RaRE: Social Rank Regulated Large-scale Network Embedding

Yupeng Gu
University of California, Los Angeles
Los Angeles, CA
ypgu@cs.ucla.edu

Yizhou Sun
University of California, Los Angeles
Los Angeles, CA
yzsun@cs.ucla.edu

Yanen Li
Snapchat Inc.
Venice, CA
yanen.li@snap.com

Yang Yang
Zhejiang University
Hangzhou, China
yangya@zju.edu.cn

ABSTRACT

Network embedding algorithms that map nodes in a network into a low-dimensional vector space are prevalent in recent years, due to their superior performance in many network-based tasks, such as clustering, classification, and link prediction. The main assumption of existing algorithms is that the learned latent representation for nodes should preserve the structure of the network, in terms of first-order or higher-order connectivity. In other words, nodes that are more similar will have higher probability to connect to each other. This phenomena is typically explained as homophily in network science. However, there is another factor usually neglected by the existing embedding algorithms, which is the popularity of a node. For example, celebrities in a social network usually receive numerous followers, which cannot be fully explained by the similarity of the two users. We denote this factor with the terminology "social rank". We then propose a network embedding model that considers both of the two factors in link generation, and learn proximity-based embedding and social rank-based embedding separately. Rather than simply treating these two factors independent with each other, a carefully designed link generation model is proposed, which explicitly models the interdependency between these two types of embeddings. Experiments on several real-world datasets across different domains demonstrate the superiority of our novel network embedding model over the state-of-the-art methods.

KEYWORDS

Network embedding, social rank, representation learning

ACM Reference Format:
Yupeng Gu, Yizhou Sun, Yanen Li, and Yang Yang. 2018. RaRE: Social Rank Regulated Large-scale Network Embedding. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3178876.3186102

1 INTRODUCTION

Studying the latent representation of nodes in information networks has been a prevalent topic recently. Latent representations, also known as latent features or embeddings, often reside in a lower dimensional continuous vector space, and are especially helpful in terms of understanding the nodes. A wide variety of applications can be achieved as a result, including classification, visualization, community detection and so on.

Early methods include mapping nodes onto a lower dimensional manifold by finding the intrinsic dimensionality using spectral methods on graph adjacency matrix, such as locally linear embedding (LLE) [35], Isomap [41], multidimensional scaling [23] and so on. Most of these methods do not scale for large networks. Later on more principled statistical models have been developed where node parameters are deduced by optimizing some global objective function. The intuition is rather straightforward: similarity in the graph should be preserved in the lower dimensional space. Proximity in the graph is usually embodied as neighbors: according to the homophily assumption [29], entities that are connected in the graph represent some sort of similarity. Neighbors are also essential in random walk based methods [16, 34], where information and label propagate. As a result, entities that are connected in the original graph are often adjacent to each other in the latent space.

However, this seemingly plausible approach has some severe drawbacks and naturally triggers several open questions. First of all, *is it consistently true that all links occur between similar nodes*? As per the preferential attachment process in network generation [4], nodes are believed to have a higher chance to connect to (e.g. follow) high-degree nodes (e.g. celebrities) in general. For example, a new Twitter user may first choose to follow some well-known politicians or movie stars, regardless of being a fan or not. Those links are generated due to the high exposure and popularity of certain accounts, rather than their similar tastes (i.e. proximity of latent representation). For instance, famous politicians usually have a fair amount of followers on social networks, but certainly not 100% of the followers should be considered as similar to them in terms of political opinions. On the other hand, everyone has limited amount of energy and resources, which prevents many actually similar pairs from being present. During a literature review, famous groups of scholars in the corresponding field of study are usually considered first, while a vast majority of junior groups may be ignored, even though they are working on very similar topics. With all being said, node features inferred according to the homophily assumption will be a mixture of popularity and proximity factors, and thus are not desirable for clustering or classification tasks. In this work, we use a specific terminology "social rank" to denote the popularity factor, namely the position where an entity is ranked

WWW 2018, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5639-8/18/04..
https://doi.org/10.1145/3178876.3186102

among the network, and "proximity-based representation" to denote the general embedding vector which denotes the opinions and preferences of an entity.

The second question is, *should these two factors be considered as totally independent with each other*? From the case studies in the above question, we notice that how much proximity will contribute to link formation depends on the relative social status/rank between a pair of nodes. On one hand, when a link from a node to a more popular one is observed, it is somewhat likely to be explained by the popularity of the latter. On the other hand, when a link from a node to a less popular one is present, proximity factor should account for most of the intentions. For example, in a bibliography citation network, it is often more common for papers to cite very famous papers due to their substantial public attention. However, when a less popular paper is cited, it is almost certainly the case that it is essentially very relevant to the work. In sum, homophily itself does not suffice to explain the reason behind link generation, and we should decide *the extent* to which homophily is trusted considering the social rank of nodes. A principled methodology is desired to balance the effect of social rank and proximity factor in terms of network generation.

In this paper we propose a Social Rank Regulated Network Embedding (RaRE) model which incorporates both latent *rank factor* and latent *proximity-based factor* to interpret the network generation process. Our unified Bayesian framework models the probabilistic relationship between social rank, proximity-based representation and existence of a link. We discuss what portion is truly justified by the proximity between nodes given their social rank difference, which explains the generation of a link from a brand new perspective. Our method is also scalable to large networks. The contribution of our method can be summarized as follows:

- We propose to solve the network embedding problem from a novel Bayesian perspective, which integrates both social rank and proximity-based embedding.
- A brand new probabilistic link formation model is formulated that explicitly models the extent of contribution of proximity-based embedding under different relative social rank difference.
- Our method is easily scalable to real-world large-scale networks which consist of millions of nodes.

2 PROBLEM DEFINITION

2.1 Background

Extracting latent representation of nodes (also known as embedding) in an information network is essential in understanding the relative position of each node in the network. A natural embedding is the row vector in the adjacency matrix where each dimension denotes the link status between a pair of nodes. Nevertheless, this plain strategy is seldom applicable to real-world tasks due to the computational complexity brought by its high dimensional representation, as well as its low representation power in terms of preserving network structure. In order to tackle this problem, traditional approaches extract dimensions with the biggest contribution to the data (e.g. PCA, SVD, IsoMap), or find lower dimensional representation of nodes by factorizing the adjacency matrix [1, 28, 30, 42].

More recent approaches introduce the notion of "node embedding", a low-dimensional vector representation of node, which embodies the latent merits and characteristics of an individual. The concept of embedding is very similar to word embeddings [31], where every word is represented by a low dimensional vector. These embedding vectors are learned by preserving similarity in the corpus (i.e. between every word and its context) and similarity in the latent space (i.e. vector dot product). Levy and Goldberg [24] also reveal the connection between matrix factorization and word embedding, arguing that estimating word embedding is equivalent to factorizing a pointwise mutual information matrix. In the realm of information networks, the concept of "context" no longer exists, and many researchers have proposed ways to define similar nodes in the network, such as n-hop neighbors or the nodes reachable from a random walk [13, 16, 34, 40, 43]. Besides, the generation process also allows great flexibility in the modeling part, and different link generation approaches have been proposed [13, 14, 43]. In sum, the lower dimensional representations are much more succinct while keeping the majority of information in the graph.

However, none of the network embedding approaches explicitly interpret the meaning of the representation, and assume the link is generated based on the proximity of representations, or that proximity can propagate through links. We argue that, there is another essential factor ("social rank") other than proximity that leads to the formation of a link, which is not homogeneous for connected nodes and should be detached from the general proximity-based embedding. Some matrix factorization methods model popularity in terms of bias (e.g. [6, 20, 22]), but they simply treat proximity and popularity as two independent factors. Instead, we model the interdependency of the two factors in link formation explicitly. We will explain the necessity of modeling their interdependency in Section 3 and define our problem formally in the next paragraph.

2.2 Problem Definition

We define our embedding problem as follows. An information network can be formatted as $G = (V, E)$, where $V = \{u_n\}_{n=1}^{N}$ is the set of vertices and $E \subset V^2$ is the set of edges. We use e_{ij} to denote the binary status of the link from u_i to u_j for unweighted networks, or the multiplicity of the link for weighted networks. Our goal is to infer both latent *proximity-based* representation $\{z_v | v \in V\} \subset \mathbb{R}^K$ and latent *rank* representation $\{r_v | v \in V\} \subset \mathbb{R}^+$ for nodes in the network. Similar to the ordinal numbers, our ranking assumes a higher social rank for a smaller value of r, and requires that all social ranks are positive. The rank r can also be considered as radius of a node in the visualization, where center nodes (smaller radius) are most influential.

3 APPROACH

In this section, we illustrate our approach by introducing a general form of network generation prototype for unweighted networks, investigate the details, instantiate our final model using mathematical derivations, and extend it to general networks. We will discuss its scalability and relationship to existing models as well.

3.1 Base Model

The most general model for the binary status of a link e_{ij} (1 if present, 0 otherwise) is a Bernoulli event with parameter

$$p(e_{ij} = 1 | r_i, r_j, z_i, z_j) = f(r_i, r_j, z_i, z_j) \tag{1}$$

where $f()$ is a probability function to be designed. First, we would like $f()$ to encode the difference of the social ranks, and the most natural measure is the difference $r_i - r_j$. Besides, a recent work [2] has also found that the probability of a friendship is a function of the difference of two users' social ranks (independent of their absolute values), meaning that the function should be translation-invariant. We adopt their assumption and thus always study the difference of two entities' ranks (denoted by $dr = r_i - r_j$) in function $f()$. For the proximity-based representation, the homophily assumption has been widely used and accepted in related studies [18, 29], which posits that in information networks, most interactions occur between nodes with similar merits and characteristics. Therefore, with the hope that the representation z is a reflection of people's hidden characteristics, we design the probability to be a function of their Euclidean distance $dz = ||z_i - z_j||_2$ in the lower dimensional space. In sum, Equation 1 becomes:

$$p(e_{ij} = 1 | r_i, r_j, z_i, z_j) = f(dr, dz) \tag{2}$$

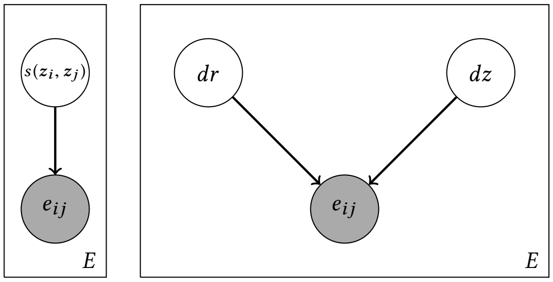


Figure 1: Left: traditional embedding models (e.g. [30, 40]), where $s(\cdot, \cdot)$ measures vector similarity. Right: graphical model representation of our model. Shadowed unit e_{ij} represents the observed variable (i.e. the status of the link).

A graphical model representation is illustrated in Figure 1. It would be controversial about the exact form of $f()$ at this stage, therefore in order to have a more convincing formulation, we will investigate the distribution of parameters under different circumstances (i.e. when the link is present/absent) and derive $f()$ from a Bayesian perspective.

To reach the concrete form of the probability function, we define the conditional distribution $p(dr, dz | e_{ij}) = p(dr | dz, e_{ij}) \cdot p(dz | e_{ij})$ in Section 3.2 and the prior distribution $p(r)$ and $p(z)$ in Section 3.3. Finally, the exact form of $f()$ is determined by Bayes' rule in Section 3.4.

3.2 Conditional Distributions

When the link is present, it is generally due to two reasons: (1) the link-receiver is famous (or at least more famous than the link-sender); or (2) the two individuals are similar (homophily). In other words, a majority of the links occur between pairs of nodes (u_i, u_j) where $dr = r_i - r_j$ is positive or $dz = ||z_i - z_j||_2$ is small. From the graphical model in Figure 1 (right), dr and dz are no longer independent when we have knowledge about the link e_{ij}, which is referred to as "explaining away" in Bayesian networks. In other words, $p(dr, dz | e_{ij}) \neq p(dr | e_{ij}) \cdot p(dz | e_{ij})$ because they are conditionally dependent on e_{ij}. In particular, given the presence of a link, dissimilarity of two users (i.e. large dz) will increase our belief that j is more popular than i (i.e. positive dr). For example, sportsmen followed by a non-sporty person are likely to be very influential. On the other hand, proximity of z (i.e. small dz) will eliminate some possibility that j is popular (i.e. positive dr), thus shifting the mean of dr towards left. For example, followees of a sportholic might as well be some unspectacular players in his/her home team. Therefore, we assume the distribution of dr is a Gaussian with a positive mean, and dz follows a (truncated) Gaussian distribution where the peak is at $dz = 0$:

$$p(dr | dz, e_{ij} = 1) = \mathcal{N}(\mu \cdot h(dz), \sigma_R^2) \tag{3}$$

$$p(dz | e_{ij} = 1) = \frac{1}{\mathcal{Z}} \cdot I_{\mathbb{R}^+}(dz) \cdot \mathcal{N}(0, \sigma_1^2) \tag{4}$$

where $\mu \cdot h(dz)$ ($\mu > 0$) is the mean of dr conditioned on dz and $e_{ij} = 1$, $\mathcal{Z}$ is the normalization term ($\mathcal{Z} = 1/2$ obviously) and $I_S(x)$ is the indicator function (takes value 1 if $x \in S$ and 0 otherwise). The hyper-parameter μ controls the scale of the mean, while $h(dz)$ adjusts the mean with respect to different values of dz.

Ideally, dr and dz in the above equations should follow our previous intuition exactly. Therefore, function $h()$ should obey the following properties:

- Non-negative, bounded and supported on $\mathbb{R}^+$. In general, most of existing links are likely to occur from a node to a more influential one (i.e., $dr = r_i - r_j \geq 0$). In other words, no matter what value dz takes, the mean of the rank difference (given the link is present) $\mathbb{E}[dr] = \mu \cdot h(dz)$ should always be non-negative.
- Non-decreasing monotonic. Given the existence of a link, the relative rank difference between two nodes should be more significant as their dissimilarity increases. This corresponds to the "explaining away" effect above.
- Concave on some right-unbounded interval. From our intuition, as dz increases, the marginal gain of $\mathbb{E}[dr] = \mu \cdot h(dz)$ should be diminishing. In other words, h should be a concave function when dz exceeds some threshold.

Lots of families of functions have these properties, but a simple one of them we find to work well is $h(dz) = \frac{dz^2}{1+dz^2}$. An illustration of the conditional probability distributions of dr (given dz and $e_{ij} = 1$) and dz (given $e_{ij} = 1$) is shown in Figure 2 (left).

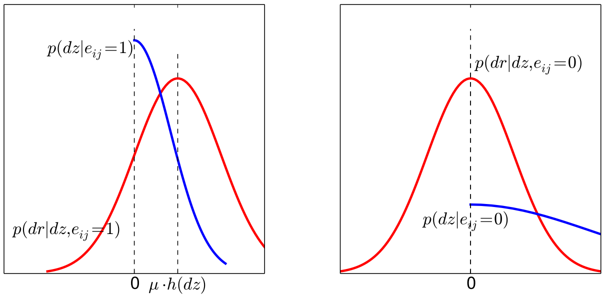
When the link is absent, it does not necessarily mean that the two individuals are dissimilar, or one of them is not popular enough to be observed. In other words, there are various reasons why no link is observed between two nodes. Therefore, without much confidence in claiming any special characteristic of an absent link, we assume a Gaussian distribution on dr and dz, both centered at 0. In other words, dr and dz will have equal chance of being positive and negative. Meanwhile, the variance of dz should be large as we expect the distribution to be more uniform.

$$p(dr|dz, e_{ij} = 0) = \mathcal{N}(0, \sigma_R^2) \tag{5}$$

$$p(dz|e_{ij} = 0) = \frac{1}{\mathcal{Z}} \cdot I_{\mathbb{R}^+}(dz) \cdot \mathcal{N}(0, \sigma_2^2) \tag{6}$$

where $\mathcal{Z} = 1/2$ is the normalization term.

Note that we assume $\sigma_2^2 > \sigma_1^2$ (Equation 4 and 6), as the probability distribution function of dz should look much more flat when the link is absent (illustrated in Figure 2 (b)). For simplicity, we assume the variance of dr remains the same as σ_R^2 for the two scenarios.



(a) When the link is *present* ($e_{ij} = 1$).

(b) When the link is *absent* ($e_{ij} = 0$).

Figure 2: Conditional probability distributions of variable dr given dz and e_{ij}, and variable dz given e_{ij}.

3.3 Prior Distributions

Social rank r reflects the ranking of actors in a network, and the most popular entities are assumed to have the smallest r values. Intuitively, influential nodes should always be fewer than ordinary nodes, which means there should be more nodes with large r values. Since power law is usually utilized to model a node's characteristics (e.g. ranks of individuals, foraging patterns of many species, etc.), we use the (truncated) long-tailed power law distribution as the prior distribution of the inverse of social rank, namely $p(r) \sim (1/r)^{-k_R}$ ($k_R > 0$). Mathematically, a power law cannot be a well-defined probability distribution, but a distribution that is a truncated power law is possible: $p(r) = C \cdot (1/r)^{-k_R}$ when $1/r > r_{min}$. It is easy to reveal that the normalization factor $C = (k_R + 1) \cdot r_{min}^{k_R+1}$, and thus $p(r) = (k_R + 1) \cdot r_{min}^{k_R+1} \cdot (1/r)^{-k_R}$ $(0 < r < 1/r_{min})$.

Proximity-based representation z should generally lie more uniformly on the K dimensional space, preferably not too far away from the origin. Therefore, a Gaussian prior is assumed on z: $p(z) = \mathcal{N}(0, \sigma_Z^2 \cdot I_K)$, where I_K denotes the K dimensional identity matrix. An illustration of the prior distributions is shown in Figure 3.

3.4 Objective Function and Optimization

Given the conditional and prior distribution of variables, we can formalize the function $f()$ in the beginning (Equation (2)) using Bayes' rule, and thus finalize our objective.

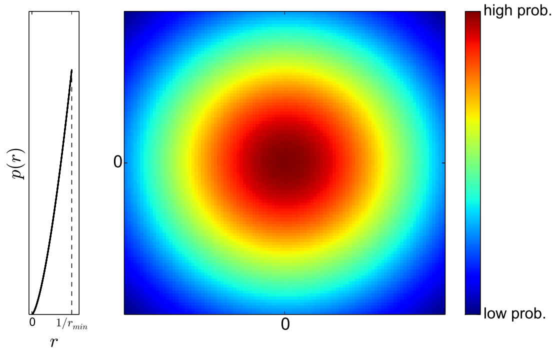


Figure 3: Left: prior of r. Right: prior of z (2D is used for visualization purpose).

$$\begin{aligned} p(e_{ij} = 1|dr, dz) &= \frac{p(dr, dz, e_{ij} = 1)}{\sum_{e=0}^{1} p(dr, dz, e_{ij} = e)} \\ = &\frac{p(dr|dz, e_{ij} = 1) \cdot p(dz|e_{ij} = 1) \cdot p(e_{ij} = 1)}{\sum_{e=0}^{1} p(dr|dz, e_{ij} = e) \cdot p(dz|e_{ij} = e) \cdot p(e_{ij} = e)} = sigmoid(f_{ij}) \end{aligned} \tag{7}$$

where

$$f_{ij} = \log \frac{p(dr|dz, e_{ij} = 1) \cdot p(dz|e_{ij} = 1) \cdot p(e_{ij} = 1)}{p(dr|dz, e_{ij} = 0) \cdot p(dz|e_{ij} = 0) \cdot p(e_{ij} = 0)} \tag{8}$$

and $sigmoid(x) = 1/(1 + e^{-x})$. Combining Equations (3)-(6), we have

$$\begin{aligned} f_{ij} = &\frac{\mu}{\sigma_R^2} \cdot dr \cdot h(dz) - \frac{\mu^2}{2\sigma_R^2} \cdot h^2(dz) - (\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_2^2})(dz)^2 \\ &+ \log \frac{\sigma_2 \cdot p(e_{ij} = 1)}{\sigma_1 \cdot p(e_{ij} = 0)} \end{aligned} \tag{9}$$

Since $h(dz) < 1$, we ignore the insignificant second order term $h^2(dz)$ as an approximation for now. For short, we use λ_R to denote $\frac{\mu}{\sigma_R^2}$ ($\mu > 0 \Rightarrow \lambda_R > 0$), λ_Z to denote $\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_2^2}$ ($\sigma_1^2 < \sigma_2^2 \Rightarrow \lambda_Z > 0$) and λ_0 to denote $\log \frac{\sigma_2 \cdot p(e_{ij}=1)}{\sigma_1 \cdot p(e_{ij}=0)}$ (for sparse networks, $p(e_{ij} = 1) \ll p(e_{ij} = 0) \Rightarrow \lambda_0 < 0$). Then the equation above is altered to the following:

$$f_{ij} = \lambda_R \cdot dr \cdot h(dz) - \lambda_Z \cdot (dz)^2 + \lambda_0 \tag{10}$$

where $\lambda_R, \lambda_Z, \lambda_0$ are hyper-parameters that need to be pre-assigned. Intuitively, large λ_R indicates the rank factor is more important in the network; while a large λ_Z indicates the proximity-based factor is more important. λ_0 reflects the sparsity of the network. We find that our model is not sensitive to λ_0, and the optimal values of λ_R and λ_Z can be found using cross validation.

As an extension to weighted graphs (denoting the weight of a link by w_{ij}), we treat the weight as multi-edges, and consider each edge independently. Thus the probability of observing a weighted edge is simply generalized to

$$p(e_{ij} = w_{ij}|dr, dz) := p(e_{ij} = 1|dr, dz)^{w_{ij}}. \tag{11}$$

The model parameters are inferred using maximum a posteriori (MAP) estimation, i.e.,

$$\begin{aligned}\{r^*, z^*\} &= \operatorname*{argmax}_{r,z} \log p(r, z|G) \\ &= \operatorname*{argmax}_{r,z} \log p(G|r, z) + \log p(r, z) \\ &= \operatorname*{argmax}_{r,z} \log p(G|r, z) + \sum_{i=1}^{N} \log p(r_i) + \sum_{i=1}^{N} \log p(\mathbf{z_i}).\end{aligned} \tag{12}$$

Since each entity generally has more freedom to issue links (e.g. a webpage can contain arbitrary many hyperlinks; a person can follow/retweet as many people/tweets as she wants), in order to eliminate the effect of size, we define the weight w_{ij} as the number of occurrence from i to j normalized by i's out-degree: $w_{ij} = \#(i \to j)/deg_{out}(i)$.

Edges in the graph are assumed to be generated independently. Therefore the likelihood of the graph is simply the product of the probabilities of all edges. Negative sampling is adopted in consideration of the efficiency issue [31]; in other words, for each existing link e_{ij}, k non-existing links are sampled $\{e_{il} = 0\}_{l \sim \mathcal{P}_0}$. According to [31], the background probability $\mathcal{P}_0$ is set to $\mathcal{P}_0(n) \propto deg_{out}(n)^{0.75}$. Denoting the sampled negative links by $S_0 = \cup_i \{(i, l)|e_{il} = 0\}_{l \sim \mathcal{P}_0}$, Equation (12) becomes

$$\begin{aligned}\{r^*, z^*\} &= \operatorname*{argmax}_{r,z} \Bigg(\sum_{(i,j):e_{ij}>0} w_{ij} \cdot \log p(e_{ij} = 1|dr, dz) \\ &+ \sum_{(i,j)\in S_0} \log \Big(1 - p(e_{ij} = 1|dr, dz)\Big) + \sum_{i=1}^{N} \log p(r_i) + \sum_{i=1}^{N} \log p(\mathbf{z_i})\Bigg) \\ &= \operatorname*{argmax}_{r,z} \Bigg(\sum_{(i,j):e_{ij}>0|dr,dz} w_{ij} \cdot \log sigmoid(f_{ij}) \\ &+ \sum_{(i,j)\in S_0} \log sigmoid(-f_{ij}) + \sum_{i=1}^{N} \log p(r_i) + \sum_{i=1}^{N} \log p(\mathbf{z_i})\Bigg).\end{aligned} \tag{13}$$

The optimization is done using stochastic gradient ascent.

3.5 Complexity

The first two summations in Equation (13) consist of $O(E \cdot K)$ terms, where E is the number of edges in the graph, and K is the dimension of the proximity-based factor. The last two summations in Equation (13) consist of $O(N \cdot K)$ terms, where N is the number of nodes. Therefore the computational complexity for each epoch of the data is $O((E + N) \cdot K)$. In sum, the running time is linear to the number of edges and nodes, therefore is scalable to large networks. In practice, it takes only a few epoches to reach convergence.

3.6 Relation to Other Models

Our model can be treated as a generalization of the well established Bradley-Terry model [10] in the realm of ranking and pairwise comparison, which has been successfully applied in learning to rank [12]. One parametrization of the Bradley-Terry model estimates the probability that the pairwise comparison $i \succ j$ (interpreted as "i is preferred to j", or "i ranks higher than j") is true as

$$P(i \succ j) = \frac{e^{s_i}}{e^{s_i} + e^{s_j}} = sigmoid(s_i - s_j) \tag{14}$$

where s_i is a real-valued score assigned to i and will be inferred. s_i can be treated as the rank of i; for example, if $s_i > s_j$, then $P(i \succ j) > 0.5$. This is essentially the same as the probability of a link $p(e_{ji} = 1)$ in a special case of our model, where $\lambda_R = 1$, $\lambda_Z = \lambda_0 = 0$ and $h(dz) = 1$ (constant).

4 EXPERIMENTS

4.1 Datasets

We use the following real-world datasets from different domains to test our new embedding algorithm:

- Snapchat Friendship. Snapchat is a US based ephemeral photo-messaging application developed by camera company Snap Inc. Users can make bi-directional friend links with others. We extract all friendship relations from a relatively small and isolated country, resulting in a network with about 1.5 million total nodes (users) and about 66 millions edges (bi-directional friendship links).
- Tencent Weibo Retweet [48]. Tencent weibo is a Chinese microblogging service where users can post tweets and follow/retweet from others. We extract the complete retweet relationships on November 1st, 2011. Tweets that have been retweeted for less than 5 times are excluded. A directed edge (u_A, u_B) is added when a user u_A retweets from user u_B.
- Venue Citation: We extract the paper citation links in the computer science domain and build a venue citations network from the Microsoft Academic Graph data [36]. A link from $venue_A$ to $venue_B$ indicates a citation from a paper published in $venue_A$ to a paper published in $venue_B$. We do not differentiate each proceeding of the venues. Links are aggregated over a 10-year period of time (2007-2016) and it naturally becomes a weighted graph. Venues are manually labeled according to their field of study, and we look into the following eight categories: AI (artificial intelligence and machine learning), NET (network), SE (software engineering), CT (computer theory), CV (computer vision and graphics), DB (database), PL (programming languages) and DM (data mining). Some venues may have multiple labels.
- Wikipedia Hyperlink[1]. An edge from i to j represents a hyperlink from wikipage i to wikipage j. Wikipages and their categories are structured in a collaborative hierarchical framework, and the folksonomy information is further cleaned according to [7]. We only keep wikipages with clear hierarchical labels in the previous step. Top six categories that contain most webpages are picked for our classification task: *sports*, *politics*, *science*, *Christian*, *geography* and *musician*. Some wikipages may have multiple labels. Edge multiplicity is not available for this dataset.
- Wikipedia Clickstream[2]. This dataset contains counts of (referer, resource) pairs extracted from the HTTP request logs of Wikipedia during Jan. 2017, where people navigate

[1] https://snap.stanford.edu/data/enwiki-2013.html
[2] https://meta.wikimedia.org/wiki/Research:Wikipedia_clickstream

from one wikipage (i.e. referer) to another (i.e. resource). Node labels are obtained in the same way as the Wikipedia Hyperlink dataset.

The details about the above datasets can be found in Table 1.

Dataset	#Nodes	#Edges	Weighted?
Snapchat Friendship	1.5M	66M	No
Tencent Weibo Retweet	842K	1.9M	Yes
Venue Citation	1.2K	91K	Yes
Wikipedia Hyperlink	488K	5.5M	No
Wikipedia Clickstream	2.4M	15M	No

Table 1: Dataset Statistics

In all experiments, 90% of the links are randomly sampled as the training dataset. Hyper-parameters for prior distribution (introduced in Section 3.3) are set to $k_R = 1.5$, $r_{min} = 0.1$ and $\sigma_Z^2 = 10^6$. We do not observe significance in terms of the evaluation metrics for different settings of $k_R \in [1, 2]$, $r_{min} \leq 0.2$ and $\sigma_Z^2 \geq 10^4$.

4.2 Applications

4.2.1 Classification. We demonstrate the advantage of our embedding in terms of multi-label classification results. We compare with the following baseline results.

- Matrix factorization techniques for recommender systems (MF) [22]. Although this method is designed for recommender systems, we apply their method on the user-user affinity matrix and treat the low dimensional feature for users as embedding. Note that, the popularity of a user is explicitly modeled by an additional bias factor.
- Graph Factorization (GF) [1]. This is another matrix factorization based approach that factorizes the graph adjacency matrix in order to obtain the low dimensional vector representation for nodes.
- Large-scale information network embedding (LINE) [40]. It embeds large information networks into lower dimensional vector spaces. The authors propose two measures of node proximity (i.e. 1st and 2nd order), and we will compare with both of them.
- Node2vec [16]. This approach learns low-dimensional representations for nodes in a graph by optimizing a neighborhood preserving objective.

The lower dimensional feature vector for each entity is then used to predict its label. For fair comparison, we only use the proximity-based representation z as the feature in our model. We also try treating the concatenation of both r and z as the feature vector; however, since there is little correlation between rank r and label (e.g. the popularity of a person does not indicate her occupation), we report the performance using z only. The state-of-the-art Conditional Bernoulli Mixtures (CBM) model [25] is adopted as the multi-label classifier, where the authors kindly make their code available online[3]. 90% of the nodes (labels as well as their vector representations) are randomly sampled for training. We use the following metrics for evaluation (denoting $Pred$ as the set of predicted labels, and $True$ as the set of ground truth labels for each entity):

- Jaccard Index: the number of correctly predicted labels divided by the union of predicted and true labels: $J(Pred, True) = \frac{|Pred \cap True|}{|Pred \cup True|}$. Larger values indicate better performance.
- Hamming Loss: the fraction of the wrong labels to the total number of labels: $\frac{1}{L}\sum_{i=1}^{L} xor(Pred_i, True_i)$, where L is the number of total labels. Smaller values indicate better performance.
- F1 score: the harmonic mean of precision $\frac{|Pred \cap True|}{|Pred|}$ and recall $\frac{|Pred \cap True|}{|True|}$. Larger values indicate better performance.

The classification results (average of the above metrics for each entity) are shown in Figures 4-6. We only evaluate datasets where ground truth user labels are available (i.e., Venue Citation, Wikipedia Hyperlink and Wikipedia Clickstream).

4.2.2 Link Prediction. Although not all baseline methods explicitly mention the application in link prediction, they all assign a probability score to every pair of nodes, which can be sorted and evaluated using the area under the ROC curve (AUC) score. Specifically, 10% of the existing edges and non-existing edges are hidden from the training set, and their probabilities are examined by the model. For methods designed for undirected networks, the probability of a directed link $u_i \to u_j$ is simply regarded as that of the undirected dyad (u_i, u_j). The evaluation results are reported in Table 2.

Particularly, a significant improvement over the baseline methods is observed on Snapchat and Weibo dataset, as users tend to interact with others (especially celebrities) due to their popularity instead of proximity, which is never captured by most baseline methods. This observation agrees with our intuition. Note that by using our embedding algorithm (RaRE), a 2-dimensional proximity-based embedding can already beat much higher-dimensional embedding (e.g., $K = 32$) for all the other baselines in almost all the datasets.

4.2.3 Embedding as Additional Features for Classification. The lower dimensional feature vector for each entity can also serve as additional features for real-world applications. For example, in Snapchat, the gender information is not required at the time of registration, however, knowledge of gender is crucial for better user understanding, ADs targeting and content recommendation. Currently, gender information is predicted by a Gradient Boosted Decision Trees model with a few highly engineered features, which already has an accuracy of 93.5% (the true gender information is obtained from Bitmoji user avatar). The features are derived from first name of a user, country, historical usage behavior of the Snapchat app products such as Lens, User Story, Discover etc. In this classification task, we collected 258,014 labeled examples, and we split it into training set and test set using 0.7 to 0.3 ratio. We concatenate the learned embedding vector as additional features to the basic features we used for gender prediction, and report the accuracy in Table 3 (accuracy = 1.0 - error rate). Given the high accuracy of the baseline model (93.5%), absolute accuracy lift of greater than 1% is considered very challenging and significant. As we can see from Table 3 that adding embedding vectors produced by our proposed method RaRE significantly outperforms other baselines for the gender prediction task. For example, when using 32 dimensional embeddings trained by RaRE, we observed 1.5%

[3] https://github.com/cheng-li/pyramid

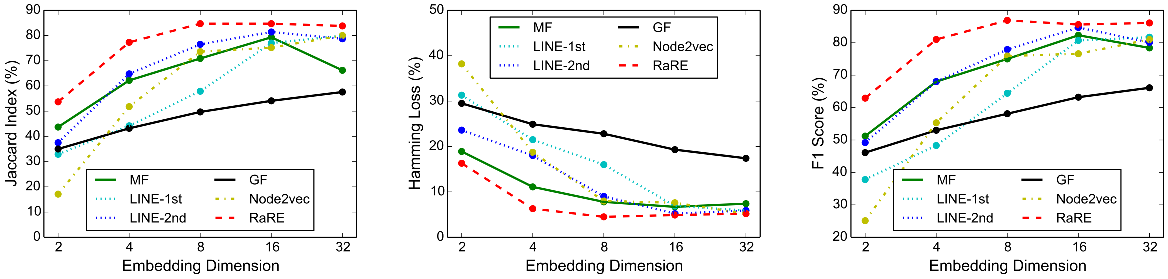


Figure 4: Multilabel classification results on Venue Citation dataset

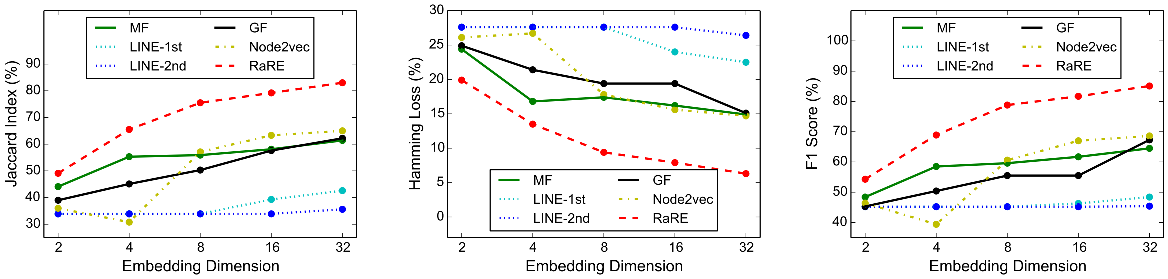


Figure 5: Multilabel classification results on Wikipedia Hyperlink dataset

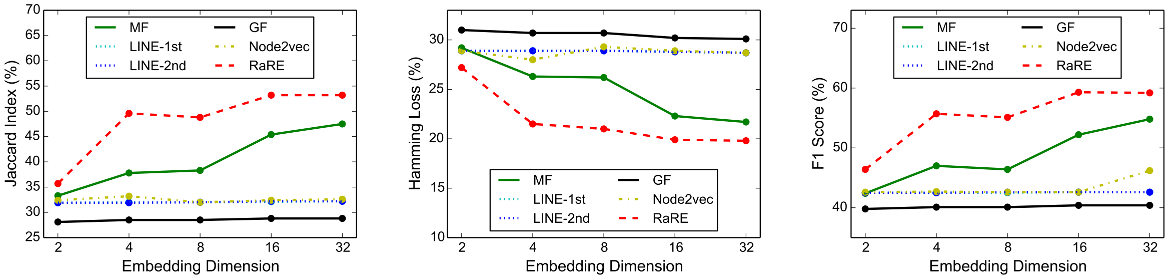


Figure 6: Multilabel classification results on Wikipedia Clickstream dataset

lift on the gender prediction accuracy over the current production model. The results show that the embedding information of a user carries useful signals for predicting the basic profile of a user such as gender.

4.2.4 A Novel Polar Coordinate-based Visualization. Visualization is another way to demonstrate the effectiveness of learned representation. A good embedding algorithm should be able to distinguish nodes of different labels by separating them in the vector representation space. Conventionally, a 2D or 3D vector representation is learned for each individual, which is treated as his/her coordinates and thus can be displayed in a scatter plot. We list the results of a few visualization methods in Figure 7, and it is very clear that the proximity-based embedding of RaRE does well in detecting different computer science research communities.

It is also interesting to reveal the visualization of the nodes by combining social rank and proximity-based representation in a unified plot. While many of the visualization approaches are capable of capturing much of the local structure (e.g. neighbors) as well as the global structure (e.g. clusters), they fail to identify the influential entities in the plot. In our method, we depict the coordinate of a node using the polar system, where the radius is simply its social rank r, and the angle θ is obtained from its proximity-based representation z by a simple transformation:

$$\{\theta\}^* = \underset{\theta}{\operatorname{argmin}} \sum_{(i,j)\in E_Z} \log\left(1+e^{-\cos(\theta_i-\theta_j)}\right) + \sum_{(i,j)\in \hat{E_Z}} \log\left(1+e^{\cos(\theta_i-\theta_j)}\right) \tag{15}$$

		Dimension of embedding K				
		2	4	8	16	32
Snapchat	MF	54.0	54.2	54.3	54.3	54.3
	GF	63.1	66.7	69.5	72.1	73.5
	LINE-1st	54.6	57.8	59.1	60.3	58.8
	LINE-2nd	55.8	56.0	56.0	56.0	56.0
	Node2vec	57.3	56.9	56.1	53.9	65.7
	RaRE	**86.7**	**92.1**	**94.0**	**94.8**	**94.6**
Tencent Weibo	MF	92.7	92.9	93.0	**93.1**	93.1
	GF	71.7	76.9	76.7	76.3	76.8
	LINE-1st	62.3	64.5	70.9	74.7	75.8
	LINE-2nd	57.1	60.4	60.2	60.9	61.6
	Node2vec	68.3	71.0	71.9	72.3	72.5
	RaRE	**95.3**	**95.2**	**94.2**	**93.1**	**96.6**
Venue Citation	MF	85.5	90.2	92.3	91.8	91.6
	GF	78.0	87.1	92.6	93.7	**94.4**
	LINE-1st	55.0	56.4	63.0	79.8	80.0
	LINE-2nd	64.4	74.2	80.0	81.2	81.5
	Node2vec	81.0	85.3	89.4	90.9	91.2
	RaRE	**91.4**	**93.7**	**94.0**	**94.3**	94.2
Wikipedia Hyperlink	MF	84.9	87.8	89.6	90.8	91.5
	GF	80.4	88.7	93.5	95.6	96.6
	LINE-1st	52.8	55.8	63.7	69.6	77.7
	LINE-2nd	50.0	50.1	50.2	50.7	51.7
	Node2vec	77.2	84.9	88.7	89.1	89.4
	RaRE	**97.5**	**97.6**	**97.7**	**97.8**	**97.8**
Wikipedia Clickstream	MF	63.4	68.6	72.1	74.5	76.9
	GF	76.1	82.5	86.1	86.7	86.7
	LINE-1st	73.8	78.1	78.6	78.8	78.8
	LINE-2nd	69.3	71.2	72.4	73.0	73.5
	Node2vec	82.9	87.2	88.1	89.6	89.0
	RaRE	**90.7**	**94.4**	**94.7**	**94.3**	**93.7**

Table 2: Link prediction AUC (%) on all datasets

Method	$K = 2$	$K = 4$	$K = 8$	$K = 16$	$K = 32$
MF	93.3	93.4	93.5	93.7	94.0
GF	93.6	93.6	93.8	94.0	94.3
LINE-1st	93.5	93.5	93.6	93.7	93.7
LINE-2nd	93.5	93.5	93.7	93.7	93.8
Node2vec	94.2	94.2	94.3	94.3	94.5
RaRE	**94.5**	**94.6**	**94.7**	**94.9**	**95.0**

Table 3: Gender prediction accuracy (%) on Snapchat dataset

where $\cos(\theta_i - \theta_j)$ reflects the similarity on the 2D sphere, E_Z is the new set of edges defined in the space of proximity-based representation: $E_Z = \{(i, j) : ||z_i - z_j||^2 < t\}$ and $\hat{E_Z}$ is the corresponding non-existing pairs of nodes sampled using the same strategy as Section 3.4. The equation above can be considered as preserving the proximity in both spaces (minimizing the logit loss), as similar to various dimension reduction approaches. Here we pick $t = 0.5$ and $z \in \mathbb{R}^K$ where $K = 2$, and we do not observe significant variance in terms of these parameters.

In the bottom right figure of Figure 7 (polar coordinates from RaRE), we can clearly observe the most influential venues around the center, among which top conferences in different areas (e.g. CHI, WWW, ICSE, CVPR, SIGGRAPH, SIGMOD, INFOCOM, AAAI, KDD, VLDB, ICML, STOC) are successfully identified.

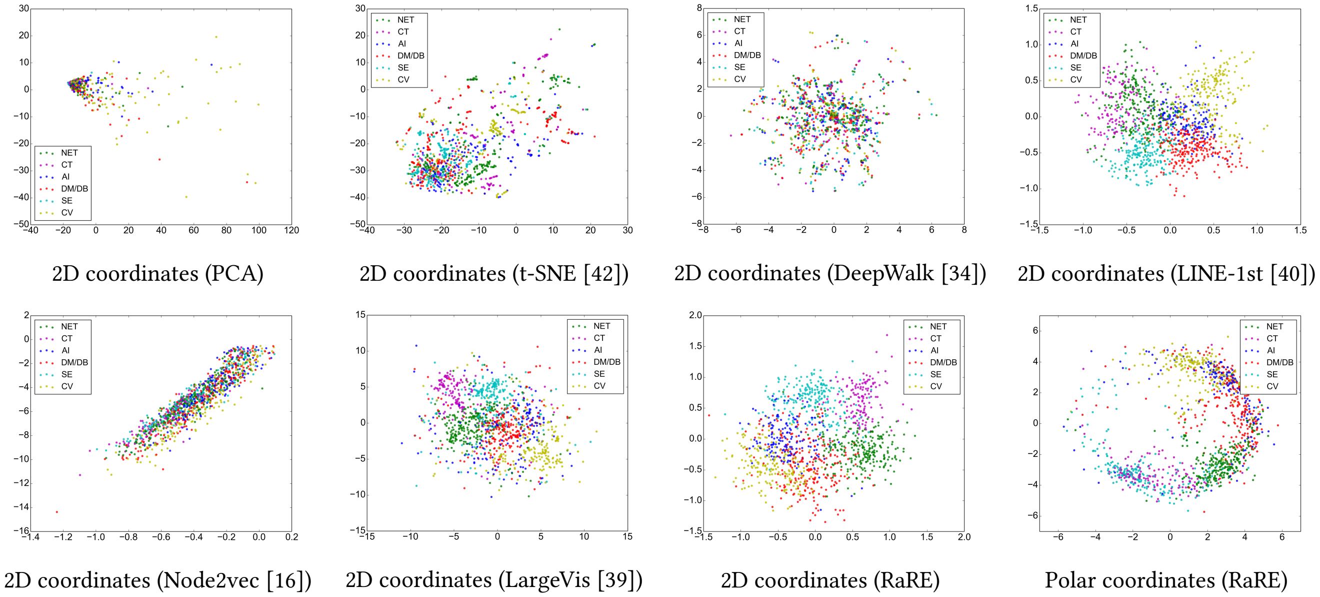
5 RELATED WORK

5.1 Ranking

Mapping entities in the network to a spectrum of importance scores has been a very popular research topic for decades. The notion of influential nodes emerged from the large-scale World Wide Web (WWW), where it is vital for crawlers to start from important pages first [15], or generally, an objective method is desired to measure human's interest in a collections of webpages. Various algorithms have been proposed, e.g. PageRank [32], HITS [21], and have been widely generalized towards various needs [3, 11, 17, 27, 45]. These algorithms have an assumption in common: a random surfer is assumed to browse the web and click hyperlinks randomly, and the probability distribution of a webpage being visited will converge to a score related to its rank. Similar ideas are also applicable to information networks by defining the weights between entities [45, 46]. In addition to random walk-based approaches, many ranking methods are proposed based on Bayesian network and inference. Pal and Counts [33] generate a list of features for microblogs based on their followers, number of hashtags and so on. Clusters are then revealed using a Gaussian mixture model, and the rank of a microblog is an aggregation of the rank of its features. Ball and Newman [2] study several friendship networks, and find that unreciprocated friendships often consist of a lower-ranked individual claiming friendship with a higher-ranked one. Based on this assumption, they deduce such social rankings using maximum likelihood estimation.

All the above approaches provide a global ranking for each entity, however, the ranking makes little sense when entities in several categories are mingled together without distinguishing the clustering information. For example, we seldom mention comparisons such as "a computer system conference is ranked higher than a database conference". For heterogeneous networks where multiple types of entities or relations may exist, determining the network becomes even tricker and multiple ranking systems may occur according to different schemas or topics. Sun et al. [37, 38] associate heterogeneous network clustering and ranking together, and assume several rank distributions conditional on different cluster structures. Liu et al. [26] propose a probabilistic generative model, which explains the network generation process from users and documents and is able to reveal the most related nodes with a given topic (query). However, additional information other than cluster labels is usually desired to understand the network. Our method, on the other hand, provides the lower dimensional proximity-based representation for each entity, which has wide applications including classification, clustering, visualization and so on.

[4] Labels (colors) are identified by investigating the topic of venues within each class of KNN on the original one-hot encoding vector.



2D coordinates (PCA)

2D coordinates (t-SNE [42])

2D coordinates (DeepWalk [34])

2D coordinates (LINE-1st [40])

2D coordinates (Node2vec [16])

2D coordinates (LargeVis [39])

2D coordinates (RaRE)

Polar coordinates (RaRE)

Figure 7: Visualization on Venue Citation dataset. These plots are best viewed in color[4].

5.2 Network Embedding

Detecting latent representation (embedding) of nodes in a network is essential in understanding the opinions of individuals and desired for various machine learning tasks. Traditional approaches usually utilize the adjacency matrix in order to extract essential dimensions of the data [5, 23, 35, 41, 47], which involve finding the eigenvalues of a matrix and thus not scalable for large networks. Matrix factorization finds an approximation of a matrix by the product of two lower rank matrices, and this technique has been popular especially in recommender systems [6, 19, 20, 22]. In terms of graph data, matrix factorization can be applied on the affinity matrix [1, 30], and each row of the low dimensional matrix naturally becomes the vector representation of the corresponding node.

More recent approaches introduce embedding for nodes in a network, which is a low-dimensional vector that represents the latent characteristics of a node. These embedding vectors are learned by preserving similarity in the network and similarity in the latent Euclidean space. The notion of embedding originates from word embedding [31], and Levy and Goldberg [24] establishes the connection between matrix factorization and word embedding, arguing that estimating word embedding is equivalent to factorizing a pointwise mutual information matrix. Later on, researchers have discovered strategies to explain the generation of links from a probabilistic perspective, with the assumption that the likelihood of a link should be proportional to the similarity of both nodes (neighbors) [1, 28, 30, 42] or entities [8, 9, 14, 44]. More recent approaches generalize the notion of similar nodes to n-hop neighbors [13, 16, 34, 40, 43]. Generally, links are assumed to be explained as the proximity between the representation of two actors (i.e. the "homophily" assumption [29]). However, we often observe links to highly-ranked nodes (e.g. many users follow *celebrities* on Twitter, and scholars tend to cite *popular* works and authors in a bibliographic network). As a result, some nodes are poorly modeled by the homophily assumption. Embedding-based approaches ignore this seminal factor in link generation, which may lead to inaccurate estimation as a result. Some matrix factorization methods consider the popularity factor by introducing a bias term [6, 20, 22], but they model these two factors independently, while neglecting the fact that the knowledge of one can affect the distribution of the other.

6 CONCLUSION

In this paper we present a novel approach for information network embedding with consideration of individuals' social ranks. From the graph generation perspective, we refine the latent representation of nodes on information network by analyzing the role of individuals in terms of their social rank. Moreover, we provide solid derivations on the reason behind a link in terms of both latent proximity-based representation as well as social rank of a node, which provides a brand new insight of the problem. We carefully design a framework that explicitly models the interdependency between these two types of embeddings. Finally, we evaluate our model on several real-world large-scale datasets, and the results on classification, link prediction and visualization demonstrate our advantage over the state-of-the-art network embedding methods.

ACKNOWLEDGEMENT

We would like to thank Snapchat Inc. for its gift funding. We would also like to share our gratitude to the anonymous reviewers for their precious comments. This work is partially supported by NSF CAREER Award #1741634 and #1453800.

REFERENCES

[1] A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola. Distributed large-scale natural graph factorization. In *Proceedings of the 22nd international conference on World Wide Web*, pages 37–48. ACM, 2013.

[2] B. Ball and M. E. Newman. Friendship networks and social status. *Network Science*, 1(01):16–30, 2013.

[3] A. Balmin, V. Hristidis, and Y. Papakonstantinou. Objectrank: Authority-based keyword search in databases. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 564–575. VLDB Endowment, 2004.

[4] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[5] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*, pages 585–591, 2002.

[6] R. M. Bell, Y. Koren, and C. Volinsky. The bellkor 2008 solution to the netflix prize. *Statistics Research Department at AT&T Research*, 2008.

[7] P. Boldi and C. Monti. Cleansing wikipedia categories using centrality. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 969–974. International World Wide Web Conferences Steering Committee, 2016.

[8] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795, 2013.

[9] A. Bordes, J. Weston, R. Collobert, Y. Bengio, et al. Learning structured embeddings of knowledge bases. In *AAAI*, volume 6, page 6, 2011.

[10] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[11] U. Brandes and S. Cornelsen. Visual ranking of link structures. *J. Graph Algorithms Appl.*, 7(2):181–201, 2003.

[12] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96. ACM, 2005.

[13] S. Cao, W. Lu, and Q. Xu. Grarep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 891–900. ACM, 2015.

[14] S. Chang, W. Han, J. Tang, G.-J. Qi, C. C. Aggarwal, and T. S. Huang. Heterogeneous network embedding via deep architectures. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 119–128. ACM, 2015.

[15] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through url ordering. *Computer Networks and ISDN Systems*, 30(1):161–172, 1998.

[16] A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864. ACM, 2016.

[17] T. H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526. ACM, 2002.

[18] P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

[19] Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 263–272. Ieee, 2008.

[20] C. C. Johnson. Logistic matrix factorization for implicit feedback data. *Advances in Neural Information Processing Systems*, 27, 2014.

[21] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[22] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[23] J. B. Kruskal and M. Wish. *Multidimensional scaling*, volume 11. Sage, 1978.

[24] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185, 2014.

[25] C. Li, B. Wang, V. Pavlu, and J. Aslam. Conditional bernoulli mixtures for multi-label classification. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2482–2491, 2016.

[26] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang. Mining topic-level influence in heterogeneous networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 199–208. ACM, 2010.

[27] L. Lü, Y.-C. Zhang, C. H. Yeung, and T. Zhou. Leaders in social networks, the delicious case. *PloS one*, 6(6):e21202, 2011.

[28] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[29] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[30] A. Menon and C. Elkan. Link prediction via matrix factorization. *Machine Learning and Knowledge Discovery in Databases*, pages 437–452, 2011.

[31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[32] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[33] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 45–54. ACM, 2011.

[34] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[35] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[36] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu, and K. Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246. ACM, 2015.

[37] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 565–576. ACM, 2009.

[38] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 797–806. ACM, 2009.

[39] J. Tang, J. Liu, M. Zhang, and Q. Mei. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th International Conference on World Wide Web*, pages 287–297. International World Wide Web Conferences Steering Committee, 2016.

[40] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077. ACM, 2015.

[41] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[42] L. Van Der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of machine learning research*, 15(1):3221–3245, 2014.

[43] D. Wang, P. Cui, and W. Zhu. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1225–1234. ACM, 2016.

[44] Z. Wang, J. Zhang, J. Feng, and Z. Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, pages 1112–1119, 2014.

[45] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.

[46] Y. Yamaguchi, T. Takahashi, T. Amagasa, and H. Kitagawa. Turank: Twitter user ranking based on user-tweet graph analysis. In *International Conference on Web Information Systems Engineering*, pages 240–253. Springer, 2010.

[47] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, 29(1):40–51, 2007.

[48] Y. Yang, J. Tang, C. W.-k. Leung, Y. Sun, Q. Chen, J. Li, and Q. Yang. Rain: Social role-aware information diffusion. In *AAAI*, pages 367–373, 2015.