ExplaiNE: An Approach for Explaining Network Embedding-based Link Predictions

Bo Kang [1] **Jefrey Lijffijt** [1] **Tijl De Bie** [1]

Abstract

Networks are powerful data structures, but are challenging to work with for conventional machine learning methods. Network Embedding (NE) methods attempt to resolve this by learning vector representations for the nodes, for subsequent use in downstream machine learning tasks.

Link Prediction (LP) is one such downstream machine learning task that is an important use case and popular benchmark for NE methods. Unfortunately, while NE methods perform exceedingly well at this task, they are lacking in transparency as compared to simpler LP approaches.

We introduce ExplaiNE, an approach to offer counterfactual explanations for NE-based LP methods, by identifying existing links in the network that explain the predicted links. ExplaiNE is applicable to a broad class of NE algorithms. An extensive empirical evaluation for the NE method 'Conditional Network Embedding' in particular demonstrates its accuracy and scalability.

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 (LP): the task to predict whether nodes are likely to be(come) connected in incomplete 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., Martínez et al., 2017), recently proposed NE-based methods typically outperform those heuristic approaches (e.g., Grover & Leskovec, 2016; Kang et al., 2019).

While the superior performance of NE-based LP methods is an advantage, a major disadvantage is that they do not easily allow for human-intelligible explanations of the predicted links. Yet, the ability to understand link 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.

We present ExplaiNE, a mathematically principled counterfactual reasoning approach for explaining NE-based link predictions. In its simplest form, ExplaiNE quantifies how the probability of a predicted link $\{i, j\}$ would be affected by weakening an existing link $\{i, k\}$. Links $\{i, k\}$ that after weakening most strongly reduce the probability of the predicted link $\{i, j\}$ then serve as counterfactual explanations.

Example. We show the idea of ExplaiNE on Zachary's karate club network (Zachary, 1977). The network consists of 34 karate club members (nodes), with 78 friendship links. An NE-based LP method[1] to predict a link for node $i = 33$ (green pentagon) indicates a high probability link to node $j = 24$ (green square). Figure 1 visualizes the embedding and highlights which existing links incident to i ExplaiNE deems explanatory for this prediction in a positive (orange circle, dotted edge) or negative sense (blue circle, dashed edge). 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. Note that these effects are quite intuitive given the geometry of the embedding: the orange nodes 'pull' node 33 closer to 24, while the blue nodes pull node 33 away from 24.

ExplaiNE is first derived as generically as possible, 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-

[1]Department of Electronics and Information Systems (ELIS), IDLab, Ghent University, Ghent, Belgium. Correspondence to: Bo Kang <bo.kang@ugent.be>, Jefrey Lijffijt <jefrey.lijffijt@ugent.be>, Tijl De Bie <tijl.debie@UGent.be>.

[1]The embedding used is 2-dimensional and derived using Conditional Network Embedding (Kang et al., 2019).

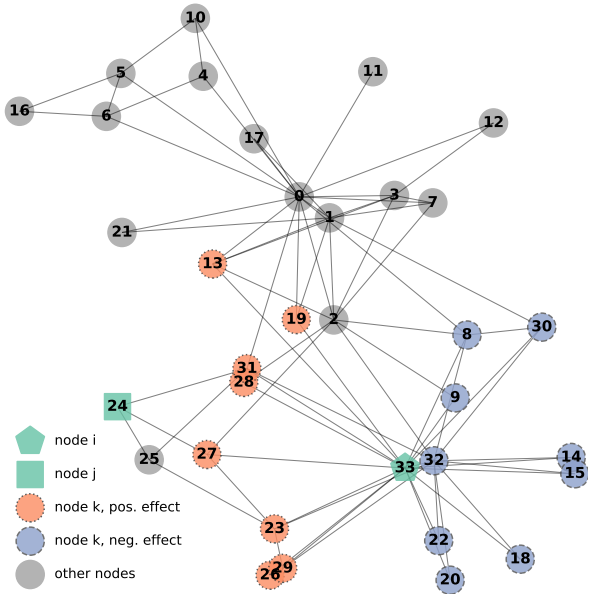

Figure 1. In Zachary's karate club network, we explain the predicted link between $i = 33$ and $j = 24$. The colored nodes with dashed or dotted edges are the neighbors of node 33. According to ExplaiNE, the links to the orange nodes with dotted edge have a positive effect on the probability of link $\{i, j\}$ to exist, while the effect of the links to blue nodes with dashed edge is negative.

links. We then reduce its scope to explanations of the type used in the example above (i.e., only incident links), and make an approximation (which we justify empirically), in order to obtain a still generic but highly scalable approach.

Next we apply ExplaiNE to Conditional Network Embedding (CNE; Kang et al., 2019), a recent state-of-the-art NE method. The application of ExplaiNE to CNE is particularly transparent, thanks to the mathematical elegance of the CNE model and 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 (Tang et al., 2015b), DeepWalk (Perozzi et al., 2014), PTE (Tang et al., 2015a), and node2vec (Grover & Leskovec, 2016).

Contributions. The main contributions are:

- ExplaiNE, a mathematically principled counterfactual reasoning approach for explaining link predictions based on network embeddings (Sec. 2.2).
- A scalable tight approximation of ExplaiNE (Sec. 2.3).
- A detailed application of ExplaiNE to CNE (Sec. 2.4)
- An outline of how to apply ExplaiNE to NE methods based on skip gram with negative sampling (Sec. 2.5)
- Quantitative and run time analyses showing the stability and scalability of the approximation. (Sec. 3.1,3.4)
- Qualitative and quantitative realistic case studies confirming the usefulness of ExplaiNE. (Sec. 3.2,3.3)

2. Methods

To introduce ExplaiNE in full generality, we first provide a simple but generic description of NE-based link prediction methods in Section 2.1. We then formalize ExplaiNE in a generic manner in Section 2.2, before describing a scalable approximation in Section 2.3. In Section 2.4 we develop ExplaiNE in detail for CNE. In Section 2.5 we outline how ExplaiNE can also be applied to other popular NE methods. But before all that, we first introduce some notation.

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 (also known as edges). A link is denoted by an unordered node pair $\{i, j\} \in E$. Let $\boldsymbol{A}$ denote the 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 $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)' \in \mathbb{R}^{n \times d}$, with $\boldsymbol{X}^*$ denoting an optimal embedding for adjacency matrix $\boldsymbol{A}$ (suppressing the dependency of $\boldsymbol{X}$ on $\boldsymbol{A}$ for conciseness—see below), and similarly $\hat{\boldsymbol{X}}^*$ optimal for $\hat{\boldsymbol{A}}$.

2.1. Network Embedding-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[2] objective function $\mathcal{L}(\boldsymbol{A}, \boldsymbol{X})$ for the given adjacency matrix $\boldsymbol{A}$. Thus $\boldsymbol{X}^*$ must satisfy the following necessary condition for optimality:

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

Defining $\boldsymbol{F}(\boldsymbol{A}, \boldsymbol{X}) \triangleq \nabla_{\boldsymbol{X}} \mathcal{L}(\boldsymbol{A}, \boldsymbol{X})$, the 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. 2.5), 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 optimal w.r.t. $\boldsymbol{A}$. I.e., g^*_{ij} directly computes the link probability w.r.t. an optimal embedding for a specified adjacency matrix.

[2]Note that, although NE methods are often described for unweighted networks (i.e., a binary adjacency matrix), the objective $\mathcal{L}(\boldsymbol{A}, \boldsymbol{X})$ 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.

2.2. 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 $\{k, l\} \in E$. Then, if removing the link between them 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 $\{k, l\} \notin E$. Then, if adding a link between them 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 this predicted link.

Intuitively, adding or removing an existing link 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. 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 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. As there are potentially many local optima, this can change the optimal embedding entirely (even if initialized with the 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 explanations as 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, with a positive sign 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 with a negative sign 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:

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

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., Chiang, 1984). Rephrased for our specific setting, this theorem states (note that we are overloading the symbol $\boldsymbol{X}^*$ here to also signify a function):

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:*

$$\boldsymbol{X}^*(\hat{\boldsymbol{A}}) = \hat{\boldsymbol{X}}^*, \text{ and}$$
$$\boldsymbol{F}(\boldsymbol{A}, \boldsymbol{X}^*(\boldsymbol{A})) = \boldsymbol{0} \text{ for all } \boldsymbol{A} \in S,$$

and:

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

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 $\mathcal{L}$ 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 $\mathcal{L}$), this situation can be avoided by adding a regularizer to $\mathcal{L}$ 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} as follows:

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

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

2.3. 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 (see supplement Sec. 4.1)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 $\boldsymbol{x}^*_j$.[3] Thus, Eq. (2) can be written as:

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

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 $\boldsymbol{x}^*_j$. This means that the second term in Eq. (5) can be neglected.

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 optimal embeddings 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 can be written as:

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

For conciseness, let us define $\hat{\boldsymbol{F}}_i(\boldsymbol{A}, \boldsymbol{x}_i) \triangleq \nabla_{\boldsymbol{x}_i} \mathcal{L}(\boldsymbol{A}, \boldsymbol{x}_i, \hat{\boldsymbol{X}}^*_{(i)})$. 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 now use the implicit function theorem on this optimality condition to approximate $\frac{\partial \boldsymbol{x}^*_i}{\partial a_{ik}}$ as:

$$\frac{\partial \boldsymbol{x}^*_i}{\partial a_{ik}}(\hat{a}_{ik}) = -\boldsymbol{H}_i^{-1} \cdot \frac{\partial \hat{\boldsymbol{F}}_i}{\partial a_{ik}}. \tag{6}$$

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 $\mathcal{L}$ w.r.t. $\boldsymbol{x}_i$, evaluated at $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{X}}^*)$.

[3] This is true for all NE methods we are aware of, and thus hardly a limitation at all.

Putting Eqs. (6) and (5) (neglecting the second term as discussed) together, this yields:

$$\frac{\partial g^*_{ij}}{\partial a_{ik}}(\hat{\boldsymbol{A}}) = -\nabla_{\boldsymbol{x}_i} g_{ij} \left(\hat{\boldsymbol{X}}^*\right)^T \cdot \boldsymbol{H}_i^{-1} \cdot \frac{\partial \hat{\boldsymbol{F}}_i}{\partial a_{ik}}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{x}}^*_i). \tag{7}$$

Comparing Eq. (4) with Eq. (7) reveals the dramatic complexity reduction achieved: Inverting $\boldsymbol{H}_i \in \mathbb{R}^{d \times d}$ has a practical complexity of only $\mathcal{O}(d^3)$, which is entirely feasible given common dimensionalities used in the literature (often 128). The experiments will validate that the approximations made are entirely justified in practice.

2.4. ExplaiNE for Conditional Network Embedding

We now apply the generic ExplaiNE approach to Conditional Network Embedding (CNE), a specific NE method. Detailed derivations are deferred to the supplement Sec. 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 $\mathcal{L}$ in CNE is the log-probability of the network conditioned on the embedding:

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

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

$$P_{ij}(a_{ij} = 1|\boldsymbol{X}) = 1 - P_{ij}(a_{ij} = 0|\boldsymbol{X}) = \frac{P_{\hat{\boldsymbol{A}},ij}\mathcal{N}_{+,\sigma_1}(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|)}{P_{\hat{\boldsymbol{A}},ij}\mathcal{N}_{+,\sigma_1}(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|) + (1 - P_{\hat{\boldsymbol{A}},ij})\mathcal{N}_{+,\sigma_2}(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|)}, \tag{8}$$

where $\mathcal{N}_{+,\sigma}$ denotes a half-Normal distribution (Leone et al., 1961) 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)—see, e.g., Adriaens et al. (2017); van Leeuwen et al. (2016).

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. (8). 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.[4] First, we derive the optimality condition:

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

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. (7):[5]

$$\nabla_{\boldsymbol{x}_i} g_{ij}\left(\hat{\boldsymbol{X}}^*\right) = -\gamma(\boldsymbol{x}_i^* - \boldsymbol{x}_j^*)\hat{P}_{ij}^*(1 - \hat{P}_{ij}^*).$$

$$\begin{aligned}\boldsymbol{H}_i &= \nabla_{\boldsymbol{x}_i} \hat{\boldsymbol{F}}_i(\hat{\boldsymbol{A}}, \hat{\boldsymbol{x}}_i^*) \\ &= \gamma \boldsymbol{I} \sum_{l \neq i} (P_{il}^* - \hat{a}_{il}) \\ &\quad - \gamma^2 \sum_{l \neq i} (\boldsymbol{x}_i^* - \boldsymbol{x}_l^*)(\boldsymbol{x}_i^* - \boldsymbol{x}_l^*)' \hat{P}_{il}^*(1 - \hat{P}_{il}^*).\end{aligned}$$

$$\frac{\partial \hat{\boldsymbol{F}}_i}{\partial a_{ik}}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{x}}_i^*) = \gamma(\boldsymbol{x}_k^* - \boldsymbol{x}_i^*).$$

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 located 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 where there are 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.

2.5. 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 Sec. 3 of the supplement, we derive a concrete example for LINE (Tang et al., 2015b).

[4]Due to the limited space, here we only show how to apply approximated ExplaiNE to CNE, as the exact version is not used in the experiments except for the experiment validating the approximated version. For the application of exact ExplainNE to CNE, we refer the reader to the supplement Sec. 2. From now on, we drop the modifier 'approximated' when the context is clear.

[5]Detailed derivations are provided in the supplementary material Sec. 1 due to space constraints.

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.

Levy & Goldberg (2014) and Qiu et al. (2018) 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, Qiu et al. (2018) 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? **Q2** Does ExplaiNE give sensible explanations? **Q3** Does the proposed method scale?

All experiments are based on CNE with parameters $\sigma_1 = 1$, $\sigma_2 = 2$. Any weights associated to the links in the networks are ignored. We used the following networks.

Game of Thrones' (GoT) network.[6] 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.

DBLP co-authorship network (Tang et al., 2008).[7] Containing 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 32-dimensional embedding of this network.

MovieLens dataset (Harper & Konstan, 2016).[8] Containing 100,000 ratings by 943 users on 1,682 movies. The

[6]https://github.com/mathbeveridge/asoiaf

[7]DBLP dataset V10: https://aminer.org/citation

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

network is thus bipartite and consists of 943+1,682 nodes and 100,000 edges. The 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 16-dimensional embedding of this network.

In Sec. 3.1 we analyze the quality of the approximation. In Sec. 3.2, we conduct a qualitative analysis of explanations on the DBLP and MovieLens networks. In Sec. 3.3 we quantitatively analyze the quality of the explanations. Finally, in Sec. 3.4 we consider the scalability of ExplaiNE.

3.1. Quality of the ExplaiNE approximation

Before applying approximated ExplaiNE to real world dataset, 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 subset 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 2,3,4, and 7 differences out of 796 for rank 5, see supplement Sec. 4.2).

In the supplement Sec. 4.3 we also compared the complete ranking of the neighbors between the approximated and exact ExplaiNE versions, and this for the most probable link for every node (i.e., seeking explanations for links that are actually present in the network). We computed the normalized Kendall tau distance[9] between the ranked explanations given by approximated and exact ExplaiNE. The average normalized Kendall tau distance is 0.05 ± 0.08. For comparison, the average Kendall tau distance between a random ranking and exact ExplaiNE is 0.51 ± 0.15.

Now confident in its accuracy, we can now evaluate the behavior of approximated ExplaiNE on two realistic networks.

[9] https://en.wikipedia.org/wiki/Kendall_tau_distance

Table 1. Predicted/recommended collaborations for Eric P. Xing. The top link (author: Adams Wei Yu) predicted by CNE are explained through co-authors of Eric P. Xing that are also colleagues or co-authors of Adams Wei Yu. The most relevant five co-authors of Eric P. Xing also cover major parts of Adams Wei Yu's research interests: large scale optimization and deep learning.

Rank	Recommendations	Explain: 'Adams Wei Yu'
1	Adams Wei Yu	Hao Su
2	Jure Leskovec	Li Fei-Fei
3	Sunita Sarawagi	Suvrit Sra
4	Tong Zhang	Fan Li
5	Soumen Chakrabarti	Wei Dai

3.2. Qualitative evaluation

Here we apply ExplaiNE to explain the predicted links in two real world networks (the DBLP co-authorship and the MovieLens rating networks) to assess whether ExplaiNE gives sensible explanations to the predicted links (**Q2**).

DBLP network. In the co-authorship network, a predicted link between authors i and j suggests a collaboration between them. 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 ICML'19 general chair Eric P. Xing (node i), and compute the explanations for his top recommendation (node j): Adams Wei Yu. It turns out that the existing co-authors of Eric P. Xing identified by ExplaiNE as top-5 explanations for this recommendation (see Table 1) are either colleagues or co-authors of Adams Wei Yu, with a shared interest in large scale optimization and deep learning.

MovieLens network. In the rating network, a predicted link between a user i and movie j amounts to a recommendation of movie j to user i. In making this recommendation CNE did not have access to any 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. As an example, we computed the recommendation for the first user (uid=0) in the user list (See Table. 2). The top recommended movie is 'Batman' with genre tags 'Action', 'Adventure', 'Crime', and 'Drama'. The genres of the top explanations given by ExplainNE arguably have strongly overlapping genre tags (e.g., all top-5 are tagged with 'Action'). Moreover, the second-highest ranked explanation is 'Batman Forever'.

More case studies are given in the supplement Sec. 5. These results suggest that ExplaiNE gives sensible explanations. The next subsection aims to quantify these findings.

Table 2. Recommended movie to user uid=0. The top movie recommended by CNE (Batman) 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	Batman	Action, Adventure, Crime, Drama
2	E.T. the Extra-Terrestrial	Children's, Drama, Fantasy, Sci-Fi
3	The Secret of Roan Inish	Adventure

k	Explanations for 'Batman'	Genres
1	Supercop	Action, Thriller
2	Batman Forever	Action, Adventure, Comedy, Crime
3	The Crow	Action, Romance, Thriller
4	Full Metal Jacket	Action, Drama, War
5	Young Guns	Action, Comedy, Western

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 in this section, it is possible to use metadata to derive reasonable ground truth explanations, and compare with those.

DBLP network. Here, 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 perfectly legitimate 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 will of course not always be true, we postulate that it is sufficiently common for ExplaiNE—providing it works well—to highlight the other co-authors as explanations for the observed link $\{i, j\}$.

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. 7). We then took the top-r of this ranked list as predicted co-authors on the paper i co-authored with j. Based on this, we 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. Figure 2 shows the ROC curves for Eric P. Xing as node i and three author-list lengths. For comparison, also ROC curves computed based on a randomly ranked list is shown (as the size of the data is rather small, these are not always close to the diagonal). ROC curves for other nodes i as well as Precision-Recall curves can be found in the supplement. All results indicate that the explanations are remarkably effective at this task, indicating that ExplaiNE performs well.

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: 13.1 Gb for MovieLens and on DBLP we 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	time exact	time approx
Karate	34	2	0.03	1.8e−4
GoT	796	2	64.1	4.1e−4
GoT	796	8	1490.2	9.8e−4
MovieLens	2625	16	$\sim$ 1.63e6	6.8e−3
DBLP	23359	32	—	0.02

MovieLens network. A good explanation k of a predicted link between a movie-user pair $\{i, j\}$ should arguably 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—indicating the null hypothesis is false. A Kolmogorov-Smirnoff test indeed shows an extremely high significance (p-value numerically 0).

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 (7), on average over random pairs of nodes $\{i, j\}$. The runtime was measured on a PC with quad-core 2.7GHz Intel Core i5 and 16GB 1600MHz DDR3 RAM. Table 3, shows that approximated ExplaiNE is efficient and applicable to large networks with higher dimensionality, while exact ExplaiNE is not.

4. Related Work

LP, as an important network analysis task, has recently been extensively studied in the NE literature (Hamilton et al.,

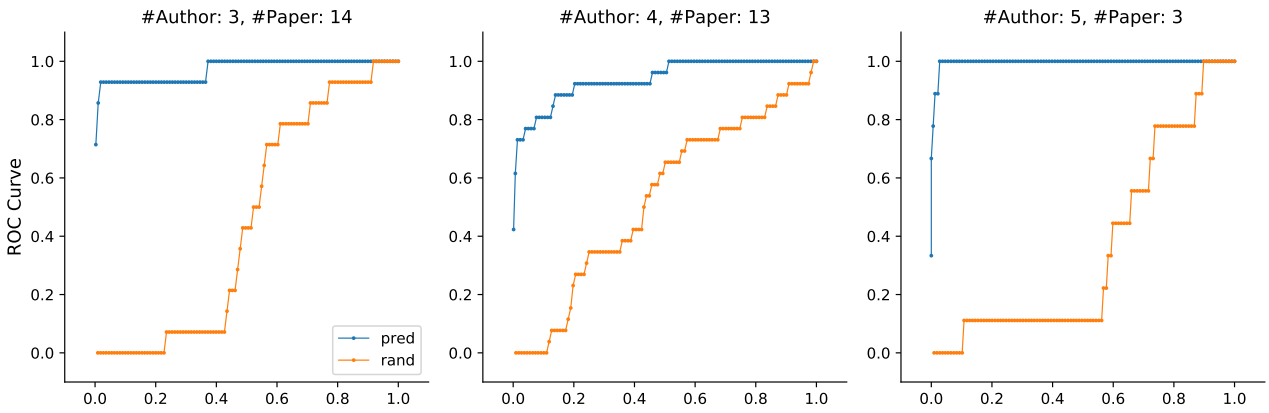

Figure 2. ROC curves of co-author predictions for $i =$'Eric P. Xing', with author-list lengths 3, 4, and 5 (orange=rand., blue=ExplaiNE).

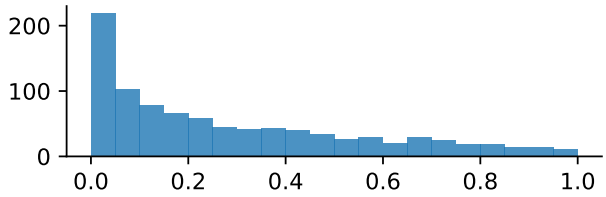

Figure 3. P-values that indicates the significance of the correlation between the genre recommended and the genres in the explanation. Each p-value is computed against 50 random explanations. Those explanations are drawn from user's watched movies. The empirical distribution has Kolmogorov-Smirnov test statistic 0.32 and a p-value that is numerically 0.0 against uniform distribution. This shows the significance of positive correlation between the recommended movies and the explanations made by ExplaiNE.

2017; Cui et al., 2018). By embedding the nodes in a vector space, the link prediction task can be addressed using traditional machine learning (ML) methods. This has led to new and accurate approaches for LP (Grover & Leskovec, 2016; Kang et al., 2019), but at the expense of explainability.

In parallel, the importance of accountability of AI has sparked growing research interest in interpretable ML. Approaches to interpretable machine learning research can be categorized into model-based and post-hoc approaches (Du et al., 2018; Murdoch et al., 2019). The first category focuses on incorporating interpretability (e.g., sparsity) 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). The most strongly related work (although not for LP) are Ribeiro et al. (2016) and Lundberg & Lee (2017), who provide a model-agnostic explanation via local approximation of the target model. More closely related, Simonyan et al. (2013) and Koh & Liang (2017) propose to compute the gradient of the loss function of a (black-box) model with respect to the input to gauge the relevance of the input features. The first of these computes the gradient using back-propagation, while the second approximates the gradient using a Taylor series expansion.

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. Moreover, to the best of our knowledge, ExplaiNE is the first method that uses the implicit function theorem 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) in networks is an important task, with applications to social networks, recommenders, and knowledge graphs. State-of-the-art approaches are based on first embedding the nodes in a vector space, followed by a LP step. Unfortunately, while accurate, these approaches offer no insight in their predictions. To remedy this, we introduced ExplaiNE, a generic approach to explain LPs based on Network Embeddings (NEs) in terms of existing links in the network. ExplaiNE is applicable for a wide range of NE methods. We applied it to CNE, a state-of-the-art NE method, and outlined how it can be applied for a wide range of other NE methods. Extensive qualitative and quantitative evaluations show the usefulness of ExplaiNE, and its ability to scale to large networks.

In the future we aim to develop ExplaiNE for other NE methods, apply it to recommender systems, and extend it to offer explanations in terms of the presence of dense communities or other larger substructures in the network.

ACKNOWLEDGMENTS

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, from the FWO (project no. G091017N, G0F9816N), and from the European Union's Horizon 2020 research and innovation programme and the FWO under the Marie Sklodowska-Curie Grant Agreement no. 665501.

References

Adriaens, F., Lijffijt, J., and De Bie, T. Subjectively interesting connecting trees. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 53–69. Springer, 2017.

Chiang, A. C. *Fundamental methods of mathematical economics*. Aukland (New Zealand) McGraw-Hill, 1984.

Cui, P., Wang, X., Pei, J., and Zhu, W. A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

Du, M., Liu, N., and Hu, X. Techniques for interpretable machine learning. *arXiv preprint arXiv:1808.00033*, 2018.

Grover, A. and Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864. ACM, 2016.

Hamilton, W. L., Ying, R., and Leskovec, J. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017.

Harper, F. M. and Konstan, J. A. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):19, 2016.

Kang, B., Lijffijt, J., and De Bie, T. Conditional network embeddings. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryepUj0qtX.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pp. 1885–1894, 2017.

Leone, F., Nelson, L., and Nottingham, R. The folded normal distribution. *Technometrics*, 3(4):543–550, 1961.

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

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

Martínez, V., Berzal, F., and Cubero, J.-C. A survey of link prediction in complex networks. *ACM Computing Surveys (CSUR)*, 49(4):69, 2017.

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

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019.

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

Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K., and Tang, J. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 459–467. ACM, 2018.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. ACM, 2016.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 990–998. ACM, 2008.

Tang, J., Qu, M., and Mei, Q. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1165–1174. ACM, 2015a.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 1067–1077. International World Wide Web Conferences Steering Committee, 2015b.

van Leeuwen, M., De Bie, T., Spyropoulou, E., and Mesnage, C. Subjective interestingness of subgraph patterns. *Machine Learning*, 105(1):41–75, 2016.

Zachary, W. W. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.

Supplementary materials for ExplaiNE

1. Approximated ExplaiNE for Conditional Network Embedding

We now apply approximated ExplaiNE approach to Conditional Network Embedding. Our goal is to compute gradient of predicted link probability g^*_{ij} with respect to the change of link a_{ik} (main paper Eq.7), namely:

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

In order to compute the derivative, we first drive the optimality conditions for CNE. Recall the objective function $\mathcal{L}$ in CNE is log-probability:

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

where the link probabilities conditioned on the embedding are defined as:

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

where $\mathcal{N}_{+,\sigma}$ denotes a half-Normal distribution (Leone et al., 1961) with spread parameter σ, where $\sigma_2 > \sigma_1 = 1$, 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 types of information about the structure of the network), derived as explained in Adriaens et al. (2017); van Leeuwen et al. (2016).

Denoting $\gamma = \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}$, and $\hat{P}^*_{ij} \triangleq g^*_{ij}(\hat{\boldsymbol{A}}) = P(a_{ij}=1|\hat{\boldsymbol{X}}^*)$ (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 terms in Eq. 1.

First term.. First term is derived by taking the gradient of predicted link probability $g_{ij}\left(\hat{\boldsymbol{X}}^*\right)$ with respect to $\boldsymbol{x}_i$:

$$
\begin{aligned}
\nabla_{\boldsymbol{x}_i} g_{ij}\left(\hat{\boldsymbol{X}}^*\right) &= \nabla_{\boldsymbol{x}_i} P(a_{ij}=1|\boldsymbol{X}) \\
&= \nabla_{\boldsymbol{x}_i}\left(\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^*\|)}\right) \\
&= \frac{\nabla_{\boldsymbol{x}_i}\left(P_{\hat{\boldsymbol{A}},ij}\mathcal{N}_{+,\sigma_1}(\|\boldsymbol{x}_i^*-\boldsymbol{x}_j^*\|)\right)\left(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^*\|)\right)}{\left(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^*\|)\right)^2} \\
&- \frac{\left(P_{\hat{\boldsymbol{A}},ij}\mathcal{N}_{+,\sigma_1}(\|\boldsymbol{x}_i^*-\boldsymbol{x}_j^*\|)\right)\nabla_{\boldsymbol{x}_i}\left(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^*\|)\right)}{\left(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^*\|)\right)^2} \\
&= \frac{\frac{-1}{\sigma_1^2}(\boldsymbol{x}_i^*-\boldsymbol{x}_j^*)\left(P_{\hat{\boldsymbol{A}},ij}\mathcal{N}_{+,\sigma_1}(\|\boldsymbol{x}_i^*-\boldsymbol{x}_j^*\|)\right)\left(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^*\|)\right)}{\left(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^*\|)\right)^2} \\
&- \frac{\left(P_{\hat{\boldsymbol{A}},ij}\mathcal{N}_{+,\sigma_1}(\|\boldsymbol{x}_i^*-\boldsymbol{x}_j^*\|)\right)\left(\frac{-1}{\sigma_1^2}(\boldsymbol{x}_i^*-\boldsymbol{x}_j^*)\left(P_{\hat{\boldsymbol{A}},ij}\mathcal{N}_{+,\sigma_1}(\|\boldsymbol{x}_i^*-\boldsymbol{x}_j^*\|)\right)+\frac{-1}{\sigma_2^2}(1-P_{\hat{\boldsymbol{A}},ij})\mathcal{N}_{+,\sigma_2}(\|\boldsymbol{x}_i^*-\boldsymbol{x}_j^*\|)\right)}{\left(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^*\|)\right)^2} \\
&= -\gamma\frac{P_{\hat{\boldsymbol{A}},ij}(1-P_{\hat{\boldsymbol{A}},ij})\mathcal{N}_{+,\sigma_1}(\|\boldsymbol{x}_i^*-\boldsymbol{x}_j^*\|)\mathcal{N}_{+,\sigma_2}(\|\boldsymbol{x}_i^*-\boldsymbol{x}_j^*\|)}{\left(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^*\|)\right)^2} \\
&= -\gamma(\boldsymbol{x}_i^*-\boldsymbol{x}_j^*)P(a_{ij}=1|\hat{\boldsymbol{X}}^*)P(a_{ij}=0|\hat{\boldsymbol{X}}^*) \\
&= -\gamma(\boldsymbol{x}_i^*-\boldsymbol{x}_j^*)\hat{P}_{ij}^*(1-\hat{P}_{ij}^*)
\end{aligned} \tag{2}
$$

Second term. The Hessian $\boldsymbol{H}_i$ the gradient of $\hat{\boldsymbol{F}}_i(\hat{\boldsymbol{A}},\hat{\boldsymbol{x}}_i^*)$ w.r.t. $\boldsymbol{x}_i^*$. Recall the optimality condition: The optimality condition in approximated ExplaiNE can be derived as:

$$
\begin{aligned}
\hat{\boldsymbol{F}}_i(\hat{\boldsymbol{A}},\hat{\boldsymbol{x}}_i^*) &= \nabla_{\boldsymbol{x}_i^*}\log(P(\hat{\boldsymbol{A}}|\hat{\boldsymbol{X}}^*)) \\
&= \gamma\sum_{j\neq i}(\hat{\boldsymbol{x}}_i^*-\hat{\boldsymbol{x}}_j^*)\left(P\left(a_{ij}=1|\hat{\boldsymbol{X}}^*\right)-\hat{a}_{ij}\right) \\
&= \mathbf{0}.
\end{aligned} \tag{3}
$$

Based on this condition we derive the Hessian:

$$
\begin{aligned}
\boldsymbol{H}_i &= \nabla_{\boldsymbol{x}_i}\hat{\boldsymbol{F}}_i(\hat{\boldsymbol{A}},\hat{\boldsymbol{x}}_i^*) \\
&= \nabla_{\boldsymbol{x}_i}\left(\gamma\sum_{j\neq i}(\hat{\boldsymbol{x}}_i^*-\hat{\boldsymbol{x}}_j^*)\left(P\left(a_{ij}=1|\hat{\boldsymbol{X}}^*\right)-\hat{a}_{ij}\right)\right) \\
&= \gamma\left(\boldsymbol{I}\sum_{j\neq i}\left(P_{ij}^*-\hat{a}_{ij}\right)-\gamma\sum_{j\neq i}(\boldsymbol{x}_i^*-\boldsymbol{x}_j^*)(\boldsymbol{x}_i^*-\boldsymbol{x}_j^*)'P_{ij}^*(1-P_{ij}^*)\right).
\end{aligned} \tag{4}
$$

The last line follows similar derivation of Eq. 2.

Third term. Now compute the gradient with respect to a_{ik} from optimality condition (Eq. 3):

$$
\begin{aligned}
\frac{\partial\hat{\boldsymbol{F}}_i}{\partial\hat{a}_{ik}}(\hat{\boldsymbol{A}},\hat{\boldsymbol{x}}_i^*) &= \frac{\partial}{\partial\hat{a}_{ik}}\left(\gamma\sum_{j\neq i}(\hat{\boldsymbol{x}}_i^*-\hat{\boldsymbol{x}}_j^*)\left(P\left(a_{ij}=1|\hat{\boldsymbol{X}}^*\right)-\hat{a}_{ij}\right)\right) \\
&= \gamma(\boldsymbol{x}_k^*-\boldsymbol{x}_i^*).
\end{aligned} \tag{5}
$$

Putting things together, we have

$$\begin{aligned}&\frac{\partial g_{ij}^*}{\partial a_{ik}}(\hat{\boldsymbol{A}})\\&= (\boldsymbol{x}_i^* - \boldsymbol{x}_j^*)' \cdot \left(\frac{-\boldsymbol{H}_i}{\gamma^2 \hat{P}_{ij}^*(1-\hat{P}_{ij}^*)}\right)^{-1} \cdot (\boldsymbol{x}_i^* - \boldsymbol{x}_k^*)\end{aligned} \tag{6}$$

2. Exact ExplaiNE for Conditional Network Embedding

We now apply exact ExplaiNE approach to Conditional Network Embedding (CNE). Our goal is to compute the change (i.e., gradient) of link probability of a node paire $\{i, j\}$ with respect to node-pairs $\{k, l\}$ ($k \neq l$ and $\{k, l\} \neq \{i, j\}$), i.e., the explanations. Applying the chain rule and implicit function theorem, the exact ExplaiNE computes this gradient as (main paper Eq.4):

$$\frac{\partial g_{ij}^*}{\partial a_{kl}}(\hat{\boldsymbol{A}}) = -\nabla_{\boldsymbol{X}} g_{ij}\left(\hat{\boldsymbol{X}}^*\right)^T \cdot \boldsymbol{H}^{-1} \cdot \frac{\partial \boldsymbol{F}}{\partial a_{kl}}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{X}}^*). \tag{7}$$

Now let's derive each term in Eq. 7 for CNE.

First term. Because in CNE the link function $g_{ij}^*(\hat{\boldsymbol{A}})$ only depends on $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, the gradient $\nabla_{\boldsymbol{X}} g_{ij}\left(\hat{\boldsymbol{X}}^*\right)$ is a $nd \times 1$ vector with $2d$ non-zero entries. Formally:

$$\nabla_{\boldsymbol{X}} g_{ij}\left(\hat{\boldsymbol{X}}^*\right) = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \nabla_{\boldsymbol{x}_i} g_{ij}\left(\hat{\boldsymbol{X}}^*\right) \\ \vdots \\ \nabla_{\boldsymbol{x}_j} g_{ij}\left(\hat{\boldsymbol{X}}^*\right) \\ \vdots \\ \mathbf{0} \end{bmatrix} \tag{8}$$

where

$$\nabla_{\boldsymbol{x}_i} g_{ij}\left(\hat{\boldsymbol{X}}^*\right) = -\gamma(\boldsymbol{x}_i^* - \boldsymbol{x}_j^*)\hat{P}_{ij}^*(1-\hat{P}_{ij}^*),$$

and

$$\nabla_{\boldsymbol{x}_j} g_{ij}\left(\hat{\boldsymbol{X}}^*\right) = -\gamma(\boldsymbol{x}_j^* - \boldsymbol{x}_i^*)\hat{P}_{ij}^*(1-\hat{P}_{ij}^*).$$

Second term. We drive the Hessian $\boldsymbol{H}$ ($nd \times nd$) as:

$$\begin{aligned}\boldsymbol{H} &= \nabla_{\boldsymbol{X}} \boldsymbol{F}(\hat{\boldsymbol{A}}, \boldsymbol{X}^*)\\ &= \begin{bmatrix} \ddots & & & & \\ & \nabla_{\boldsymbol{x}_i^*}\boldsymbol{F}_i(\hat{\boldsymbol{A}}, \boldsymbol{x}_i^*) & \dots & \nabla_{\boldsymbol{x}_i^*}\boldsymbol{F}_j(\hat{\boldsymbol{A}}, \boldsymbol{x}_j^*) & \\ & \vdots & \ddots & \vdots & \\ & \nabla_{\boldsymbol{x}_j^*}\boldsymbol{F}_i(\hat{\boldsymbol{A}}, \boldsymbol{x}_i^*) & \dots & \nabla_{\boldsymbol{x}_j^*}\boldsymbol{F}_j(\hat{\boldsymbol{A}}, \boldsymbol{x}_j^*) & \\ & & & & \ddots \end{bmatrix},\end{aligned} \tag{9}$$

where

$$\nabla_{\boldsymbol{x}_i^*}\boldsymbol{F}_i(\hat{\boldsymbol{A}}, \boldsymbol{x}_i^*) = \gamma \boldsymbol{I} \sum_{j \neq i}(\hat{P}_{ij}^* - \hat{a}_{ij}) - \gamma^2 \sum_{j \neq i}(\boldsymbol{x}_i^* - \boldsymbol{x}_j^*)(\boldsymbol{x}_i^* - \boldsymbol{x}_j^*)'\hat{P}_{ij}^*(1-\hat{P}_{ij}^*)$$

and

$$\nabla_{\boldsymbol{x}_j^*}\boldsymbol{F}_j(\hat{\boldsymbol{A}}, \boldsymbol{x}_i^*) = -\gamma \boldsymbol{I}(\hat{P}_{ij}^* - \hat{a}_{ij}) + \gamma^2(\boldsymbol{x}_i^* - \boldsymbol{x}_j^*)(\boldsymbol{x}_i^* - \boldsymbol{x}_j^*)'\hat{P}_{ij}^*(1-\hat{P}_{ij}^*)$$

Third term. Finally, we can compute derivative $\frac{\partial \boldsymbol{F}}{\partial a_{kl}}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{X}}^*)$ $(nd \times 1)$ as:

$$\frac{\partial \boldsymbol{F}}{\partial a_{kl}}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{X}}^*) = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \gamma(\boldsymbol{x}_k^* - \boldsymbol{x}_l^*) \\ \vdots \\ \gamma(\boldsymbol{x}_l^* - \boldsymbol{x}_k^*) \\ \vdots \\ \mathbf{0} \end{bmatrix} \tag{10}$$

Putting all together, we have gradient of link probability $g_{ij}^*(\hat{\boldsymbol{A}})$ with respect to a link a_{kl}:

$$\begin{aligned} \frac{\partial g_{ij}^*}{\partial a_{kl}}(\hat{\boldsymbol{A}}) &= -\nabla_{\boldsymbol{X}} g_{ij}\left(\hat{\boldsymbol{X}}^*\right)^T \cdot \boldsymbol{H}^{-1} \cdot \frac{\partial \boldsymbol{F}}{\partial a_{kl}}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{X}}^*) \\ &= \begin{bmatrix} \mathbf{0} \\ \vdots \\ \nabla_{\boldsymbol{x}_i} g_{ij}\left(\hat{\boldsymbol{X}}^*\right) \\ \vdots \\ \nabla_{\boldsymbol{x}_j} g_{ij}\left(\hat{\boldsymbol{X}}^*\right) \\ \vdots \\ \mathbf{0} \end{bmatrix}' \cdot \begin{bmatrix} \ddots & & & & \\ & -\nabla_{\boldsymbol{x}_i^*} \boldsymbol{F}_i(\hat{\boldsymbol{A}}, \boldsymbol{x}_i^*) & \dots & -\nabla_{\boldsymbol{x}_i^*} \boldsymbol{F}_j(\hat{\boldsymbol{A}}, \boldsymbol{x}_j^*) & \\ & \vdots & \ddots & \vdots & \\ & -\nabla_{\boldsymbol{x}_j^*} \boldsymbol{F}_i(\hat{\boldsymbol{A}}, \boldsymbol{x}_i^*) & \dots & -\nabla_{\boldsymbol{x}_j^*} \boldsymbol{F}_j(\hat{\boldsymbol{A}}, \boldsymbol{x}_j^*) & \\ & & & & \ddots \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} \\ \vdots \\ \gamma(\boldsymbol{x}_k^* - \boldsymbol{x}_l^*) \\ \vdots \\ \gamma(\boldsymbol{x}_l^* - \boldsymbol{x}_k^*) \\ \vdots \\ \mathbf{0} \end{bmatrix} \end{aligned}$$

3. Approximated ExplaiNE for LINE

LINE (Tang et al., 2015b) computes embeddings that approximate both 1st and 2nd order proximity. The model consists of two objective functions. The first one measures the distance (KL) between the model and empirical distribution of the 1st proximity measure. Similarly the second objective function measures the distance (KL) between the model and empirical distribution of the 2nd proximity measure. Optimizing over the objective functions thus gives two embeddings for an input word, one for each type of proximities. Then the embeddings are simply concatenated and used as the final node embedding. Usually the context embedding in the 2nd proximity approximation are not used in the practice, similar to the situation in the language modeling (e.g., skip-gram with negative sampling a.k.a SGNS).

Similar to CNE, we assume the given network is unweighted, undirected. Explaining the first order proximity embedding (LINE 1st) is simple, thus our derivation assumes the link prediction is based solely on the LINE 2nd embedding, adopt the reasoning in the works (Levy & Goldberg, 2014; Qiu et al., 2018). For LINE 2nd, the probability of encountering "context" node v_j given node v_i is defined as,

$$p(v_j|v_i) = \frac{\exp(\boldsymbol{x}_i' \boldsymbol{y}_j)}{\sum_k \exp(\boldsymbol{x}_i' \boldsymbol{y}_k)}$$

where $\boldsymbol{x}_i$ denotes the embedding of a target node i and $\boldsymbol{y}_j$ denotes the embedding of context node j.

Then the 2nd proximity objective reads:

$$O = \sum_{i \in V} d_i \mathrm{KL}\left(\hat{p}(\cdot|v_i) \| p(\cdot|v_i)\right)$$

where $d_i = \sum_j a_{ij}$ weights the importance of i-th node, and empirical distribution $\hat{p}(\cdot|v_i)$ is defined as $\hat{p}(v_j|v_i) = \frac{a_{ij}}{d_i}$. Carrying out the KL-divergence, the objective function can be further expressed as:

$$O = \sum_{(i,j) \in E} a_{ij} \log p(v_j|v_i)$$

Because the normalization factor of the distribution is computationally expensive, the objective function is approximated using negative sampling (Mikolov et al., 2013):

$$\mathcal{L} = \sum_{(i,j)\in E} a_{ij} \log g(\boldsymbol{x}_i' \boldsymbol{y}_j) + b\mathbb{E}_{j'\sim P_N} \left[\log g\left(-\boldsymbol{x}_i' \boldsymbol{y}_{j'}\right)\right]$$

where b is the number of negative samples, $g(z)$ is the sigmoid function. The expectation term can be expressed as

$$\mathbb{E}_{j'\sim P_N} \left[\log g\left(-\boldsymbol{x}_i' \boldsymbol{y}_{j'}\right)\right] = \sum_{j'} \frac{d_{j'}}{\text{vol}(G)} \log g\left(-\boldsymbol{x}_i' \boldsymbol{y}_{j'}\right)$$

Then the objective function of LINE (2nd) can be expressed as:

$$\mathcal{L} = \sum_{(i,j)\in E} a_{ij} \log g(\boldsymbol{x}_i' \boldsymbol{y}_j) + b\frac{d_i d_j}{\text{vol}(G)} \log g(-\boldsymbol{x}_i' \boldsymbol{y}_j)$$

Denote $g_{ij}^+ = g(\boldsymbol{x}_i' \boldsymbol{y}_j)$ and $g_{ij}^- = g(-\boldsymbol{x}_i' \boldsymbol{y}_j)$ we can compute the gradient of $\mathcal{L}$ with respect to $\boldsymbol{x}_i$ as:

$$\nabla_{\boldsymbol{x}_i} \mathcal{L} = (\sum_j a_{ij} g_{ij}^- - b\frac{d_i d_j}{\text{vol}(G)} g_{ij}^+) \cdot \boldsymbol{y}_j$$

Equating the gradient to zero we have implicit function given by optimal embedding $\boldsymbol{x}_i^*$ and $\boldsymbol{y}_j^*$:

$$\begin{aligned} \boldsymbol{F}_i(\hat{\boldsymbol{A}}, \hat{\boldsymbol{x}}_i) = \nabla_{\boldsymbol{x}_i} L_{ij} &= (\sum_j \hat{a}_{ij} g_{ij}^- - b\frac{d_i d_j}{\text{vol}(G)} g_{ij}^+) \cdot \boldsymbol{y}_j^* \\ &= \boldsymbol{0} \end{aligned}$$

Our goal is to explain the predicted link probability g_{ij}^* with respect the change of an edge $\hat{a}_{kl}$. Assume the function function g to be the logistic regression classifier (with parameter $\boldsymbol{w}$ and b) trained on edge embedding that combines two node embedding using Hadamard operator (i.e., element wise product denoted as $\boldsymbol{x}_i^* \circ \boldsymbol{x}_j^*$) (Grover & Leskovec, 2016):

$$g_{ij}^*(\hat{\boldsymbol{X}}^*) = \frac{1}{1 + e^{-(\boldsymbol{w}'(\boldsymbol{x}_i^* \circ \boldsymbol{x}_j^*) + b)}} \tag{11}$$

Our goal is compute gradient:

$$\frac{\partial g_{ij}^*}{\partial a_{ik}}(\hat{\boldsymbol{X}}^*) = -\nabla_{\boldsymbol{x}_i} g_{ij} \left(\hat{\boldsymbol{X}}^*\right)^T \cdot \boldsymbol{H}_i^{-1} \cdot \frac{\partial \hat{\boldsymbol{F}}_i}{\partial a_{ik}}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{x}}_i^*). \tag{12}$$

First term in the derivative:

$$\nabla_{\boldsymbol{x}_i} g_{ij}^*(\hat{\boldsymbol{X}}^*) = g_{ij}^*(1 - g_{ij}^*) \cdot (\boldsymbol{w} \circ \boldsymbol{x}_j^*)$$

Second term in the derivative:

$$\begin{aligned} \boldsymbol{H}_i &= \nabla_{\boldsymbol{x}_i} \boldsymbol{F}_i(\hat{\boldsymbol{A}}, \hat{\boldsymbol{x}}_i^*) \\ &= \nabla_{\boldsymbol{x}_i} \left((\sum_j \hat{a}_{ij} g_{ij}^- - b\frac{d_i d_j}{\text{vol}(G)} g_{ij}^+) \cdot \boldsymbol{y}_j^* \right) \\ &= \sum_j -\hat{a}_{ij} g_{ij}^- (1 - g_{ij}^-) \boldsymbol{y}_j^* (\boldsymbol{y}_j^*)' - b\frac{d_i d_j}{\text{vol}(G)} g_{ij}^+ (1 - g_{ij}^+) \boldsymbol{y}_j^* (\boldsymbol{y}_j^*)' \\ &= -\sum_j \left(\hat{a}_{ij} + \frac{d_i d_j}{\text{vol}(G)} \right) g_{ij}^+ g_{ij}^- \boldsymbol{y}_j^* (\boldsymbol{y}_j^*)' \end{aligned}$$

The third term reads:

$$\begin{aligned}\frac{\partial \boldsymbol{F}_i}{\partial \hat{a}_{ik}}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{x}}_i^*) &= \frac{\partial}{\partial \hat{a}_{ik}}\left(\sum_j \hat{a}_{ij} g_{ij}^- \boldsymbol{y}_j^* - b\frac{d_i d_j}{\text{vol}(G)} g_{ij}^+ \cdot \boldsymbol{y}_j^*\right) \\ &= g_{ik}^- \boldsymbol{y}_k^* - b\left(\frac{\partial}{\partial \hat{a}_{ij}}\frac{d_i}{\text{vol}(G)} \cdot \sum_j d_j g_{ij}^+ \boldsymbol{y}_j^*\right) \\ &= g_{ik}^- \boldsymbol{y}_k^* - b\left(\frac{\text{vol}(G) - d_i}{\text{vol}(G)^2} \cdot \left(\sum_j d_j g_{ij}^+ \boldsymbol{y}_j^*\right) + \frac{d_i}{\text{vol}(G)} g_{ik}^+ \boldsymbol{y}_k^*\right) \\ &= \left(g_{ik}^- - \frac{b d_i}{\text{vo}(G)} g_{ik}^+\right)\boldsymbol{y}_k^* - \frac{b(\text{vol}(G) - d_i)}{\text{vol}(G)^2} \cdot \sum_j d_j g_{ij}^+ \boldsymbol{y}_j^*\end{aligned}$$

where $\frac{\partial d_i}{\partial a_{ik}} = 1$, $\frac{\partial \text{vol}(G)}{\partial a_{ik}} = 1$, $\frac{\partial d_j}{\partial a_{ik}} = \begin{cases} 1, \text{for } j = k \\ 0, \text{otherwise} \end{cases}$.

More General, the loss function based on term $\frac{\#(w,c)}{\#w\frac{\#c}{|C|}}$ is applicable for all skip gram with negative sampling (SGNS) model based methods (LINE, PTE, DeepWalk, node2vec) (Levy & Goldberg, 2014; Qiu et al., 2018). Based on different sampling strategy, $\frac{\#(w,c)}{\#w\frac{\#c}{|C|}}$ becomes different functions of $\hat{a}_i j$. For example, in LINE this is ratio $\frac{\sum_i a_{ij} \cdot \sum_j a_{ij}}{\sum_{ij} a_{ij}}$. PTE is similar to but extended to multiple networks. In DeepWalk and node2vec, this expression converges probabilistically to entries of function of random walk matrix $\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{A}$. So in theory, the idea of ExplaiNE can be extended to all SGNS based methods.

4. Quality of the ExplaiNE approximation

In this section we describe the detailed evaluation about the quality of the ExplaiNE approximation. The results are obtained on GoT network with embedding dimensionality 2.

4.1. Compare the gradient of incident nodes v.s. non-incident nodes

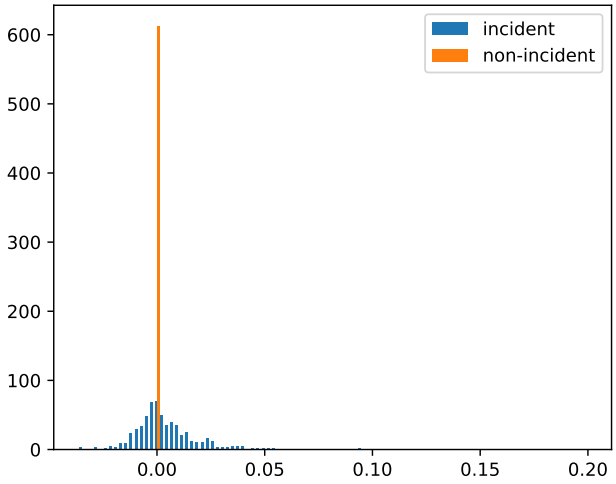

Figure 4. Gradient values of (top) predicted link probability of 100 randomly sampled nodes with respect to incident nodes (blue bars) versus non-incident nodes (orange bars). The gradient values of non-incident node pairs (on average $-2.13\text{e}{-7} \pm 6.96\text{e}{-5}$)are mush smaller than the gradient values of the incident node pairs (on average 0.005 ± 0.02). This validates our assumption that the derivative of the predicted link probability is higher w.r.t incident node pairs $\{i, k\}$ than w.r.t non-incident node pairs $\{k, l\}$..

Our first goal is to validate the assumption (main paper Sec.2.3) that the derivative of the predicted link probability is

higher w.r.t incident node pairs $\{i, k\}$ than w.r.t non-incident node pairs $\{k, l\}$. We randomly sampled 100 nodes, for each node i we compute (using exact ExplaiNE) the gradient of the link probability of i's top predicted link $\{i, j\}$ with respect to $i's$ incident pairs $\{i, k\}$. These values summarized in the histogram (Fig.4) with blue bars. As comparison, we also compute (using exact ExplaiNE) for each node i the gradient of predicted probability of link $\{i, j\}$ with respect to a random sample (same number as i's incident nodes) of non-incident pairs $\{k, l\}$ ($k \neq l$ and $\{k, l\} \neq \{i, j\}$). These values are also summarized in the histogram (Fig.4) with orange bars. The plot shows the gradient values of non-incident node pairs (on average $-2.13\text{e}{-7} \pm 6.96\text{e}{-5}$)are mush smaller than the gradient values of the incident node pairs (on average 0.005 ± 0.02). This validates our assumption.

4.2. Evaluate ExplaiNE approximation on predicted links

In this section, we 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.

The 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. 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 subset of K neighbors would share l elements with probability $\frac{\binom{K}{l}\binom{m-K}{K-l}}{\binom{m}{K}}$, which is large for large l if m is not much larger than K.

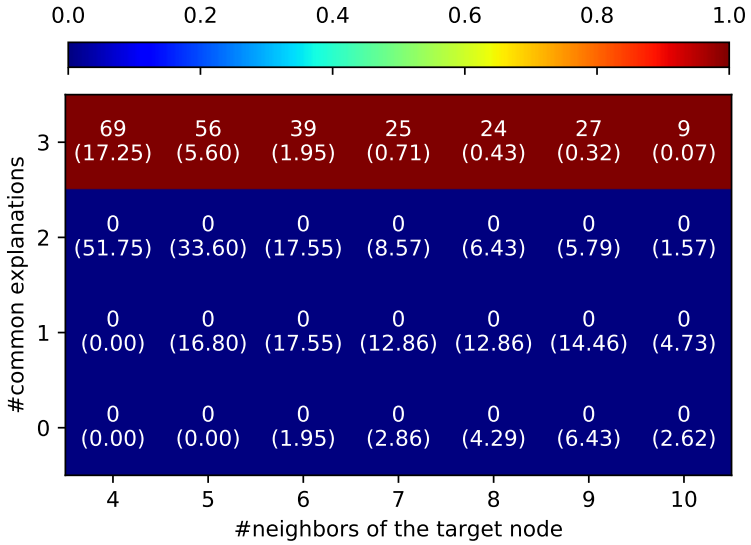

Figure 5. Histogram for $K = 3$ degrees up to 10 (668 out of 796 nodes have smaller or equal degrees). It reveals that the top-3 explanations given by approximated ExplaiNE in almost all cases completely overlaps with the explanation given by exact ExplaiNE.

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 plot the histogram for $K = 3$ degrees up to 10 (668 out of 796 nodes have smaller or equal degrees) (Fig. 5) and for $K = 5$ degrees up to 12 (691 out of 796 nodes have smaller or equal degrees) (Fig. 6). Both histogram revealed that the top-K explanations given by approximated ExplaiNE in almost all cases completely overlaps with the explanation given by exact ExplaiNE.

For completeness, we counted the cases of perfect overlapping up to the largest node degree 122 for $K = 1, \ldots, 5$, summarized in Table. 4. The result agains shows the almost perfect overlapping between the explanations given by approximated ExplainNE and exact ExplainNE. Note for larger neighborhood size, the more probable the top explanations contains more noise, thus the top ranked explanations have lower similarities (e.g. in $K = 5$ case) with the exact version.

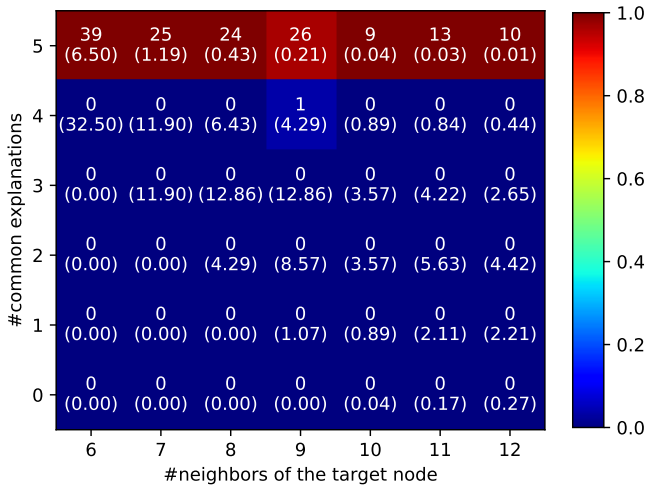

Figure 6. Histogram for $K = 5$ degrees up to 12 (691 out of 796 nodes have smaller or equal degrees). It reveals that the top-5 explanations given by approximated ExplaiNE in almost all cases completely overlaps with the explanation given by exact ExplaiNE.

Table 4. Number of perfect overlapping between the explanations given by approximated ExplainNE and exact ExplainNE. Count up to the largest node degree 122 for $K = 1, \ldots, 5$.

	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$
Number of perfect overlaps out of 796 cases	796	794	793	794	789

4.3. Evaluate ExplaiNE approximation on existing links

Although the top-K explanation are arguably more relevant than the tail of the ranking, for completeness we also compared the total ranking of the neighbors between the approximated and exact ExplaiNE versions, and this for the top predicted link as well as for all existing links (i.e. seeking explanations for links that are actually present in the network). More specifically, for each node's top existing link (according to the link probability) we computed the normalized Kendall tau distance between the ranked explanation given by approximated ExplaiNE and exact ExplaiNE. We compared this with the normalized Kendall tau distance measured between a random ranking and the ranking given by exact ExplaiNE (See Fig. 7).The average normalized Kendall tau distance between explanations given by approximated and exact ExplaiNE is 0.008 ± 0.04 and, for comparison, the average between a random ranking and exact ExplaiNE is 0.49 ± 0.29. These results again confirm the accuracy (here, in terms of ranking distance) of approximated ExplaiNE well approximates exact version.

5. Qualitative study

DBLP network. In the co-authorship network, a predicted link between authors i and j suggests a collaboration between them. 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 first example, we predict links for ICML'19 general chair Eric P. Xing (node i), and compute the explanations for his top recommendation (node j): Adams Wei Yu. It turns out that the existing co-authors of Eric P. Xing identified by ExplaiNE as top-5 explanations for this recommendation (see Table 5) are either colleagues or coauthors of Adams Wei Yu, with a shared interest in large scale optimization and deep learning.

As second example, we predict links for ICML'19 program chair Kamalika Chaudhuri (node i), and compute the explanations for his top recommendation (node j): Matus Telgarsky. It turns out that the existing co-authors of Kamalika Chaudhuri identified by ExplaiNE as top-5 explanations for this recommendation (see Table 6) are coauthors, advisors, or colleagues of Matus Telgarsky, with a shared interest in deep learning.

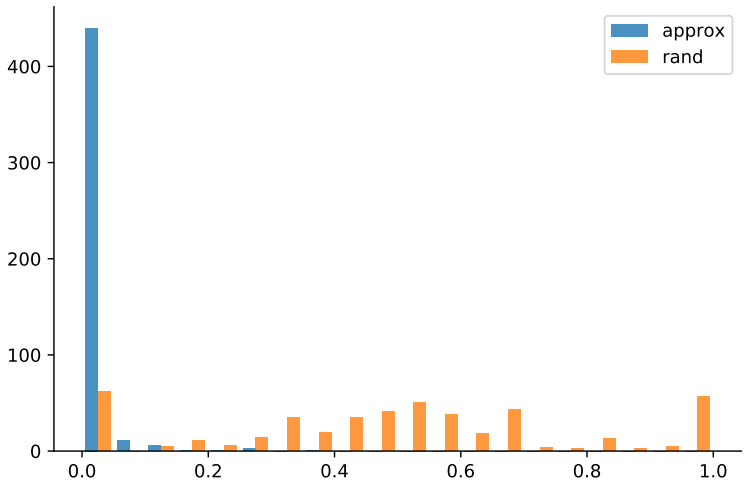

Figure 7. For each node's top existing link (according to the link probability) we computed the normalized Kendall tau distance between the ranked explanation given by approximated ExplaiNE and exact ExplaiNE. We compared this with the normalized Kendall tau distance measured between a random ranking and the ranking given by exact ExplaiNE. The average normalized Kendall tau distance between explanations given by approximated and exact ExplaiNE is 0.008 ± 0.04 and, for comparison, the average between a random ranking and exact ExplaiNE is 0.49 ± 0.29. This again confirms the accuracy (here, in terms of ranking distance) of approximated ExplaiNE well approximates exact version.

Table 5. Predicted/recommended collaborations for Eric P. Xing. The top link (author: Adams Wei Yu) predicted by CNE are explained through co-authors of Eric P. Xing that are also colleagues or co-authors of Adams Wei Yu. The most relevant five co-authors of Eric P. Xing also cover major parts of Adams Wei Yu's research interests: large scale optimization and deep learning.

Rank	Recommendations	Explain: 'Adams Wei Yu'
1	Adams Wei Yu	Hao Su
2	Jure Leskovec	Li Fei-Fei
3	Sunita Sarawagi	Suvrit Sra
4	Tong Zhang	Fan Li
5	Soumen Chakrabarti	Wei Dai

As thrid example, we predict links for ICML'19 program chair Ruslan Salakhutdinov (node i), and compute the explanations for his top recommendation (node j): Rich Zemel. It turns out that the existing co-authors of Ruslan Salakhutdinov identified by ExplaiNE as top-5 explanations for this recommendation (see Table 7) are students, coauthors, or colleagues of Rich Zemel, with a shared interest in deep learning. As fourth example, we predict links for Prof. Yann LeCun (node i), and compute the explanations for his top recommendation (node j): Tomas Mikolov. It turns out that the existing co-authors of Yann LeCun identified by ExplaiNE as top-5 explanations for this recommendation (see Table 8) are either colleagues or coauthors of Tomas Mikolov, with a shared interest in deep learning. As fifth example, we predict links for Prof. Michael I. Jordan (node i), and compute the explanations for his top recommendation (node j): Christos Faloutsos. It turns out that the existing co-authors of Michael I. Jordan identified by ExplaiNE as top-5 explanations for this recommendation (see Table 9) are either colleagues or coauthors of Christos Faloutsos, with a shared interest mainly in data mining and database systems.

MovieLens network. In the rating network, a predicted link between a user i and movie j amounts to a recommendation of movie j to user i. In making this recommendation CNE did not have access to any 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. As our first example, we computed the recommendation for the first user (uid=0) in the user list (See Table. 10). The top recommended movie is 'Batman' with genre tags 'Action', 'Adventure', 'Crime', and 'Drama'. The genres of the top explanations given by explainNE arguably have strongly overlapping genre tags (e.g. all top-5 are tagged with 'Action'). Moreover, the second-highest ranked explanation is 'Batman Forever'. As second example, we computed the recommendation for the first user (uid=1) in the user list (See Table. 11). The top recommended movie is 'The Devil's Own'

Table 6. Predicted/recommended collaborations for Kamalika Chaudhuri. The top link (author: Matus Telgarsky) predicted by CNE are explained through co-authors of Kamalika Chaudhuri that are also coauthors, advisors, or colleagues of Matus Telgarsky. The most relevant five co-authors of Kamalika Chaudhuri also cover major parts of Matus Telgarsky's research interests: deep learning.

Rank	Recommendations	Explain: 'Matus Telgarsky'
1	Matus Telgarsky	Animashree Anandkumar
2	Elad Hazan	Chicheng Zhang
3	Majid Janzamin	Yoav Freund
4	Rong Ge	Sanjoy Dasgupta
5	Shai Shalev-Shwartz	Sham M. Kakade

Table 7. Predicted/recommended collaborations for Ruslan Salakhutdinov. The top link (author: Rich Zemel) predicted by CNE are explained through co-authors of Ruslan Salakhutdinov that are also students, coauthors, or colleagues of Rich Zemel. The most relevant five co-authors of Ruslan Salakhutdinov also cover major parts of Rich Zemel's research interests: deep learning.

Rank	Recommendations	Explain: 'Rich Zemel'
1	Rich Zemel	Ryan P. Adams
2	Ruslan Salakhudinov	Jimmy Ba
3	Russell Greiner	Kevin Swersky
4	Dale Schuurmans	Kyunghyun Cho
5	Aaron C. Courville	Ryan Kiros

with genre tags 'Action', 'Drama', 'Thriller',' and 'War'. The genres of the top explanations given by explainNE arguably have strongly overlapping genre tags.

As thrid example, we computed the recommendation for the first user (uid=1) in the user list (See Table. 12). The top recommended movie is 'The Replacement Killers' with genre tags 'Action' and 'Thriller'. The genres of the top explanations given by explainNE arguably have descent amount of overlapping genre tags. As fourth example, we computed the recommendation for the first user (uid=1) in the user list (See Table. 13). The top recommended movie is ' Murder at 1600' with genre tags 'Mystery' and 'Thriller'. The genres of the top explanations given by explainNE arguably have strongly overlapping genre tags.

As fifth example, we computed the recommendation for the first user (uid=1) in the user list (See Table. 14). The top recommended movie is 'Dumbo' with genre tags 'Animation', 'Childrens', and 'Musical'. The genres of the top explanations given by explainNE arguably have strongly overlapping genre tags (e.g. all top-4 are tagged with 'Childrens').

6. Quantitative study

DBLP network. Here, 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 perfectly legitimate 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 will of course not always be true, we postulate that it is sufficiently common for ExplaiNE—providing it works well—to highlight the other co-authors as explanations for the observed link $\{i, j\}$.

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. We then took the top-r of this ranked list as predicted co-authors on the paper i co-authored with j. Based on this, we 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.

Figure 8 shows the PR and ROC curves for Eric P. Xing as node i and three author-list lengths. For comparison, also PR and

Table 8. Predicted/recommended collaborations for Prof. Yann LeCun. The top link (author: Tomas Mikolov) predicted by CNE are explained through co-authors of Yann LeCun that are also students, coauthors, or colleagues of Tomas Mikolov. The most relevant five co-authors of Yann LeCun also cover major parts of Tomas Mikolov's research interests: deep learning.

Rank	Recommendations	Explain: 'Tomas Mikolov'
1	Tomas Mikolov	Andrew Caplin
2	Hans Peter Graf	Sumit Chopra
3	Graham W. Taylor	Ido Kanter
4	Volodymyr Mnih	Clement Farabet
5	Rodolfo A. Milito	Wojciech Zaremba

Table 9. Predicted/recommended collaborations for Prof. Michael I. Jordan. The top link (author: Christos Faloutsos) predicted by CNE are explained through co-authors of Michael I. Jordan that are also students, coauthors, or colleagues of Christos Faloutsos. The most relevant five co-authors of Michael I. Jordan also cover major parts of Christos Faloutsos's research interests: data mining and database systems

Rank	Recommendations	Explain: 'Christos Faloutsos'
1	Christos Faloutsos	Pinar Duygulu
2	Carlos Guestrin	Deepayan Chakrabarti
3	Wolfgang Maass	Jennifer G. Dy
4	Andrew Zisserman	Richard M. Karp
5	Kotagiri Ramamohanarao	Stephen Tu

ROC curves computed based on a randomly ranked list is shown (as the size of the data is rather small, these are not always close to the diagonal). Figure 9 shows the ROC curves for Kamalika Chaudhuri as node i and three author-list lengths. Figure 10 shows the PR and ROC curves for Ruslan Salakhutdinov as node i and three author-list lengths. Figure 11 shows the PR and ROC curves for Yann LeCun as node i and three author-list lengths. Figure 12 shows the PR and ROC curves for Michael I. Jordan as node i and three author-list lengths.

MovieLens network. A good explanation k of a predicted link between a movie-user pair $\{i, j\}$ should arguably 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. We also computed the empirical p-values by randomly sampling 50 sets of 5 'explanations' drawn from the all movies for each i. A histogram of these p-values is shown in Fig. 13a,b. 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—indicating the null hypothesis is false. In both settings, a Kolmogorov-Smirnoff test indeed shows an extremely high significance (p-value numerically 0).

Table 10. Recommended movie to user uid=0. The top movie recommended by CNE (Batman) 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	Batman	Action, Adventure, Crime, Drama
2	E.T. the Extra-Terrestrial	Children's, Drama, Fantasy, Sci-Fi
3	The Secret of Roan Inish	Adventure

k	Explain: 'Batman'	Genres
1	Supercop	Action, Thriller
2	Batman Forever	Action, Adventure, Comedy, Crime
3	The Crow	Action, Romance, Thriller
4	Full Metal Jacket	Action, Drama, War
5	Young Guns	Action, Comedy, Western

Table 11. Recommended movie to user uid=1. The top movie recommended by CNE (The Devil's Own) is explained through movies already seen by user uid=1. The top-ranked explanations have genres that overlap with the recommended movie.

j	Recommendations	Genres
1	The Devil's Own	Action, Drama, Thriller, War
2	Everyone Says I Love You	Comedy, Musical, Romance
3	Lone Star	Drama, Mystery

k	Explain: 'The Devil's Own'	Genres
1	Heat	Action, Crime, Thriller
2	Midnight in the Garden of Good and Evil	Comedy, Crime, Drama, Mystery
3	A Time to Kill	Action, Drama
4	Liar Liar	Comedy
5	Up Close & Personal	Drama, Romance

Table 12. Recommended movie to user uid=2. The top movie recommended by CNE (The Replacement Killers) is explained through movies already seen by user uid=2. The top-ranked explanations have genres that overlap with the recommended movie.

j	Recommendations	Genres
1	The Replacement Killers	Action, Thriller
2	Titanic	Action, Drama, Romance
3	The Full Monty	Comedy

k	Explain: 'The Replacement Killers'	Genres
1	Spice World	Comedy, Musical
2	Deep Rising	Action, Horror, Sci-Fi
3	Deconstructing Harry	Comedy, Drama
4	Fallen	Action, Mystery, Thriller
5	Wag the Dog	Comedy, Drama

Table 13. Recommended movie to user uid=3. The top movie recommended by CNE (Murder at 1600) is explained through movies already seen by user uid=3. The top-ranked explanations have genres that overlap with the recommended movie.

j	Recommendations	Genres
1	Murder at 1600	Mystery, Thriller
2	The Devil's Advocate	Crime, Horror, Mystery, Thriller
3	The Game	Mystery, Thriller

k	Explain: 'Murder at 1600'	Genres
1	Liar Liar	Comedy
2	Scream	Horror, Thriller
3	Cop Land	Crime, Drama, Mystery
4	Assignment	Thriller
5	Conspiracy Theory	Action, Mystery, Romance, Thriller

Table 14. Recommended movie to user uid=4. The top movie recommended by CNE (Dumbo) is explained through movies already seen by user uid=4. The top-ranked explanations have genres that overlap with the recommended movie.

j	Recommendations	Genres
1	Dumbo	Animation, Children's, Musical
2	The Lion King	Animation, Children's, Musical
3	The ox and the Hound	Animation, Children's

k	Explain: 'Dumbo'	Genres
1	Fantasia	Animation, Children's, Musical
2	Cinderella	Animation, Children's, Musical
3	The Parent Trap	Children's, Drama
4	Alice in Wonderland	Animation, Children's, Musical
5	Jack	Comedy, Drama

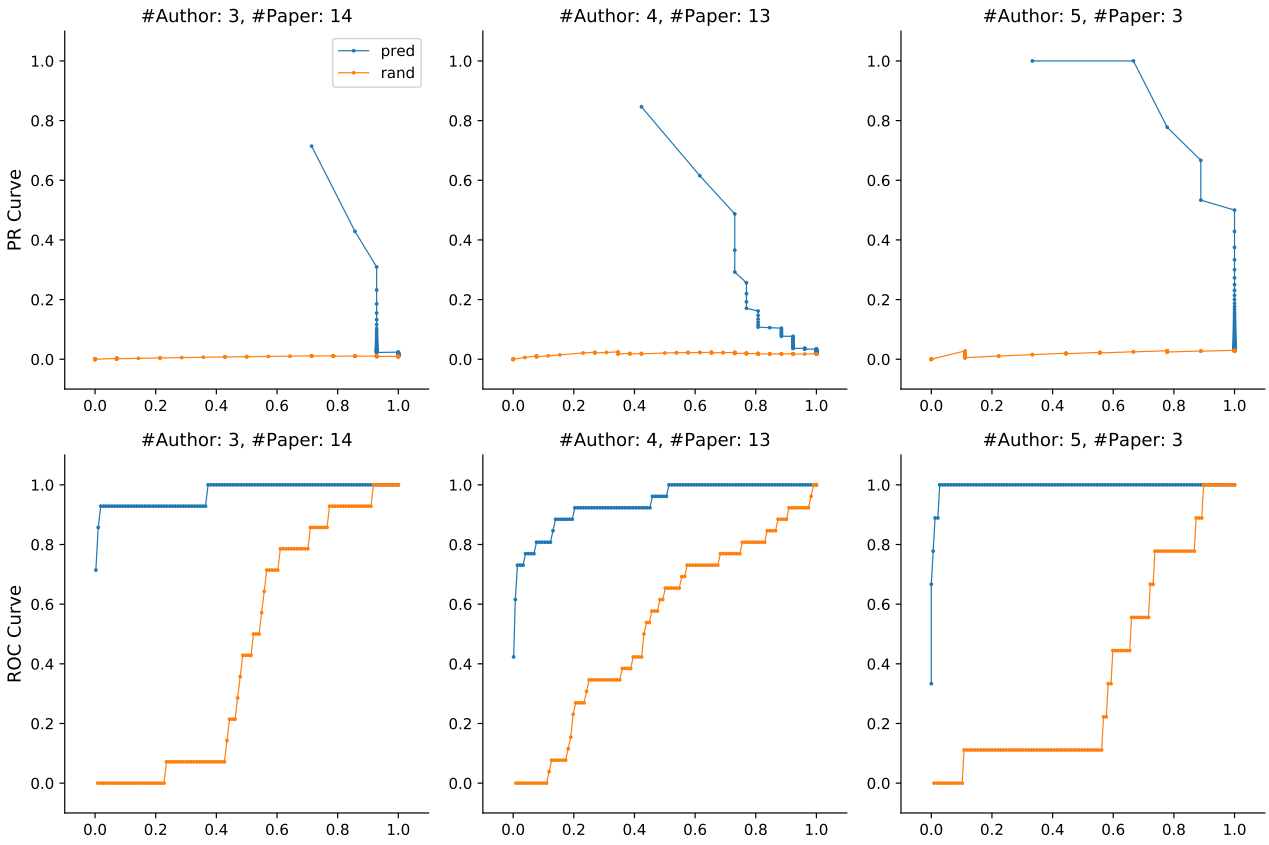

Figure 8. PR (first row) and ROC (second row) curves of co-author predictions for $i =$'Eric P. Xing', with author-list lengths 3, 4, and 5 (orange=random, blue=ExplaiNE).

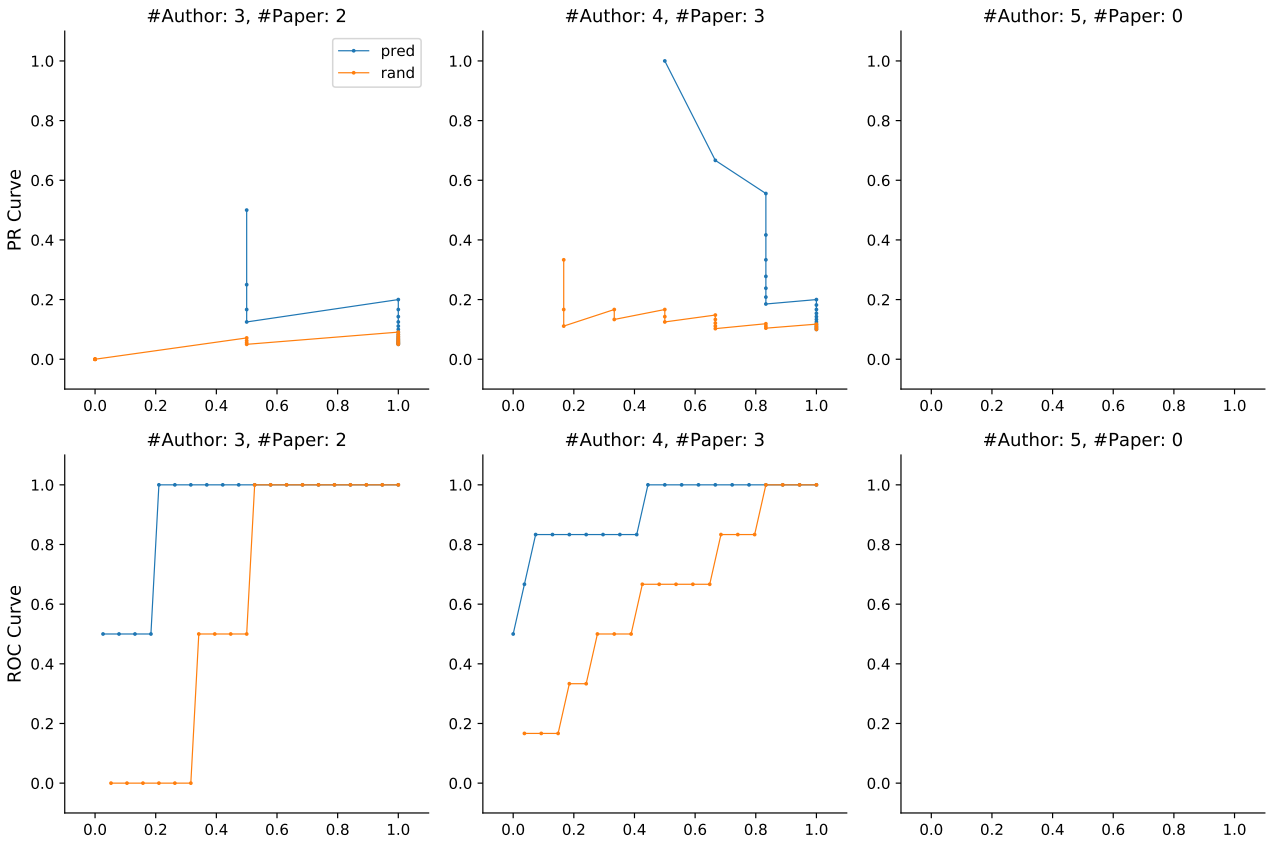

Figure 9. PR (first row) and ROC (second row) curves of co-author predictions for $i =$'Kamalika Chaudhuri', with author-list lengths 3, 4 (orange=random, blue=ExplaiNE). In this case, no paper has author list length 5.

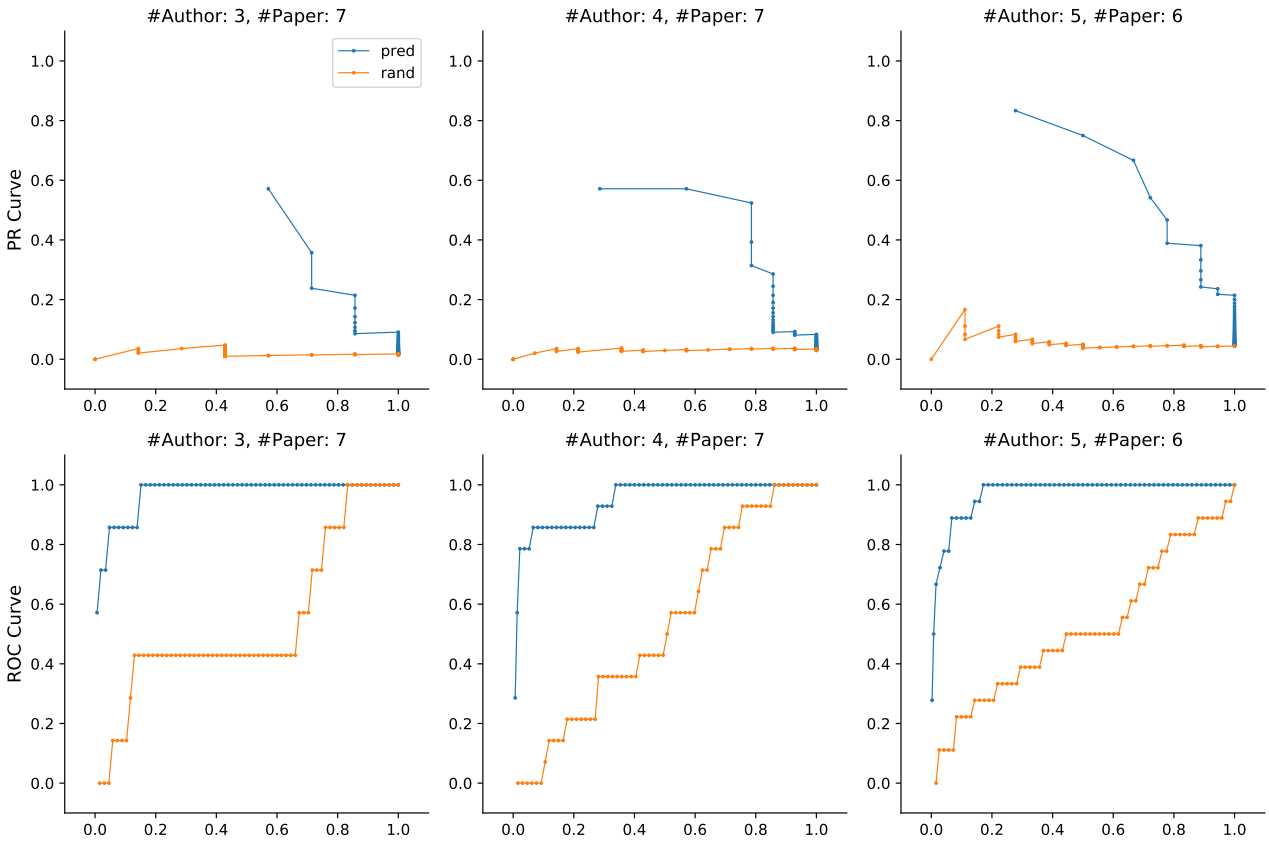

Figure 10. PR (first row) and ROC (second row) curves of co-author predictions for $i =$'Ruslan Salakhutdinov', with author-list lengths 3, 4, and 5 (orange=random, blue=ExplaiNE).

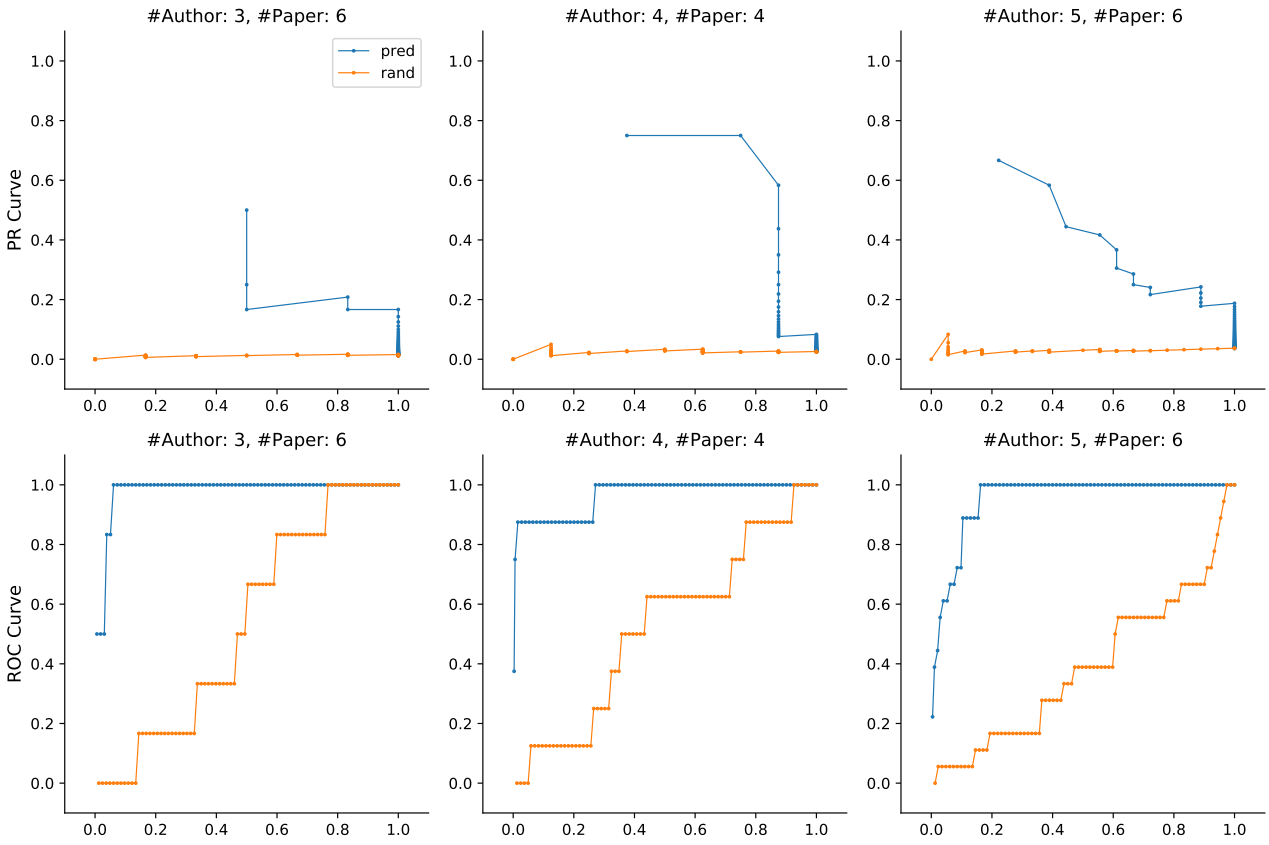

Figure 11. PR (first row) and ROC (second row) curves of co-author predictions for $i =$'Yann LeCun', with author-list lengths 3, 4, and 5 (orange=random, blue=ExplaiNE).

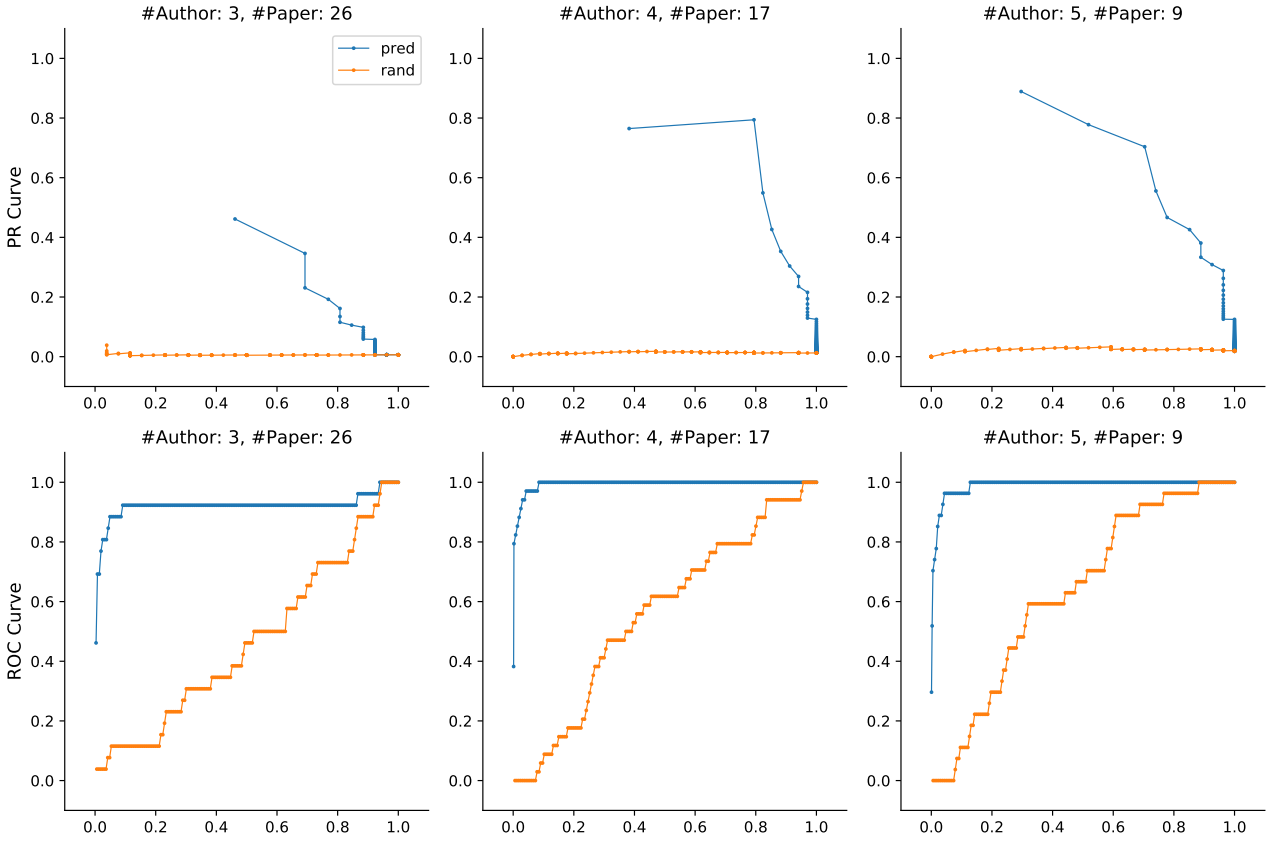

Figure 12. PR (first row) and ROC (second row) curves of co-author predictions for $i =$'Michael I. Jordan', with author-list lengths 3, 4, and 5 (orange=random, blue=ExplaiNE).

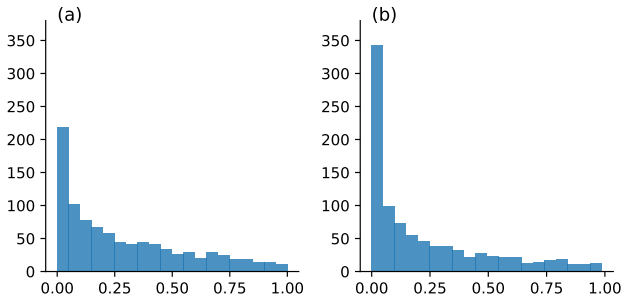

Figure 13. P-values that indicates the significance of the correlation between the genre recommended and the genres in the explanation. Each p-value is computed against 50 random explanations. (a) Explanations are drawn from user's watched movies. The empirical distribution has Kolmogorov-Smirnov test statistic 0.32 and a p-value that is numerically 0.0 against uniform distribution. (b) Explanations are drawn from all movies. The empirical distribution has Kolmogorov-Smirnov test statistic 0.42 and a p-value that is numerically 0.0 against uniform distribution. This shows the significance of positive correlation between the recommended movies and the explanations made by ExplaiNE.