Explanations for Network Embedding-based Link Predictions

Bo Kang, Jefrey Lijffijt, and Tijl De Bie

IDLab, Department of Electronics and Information Systems, Ghent University,
Technologiepark-Zwijnaarde 122, 9052 Ghent, Belgium
{firstname.lastname}@ugent.be

Abstract. Graphs (also called *networks*) are powerful data abstractions, but they are challenging to work with, as many machine learning methods may not be applied to them directly. Network Embedding (NE) methods resolve this by learning vector representations for the nodes, for subsequent use in downstream machine-learning tasks. Link Prediction is one such important downstream task, used for example in recommender systems. NE methods perform exceedingly well in accuracy for Link Prediction, but predictions following from the embeddings, whose dimensions have no intrinsic meaning, are not straightforward to understand. Explaining why predictions are made can increase trustworthiness, help understand the underlying models and give insight into what features of the network are important in light of the predictions, and answer posed regulatory requirements on the ability to explain machine-learning-based decisions. We study the problem of providing explanations for NE-based link predictions and introduce ExplaiNE, an approach to derive counterfactual explanations by identifying links in the network that explain link predictions. We show how ExplaiNE can be used generically on NE-based methods and consider ExplaiNE in more detail for Conditional Network Embedding, a particularly suitable state-of-art NE method. Extensive experiments demonstrate ExplaiNE's accuracy and scalability.

Keywords: explainability, network embedding, link prediction, XAI

1 Introduction

Network embeddings (NEs) have exploded in popularity in both the machine learning and data mining communities. By mapping a network's nodes into a vector space, NEs enable the application of a variety of machine learning methods on networks for important tasks such as link prediction and classification. Link prediction (LP) is the task to predict whether nodes are likely to be or become connected in partially observed or evolving networks. LP has wide-ranging applications, for friendship recommendations, recommender systems, knowledge graph completion, etc. While there are numerous conventional LP methods that predict links based on heuristic statistics computed over networks (e.g. based on the number of common neighbors) (see, e.g. [17]), recently proposed NE-based methods typically outperform those heuristic approaches (e.g. [9,11]).

A major disadvantage of NE-based LP methods is that they do not immediately provide intelligible explanations of the predicted links. The ability to understand link

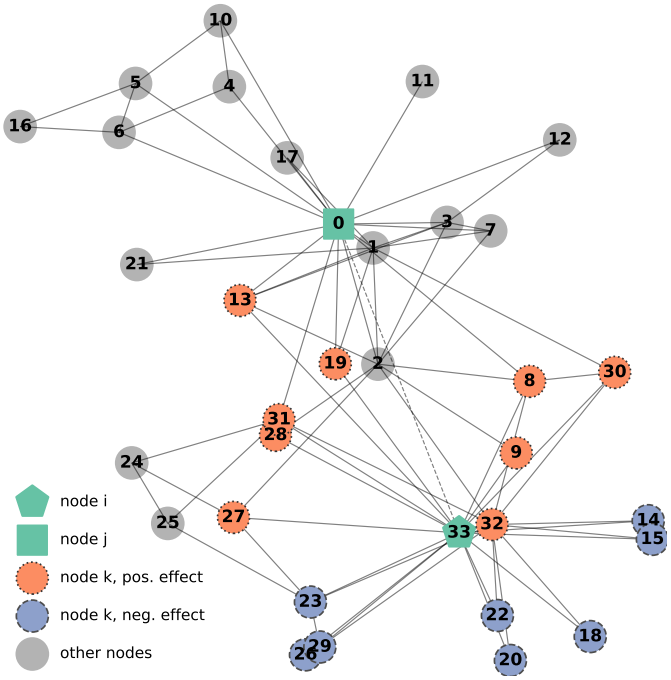

Fig. 1. Visualization of Zachary's karate club.

predictions is important and useful for several reasons: (a) recommender systems that provide explanations are more easily trusted and more effective, (b) it allows data analysts to have a better understanding of the network characteristics such as node features and network dynamics, (c) transparency of automated processing systems is required in a growing number of regulations, and explanations can increase transparency.

To address these needs, we present ExplaiNE, a principled counterfactual reasoning approach for explaining NE-based LPs. In its simplest form, ExplaiNE quantifies how the probability of a predicted link is affected by considering the non-existence (weakening) of an existing link. Links that after weakening most strongly reduce the probability of the predicted link then serve as counterfactual explanations.

Example. We illustrate the idea behind ExplaiNE with an example. Zachary's karate club network [29] is a network of 34 karate club members connected through 78 friendship links. We used a recent probabilistic NE-based LP method [11], embedded the data in two-dimensional Euclidean space (so that we can also visualize it; see Figure 1) and asked the NE-based LP method for a recommendation of a link for node $i = 33$.

The LP method suggests the most probable link for 33 is node $j = 0$. Although node 0 is not the closest unlinked node to 33, it has high degree, making a link likely under the model used by this embedding method. Figure 1 visualizes the embedding and

also shows which existing links incident to i ExplaiNE details as positive (orange circle with dotted edge) and negative (blue circle with dashed edge) contributions, towards the prediction of link $\{0, 33\}$. It concludes this because weakening links to the orange nodes would reduce the link probability $\{i, j\}$, whereas weakening links to the blue nodes would increase it. These explanations visually appear to make sense: the orange nodes 'pull' node 33 closer to 0, while the blue nodes pull node 33 away from 0.

The paper is organized as follows. In Sec. 2, we first derive ExplaiNE generically, allowing for explanations not only in terms of links incident to the predicted link, but also in terms of other links as well as non-links. We then reduce its scope to explanations in terms of only incident links (as in the example above), and make an approximation (which we justify empirically) to obtain a still generic but highly scalable approach. Next we apply ExplaiNE to Conditional Network Embedding (CNE) [11], a recent state-of-the-art NE method. The application of ExplaiNE to CNE is particularly transparent, thanks to its straightforward use in LP, requiring no training once the embedding is found. We also outline how ExplaiNE can be applied to NE methods based on skip gram with negative sampling-based such as LINE [25], DeepWalk [19], PTE [24], and node2vec [9], with further details provided in the supplementary material. Experiments covering quantitative and qualitative analysis of the performance and scalability are described in Sec. 3. Related work is discussed in Sec. 4 and the conclusions in Sec. 5.

The **main contributions** of this paper are:

- We introduce ExplaiNE, a generic counterfactual reasoning approach for explaining LPs based on NEs (Sec. 2.3).
- We provide a scalable tight approximation of ExplaiNE (Sec. 2.4).
- We present a detailed application of ExplaiNE to CNE (Sec. 2.5)
- An outline of applying ExplaiNE to skip gram with negative sampling based NE methods. (Sec. 2.6)
- We discuss quantitative and run-time analyses, finding that the approximation is stable and scalable. (Sec. 3)
- We study qualitative and quantitative empirical results from realistic case studies, which indicate ExplaiNE achieves its goal. (Secs. 3.2, 3.3)

2 Methods

2.1 Notation

We first introduce basic notation, similar to, e.g. [4] and [8]. An *undirected network* is denoted $\mathcal{G} = (V, E)$ where V is a set of $n = |V|$ *nodes* and $E \subseteq \binom{V}{2}$ is the *set of links* (or *edges*). A *link* is denoted by an unordered node pair $\{i, j\} \in E$. We use $\boldsymbol{A}$ to denote an *adjacency matrix*, with element $a_{ij} = 1$ for $\{i, j\} \in E$ and $a_{ij} = 0$ otherwise. The symbol $\hat{\boldsymbol{A}}$ will be used to denote the adjacency matrix of a particular observed network. NE methods find a mapping $f : V \to \mathbb{R}^d$ from nodes to d-dimensional real vectors. An *embedding* is denoted as $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)' \in \mathbb{R}^{n \times d}$, where $\boldsymbol{x}_i \triangleq f(i)$ for $i \in V$ is the embedding of each node. Finally, we write $\boldsymbol{X}^*$ for a locally optimal embedding of an adjacency matrix $\boldsymbol{A}$, and likewise write $\hat{\boldsymbol{X}}^*$ for $\hat{\boldsymbol{A}}$.

2.2 NE-based Link Predictions

All well-known NE methods aim to find an embedding $\boldsymbol{X}^*$ for given graph $\mathcal{G}$ (with adjacency matrix $\boldsymbol{A}$) that maximizes a continuously differentiable[1] objective function $\mathcal{O}(\boldsymbol{A}, \boldsymbol{X})$ for the given adjacency matrix $\boldsymbol{A}$. Thus $\boldsymbol{X}^*$ must satisfy the following necessary condition of local optimality:

$$\nabla_{\boldsymbol{X}} \mathcal{O}(\boldsymbol{A}, \boldsymbol{X}^*) = \boldsymbol{0}. \tag{1}$$

Defining $\boldsymbol{F}(\boldsymbol{A}, \boldsymbol{X}) \triangleq \nabla_{\boldsymbol{X}} \mathcal{O}(\boldsymbol{A}, \boldsymbol{X})$, a locally optimal embedding $\boldsymbol{X}^*$ is thus a solution to $\boldsymbol{F}(\boldsymbol{A}, \boldsymbol{X}^*) = \boldsymbol{0}$.

Based on an embedding $\boldsymbol{X}$, it is common to predict the existence of a link between any pair of nodes i and j by computing a link probability (or other score) $g_{ij}(\boldsymbol{X})$, using a differentiable function $g_{ij} : \mathbb{R}^{nd} \to \mathbb{R}$. In practice, g_{ij} often only depends on the embeddings $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ of i and j, and often it can be written as $g_{ij}(\boldsymbol{X}) = g(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for some function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. It is often found by training a classifier (e.g. logistic regression) on a set of known linked and unlinked node pairs (see Sec. 2.6), but sometimes it follows directly from the NE model (e.g. for CNE). We also introduce the function $g^*_{ij} : \mathbb{R}^{n \times n} \to \mathbb{R}$ defined as $g^*_{ij}(\boldsymbol{A}) \triangleq g_{ij}(\boldsymbol{X}^*)$ where $\boldsymbol{X}^*$ is locally optimal w.r.t. $\boldsymbol{A}$. I.e., g^*_{ij} directly computes the link probability w.r.t. the locally optimal embedding (assumed here as given) for a specified adjacency matrix.

2.3 ExplaiNE as a generic approach

ExplaiNE uses a counterfactual reasoning approach to explain link predictions based on a NE. Namely, it quantifies the change of the link probability (or other score) of a node pair $\{i, j\}$ if the presence of a link between a given pair of nodes $\{k, l\}$ were to be altered.

Consider first the situation where $\{k, l\} \in E$: If removing the link $\{k, l\}$ strongly decreases the probability of a link between i and j, the link $\{k, l\}$ is a good counterfactual explanation of this predicted link. Conversely, consider the situation where $\{k, l\} \notin E$: If adding a link $\{k, l\}$ strongly decreases the probability of a link between i and j, it is the absence of a link between k and l that is a good counterfactual explanation of the predicted link $\{i, j\}$.

Intuitively, adding or removing an existing link $\{k, l\}$ will alter the probability of a link between i and j because it will alter the optimal embedding, which in turn will change the link probability of the target pair $\{i, j\}$. For the ExplaiNE strategy to be effective, we must be able to compute and combine these two effects in an efficient manner.

A naive approach would be to recompute the embedding with a link added or removed, and to quantify how much this changes the probability of a link between i and

[1] Note that, although NE methods are often described for unweighted networks (i.e., a binary adjacency matrix), the objective $\mathcal{O}(\boldsymbol{A}, \boldsymbol{X})$ is often continuously differentiable also w.r.t. the adjacency matrix $\boldsymbol{A}$. This is required for ExplaiNE to be applicable, but as we will see this requirement is often satisfied.

j. However, recomputing the embedding is computationally demanding, and is practically impossible to do even for a moderate number of pairs $\{k, l\}$. Moreover, even adding or removing a single link can dramatically change the optimization landscape, and as there are potentially many local optima, we can end up with a very different embedding—even if initialized with the same original embedding—, making a change in link probability erratic and hard to interpret.

Instead, ExplaiNE investigates the effect of an *infinitesimal* change to a_{kl} around its observed value $\hat{a}_{kl}$, on the link probability as computed by g^*_{ij}. Specifically, ExplaiNE seeks as explanations node-pairs $\{k, l\}$ ($k \neq l$ and $\{k, l\} \neq \{i, j\}$) for which $\frac{\partial g^*_{ij}}{\partial a_{kl}}(\hat{\boldsymbol{A}})$ is large in absolute value, and positive if $\hat{a}_{kl} = 1$ (as then decreasing a_{kl} down from $\hat{a}_{kl} = 1$ by a small amount would maximally decrease g^*_{ij}), and negative if $\hat{a}_{kl} = 0$ (as then increasing a_{kl} up from $\hat{a}_{kl} = 0$ by a small amount would maximally decrease g^*_{ij}). This can be done analytically. Indeed, applying the chain rule, we find:

$$\frac{\partial g^*_{ij}}{\partial a_{kl}}(\hat{\boldsymbol{A}}) = \nabla_{\boldsymbol{X}} g_{ij} \left(\hat{\boldsymbol{X}}^*\right)^T \cdot \frac{\partial \boldsymbol{X}^*}{\partial a_{kl}}(\hat{\boldsymbol{A}}). \tag{2}$$

For many NE methods the first factor can be computed analytically from the expression for g_{ij}, as we will see in the next subsections. The second factor can be computed using the *implicit function theorem* (see, e.g. [3]). Rephrased for our specific setting and overloading the symbol $\boldsymbol{X}^*$ here to also signify a function, this theorem states:

Theorem 1 (Implicit function theorem) *Let $\boldsymbol{F} : \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ be a continuously differentiable function with arguments denoted $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{X} \in \mathbb{R}^{n\times d}$. Moreover, let $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{X}}^*$ be such that $\boldsymbol{F}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{X}}^*) = \boldsymbol{0}$. If the Jacobian matrix $\nabla_{\boldsymbol{X}} \boldsymbol{F}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{X}}^*)$ is invertible, then there exists an open set $S \subset \mathbb{R}^{n\times n}$ with $\hat{\boldsymbol{A}} \in S$ such that there exists a continuously differentiable function $\boldsymbol{X}^* : S \to \mathbb{R}^{n\times d}$ with:*

$$\begin{aligned}
&\boldsymbol{X}^*(\hat{\boldsymbol{A}}) = \hat{\boldsymbol{X}}^*, \text{ and} \\
&\boldsymbol{F}(\boldsymbol{A}, \boldsymbol{X}^*(\boldsymbol{A})) = \boldsymbol{0} \text{ for all } \boldsymbol{A} \in S, \text{ and} \\
&\frac{\partial \boldsymbol{X}^*}{\partial a_{kl}}(\boldsymbol{A}) = -(\nabla_{\boldsymbol{X}} \boldsymbol{F}(\boldsymbol{A}, \boldsymbol{X}^*(\boldsymbol{A})))^{-1} \frac{\partial \boldsymbol{F}}{\partial a_{kl}}(\boldsymbol{A}, \boldsymbol{X}^*(\boldsymbol{A})).
\end{aligned}$$

It is the latter expression, evaluated at $\hat{\boldsymbol{A}}$, that we need in in order to evaluate Eq. (2). Note that the Jacobian $\nabla_{\boldsymbol{X}} \boldsymbol{F}$ is in fact the Hessian of $\mathcal{O}$ with respect to $\boldsymbol{X}$. This means that $\nabla_{\boldsymbol{X}} \boldsymbol{F}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{X}}^*)$ is negative definite (as $\hat{\boldsymbol{X}}^*$ is optimal for $\hat{\boldsymbol{A}}$). While for some NE-methods it may not be *strictly* negative definite and thus not invertible as required by the theorem (because, e.g. any translation of $\hat{\boldsymbol{X}}^*$ may be equally optimal according to $\mathcal{O}$), this situation can be avoided by adding a regularizer to $\mathcal{O}$ on, e.g. the Frobenius norm of $\hat{\boldsymbol{X}}^*$ with very small weight. Without going into detail, we note that as this regularization constant approaches zero, this becomes equivalent with using the pseudo-inverse of the Hessian, instead of its inverse. This is the approach we have taken whenever this situation arose. Denoting this Hessian evaluated at $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{X}}^*$ as $\boldsymbol{H}$, can thus write:

$$\frac{\partial \boldsymbol{X}^*}{\partial a_{kl}}(\hat{\boldsymbol{A}}) = -\boldsymbol{H}^{-1} \cdot \frac{\partial \boldsymbol{F}}{\partial a_{kl}}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{X}}^*). \tag{3}$$

Putting Eqs. (2) and (3) together, we now can compute the derivative of g^*_{ij} with respect to a_{kl}:

$$\frac{\partial g^*_{ij}}{\partial a_{kl}}(\hat{\boldsymbol{A}}) = -\nabla_{\boldsymbol{X}} g_{ij} \left(\hat{\boldsymbol{X}}^*\right)^T \cdot \boldsymbol{H}^{-1} \cdot \frac{\partial \boldsymbol{F}}{\partial a_{kl}}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{X}}^*). \tag{4}$$

For efficiency, one can compute the partial derivatives for a given predicted link $\{i, j\}$ and for all pairs $\{k, l\}$ by pre-computing the vector $\nabla_{\boldsymbol{X}} g_{ij} \left(\hat{\boldsymbol{X}}^*\right)^T \cdot \boldsymbol{H}^{-1}$: solve a linear system with nd variables and equations, and right multiplying it with the vectors $\frac{\partial \boldsymbol{F}}{\partial a_{kl}}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{X}}^*)$ that depend on k and l. Unfortunately, the computational cost of solving this linear system is $O((nd)^3)$, limiting scalability both in network size and dimensionality. Thus, while this is a clear improvement over the naive approach, it is not sufficient for realistic network sizes. The next subsection describes how to make ExplaiNE tractable also for large networks and dimensionalities.

2.4 Making ExplaiNE scalable

First, we choose to focus on explanations in terms of linked pairs $\{k, l\}$, rather than in terms of unlinked pairs. Such positive explanations are arguably more insightful than negative ones, and especially in sparse networks.

Second, experiments show that the best explanation for a predicted link $\{i, j\}$ for a node i, tends to be a link $\{k, l\}$ that is incident to node i, i.e., for which $l = i$. This is arguably because links adjacent to node i affect the link probability $g^*_{ij}(\hat{\boldsymbol{A}})$ by directly affecting the embedding $\boldsymbol{x}^*_i$, whereas links not incident to i are likely to have a secondary effect only. Besides this, we also believe that nodes incident to i are likely to be more meaningful from node i's perspective than other links, in practical applications. Thus, we can restrict ourselves to seeking an explanation for a predicted link from node i to node j in terms of an existing link $\{i, k\}$ for which $\frac{\partial g^*_{ij}}{\partial \hat{a}_{ik}}(\hat{\boldsymbol{A}})$ is large and positive.

Third, we consider only NE methods where $g_{ij}(\boldsymbol{X}^*)$ only depends on $\boldsymbol{x}^*_i$ and $\boldsymbol{x}^*_j$ (true for all NE methods we are aware of). Thus, Eq. (2) can be written as:

$$\begin{aligned}\frac{\partial g^*_{ij}}{\partial a_{ik}}(\hat{\boldsymbol{A}}) =& \nabla_{\boldsymbol{x}_i} g_{ij} \left(\hat{\boldsymbol{X}}^*\right)^T \cdot \frac{\partial \boldsymbol{x}^*_i}{\partial a_{ik}}(\hat{\boldsymbol{A}}) + \\ & \nabla_{\boldsymbol{x}_j} g_{ij} \left(\hat{\boldsymbol{X}}^*\right)^T \cdot \frac{\partial \boldsymbol{x}^*_j}{\partial a_{ik}}(\hat{\boldsymbol{A}}).\end{aligned} \tag{5}$$

Finally, we make an approximation inspired by the fact that changing a_{ik} will have a direct effect on the optimal embeddings $\boldsymbol{x}^*_i$ and $\boldsymbol{x}^*_k$, but only indirectly (and thus typically less so) on the embedding of the other nodes—including on $\boldsymbol{x}^*_j$. Hence, we ignore the second term in Eq. (5).

What remains to be computed is thus $\frac{\partial \boldsymbol{x}^*_i}{\partial a_{ik}}(\hat{\boldsymbol{A}})$. To do so, we consider the optimality condition of the embedding w.r.t. $\boldsymbol{x}^*_i$ alone, considering all other node embeddings fixed to their optimum in $\hat{\boldsymbol{X}}^*$ for the observed $\hat{\boldsymbol{A}}$. Letting $\hat{\boldsymbol{X}}^*_{(i)}$ denote the set of $\hat{\boldsymbol{x}}_l$ with $l \neq i$, this optimality condition is:

$$\nabla_{\boldsymbol{x}_i} \mathcal{O}(\hat{\boldsymbol{A}}, \boldsymbol{x}_i, \hat{\boldsymbol{X}}^*_{(i)}) = \boldsymbol{0} \tag{6}$$

Writing $\hat{\boldsymbol{F}}_i(\boldsymbol{A}, \boldsymbol{x}_i) \triangleq \nabla_{\boldsymbol{x}_i} \mathcal{O}(\boldsymbol{A}, \boldsymbol{x}_i, \hat{\boldsymbol{X}}^*_{(i)})$ for conciseness, optimality of $\hat{\boldsymbol{x}}^*_i$ given the observed network $\hat{\boldsymbol{A}}$ then requires that $\hat{\boldsymbol{F}}_i(\hat{\boldsymbol{A}}, \hat{\boldsymbol{x}}^*_i) = \boldsymbol{0}$. We can use the implicit function theorem on this optimality condition to approximate $\frac{\partial \boldsymbol{x}^*_i}{\partial a_{ik}}$ as:

$$\frac{\partial \boldsymbol{x}^*_i}{\partial a_{ik}}(\hat{a}_{ik}) = -\boldsymbol{H}_i^{-1} \cdot \frac{\partial \hat{\boldsymbol{F}}_i}{\partial a_{ik}}. \tag{7}$$

Here, $\boldsymbol{H}_i = \nabla_{\boldsymbol{x}_i} \hat{\boldsymbol{F}}_i(\hat{\boldsymbol{A}}, \hat{\boldsymbol{x}}^*_i)$ is the Jacobian of $\hat{\boldsymbol{F}}_i$ or equivalently the Hessian of $\mathcal{O}$ w.r.t. $\boldsymbol{x}_i$, evaluated at $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{X}}^*)$. Putting Eqs. (7) and (5) (ignoring the second term as discussed) together, this yields:

$$\frac{\partial g^*_{ij}}{\partial a_{ik}}(\hat{\boldsymbol{A}}) = -\nabla_{\boldsymbol{x}_i} g_{ij} \left(\hat{\boldsymbol{X}}^*\right)^T \cdot \boldsymbol{H}_i^{-1} \cdot \frac{\partial \hat{\boldsymbol{F}}_i}{\partial a_{ik}}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{x}}^*_i). \tag{8}$$

Comparing Eq. (4) with Eq. (8) reveals the dramatic complexity reduction achieved: Inverting $\boldsymbol{H}_i \in \mathbb{R}^{d \times d}$ has a practical complexity of only $O(d^3)$, which is entirely feasible given common dimensionalities used in the literature (often 128).

2.5 ExplaiNE for CNE

We now apply the generic ExplaiNE approach to the Conditional Network Embedding method (CNE) [11]. Detailed derivations are deferred to the supplementary Section.1. CNE proposes a probability distribution for the network conditional on the embedding, and finds the optimal embedding by maximum likelihood estimation. Specifically, the objective function $\mathcal{O}$ in CNE is the log-probability of the network conditioned on the embedding:

$$\begin{aligned}\mathcal{O}(\hat{\boldsymbol{A}}, \boldsymbol{X}) &= \log(P(\hat{\boldsymbol{A}}|\boldsymbol{X})) \\ &= \sum_{\{i,j\}:\hat{a}_{ij}=1} \log P_{ij}(a_{ij} = 1|\boldsymbol{X}) + \\ &\qquad \sum_{\{i,j\}:\hat{a}_{ij}=0} \log P_{ij}(a_{ij} = 0|\boldsymbol{X}).\end{aligned}$$

Here, the link probabilities P_{ij} conditioned on the embedding are defined as follows:

$$\begin{aligned}&P_{ij}(a_{ij} = 1|\boldsymbol{X}) = 1 - P_{ij}(a_{ij} = 0|\boldsymbol{X}) \\ &= \frac{P_{\hat{\boldsymbol{A}},ij} \mathcal{N}_{+,\sigma_1}(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|)}{P_{\hat{\boldsymbol{A}},ij} \mathcal{N}_{+,\sigma_1}(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|) + (1 - P_{\hat{\boldsymbol{A}},ij}) \mathcal{N}_{+,\sigma_2}(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|)},\end{aligned} \tag{9}$$

where $\mathcal{N}_{+,\sigma}$ denotes a half-Normal distribution [14] with spread parameter σ, $\sigma_2 > \sigma_1 = 1$, and where $P_{\hat{\boldsymbol{A}},ij}$ is a prior probability for a link to exist between nodes i and j as inferred from the degrees of the nodes (or based on other information about the structure of the network [1,13]). CNE, being based on a probabilistic model for the graph conditioned on the embedding, naturally allows for LP using the probabilities $P_{ij}(\hat{a}_{ij} = 1|\boldsymbol{X})$. In other words, $g_{ij}(\boldsymbol{X}) = P_{ij}(a_{ij} = 1|\boldsymbol{X})$ as shown in Eq. (9). Note

that it depends on $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ alone, as required for the approximate version of ExplaiNE to be applicable (third assumption).

Next we show how to apply approximated ExplaiNE to CNE (in the remainder of the text, we drop the modifier 'approximated'; exact ExplainNE applied to CNE is derived in supplementary Section.2. First, we derive the optimality condition:

$$\begin{aligned}\hat{\boldsymbol{F}}_i(\hat{\boldsymbol{A}}, \hat{\boldsymbol{x}}_i^*) &= \nabla_{\boldsymbol{x}_i^*} \log(P(\hat{\boldsymbol{A}}|\hat{\boldsymbol{X}}^*)) \\ &= \gamma \sum_{j \neq i} (\hat{\boldsymbol{x}}_i^* - \hat{\boldsymbol{x}}_j^*) \left(P\left(a_{ij} = 1 | \hat{\boldsymbol{X}}^*\right) - \hat{a}_{ij} \right) \\ &= \mathbf{0}.\end{aligned}$$

Denoting $\gamma = \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}$, and $\hat{P}_{ij}^* \triangleq g_{ij}^*(\hat{\boldsymbol{A}}) = P_{ij}(a_{ij} = 1|\hat{\boldsymbol{X}}^*)$ (the probability of a link between i and j given the optimal embedding $\hat{\boldsymbol{X}}^*$ for $\hat{\boldsymbol{A}}$), we can now derive the three factors in Eq. (8) (see Supplementary Section.1 for detailed derivations):

$$\begin{aligned}\nabla_{\boldsymbol{x}_i} g_{ij}\left(\hat{\boldsymbol{X}}^*\right) &= -\gamma(\boldsymbol{x}_i^* - \boldsymbol{x}_j^*)\hat{P}_{ij}^*(1 - \hat{P}_{ij}^*), \\ \boldsymbol{H}_i &= \nabla_{\boldsymbol{x}_i} \hat{\boldsymbol{F}}_i(\hat{\boldsymbol{A}}, \hat{\boldsymbol{x}}_i^*) \\ &= \gamma \boldsymbol{I} \sum_{l \neq i} (P_{il}^* - \hat{a}_{il}) \\ &\quad - \gamma^2 \sum_{l \neq i} (\boldsymbol{x}_i^* - \boldsymbol{x}_l^*)(\boldsymbol{x}_i^* - \boldsymbol{x}_l^*)' \hat{P}_{il}^*(1 - \hat{P}_{il}^*), \\ \frac{\partial \hat{\boldsymbol{F}}_i}{\partial a_{ik}}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{x}}_i^*) &= \gamma(\boldsymbol{x}_k^* - \boldsymbol{x}_i^*),\end{aligned}$$

This means:

$$\frac{\partial g_{ij}^*}{\partial a_{ik}}(\hat{\boldsymbol{A}}) = (\boldsymbol{x}_i^* - \boldsymbol{x}_j^*)^T \left(\frac{-\boldsymbol{H}_i}{\gamma^2 \hat{P}_{ij}^*(1 - \hat{P}_{ij}^*)} \right)^{-1} (\boldsymbol{x}_i^* - \boldsymbol{x}_k^*).$$

Note that the Hessian should be invertible and negative definite, if $\hat{\boldsymbol{x}}_i^*$ is indeed a local maximum. Interestingly, this expression has an intuitive interpretation: without the inverted Hessian, it would be an inner product between the distance of $\boldsymbol{x}_i^*$ to the embeddings of both nodes $\boldsymbol{x}_j^*$ and $\boldsymbol{x}_k^*$, indicating that the best explanation is as far as possible in the direction of $\boldsymbol{x}_j^*$ as seen from $\boldsymbol{x}_i^*$. Yet, the Hessian modulates the metric and reduces the explanatory power in directions with lots of embedded nodes l for which $\hat{P}_{il}^*(1 - \hat{P}_{il}^*)$ is large, i.e. for which the model is undecided whether there should be a link.

2.6 ExplaiNE for other NE methods

Here we illustrate the generic applicability of ExplaiNE by outlining the steps of applying it to NE methods based on skip gram with negative sampling (SGNS) (e.g. LINE, PTE, DeepWalk, node2vec). In supplementary Section.3, we derive a concrete example for LINE [25].

In those methods, $g_{i,j}(\boldsymbol{X}) = g(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where $g \triangleq \sigma \circ h$ with $\sigma : \mathbb{R}^d \to \mathbb{R}$ a linear classifier (often logistic regression) applied to edge embeddings, whereby the embedding $h(\boldsymbol{x}_i, \boldsymbol{x}_j)$ of an edge $\{i, j\}$ is computed by applying an edge embedding operator $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ (e.g. element-wise product) to the embeddings of the nodes at its end-points.

[15] and [21] found that SGNS-based NE methods all share the same objective:

$$\mathcal{L} = \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} \log \sigma(\boldsymbol{x}_i \cdot \boldsymbol{y}_j) + b \sum_{i=1}^{|V|} \mathbb{E}_{j' \sim P_N} \left[\log \sigma(-\boldsymbol{x}_i \cdot \boldsymbol{y}_{j'}) \right],$$

where $\boldsymbol{x}_i$ is the target embedding of node i, $\boldsymbol{y}_i$ is the embedding of node j as context (usually discarded, node2vec does not differentiate target and context), $\sigma(\cdot)$ is a sigmoid function, P_N is known as the noise that generates negative samples, and b is the number of negative samples. Moreover, [21] showed that $\mathcal{L}$ often has a closed form representation (or converges to one in probability). This makes it possible to obtain an analytical expression of the NE optimality condition, and thus of the function $\boldsymbol{F}(\boldsymbol{A}, \boldsymbol{X})$. Given this, both exact and approximated ExplaiNE can be derived.

3 Experiments

We investigated the following questions: **Q1** How does the approximation compare to the exact version (Sec. 3.1)? **Q2** Does ExplaiNE give sensible explanations (Sec. 3.2–Sec. 3.3)? **Q3** Does the proposed method scale (Sec. 3.4)? . All experiments are based on CNE with parameters $\sigma_1 = 1$, σ_2 is tuned on a $90\% - 10\%$ train-validation split with values from $\{2, 8, 16, 32, 64\}$. Any weights associated to links are ignored. Due to the lack of space, we summarize the experimental results in this section, and discuss the results more extensively in supplementary material. Code to reproduce all the experiments as well as supplementary material are available at: `https://github.com/aida-ugent/ExplaiNE`. We used the following networks.

The **Game of Thrones' (GoT) network**[2] consisting of 796 characters (nodes) and 2823 links between characters that are mentioned within 15 words of one another in books 1-5. We used a 2-dimensional embedding of this network to assess the quality of the approximated ExplaiNE approach.

The **DBLP co-authorship network** [26] (DBLP dataset V10 [3]) with papers published up to year 2017, from which we selected all papers published at ICML, NeurIPS, ICLR, JMLR, MLJ, KDD, ECML-PKDD, and DMKD. This results in 23,359 authors (nodes) and 20,545 papers, converted into 66,597 links between authors who co-authored at least one paper. We conducted both qualitative and quantitative evaluations on a 8-dimensional embedding of this network.

The **MovieLens dataset**[4] [10] with 100,000 ratings by 943 users on 1,682 movies. The network is thus bipartite and consists of 943+1,682 nodes and 100,000 edges. The

[2] `https://github.com/mathbeveridge/asoiaf`

[3] `https://aminer.org/citation`

[4] `https://grouplens.org/datasets/movielens/100k/`

dataset also contains metadata such as title and genre, which we have used as external validation sources. We conducted qualitative and quantitative experiments on a 8-dimensional embedding of this network.

3.1 Quality of the ExplaiNE approximation

Before applying approximated ExplaiNE to the real-world datasets, we first evaluate the quality of the approximation (**Q1**). We will assess the extent to which the top K explanations for a predicted link $\{i, j\}$ incident to a given node i, as given by approximated ExplaiNE, overlap with the top-K explanations given by exact ExplaiNE. Relevant parameters here are (1) the value of K and (2) the number of neighbors. As we consider only links to neighbors as candidate explanations, K must be smaller than the number of neighbors of i. Moreover, if the number of neighbors is not much larger than K, a substantial overlap in the top-K explanations of the exact and approximate method is not surprising. Indeed, if i has m neighbors, two random subsets of K neighbors would share l elements with probability $\binom{K}{l}\binom{m-K}{K-l}/\binom{m}{K}$, which is large for large l if m is not much larger than K.

Thus, we performed a stratified analysis, computing the size of the overlap of the top-K explanations, aggregated in a histogram over nodes with a specific degree. We did this on the GoT dataset for K from 1 to 5. This experiment revealed that the top-1 is always identical between the approximated and exact versions, while the elements further in the ranked list very rarely swapped positions (2 to 3 differences out of 796 on ranks 2–4, and 7 differences out of 796 for rank 5).

We also compared the complete ranking of the neighbors between the approximated and exact ExplaiNE versions, and this simply for the most probable link for every node (which empirically meant seeking explanations for links that are also present in the network). We computed the normalized Kendall tau distance between the ranked explanations given by approximated and exact ExplaiNE. The average normalized Kendall tau distance is 0.008 ± 0.03. For comparison, the average Kendall tau distance between a random ranking and exact ExplaiNE is 0.47 ± 0.28, so the observed rank distance between the exact and approximate ranking is indeed very small. Having established confidence in the accuracy of the approximation, we can now evaluate the behavior of approximated ExplaiNE on two larger networks.

3.2 Qualitative evaluation

As a qualitative evaluation, we applied ExplaiNE to two real world networks to assess whether ExplaiNE gives sensible explanations for predicted links (**Q2**).

DBLP network. In this co-authorship network, a predicted link between authors i and j suggests two non-linked authors would network-wise form a sensible collaboration.

While ExplaiNE uses no external information to provide its explanations for such suggested collaborations, our experiments indicate that such explanations tend to be existing collaborators working on a topic on which the suggested collaborator is active as well. As an example, we predict links for Eric P. Xing (node i), and compute the

Table 1. Predicted/recommended collaborations for Eric P. Xing.

Rank	Recommendations	Explain: 'Zoubin Ghahramani'
1	Zoubin Ghahramani	Yee Whye Teh
2	John D. Lafferty	Mário A. T. Figueiredo
3	Tong Zhang	Willie Neiswanger
4	Jeff G. Schneider	Andrew Gordon Wilson
5	Dale Schuurmans	Ruslan Salakhutdinov

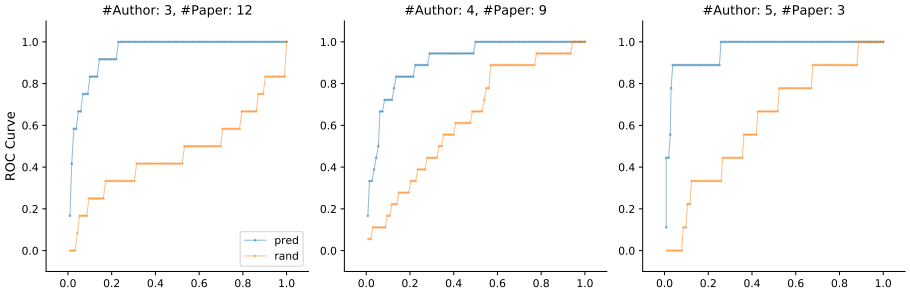

Fig. 2. ROC curves of co-author predictions for $i =$'Eric P. Xing', with author-list lengths 3, 4, and 5 (orange=rand., blue=ExplaiNE).

explanations for his top recommendation[5]: Zoubin Ghahramani (node j). We find that the existing co-authors of Eric P. Xing identified by ExplaiNE as top-5 explanations for this recommendation are either colleagues or co-authors of Zoubin Ghahramani (see Table 1), with a shared interest in graphical models and Bayesian machine learning.

MovieLens network. In this network, a predicted link between a user i and movie j amounts to recommending movie j to user i. To do this, CNE is not given access to meta-data of the users or movies, and neither does ExplaiNE to identify explanations. Yet, we can make use of this meta-data to qualitatively assess whether the explanations make sense: We computed the recommendation for the first user (uid=0) in the user list. The top recommended movie is 'Mission: Impossible' with genre tags 'Action, Adventure, Mystery'. The genres of the top explanations given by ExplainNE arguably have strongly overlapping genre tags.

More results from these two data sets are given in the supplementary material. Based on these and other qualitative experiments (not included), our general impression is that ExplaiNE gives sensible explanations. The next subsection aims to quantify these findings.

[5] The recommended authors are not co-authors of the querying author according to the constructed co-author network.

Table 2. Recommended movie to user uid=0. The top movie recommended by CNE (Mission: Impossible) is explained through movies already seen by user uid=0. The top-ranked explanations have genres that overlap with the recommended movie.

j	Recommendations	Genres
1	Mission: Impossible	Action, Adventure, Mystery
2	E.T. the Extra-Terrestrial	Children's, Drama, Fantasy, Sci-Fi
3	Scream	Horror, Thriller
4	Liar Liar	Comedy
5	Schindler's List	Drama, War

k	Explanations for 'Mission: Impossible'	Genres
1	Twister	Action, Adventure, Thriller
2	Independence Day	Action, Sci-Fi, War
3	The Rock	Action, Adventure, Thriller
4	Mr. Holland's Opus	Drama
5	Phenomenon	Drama, Romance

3.3 Quantitative evaluation

Objectively evaluating the quality of an explanation is conceptually non-trivial, due to a lack of datasets with ground-truth explanations for LP. Yet, as we show here, it is possible to use metadata to derive reasonable ground-truth proxy explanations, and compare with those.

DBLP network. For this data, we can construct ground truth explanations for *existing* links (as opposed to *predicted* ones). While this is not the intended use case of ExplaiNE, it is a possible use case and justified here given our intention to objectively validate the quality of the explanations. Our approach is based on the intuition that a one-time co-author j of a given author i could have been introduced to that author i by another co-author k on the same paper, thus explaining the link $\{i, j\}$. While this is not always true, we postulate it is sufficiently common.

Given an author i and a one-time co-author j of i, we used ExplaiNE to rank the other co-authors of i, from more to less explanatory (according to Eq. 8). We then took the top-r of this ranked list as predicted co-authors on the paper i co-authored with j, and created a confusion matrix. Clearly, the hardness of this prediction task is different for papers with different numbers of authors. Thus, in order to get a more aggregate assessment, we summed the top-r confusion matrices for all one-time co-authors of node i on papers with a given number of co-authors L, and this for different L between 3 and 5. For a given author-list length, the confusion matrices with different r were then used to create precision-recall curves or ROC curves. Fig. 2 shows the ROC curves for Eric P. Xing as node i for each author-list length. For comparison, also ROC curves computed based on randomly ranked lists are shown (as the size of the data is rather small, these are not always close to the diagonal). All results indicate that the explanations are remarkably effective, indicating that ExplaiNE performs well. Results for other nodes are similar.

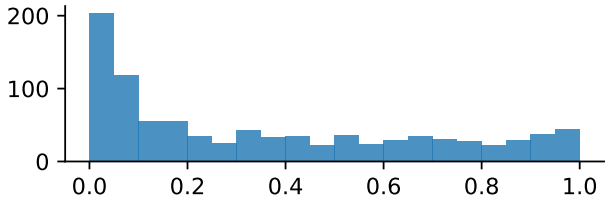

Fig. 3. Histogram of p-values that indicates the significance of the correlation between the genre recommended and the genres in the explanation. Each p-value is computed against 50 random explanations. Those explanations are drawn from user's watched movies. The empirical distribution has Kolmogorov-Smirnov test statistic 0.26 and a p-value against uniform distribution that is smaller than the numpy floating point machine accuracy $5.5e-58$. This shows the significance of positive correlation between the recommended movies and the explanations made by ExplaiNE.

MovieLens network. The idea behind our second quantitative evaluation is that a good explanation k of a predicted link between a movie-user pair $\{i, j\}$ should often have a similar list of genres as j. To test this, we computed the top-5 explanations for user i and her top recommended movie j. Then we averaged the Jaccard similarity between the set of genres for movie j and the set of genres of each of the 5 explanations. To assess the significance of this average, we computed an empirical p-value for it by randomly sampling 50 sets of 5 'explanations' drawn from the watched movies of i, resulting in 50 random average Jaccard similarities to compare with the one obtained by ExplaiNE. Thus we obtained an empirical p-value for each user i, indicating the significance of the overlap between the set of genres of the recommended movie j and the top-5 explanations.

A histogram of these p-values is shown in Fig. 3. While p-values are uniformly distributed under the null hypothesis—that the explanations have genres unrelated to those of j—, here this is not the case, which indicates the null hypothesis is false. The Kolmogorov-Smirnoff test indeed rejects the null (with test statistic 0.26 and p-value is smaller than the numpy floating point machine accuracy $5.5e-58$).

3.4 Scalability and runtime

To address **Q3**, we measured the runtime of exact and approximated ExplaiNE when computing $\frac{\partial g^*_{ij}}{\partial a_{ik}}(\hat{\boldsymbol{A}})$ for all $k \notin \{i, j\}$, as per Eqs. (4) and (8), on average over random pairs of nodes $\{i, j\}$. The runtime was measured on a PC with Intel quad Core i5 2.7GHz and 16GB RAM. Results shows that approximated ExplaiNE is efficient and applicable to large networks with higher dimensionality (e.g. on average 0.02s on DBLP network with dim $= 8$), while exact ExplaiNE is not (exact method ran out of memory).

Table 3. Average runtime (in sec., 10 trials) of exact and approximated ExplaiNE in computing the explanations for a random pair of nodes $\{i, j\}$. Note that the exact method also has substantial memory cost: 10.4 Gb for MovieLens and on DBLP it went out of memory. On MovieLens, the time was computed only for one k, and multiplied by $n - 2$ to get an estimated total time for all k.

Network	#nodes	dim	exact	approx.
Karate	34	2	0.03	1.8e−4
GoT	796	2	64.1	4.1e−4
GoT	796	8	1490	9.8e−4
MovieLens	2625	8	3.4e5	6.3e−4
DBLP	23359	8	—	0.02

4 Related Work

There are a few existing works that aim to explain link predictions (see, e.g. [7,2,27]). More specifically, they explicitly encode specific edge types (e.g. topological roles, whether it is social or topic link) in the model, thus the explanations are limited to the edge types that the model induced.

In parallel, the importance of accountability of AI has sparked growing research interest in interpretable ML. Interpretable ML research can be categorized into model-based and post-hoc approaches [6,18]. The first category focuses on incorporating interpretability while constructing the ML model. ExplaiNE belongs to the second category of interpretable ML methods: it is a post-hoc method that focuses on interpreting the local structure of ML models (here, NE models). Most directly related (although not for LP) are [22] and [16], who provide a model-agnostic explanation via local approximation of the model. Closely related, [23] and [12] compute the gradient of the loss of a (black-box) model w.r.t. the input to gauge the relevance of the input features. The former computes the gradient using back-propagation, while the second approximates the gradient using a Taylor series expansion. In a more recent work [28], the explanation of a prediction made by a graph neural network is represented by the subgraph and sub-features that largely affect the prediction.

A related research line concerns adversarial attacks on graphs, as such attacks tend to search for graph modification with a large impact on a specific task. [5] and [30], for example, consider attacks that target node classifications based on graph convolutional neural networks (instead of LPs based on NE, such as ExplaiNE). [20] investigate the robustness of LP in knowledge graphs via adversarial modifications, for the specific case of knowledge graph embeddings using multiplicative models. The authors show how their approach also enables interpretations of the knowledge graph in terms of length-2 Horn rules (e.g. $\text{isMarriedTo}(a, b) \wedge \text{hasChild}(b, c) \Rightarrow \text{hasChild}(a, c)$). ExplaiNE on the other hand is derived as a generic approach, with explaining the presence of individual links as a primary focus.

ExplaiNE is the first generic approach (and, as far as we know, the first approach at all) for explaining link predictions based on a NE. Unlike the previous attempts of explaining link prediction results, ExplaiNE is agnostic to the edge types. Moreover, to the best of our knowledge, ExplaiNE is the first method that uses the implicit function theo-

rem for explainability. This proved to be a crucial element for computing the gradient of the link probability w.r.t. the network structure, as it allowed us to rigorously track the optimal embedding given an infinitesimal change in the input network. We believe this theorem can prove valuable also for other tasks, particularly those where an intermediate representation is obtained by optimizing an unsupervised objective function (e.g. an autoencoder), to be fed into a subsequent model that is trained in a supervised manner.

5 Conclusions

Link prediction (LP) is an important task, with numerous applications. State-of-the-art approaches are based on first embedding the network in a vector space, followed by a LP step. Unfortunately, these approaches offer no insight in their predictions. To remedy this, we introduced ExplaiNE, a generic approach to explain Network Embedding (NE)-based LPs based in terms of existing links in the network that contribute most to a link being predicted (or not predicted), and we presented several means to make it scalable. We applied ExplaiNE to CNE, a state-of-the-art NE method. Through extensive qualitative and quantitative evaluations, we evaluated the usefulness of ExplaiNE, and its ability to scale to large networks, indicating an affirmative answer to both. As further work, it may be considered how to exploit the principles behind ExplaiNE for detecting and mitigating adversarial modifications, and studying robustness, similar to [5,20,30].

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) (ERC Grant Agreement no. 615517), and under the European Union's Horizon 2020 research and innovation programme (ERC Grant Agreement no. 963924), from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme, and from the FWO (project no. G091017N, G0F9816N, 3G042220).

References

1. Adriaens, F., Lijffijt, J., De Bie, T.: Subjectively interesting connecting trees. In: ECML-PKDD. pp. 53–69. Springer (2017)
2. Barbieri, N., Bonchi, F., Manco, G.: Who to follow and why: link prediction with explanations. In: KDD. pp. 1266–1275. ACM (2014)
3. Chiang, A.C.: Fundamental methods of mathematical economics. Aukland (New Zealand) McGraw-Hill (1984)
4. Cui, P., Wang, X., Pei, J., Zhu, W.: A survey on network embedding. TKDE (2018)
5. Dai, H., Li, H., Tian, T., Huang, X., Wang, L., Zhu, J., Song, L.: Adversarial attack on graph structured data. In: ICML. pp. 1115–1124 (2018)
6. Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. arXiv preprint arXiv:1808.00033 (2018)
7. van Engelen, J.E., Boekhout, H.D., Takes, F.W.: Explainable and efficient link prediction in real-world network data. In: IDA. pp. 295–307. Springer (2016)

8. Goyal, P., Ferrara, E.: Graph embedding techniques, applications, and performance: A survey. Knowledge-Based Systems **151**, 78–94 (2018)
9. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: KDD. pp. 855–864. ACM (2016)
10. Harper, F.M., Konstan, J.A.: The movielens datasets: History and context. TIIS **5**(4), 19 (2016)
11. Kang, B., Lijffijt, J., De Bie, T.: Conditional network embeddings. In: ICLR (2019)
12. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: ICML. pp. 1885–1894 (2017)
13. van Leeuwen, M., De Bie, T., Spyropoulou, E., Mesnage, C.: Subjective interestingness of subgraph patterns. Machine Learning **105**(1), 41–75 (2016)
14. Leone, F., Nelson, L., Nottingham, R.: The folded normal distribution. Technometrics **3**(4), 543–550 (1961)
15. Levy, O., Goldberg, Y.: Neural word embedding as implicit matrix factorization. In: NeurIPS. pp. 2177–2185 (2014)
16. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: NeurIPS. pp. 4765–4774 (2017)
17. Martínez, V., Berzal, F., Cubero, J.C.: A survey of link prediction in complex networks. CSUR **49**(4), 69 (2017)
18. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592 (2019)
19. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: Online learning of social representations. In: KDD. pp. 701–710. ACM (2014)
20. Pezeshkpour, P., Tian, Y., Singh, S.: Investigating robustness and interpretability of link prediction via adversarial modifications. arXiv pre-print arXiv:1905.00563 (2019)
21. Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K., Tang, J.: Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In: WSDM. pp. 459–467. ACM (2018)
22. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: KDD. pp. 1135–1144. ACM (2016)
23. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
24. Tang, J., Qu, M., Mei, Q.: Pte: Predictive text embedding through large-scale heterogeneous text networks. In: KDD. pp. 1165–1174. ACM (2015)
25. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: Large-scale information network embedding. In: WWW. pp. 1067–1077 (2015)
26. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: KDD. pp. 990–998. ACM (2008)
27. Xu, L., Wei, X., Cao, J., Yu, P.S.: On exploring semantic meanings of links for embedding social networks. In: WWW. pp. 479–488 (2018)
28. Ying, Z., Bourgeois, D., You, J., Zitnik, M., Leskovec, J.: Gnnexplainer: Generating explanations for graph neural networks. In: NeurIPS. pp. 9240–9251 (2019)
29. Zachary, W.W.: An information flow model for conflict and fission in small groups. Journal of anthropological research **33**(4), 452–473 (1977)
30. Zuegner, D., Akbarnejad, A., Guennemann, S.: Adversarial attacks on neural networks for graph data. IJCAI pp. 6246–6250 (2019)