the *node2vec* model from Equation (11).

4.4.2   ASNE *versus* SVD++

The *SVD++* is one of the most effective latent factor models for collaborative filtering [54], originally proposed to model the ratings of users to items. Given a user u and an item i, the prediction model of *SVD++* is defined as

$$f_{SVD++}(u, i) = \mathbf{q}_i \cdot \left(\mathbf{p}_u + \sum_{k \in \mathcal{R}_u} \mathbf{y}_k \right),$$

where $\mathbf{p}_u$ ($\mathbf{q}_i$) denotes the embedding vector for user u (item i); $\mathcal{R}_u$ denotes the set of rated items for u, and $\mathbf{y}_k$ denotes another embedding vector for item k for modeling the item–item similarity. By treating the item as a "neighbor" of the user for estimating the proximity, we reformulate the model using the symbols of our *ASNE*

$$f_{SVD++}(u_i, u_j) = \tilde{\mathbf{u}}_j \cdot \left(\mathbf{u}_i + \mathbf{u}_i' \right),$$

where $\mathbf{u}_i'$ denotes the sum of item embedding vectors of $\mathcal{R}_u$, which corresponds to the aggregated attribute representation of u_i in *ASNE*.

To see how *ASNE* subsumes the model, we first set δ_1 to an identity function, λ to 1.0, and $\mathbf{b}^{(1)}$ to a zero vector, reducing Equation (11) to

$$f(u_i, u_j) = \tilde{\mathbf{u}}_j \cdot \mathbf{W}^{(1)} \begin{bmatrix} \mathbf{u}_i \\ \mathbf{u}_i' \end{bmatrix}.$$

By further setting $\mathbf{W}^{(1)}$ to a concatenation of two identity matrices (i.e., $\mathbf{W}^{(1)} = [\mathbf{I}, I]$), we can recover the *SVD++* model

$$f(u_i, u_j) = \tilde{\mathbf{u}}_j \cdot \left(\mathbf{u}_i + \mathbf{u}_i' \right).$$

4.4.3   ASNE *versus* TriDNR

Under the context of *TriDNR*, the link prediction model can be written as

$$f_{TriDNR}(u_i, u_j) = (\mathbf{u}_j + \lambda \mathbf{u}_j') \cdot (\mathbf{u}_i + \lambda \mathbf{u}_i'),$$

where $\mathbf{u}$ denotes the structure representation learned from *DeepWalk* model while $\mathbf{u}'$ denotes the attribute representation captured via *Doc2Vec* model. The final node representation simply comes from a series of weighted summation during the iterative learning procedure as in [31], which assumes the same vector space for structure and attribute modeling. By treating u_j as a neighbor, the representation part of *TriDNR* can even been viewed as a weighted extension for that of the *SVD++* model.

Through the connection between *ASNE* and a family of shallow models, we can see the rationality behind our design of *ASNE*. Particularly, *ASNE* deepens the shallow models so as to capture the underlying interactions between the network structure and attributes. When modeling real-world data that may have complex and non-linear inherent structure [10], [42], our *ASNE* is more expressive and can better fit on the real-world data.

5   EXPERIMENTS

In this section, we conduct experiments on four publicly accessible social network datasets to answer the following research questions.

RQ1   Can *ASNE* learn better node representations as compared to state-of-the-art network embedding methods?

RQ2   What are the key reasons that lead to better representations learned by *ASNE*?

RQ3   Are deeper layers of hidden units helpful for learning better social network embeddings?

In what follows, we first describe the experimental settings. We then answer the above three research questions.

TABLE 1
Statistics of the Datasets

Dataset	#($\mathcal{U}$)	#($\mathcal{E}$)
OKLAHOMA [9]	17,425	892,528
UNC [9]	18,163	766,800
DBLP [31]	60,744	52,890
CITESEER [31]	29,751	77,218

5.1 Experimental Setup

5.1.1 Datasets

We conduct experiments on four public datasets, which are representative of two types of social networks—social friendship networks and academic citation networks [55]. The statistics of the four datasets are summarized in Table 1.

FRIENDSHIP Networks. We use two Facebook networks constructed by [9], which contain students from two American universities: University of Oklahoma (OKLAHOMA) and University of North Carolina at Chapel Hill (UNC), respectively. Besides user ID, there are seven anonymized attributes: status, gender, major, second major, dorm/house, high school, class year. Note that not all students have the seven attributes available. For example, for the UNC dataset, 4,018 of the 18,163 users contain all attributes (as plotted in Fig. 1).

CITATION Networks. For citation networks, we use the DBLP and CITESEER[2] data used in [31]. Each node denotes a paper. The attributes are the title contents for each paper after removing stop words and the stemming process. The DBLP dataset consists of bibliography data in computer science from [56].[3] A list of conferences from four research areas are selected. The CITESEER dataset consists of scientific publications from ten distinct research areas. These research areas are treated as class labels in the node classification task.

5.1.2 Evaluation Protocols

We adopt two tasks—link prediction and node classification—which have been widely used in literature to evaluate network embeddings [3], [5]. While the link prediction task assesses the ability of node representations in reconstructing network structure [10], node classification evaluates whether the representations contain sufficient information trainable for downstream applications.

Link Prediction. We follow the widely adopted way in [5], [10]: we randomly hold out 10 percent links as the test set, 10 percent as the validation set for tuning hyper-parameters, and train *ASNE* on the remaining 80 percent links. Since the test/validation set contains only positive instances, we randomly sample the same number of non-existing links as negative instances [5], and rank both positive and negative instances according to the prediction function. To judge the ranking quality, we employ the *area under the ROC curve* (AUROC) [57], which is widely used in IR community to evaluate a ranking list. It is a summary measure that essentially averages accuracy across the spectrum of test values. A higher value indicates a better performance, and

2. http://citeseerx.ist.psu.edu/
3. http://arnetminer.org/citation (V4 version is used)

an ideal model that ranks all positive instances higher than negative instances has an AUROC value of 1.

Node Classification. We first train models on the training sets (with links and all attributes but no class labels) to obtain node representations; the hyper-parameters for each model are chosen based on the performance of link prediction. We then train classifiers using learned node representations, which is widely adopted in [3], [10]. To evaluate the node classification performance, we randomly sample a portion of labeled nodes ($\rho \in \{10\%, 30\%, 50\%\}$) as training, using the remaining labeled nodes as test. We repeat this process 10 times, and report the mean of the *Macro-F1* and *Micro-F1* scores. As only the DBLP and CITESEER datasets contain class labels for nodes, the node classification task is performed on the two datasets.

5.1.3 Comparison Methods

We compare *ASNE* with several state-of-the-art network embedding methods.

- *node2vec* [5]: It applies the Skip-Gram model [40] on the node sequences generated by biased random walk. There are two key hyper-parameters p and q that control the random walk, which we tuned them the same way as the original paper. Note that when p and q are set to 1, *node2vec* degrades to DeepWalk [3].

- *LINE* [6]: It learns two embedding vectors for each node by preserving the first-order and second-order proximity of the network, respectively. Then the embedding vectors are concatenated as the final representation for a node. We followed the hyper-parameter settings of [6] and the number of training samples S (millions) is adapted to our data size.

- *TriDNR* [31]: It learns node representations by coupling multiple neural network models to jointly exploit the network structure, node–content correlation, and label–content correspondence. This is one of the state-of-the-art network embedding methods that also uses attribute information. We searched the text weight (tw) hyper-parameter among $[0.0, 0.2, \ldots, 1.0]$.

- *AANE* [33]: It also proposes a framework to learn unified embedding representations by incorporating node attribute proximity into network embedding. However, they formulate it as a bi-convex optimization problem and decompose the complex modeling and optimization into sub-problems.

We also tried a latent variable model *SLR* [58]. Although it captures both attribute and link information, there is a 6.34 percent performance drop (in terms of *AUROC* score) on average as compared to *LINE* which only considers links. This might be because *SLR* assumes that each node takes roles following certain distributions such as Bernoulli, while methods like *LINE* do not have such assumptions. In addition, *SLR* works on triangular motifs while methods like *LINE* capture both first-order and second-order proximities in the network. As there exists a large performance gap, we omit the results of *SLR* for clear comparisons in Fig. 5.

For all baselines, we use the implementation released by the original authors. Note that although *node2vec* and *LINE* are state-of-the-art methods for embedding networks, they are designed to use only the structure information. For a fair comparison with *ASNE* that additionally exploits attributes, we further extend them to include attributes by concatenating the learned node representation with the
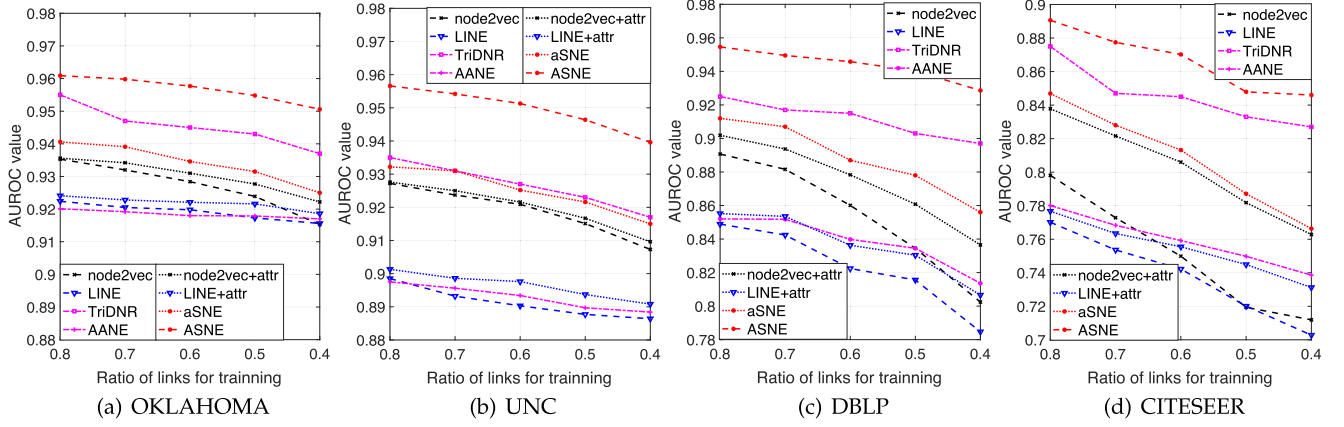
Fig. 5. Performance of link prediction on social networks *w.r.t.* different network sparsity (RQ1).

attribute feature vector. We dub the variants *node2vec+attr* and *LINE+attr*. In addition, we compare with the pure multi-layer structure modeling part denoted as *sSNE* to check the effect of going deep. Moreover, to validate the effectiveness of early attribute fusion, we design a late fusion variant *aSNE*, which directly concatenates separate embeddings learned from the multi-layer structure modeling part and multi-layer attribute modeling part.

5.1.4   Parameter Settings

Our implementation of *ASNE* is based on TensorFlow.[4] Our code is public available in github page.[5] Regarding the choice of activation function of hidden layers, we have tried rectified linear unit (*ReLU*), soft sign (*softsign*) and hyperbolic tangent function (*tanh*), finding softsign leads to the best performance in general. As such, we use softsign for all experiments. We randomly initialize model parameters with a Gaussian distribution (with a mean of 0.0 and standard deviation of 0.01), optimizing the model with mini-batch Adam [50]. We tune the batch size and the learning rate via grid search. The search space of the concatenation hyper-parameter λ is the same as tw of *TriDNR*, where a value of $\lambda = 0.0$ degrades to a model that considers only the structure (*c.f.*, Section 4.1). The concatenation parameter λ is searched in same space as tw. More detailed impact of λ is studied in Section 5.2.3. The embedding dimension d is set to 128 for all methods in line with *node2vec* and *LINE*. The hyper-parameter p and q for controlling the walking procedure are set to be the same with that of *node2vec*. Without special mention, we use two hidden layers.

5.2   Quantitative Analysis (RQ1)

5.2.1   Link Prediction

Fig. 5 shows the AUROC scores of *ASNE* and baseline methods on the four datasets. To explore the robustness of embedding methods *w.r.t.* the network sparsity, we vary the ratio of training links and investigate the performance change. The key observations are as follows:

1) Our proposed *ASNE* achieves the best performance among all methods. Notably, compared to the pure structure-based methods *node2vec* and *LINE*, our *ASNE* performs significantly better with only half

links. This demonstrates the usefulness of attributes in predicting missing links, as well as the rationality of *ASNE* in leveraging attributes for learning better node representation. Moreover, we observe more dramatic performance drop of *node2vec* and *LINE* on DBLP and CITESEER, as compared to that of OKLA-HOMA and UNC. The reason is that the DBLP and CITESEER datasets contain less link information (as shown in Table 1); as such, the link sparsity problem becomes more severe when the ratio of training links decreases. On the contrary, our *ASNE* exhibits more stability when we use fewer links for training, which is credible to its effective modeling of attributes.

2) Focusing on methods that account for attributes, we find how to incorporate attributes plays a pivotal role for the performance. First, *node2vec+attr* (*LINE+attr*) slightly improves over *node2vec* (*LINE*), which reflects the value of attributes. Nevertheless, the rather modest improvements indicate that simply concatenating attributes with the embedding vector is insufficient to fully leverage the rich signal in attributes. This reveals the necessity of designing a more principled approach to incorporate attributes into the network embedding process. Moreover, by comparing the performance of early attribute fusion (*ASNE*) and late attribute fusion (*aSNE*), we validate the effectiveness of the early fusion framework. Second, we can see that *ASNE* consistently outperforms *TriDNR*—one of the most competitive baselines that also incorporates attributes into the network embedding process. Although *TriDNR* is a joint model, it separately trains the structured-based *DeepWalk* and attributed-fused *Doc2Vec* during the optimization process, which can be sub-optimal to leverage attributes. In contrast, our *ASNE* seamlessly incorporates attributes by an early fusion on the input layer, which allows the following hidden layers to capture complex structure–attribute interactions and learn more informative node representations. Third, we observe that *ASNE* also consistently outperforms *AANE*, which also incorporates attribute proximity. However, *AANE* compresses node attributes into a fixed similarity matrix by simply calculating cosine similarity, and casts structure proximity as pair-wise regularizers to form a bi-convex optimization problem. In this way, fine-grained attribute interactions as well as high order structure proximity can hardly be captured.
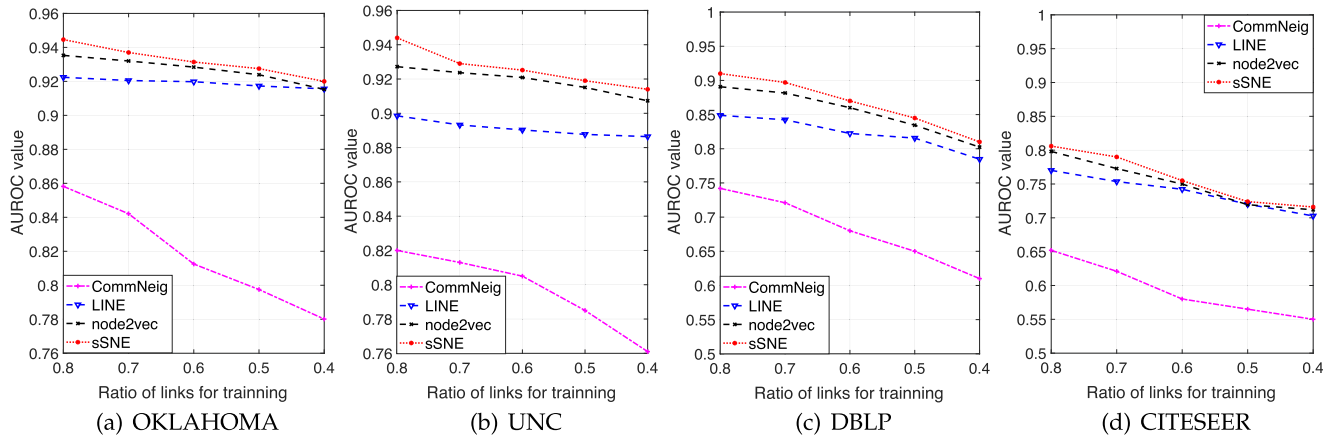
Fig. 6. Comparison of pure structure-based methods on link prediction task *w.r.t.* different structure proximity (RQ1).

3) To better show the comparison between those pure structure-based methods, we group them in Fig. 6. We observe that *node2vec* generally outperforms *LINE* across all the four datasets. This result is in consistent with Grover and Leskovec [5]'s finding. One plausible reason for *node2vec*'s superior performance might be that by performing random walks on the social network, higher-order proximity information can be captured. In contrast, *LINE* only models the first- and second-order proximities, which fails in capturing sufficient information for link prediction. To justify this, we have further explored an additional baseline that directly utilizes the second-order proximity by ranking nodes according to their common neighbors (denoted as *CommNeig*). As expected, the performance is weak for all datasets. Regarding the variant *sSNE*, since it shares the same walking procedure as *node2vec* in structure modeling, it is also capable of learning from higher-order proximities. Although without attribute information, we observe that *sSNE* manages to achieve marginal improvement over *node2vec*, which might be due to the fact that going deep enhances the model's generalization ability [41], [48].

5.2.2 Node Classification

Table 2 and Table 3 shows the *Macro-F1* and *Micro-F1* scores obtained by each method on the classification task. Upon getting the node representations, we train the classifier with different ratios of labeled data ($\rho \in \{10\%, 30\%, 50\%\}$). The performance trends are generally consistent with that of the link prediction task.

First and foremost, *ASNE* achieves the best performance among all the methods for all settings, and the one-sample paired t-test verifies that all improvements are statistically significant for $p < 0.05$. The performance of *ASNE* is followed by that of *TriDNR* and *AANE*, then followed by that of the attribute-based methods *node2vec+attr* and *LINE+attr*; *node2vec* and *LINE* which use only the network structure perform the worst. This further justifies the usefulness of attributes on social networks, and such that properly modeling them can lead to better representation learning and benefit downstream applications. Among the four attribute-based methods, *ASNE*, *AANE* and *TriDNR* demonstrate superior performance over *node2vec+attr* and *LINE+attr*, which points to the positive effects of incorporating attributes into the network embedding process rather than late fusion. Moreover, *ASNE* achieves superior performance than *AANE* and *TriDNR*. This might due to that *ASNE* fuses structure and attributes in early stage, allows close interaction between these two source of information. However, *AANE* only precalculates attribute similarity matrix of nodes and factorizes this matrix with first-order link information as regularizer. *TriDNR* just separately trains structure model and attribute model, then linearly combines the learned embeddings in an iterative way.

It is worth pointing out that the ground-truth labels of the node classification task are not involved in the network embedding process. Despite this, *ASNE* can learn effective representations that support the task well. This is attributed to *ASNE*'s modeling of network structure and attributes in a

TABLE 2
Averaged *Macro-F1*, *Micro-F1* Scores for Node Classification Task on DBLP Dataset

Dataset		DBLP						
Method		*LINE*	*node2vec*	*LINE+attr*	*node2vec+attr*	*TriDNR*	*AANE*	*ASNE*
Macro-F1	10%	0.565	0.617	0.619	0.631	0.665	0.652	**0.699**[*]
	30%	0.586	0.632	0.636	0.642	0.702	0.691	**0.725**[*]
	50%	0.628	0.677	0.692	0.695	0.715	0.708	**0.761**[*]
Micro-F1	10%	0.587	0.647	0.661	0.686	0.750	0.722	**0.763**[*]
	30%	0.632	0.665	0.678	0.749	0.778	0.758	**0.786**[*]
	50%	0.678	0.733	0.732	0.753	0.785	0.769	**0.804**[*]

[*] *denotes the statistical significance for* $p < 0.05$. *(RQ1)*

TABLE 3
Averaged *Macro-F1*, *Micro-F1* Scores for Node Classification Task on CITESEER Dataset

Dataset		CITESEER						
Method		*LINE*	*node2vec*	*LINE+attr*	*node2vec+attr*	*TriDNR*	*AANE*	*ASNE*
Macro-F1	*10%*	0.548	0.606	0.597	0.613	0.618	0.624	**0.653**[*]
	30%	0.580	0.625	0.631	0.630	0.692	0.678	**0.715**[*]
	50%	0.619	0.667	0.670	0.682	0.736	0.728	**0.752**[*]
Micro-F1	*10%*	0.573	0.623	0.607	0.628	0.644	0.665	**0.675**[*]
	30%	0.614	0.653	0.667	0.695	0.714	0.705	**0.732**[*]
	50%	0.661	0.695	0.691	0.717	0.756	0.749	**0.767**[*]

[*] denotes the statistical significance for $p < 0.05$. (RQ1)

sound way, which leads to comprehensive and informative representations for nodes.

5.2.3 Impact of λ

We further explore the impact of λ which adjusts the importance of attributes. Both the link prediction task and the node classification task are evaluated under the same evaluation protocols as Section 5.1.2. For a clear comparison, we plot the results in Figs. 7 and 8. The link prediction results are reported under training on 80 percent of links. The node classification results are obtained from training on 50 percent of labeled nodes.

Due to the fact that λ actually can be set to any real number under our learning framework, we first broadly explore the impact of λ on the range $[0, 0.01, 0.1, 1, 10, 100]$. Setting λ to 0 returns the pure structure modeling, while setting it to a large number approximates the pure attribute modeling. We found that good results are generally obtained within $[0, 1]$ across datasets. When λ becomes relatively large and the attribute part overweighs the structure part, the performance even becomes worse than pure structure modeling. Therefore, we focus our exploration on the range $[0, 1]$ at an interval of 0.2.

Generally, attributes play an important role in *ASNE* as evidenced by the improving performance when λ increases. We observe similar trends for both the link prediction and node classification across datasets. If we ignore the attribute information by setting $\lambda = 0.0$, *ASNE* degrades to pure structure modeling. Its corresponding performance is the worst for both tasks, as compared to the attributes included counterparts. Moreover, the performance improvements on DBLP and CITESEER are relatively larger. Specifically, we observe a dramatic improvement of performance on

CITESEER when λ increases from 0.0 to 0.2. As there is less link information in these two datasets as shown in Table 1, the performance improvement indicates that attributes help to alleviate the link sparsity problem.

In addition, we observe that the pure structure model ($\lambda = 0.0$) outperforms *node2vec* if we further compare the results with Fig. 5 for link prediction and Table 2 for node classification. Since the same p, q setting as *node2vec* are leveraged, we attribute the performance improvements to the non-linearity introduced by the hidden layers.

5.3 Qualitative Analysis (RQ2)

5.3.1 Case Study

To understand why *ASNE* can achieve better results than the other methods, we carry out a case study on the DBLP dataset in this section. Given the node representations learned by each method, we retrieve the three most similar papers *w.r.t.* a given query paper. Specifically, we measure the similarity using the cosine distance. For a fair comparison with the structure-based methods, the query paper we choose is a well-cited paper of KDD 2006 named "*Group formation in large social networks: membership, growth, and evolution*". According to Google Scholar by 15/1/2017, its citation number reaches 1,510. Based on the content of this query paper, we expect that relevant results should be about the structure evolution of groups or communities in social networks. The top results retrieved by different methods are shown in Table 4.

First of all, we see that *ASNE* returns rather relevant results: all the three papers are about dynamic social network analysis and community structures. For example, the first one considers the evolution of structures such as communities in large online social networks. The second result can be viewed as a follow-up work of the query, focusing on discovering temporal communities. While for *TriDNR*, the top result aims to measure social influence between
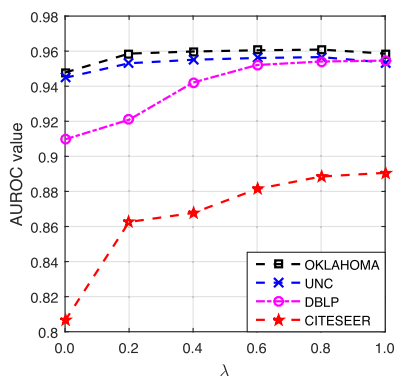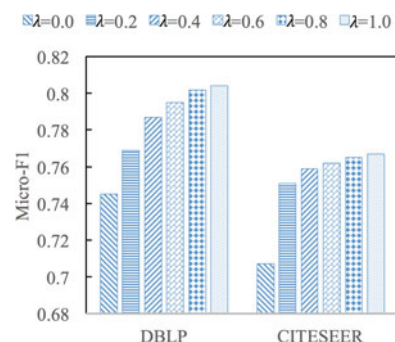

Fig. 7. Link prediction.

Fig. 8. Node classification.

TABLE 4
Top Three Results Returned by Each Method (RQ2)

Query: *Group formation in large social networks: membership, growth, and evolution*[6]

ASNE
1. Structure and evolution of online social networks
2. Discovering temporal communities from social network documents
3. Dynamic social network analysis using latent space models

TriDNR
1. Influence and correlation in social networks
2. A framework for analysis of dynamic social networks
3. A framework for community identification in dynamic social networks

AANE
1. Structure and evolution of online social networks.
2. Social influence analysis in large-scale networks.
3. Mining interesting link formation rules in social networks

node2vec
1. Latent Dirichlet Allocation
2. Maximizing the spread of influence through a social network
3. Mining the network value of customers

LINE
1. Graphs over time: densification laws, shrinking diameters and possible explanations
2. Maximizing the spread of influence through a social network
3. Relational learning via latent social dimensions

linked individuals but community structures are not of concern. In *AANE*, it tends to retieve papers with many overlapping words. This might be because the method heavily relies on the node similarity matrix caculated by cross product of bag-of-words attribute representation.

Regarding methods that only leverage structure information, the results returned by *node2vec* are less similar to the query paper. It seems that *node2vec* tends to find less related but highly cited papers. According to Google Scholar by 15/1/2017, the citation numbers for the first, second and third results are 16908, 4099 and 1815, respectively. This is because the random walk procedure can be easily biased towards the popular nodes that have more links. While *ASNE* also relies on the walking sequences, it can correct such bias to a certain extent by leveraging attributes.

Similarly, *LINE* also retrieves less relevant papers. Although the first and second results are related to dynamic social network analysis, all the three results are not concerned with group or community. It might due to the limitations of only modeling first- and second-order proximities while leaving out the abundant attributes.

Based on the above qualitative analysis, we draw the conclusion that using both network structure and attributes benefits the retrieval of similar nodes. Compared to the pure structure-based methods, the top returned results of *ASNE* are more relevant to the query paper. It is worth noting that for this qualitative study, we have purposefully chosen a popular node to migrate the sparsity issue, which actually favors the structure-based methods; even so, the structure-based methods fail at identifying relevant results. This sheds light on the limitation of solely relying on the network structure for social network embedding, and thus the importance of modeling the rich evidence sources in attributes.

5.3.2 Visualization

To better demonstrate the quality of learned representations, we visualize the DBLP citation network in two dimensional space. In this dataset, each node (or paper) may cite or be cited by other nodes (or papers), which naturally forms a citation network. We carry out experiment on the CITESEER dataset, which is a subset [59] of CiteSeerX data 2 which consist of scientific publications from 10 distinct research areas: agriculture, archaeology, biology, computer science, financial economics, industrial engineering, material science, petroleum chemistry, physics, and social science. This dataset consists of multidisciplinary 10 classes, 10,310 publications with textual information and 77,218 edges in total.

Laying out this citation network is very challenging as the ten research fields are interconnected to each other. Following [6], we first map the citation network into a low-dimensional space with different embedding approaches. We then map the low dimensional vectors of the vertices to a 2-D space with the t-SNE package [60].

Fig. 9 compares the visualization results with different embedding approaches. The visualization result of *LINE* is not well separated, in which the papers belonging to different research areas are heavily overlapped. This might due to the insufficient link information and *LINE* only captures first and second order proximities. The results of *node2vec* are better, in which a large group is formed representing the research of Petroleum Chemistry. However, limited link information still affects the performance. By adding extra attribute information, we observe that *node2vec+attr* separates different classes better. The *TriDNR*, *AANE* and *ASNE* perform well and generate meaningful layout of the network (nodes with same colors are distributed closer), which again demonstrates the usefulness of incorporating attributes. However, as discussed in Section 5.2.1, finding an effective way to incorporate attributes plays a pivotal role for obtaining meaningful representations. The clusters for *ASNE* method is more clear than those of other methods. In (f) *ASNE*, we observe that the small archaeology cluster are separated out nicely, which is largely overlapped with other clusters in other methods. The clear visualization of *ASNE* validates our design of fusing structure proximity and attribute proximity in early stage.

5.3.3 Discussion

Superior performances on link prediction as well as node classification have been achieved, which demonstrates the effectiveness of incorporating attributes as in our model. At the same time, it might be necessary to point out here that the four datasets we used show strong homophily effects, which motivates our design of modeling attribute proximity. As one of the most striking and robust regularities of social life, homophily exists in a wide variety of social networks, triggering people similar to each other to become friends. However, there might still be some networks where homophily plays a less important role. For example, in a signed network, negative links usually denote distrust or foe relationship between two nodes. In such cases, focusing on attribute similarity might lead to unsatisfactory results. Moreover, regarding some specific tasks on certain types of social networks, some kinds of attribute information may not contain useful signals, or even introduce noises. For instance, regarding the question recommendation task on Quora, attribute such as which restaurant a user went
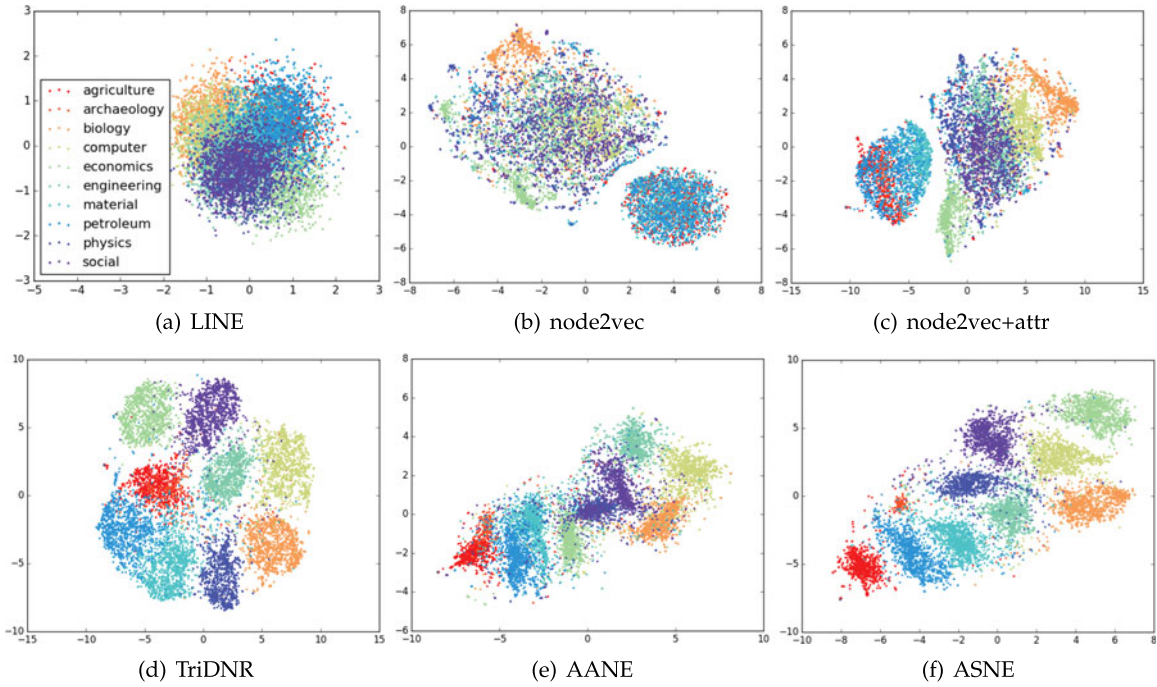
Fig. 9. Visualization of the CITESEER citation network. The nodes are mapped to the 2-D space using the t-SNE package with learned embeddings as input. Color of a node indicates the class of the paper.

yesterday or whether the user has a brother might not be informative for finding an answer. These are important things that need to be considered when applying our method, as well as some other works based on homophily.

5.4 Experiments with Hidden Layers (RQ3)

In this final section, we explore the impact of hidden layers on *ASNE*. It is known that increasing the depth of a neural network can increase the generalization ability for some models [41], [48], however, it may also degrade the performance due to optimization difficulties [61]. It is thus curious to see whether using deeper layers can empirically benefit the learning of *ASNE*.

Table 5 shows *ASNE*'s performance of the link prediction and node classification tasks *w.r.t.* different number of hidden layers on the DBLP dataset. The results on other datasets are generally similar, thus we just showcase one here. As the size of the last hidden layer determines an *ASNE* model's representation ability, we set it to the same number for all models to ensure a fair comparison. Note that for each setting (row), we have re-tuned the hyper-parameters to fully exploit the model's performance.

First, we can see the trend that with more hidden layers, the performance is improved. This indicates the positive effect of using a deeper architecture for *ASNE*, which indeed increases its generalization ability and boosts its performance. The trade-off, however, is the server CPU time

TABLE 5
Performance of Link Prediction and Node Classification on DBLP *w.r.t.* Different Number of Hidden Layers (RQ3)

Hidden layers	*AUROC*	*micro-F1*
No Hidden Layers	0.9273	0.791
128*Softsign*	0.9418	0.799
256*Softsign* → 128*Softsign*	0.9546	**0.804**
512*Softsign* → 256*Softsign* → 128*Softsign*	**0.9589**	0.802

needed for the training procedure. Specifically, on our modest commodity server (Intel Xeon CPU of 2.40 GHz), a one-layer *ASNE* takes 25.6 seconds, while a three-layer *ASNE* takes 81.9 seconds for one epoch. We stopped exploring deeper models, as the current *ASNE* uses fully connected layers, which become difficult to optimize and can be easily over-fitting and degrading with more layers [61]. The diminishing improvement of results in Table 5 also implies the potential problem. To address it, modern neural network designs shall be applied, such as the residual units and highway networks [48]. We leave this possibility for future work.

It is worth noting that when there is no hidden layer, *ASNE*'s performance is relatively weak, which is in the same level as *TriDNR*. With one more layer, the performance is significantly improved. It demonstrates the usefulness of learning structure–attribute interactions in a nonlinear way. To justify this, we have further tried to replace the softsign activation function with the identity function, i.e., using a linear function above the concatenation of structure and attribute embedding vectors. It performs worse than using the non-linear softsign function, e.g., we observe a 2.1 percent performance drop for the link prediction task.

6   CONCLUSION

To learn informative representations for social network data, it is crucial to account for both network structure and attribute information. To this end, we proposed a generic framework for embedding social networks by capturing both the *structural proximity* and *attribute proximity*. We adopted a deep neural network architecture to model the complex interrelations between structural information and attributes. Extensive experiments show that *ASNE* can learn informative representations for social networks and achieve superior performance on the tasks of link prediction and node classification comparing to other representation learning methods.

This work has tackled representation learning on social networks by leveraging both structural and attribute information. While social networks are rich sources of information containing more than links and textual attributes, we will study the following directions in future. First, we will enhance our ASNE framework by fusing data from multiple modalities. It is reported that over 45 percent tweets contain images in Weibo [62], making it urgent and meaningful to perform network embedding with multi-modal data. Second, we will develop (semi-)supervised variant for ASNE, so as to learning task-oriented embeddings to tailor for a specific task [63]. Third, we are interested in exploring how to capture the evolution nature of social networks, such as new users and social relations by using temporal-aware recurrent neural networks. Lastly, we will consider improving the efficiency of ASNE by learning to hash techniques [64] to make it suitable for large-scale industrial use.

ACKNOWLEDGMENTS

## REFERENCES

[1] X. Wang, L. Nie, X. Song, D. Zhang, and T.-S. Chua, "Unifying virtual and physical worlds: Learning toward local and global consistency," *ACM Trans. Inf. Syst.*, vol. 36, no. 1, 2017, Art. no. 4.
[2] X. He, M. Gao, M.-Y. Kan, and D. Wang, "BiRank: Towards ranking on bipartite graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 57–71, Jan. 2017.
[3] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 701–710.
[4] S. Chang, W. Han, J. Tang, G.-J. Qi, C. C. Aggarwal, and T. S. Huang, "Heterogeneous network embedding via deep architectures," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 119–128.
[5] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 855–864.
[6] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding," in *Proc. Int. Conf. World Wide Web*, 2015, pp. 1067–1077.
[7] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on Twitter," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2011, pp. 1301–1309.
[8] M. Pennacchiotti and A.-M. Popescu, "Democrats, republicans and starbucks afficionados: User classification in Twitter," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 430–438.
[9] A. L. Traud, P. J. Mucha, and M. A. Porter, "Social structure of Facebook networks," *Physica A: Statistical Mech. Appl.*, vol. 391, pp. 4165–4180, 2012.
[10] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1225–1234.
[11] G. Robins, "Exponential random graph models for social networks," *Encyclopaedia of Complexity and System Science*. Berlin, Germany: Springer, 2011.
[12] P. F. Lazarsfeld, R. K. Merton, et al., "Friendship as a social process: A substantive and methodological analysis," *Freedom Control Modern Soc.*, vol. 18, pp. 18–66, 1954.
[13] E. O. Laumann, *Prestige and Association in an Urban Community: An Analysis of an Urban Stratification System*. Indianapolis, IN, USA: Bobbs-Merrill, 1966.
[14] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annu. Rev. Sociology*, vol. 27, pp. 415–444, 2001.
[15] S. B. Kurth, "Friendships and friendly relations," *Friendship As a Social Institution*, pp. 136–170, 1970.
[16] J. M. McPherson and L. Smith-Lovin, "Homophily in voluntary organizations: Status distance and the composition of face-to-face groups," *Amer. Sociol. Rev.*, vol. 52, pp. 370–379, 1987.
[17] A. T. Fiore and J. S. Donath, "Homophily in online dating: When do you like someone like yourself?" in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2005, pp. 1371–1374.
[18] G. Kossinets and D. J. Watts, "Origins of homophily in an evolving social network1," *Amer. J. Sociology*, vol. 115, pp. 405–450, 2009.
[19] J. B. Kruskal and M. Wish, *Multidimensional Scaling*. Newbury Park, CA, USA: Sage, 1978.
[20] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Sci.*, vol. 290, pp. 2319–2323, 2000.
[21] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Sci.*, vol. 290, pp. 2323–2326, 2000.
[22] K. Nowicki and T. A. B. Snijders, "Estimation and prediction for stochastic blockstructures," *J. Amer. Statistical Assoc.*, vol. 96, pp. 1077–1087, 2001.
[23] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels," *J. Mach. Learn. Res.*, vol. 9, pp. 1981–2014, 2008.
[24] K. Palla, D. A. Knowles, and Z. Ghahramani, "Relational learning and network modelling using infinite latent attribute models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 2, pp. 462–474, Feb. 2015.
[25] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 542–550.
[26] R. Nallapati and W. W. Cohen, "Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs," in *ICWSM*, 2008, pp. 84–92.
[27] J. Chang and D. M. Blei, "Relational topic models for document networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2009, pp. 81–88.
[28] Y.-S. Cho, G. Ver Steeg, E. Ferrara, and A. Galstyan, "Latent space model for multi-modal social data," in *Proc. Int. Conf. World Wide Web*, 2016, pp. 447–458.
[29] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," *Knowl. Inf. Syst.*, vol. 42, pp. 181–213, 2015.
[30] C. Yang, Z. Liu, D. Zhao, M. Sun, and E. Y. Chang, "Network representation learning with rich text information," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 2111–2117.
[31] S. Pan, J. Wu, X. Zhu, C. Zhang, and Y. Wang, "Tri-party deep network representation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 1895–1901.
[32] Q. Mei, D. Cai, D. Zhang, and C. Zhai, "Topic modeling with network regularization," in *Proc. Int. Conf. World Wide Web*, 2008, pp. 101–110.
[33] X. Huang, J. Li, and X. Hu, "Accelerated attributed network embedding," in *Proc. SIAM Int. Conf. Data Mining*, 2017, pp. 633–641.
[34] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.
[35] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, et al., "Wide & deep learning for recommender systems," in *Proc. Workshop Deep Learn. Recommender Syst.*, 2016, pp. 7–10.
[36] J. Weston, F. Ratle, H. Mobahi, and R. Collobert, "Deep learning via semi-supervised embedding," in *Neural Networks: Tricks of the Trade*. Berlin, Germany: Springer, 2012, pp. 639–655.
[37] Z. Yang, W. Cohen, and R. Salakhudinov, "Revisiting semi-supervised learning with graph embeddings," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 40–48.
[38] S. Cao, W. Lu, and Q. Xu, "GraRep: Learning graph representations with global structural information," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 891–900.
[39] X. He, H. Zhang, M.-Y. Kan, and T.-S. Chua, "Fast matrix factorization for online recommendation with implicit feedback," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2016, pp. 549–558.
[40] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[41] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T. Chua, "Neural collaborative filtering," in *Proc. Int. Conf. World Wide Web*, 2017, pp. 173–182.

[42] D. Luo, F. Nie, H. Huang, and C. H. Ding, "Cauchy graph embedding," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 553–560.

[43] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2014, pp. 1532–1543.

[44] O. Levy, Y. Goldberg, and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings," *Trans. Assoc. Comput. Linguistics*, vol. 3, pp. 211–225, 2015.

[45] S. Rendle, "Factorization machines," in *Proc. IEEE Int. Conf. Data Mining*, 2010, pp. 995–1000.

[46] Y. Shan, T. R. Hoens, J. Jiao, H. Wang, D. Yu, and J. Mao, "Deep crossing: Web-scale modeling without manually crafted combinatorial features," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 255–262.

[47] X. He and T.-S. Chua, "Neural factorization machines for sparse predictive analytics," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2017, pp. 355–364.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[49] S. Wang, Y. Wang, J. Tang, K. Shu, S. Ranganath, and H. Liu, "What your images reveal: Exploiting visual contents for point-of-interest recommendation," in *Proc. Int. Conf. World Wide Web*, 2017, pp. 391–400.

[50] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[51] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, 2011.

[52] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop, coursera: Neural networks for machine learning," 2012.

[53] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[54] Y. Koren, "Factorization meets the neighborhood: A multifaceted collaborative filtering model," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 426–434.

[55] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 1994.

[56] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner: Extraction and mining of academic social networks," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 990–998.

[57] K. H. Zou, A. J. OMalley, and L. Mauri, "Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models," *Circulation*, vol. 115, pp. 654–657, 2007.

[58] L. Liao, Q. Ho, J. Jiang, and E.-P. Lim, "SLR: A scalable latent role model for attribute completion and tie prediction in social networks," in *Proc. IEEE Int. Conf. Data Eng.*, 2016, pp. 1062–1073.

[59] K. W. Lim and W. Buntine, "Bibliographic analysis with the citation network topic model," in *Proc. Asian Conf. Mach. Learn.*, 2015, pp. 142–158.

[60] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[61] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[62] T. Chen, X. He, and M.-Y. Kan, "Context-aware image tweet modelling and recommendation," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 1018–1027.

[63] L. Liao, X. He, Z. Ren, L. Nie, H. Xu, and T.-S. Chua, "Representativeness-aware aspect analysis for brand monitoring in social media," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 310–316.

[64] H. Zhang, F. Shen, W. Liu, X. He, H. Luan, and T.-S. Chua, "Discrete collaborative filtering," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2016, pp. 325–334.

Lizi Liao is currently working toward the PhD degree with the NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore. Her research interests include social network analysis and multimedia information retrieval. Her works have appeared in several major international conferences such as WWW, ICDE, IJCAI, and AAAI. She has served as a reviewer of international conferences such as SIGIR and WSDM.
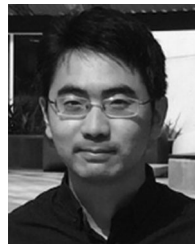
Xiangnan He received the PhD degree in computer science from the National University of Singapore (NUS). He is a senior research fellow with the School of Computing, National University of Singapore. His research interests span recommender systems, information retrieval, and multimedia processing. He has more than 30 publications appeared in several top conferences such as SIGIR, WWW, MM, CIKM, and IJCAI, and journals including the *IEEE Transactions on Knowledge and Data Engineering*, the *ACM Transactions on Information Systems*, and the *Trends in Molecular Medicine*. His work on recommender systems has received the Best Paper Award Honorable Mention of ACM SIGIR 2016. Moreover, he has served as the PC member for the prestigious conferences including SIGIR, WWW, MM, KDD, and ACL, and the regular reviewer for prestigious journals including the *IEEE Transactions on Knowledge and Data Engineering*, the *ACM Transactions on Information Systems*, the *ACM Transactions on Knowledge Discovery from Data*, the *Trends in Molecular Medicine*, etc.

Hanwang Zhang received the BEng (Hons.) degree in computer science from Zhejiang University, Hangzhou, China, in 2009, and the PhD degree in computer science from the National University of Singapore, in 2014. He is an assistant professor with Nanyang Technological University, Singapore. He was a research scientist with the Department of Computer Science, Columbia University, and a senior research fellow with the School of Computing, National University of Singapore, Singapore. His research interest includes computer vision, multimedia, and social media. He is the recipient of the Best Demo Runner-up Award in ACM MM 2012, the Best Student Paper Award in ACM MM 2013, and the Best Paper Honorable Mention in ACM SIGIR 2016. He is also the winner of Best PhD Thesis Award of the School of Computing, National University of Singapore, 2014.

Tat-Seng Chua received the PhD degree from the University of Leeds, United Kingdom. He is the KITHCT chair professor with the School of Computing, National University of Singapore. He was the acting and founding dean of the school from 1998-2000. His main research interest include multimedia information retrieval and social media analytics. In particular, his research focuses on the extraction, retrieval and question-answering (QA) of text and rich media arising from the Web and multiple social networks. He is the co-director of NExT, a joint center between NUS and Tsinghua University to develop technologies for live social media search. He is the 2015 winner of the prestigious ACM SIGMM award for Outstanding Technical Contributions to Multimedia Computing, Communications, and Applications. He is the chair of steering committee of the ACM International Conference on Multimedia Retrieval (ICMR) and Multimedia Modeling (MMM) conference series. He also was the general co-chair of ACM Multimedia 2005, ACM CIVR (now ACM ICMR) 2005, ACM SIGIR 2008, and ACM Web Science 2015. He serves on the editorial boards of four international journals. He is the co-founder of two technology startup companies in Singapore.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.