PReP: Path-Based Relevance from a Probabilistic Perspective in Heterogeneous Information Networks

Yu Shi
University of Illinois at Urbana-Champaign
yushi2@illinois.edu

Po-Wei Chan
University of Illinois at Urbana-Champaign
pchan12@illinois.edu

Honglei Zhuang
University of Illinois at Urbana-Champaign
hzhuang3@illinois.edu

Huan Gui
University of Illinois at Urbana-Champaign
huangui2@illinois.edu

Jiawei Han
University of Illinois at Urbana-Champaign
hanj@illinois.edu

ABSTRACT

As a powerful representation paradigm for networked and multi-typed data, the heterogeneous information network (HIN) is ubiquitous. Meanwhile, defining proper relevance measures has always been a fundamental problem and of great pragmatic importance for network mining tasks. Inspired by our probabilistic interpretation of existing path-based relevance measures, we propose to study HIN relevance from a probabilistic perspective. We also identify, from real-world data, and propose to model *cross-meta-path synergy*, which is a characteristic important for defining path-based HIN relevance and has not been modeled by existing methods. A generative model is established to derive a novel path-based relevance measure, which is data-driven and tailored for each HIN. We develop an inference algorithm to find the maximum a posteriori (MAP) estimate of the model parameters, which entails non-trivial tricks. Experiments on two real-world datasets demonstrate the effectiveness of the proposed model and relevance measure.

CCS CONCEPTS

•**Information systems** → **Data mining; Similarity measures;**
•**Computing methodologies** → *Maximum a posteriori modeling;*

KEYWORDS

Heterogeneous information networks, graph mining, meta-paths, relevance measures.

1 INTRODUCTION

In real-world applications, objects of various types are often interconnected with each other. These objects, together with their relationship, form numerous heterogeneous information networks (HINs) [14, 16]. Bibliographical information network is a typical example, where researchers, papers, organizations, and publication venues are interrelated. A fundamental problem in HIN analysis is to define proper measures to characterize the relevance between node pairs in the network, which also benefits various downstream applications, such as similarity search, recommendation, and community detection [14, 16].

Most existing studies derive their HIN relevance measures on the basis of *meta-path* [14, 16, 17], which is defined as a concatenation of multiple node types linked by corresponding edge types. Based on the concept of *meta-path*, researchers have proposed PathCount, PathSim [17], and path constrained random walk [10] to measure relevance between node pairs. On top of these studies, people have explored the ideas of incorporating richer information [6, 21] and more complex typed structures [4, 7, 13] to define more effective relevance scoring functions, or adding supervision to derive task-specific relevance measures [2, 19, 23].

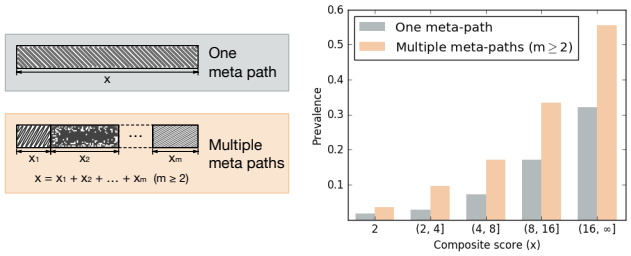
The probabilistic perspective. While building upon this powerful *meta-path* paradigm, we aim to additionally understand and model relevance from the probabilistic point of view. In this regard, we establish a probabilistic interpretation of existing HIN relevance measures, which is achieved by modeling the generating process of all path instances in an HIN and deriving the relevance of a node pair from the likelihood of observing the path instances connecting them. Relevance and likelihood can be connected by this approach because only a small portion of node pairs in an HIN are actually relevant; and a proper generating process has low likelihood to generate the path instances between each of these relevant node pairs. We will detailedly discuss this probabilistic interpretation in Sec. 3. Moreover, as a starting point for studying HIN relevance from the probabilistic perspective, we focus the scope of this paper on the basic *unsupervised* scenario. Meanwhile, we assume that the meta-paths of interest are already given. That is, we defer the study on the cases with label information and meta-path selection to future work.

In order to determine relevance between any pair of nodes, we have the key insight that a path-based HIN relevance should contain three characteristics – *node visibility*, *path selectivity*, and *cross-meta-path synergy* – which we describe in the following paragraphs.

Node visibility. One straightforward way to derive relevance in an HIN is PathCount [17]. For a meta-path $t \in \{1, \ldots, T\}$, PathCount is defined as the number of paths, P_{st} or equivalently $P_{\langle uv \rangle t}$, under this meta-path between a node pair $s = (u, v) \in \mathcal{V} \times \mathcal{V}$, *i.e.*,

KDD '17, August 13–17, 2017, Halifax, NS, Canada

978-1-4503-4887-4/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3097983.3097990



(a) (b)

Figure 1: (a) The same composite score (x) may be aggregated from different number of meta-paths, where score is represented by the length of the rectangles and each fill pattern represents a meta-path. (b) An observation made from an entity resolution task on the DBLP dataset that if linear combination is used to compute the composite score, node pairs with paths under multiple meta-paths are more likely to be relevant than those under only one meta-path. Prevalence is defined as the number of relevant node pairs divided by the total number of node pairs.

$PathCount^{(t)}(u, v) := P_{\langle uv \rangle t}$. One obvious drawback of this measure is that it favors nodes with high *node visibility*, *i.e.*, nodes with a large number of paths. To resolve this problem, [17] proposed to penalize PathCount by the arithmetic mean of the numbers of cycles attached to the two involved nodes, *i.e.*, $PathSim^{(t)}(u, v) := \frac{2 \cdot P_{\langle uv \rangle t}}{P_{\langle uu \rangle t} + P_{\langle vv \rangle t}}$. A similar design to model *node visibility* can be found in JoinSim [20], which is defined as PathCount penalized by geometric mean of the cycle numbers.

Path selectivity. Given any method defining relevance score under one meta-path, a natural question is how to combine multiple meta-paths to derive a unified relevance score – henceforth referred to as the *composite score*. To achieve this goal, Sun et al. [17] proposed to assign different weights to different meta-paths, and compute the composite score via linear combination. Let $\mathbf{w} = \{w_1, \ldots, w_T\}$ with w_t being the weight for meta-path t, the composite score of PathCount is given by $PathCount_{\mathbf{w}}(u, v) := \sum_{t=1}^{T} w_t \cdot PathCount^{(t)}(u, v)$. Similarly, one can define $PathSim_{\mathbf{w}}(u, v)$. This linear combination approach is adopted by follow-up works with multiple applications [14, 16], including personalized entity recommendation problem [22], outlier detection [9, 24], *etc.* The weights assigned or inferred in these cases specify how selective each meta-path is. The larger the *path selectivity*, the more significant this meta-path is in contributing to the composite score.

Cross-meta-path synergy. Suppose linear combination is used to find the composite score as in the previous paragraph, the two scenarios shown in Fig. 1a would receive the same composite score (x), where x_i equals to the score from the i-th meta-path multiplied by the corresponding weight. However, we have the observation that, when meta-paths do not clearly correlate, the latter scenario tends to imply a higher relevance. We take an entity resolution task on the DBLP dataset as example, which aims to merge author mentions that refer to the same entity. In this task, each node stands for an author mention, and each meta-path represents that two author mentions have both published papers in one particular research area. We label two author mentions as relevant if and only if they refer to the same entity, and we use PathCount with uniform weights as an example to compute the composite score. Results presented in Fig. 1 shows that with the same composite score, node pairs associated by paths under multiple meta-paths are more likely to be relevant than those under only one meta-path. We refer to this phenomenon as *cross-meta-path synergy*. We interpret this phenomenon as given the occurrence of one path, the happenstance of another path under the same meta-path may not be surprising, while the co-occurrence of two paths under two uncorrelated meta-paths may be a strong signal of relevance. Moreover, we should also realize that not necessarily all meta-path pairs are uncorrelated, which has been observed in a special type of HIN [15]. This implies *cross-meta-path synergy* does not necessarily exist between all pairs of meta-paths, and we deem a good relevance measure should reflect this difference.

Challenges and contributions. Regarding the three pivotal characteristics for path-based HIN relevance discussed above, the major challenge lies in how to integrate all these characteristics in a unified framework. We tackle this challenge by studying path-based relevance from a probabilistic perspective, and deriving relevance measure from a generative model. Since the model parameters are trained to fit each HIN, the derived relevance measure enjoys the property of being data-driven. That is, the derived relevance measure is tailored for each HIN. Lastly, we summarize our contributions as follows:

(1) We establish the probabilistic interpretation of existing path-based HIN relevance measures.
(2) We identify and propose to model *cross-meta-path synergy*, an important characteristic in path-based HIN relevance.
(3) We propose a novel relevance measure based on a generative model, which is data-driven and tailored for each HIN, and develop an inference algorithm with non-trivial tricks.
(4) Experiments on two real-world HINs corroborate the effectiveness of our proposed model and relevance measure.

2 PRELIMINERIES

In this section, we introduce the concepts and notations used in this paper.

Definition 2.1 (Heterogeneous Information Network). An **information network** is a directed graph $G = (\mathcal{V}, \mathcal{E})$ with a node type mapping $f : \mathcal{V} \to \mathcal{A}$ and an edge type mapping $g : \mathcal{E} \to \mathcal{R}$. Particularly, when the number of node types $|\mathcal{A}| > 1$ or the number of edge types $|\mathcal{R}| > 1$, the network is called a **heterogeneous information network** (HIN).

Due to the typed essence of HINs, paths that associate node pairs can be grouped under different meta-paths. We formally define meta-paths as follows.

Definition 2.2 (Meta-Path). A **meta-path** is a concatenation of multiple nodes or node types linked by edge types.

An example of a meta-path is [author] $\xrightarrow{\text{writes}}$ [paper] $\xrightarrow{\text{writes}^{-1}}$ [author], where a phrase in the brackets represents a node type and

a phrase above the arrow refers to an edge type. When context is clear, we simply write [author]–[paper]–[author]. In this paper, we study the relevance problem when a set of meta-paths of interest is predefined by users.

To ease presentation, we focus on unweighted HINs, and model path count defined as follows. Note that the path-based model to be proposed in this paper can be extended to the weighted case.

Definition 2.3 (Path Count). The **path count** of a meta-path $t \in \{1, \dots, T\}$ between a node pair $s = (u, v) \in \mathcal{V} \times \mathcal{V}$ is the number of concrete path instances under this meta-path that start from node u to node v, which is denoted by P_{st} or $P_{\langle uv \rangle t}$.

Note that the relevance score given by the PathCount measure [17] is exactly the path count of a meta-path between a node pair.

Lastly, we introduce the probability distributions to be used.

Definition 2.4. The probability density functions of three probability distributions used in this paper are given as follows.

(1) Exponential distribution $\mathrm{Exp}\left(\tilde{\lambda}\right)$ with rate parameter $\tilde{\lambda} > 0$:

$$p(x) = \tilde{\lambda}\, \mathrm{e}^{\tilde{\lambda} x} \quad (x > 0).$$

(2) Gamma distribution $\Gamma\left(\tilde{\alpha}, \tilde{\beta}\right)$ with shape parameter $\tilde{\alpha} > 0$ and rate parameter $\tilde{\beta} > 0$:

$$p(x) = \frac{\tilde{\beta}^{\tilde{\alpha}}}{\Gamma(\tilde{\alpha})} x^{\tilde{\alpha}-1}\, \mathrm{e}^{-\tilde{\beta} x} \quad (x > 0),$$

where $\Gamma(\tilde{\alpha}) = \int_0^\infty t^{\tilde{\alpha}-1}\, \mathrm{e}^{-t}\, dt$ is the gamma function.

(3) Symmetric Dirichlet distribution $\mathrm{Dir}_L\left(\tilde{\alpha}\right)$ of order L and concentration parameter $\tilde{\alpha}$:

$$p(x_1, \dots, x_L) = \frac{\Gamma\left(\tilde{\alpha} L\right)}{\Gamma\left(\tilde{\alpha}\right)^L} \prod_{i=1}^{L} x_i^{\tilde{\alpha}-1} \quad \left(x_i > 0 \text{ and } \sum_{i=1}^{L} x_i = 1\right),$$

where $\Gamma(\cdot)$ is the gamma function.

We denote $\mathrm{Exp}\left(x\,;\,\tilde{\lambda}\right) \coloneqq p(x)$ the probability density function of $\mathrm{Exp}\left(\tilde{\lambda}\right)$, and denote $x \sim \mathrm{Exp}\left(\tilde{\lambda}\right)$ if x is generated from $\mathrm{Exp}\left(\tilde{\lambda}\right)$. Similar notations are also used for $\Gamma\left(\tilde{\alpha}, \tilde{\beta}\right)$ and $\mathrm{Dir}_L\left(\tilde{\alpha}\right)$.

3 PROBABILISTIC INTERPRETATION OF EXISTING RELEVANCE MEASURES

In this section, we illustrate the probabilistic interpretation of existing path-based HIN relevance measures. We achieve this by studying the generating process of path counts between node pairs in an HIN, which contains a connection between relevance and the negative log likelihood. Suppose the path count under meta-path t between node pair s is generated from an exponential distribution

$$P_{st} \sim \mathrm{Exp}\left(\lambda\right),$$

with fixed rate λ, then in terms of the rank it yields, the negative log likelihood of all observed paths under meta-path t between node pair s will be equivalent to the PathCount under meta-path t

$$\begin{aligned} -LL^{(t)}(s) &= -\log(\lambda\, \mathrm{e}^{-\lambda P_{st}}) = \lambda P_{st} - \log\lambda \\ &\propto P_{st} + const = PathCount^{(t)}(s) + const. \end{aligned}$$

Further, if we assume path instances under different meta-paths are generated from exponential distribution with meta-path-specific rates $\mathbf{w} = (w_1, w_2, \dots, w_T)$, *i.e.*, $P_{st} \sim \mathrm{Exp}\left(w_t\right)$, then the negative log likelihood of all observed path counts will be equivalent to PathCount with weights $\mathbf{w}$ for linear combination

$$\begin{aligned} -LL(s) &= -\log(\prod_t w_t\, \mathrm{e}^{-w_t P_{st}}) = \sum_t w_t P_{st} - \sum_t \log w_t \\ &= \sum_t w_t P_{st} + const = PathCount_{\mathbf{w}}(s) + const. \end{aligned}$$

Moreover, if we assume each node pair s has pair-specific generating rate proportional to a parameter κ_s, *i.e.*, $P_{st} \sim \mathrm{Exp}\left(w_t/\kappa_s\right)$, then the negative log likelihood of observed path counts will be $-LL(s) = \sum_t w_t \cdot \frac{P_{st}}{\kappa_s} + T \log \kappa_s + const$. For node pair $s = (u, v)$, if we drop the logarithm term and set κ_s to be the arithmetic mean of the cycle count of the involved nodes u and v, the formula becomes

$$\sum_t w_t \cdot \frac{2 \cdot P_{\langle uv \rangle t}}{P_{\langle uu \rangle t} + P_{\langle vv \rangle t}} = PathSim_{\mathbf{w}}(s)$$

which is identical to PathSim with weights $\mathbf{w}$ for linear combination. In lieu of arithmetic mean, if we set κ_s to be the geometric mean of the same quantities, we get $\sum_t w_t \cdot \frac{P_{\langle uv \rangle t}}{\sqrt{P_{\langle uu \rangle t} \cdot P_{\langle vv \rangle t}}}$, which is identical to JoinSim with weights $\mathbf{w}$ for linear combination. Note that all the relevance measures discussed in this section are special cases of our relevance measure to be proposed in the next section.

Symbol	Definition
$\mathcal{V}$	The set of all nodes
$\mathcal{S}$	The set of all nontrivial node pairs
$T \in \mathbb{N}$	The number of meta-paths
$K \in \mathbb{N}$	The number of generating patterns
$\boldsymbol{P} \in \mathbb{R}^{\lvert\mathcal{S}\rvert \times T}$	The observed path counts between node pairs over each meta-path
$\boldsymbol{\eta} \in \mathbb{R}^T$	The *path selectivity*
$\boldsymbol{\tau} \in \mathbb{R}^{\lvert\mathcal{S}\rvert}$	The *node pair visibility*
$\boldsymbol{\rho} \in \mathbb{R}^{\lvert\mathcal{V}\rvert}$	The *node visibility*
$\boldsymbol{\Theta} \in \mathbb{R}^{\lvert\mathcal{S}\rvert \times K}$	The generating patterns over meta-paths
$\boldsymbol{\Phi} \in \mathbb{R}^{K \times T}$	The choices of generating patterns between node pairs
$\alpha \in \mathbb{R}_+$	The shape parameter of the gamma prior
$\beta \in (0, 1)$	The concentration parameter of the Dirichlet prior

Table 1: Summary of symbols

4 PROPOSED MODEL AND RELEVANCE

With the relevance–likelihood connection established in Sec. 3, we propose our **P**ath-based **Re**levance from **P**robabilistic perspective (PReP) likewise by modeling the generating process of path counts between node pairs, and further aim to model the three important characteristics. In a nutshell, the proposed generative-model-based relevance measure consists of two major parts: (i) inferring model parameters by finding the maximum a posteriori (MAP) estimate to fit the input HIN, and (ii) deriving relevance score between any node pair based on the learned model.

4.1 The PReP Model

Following the existing HIN relevance measures discussed in Sec. 3, we assume the path count, P_{st} or $P_{\langle uv \rangle t}$, between node pair $s =$

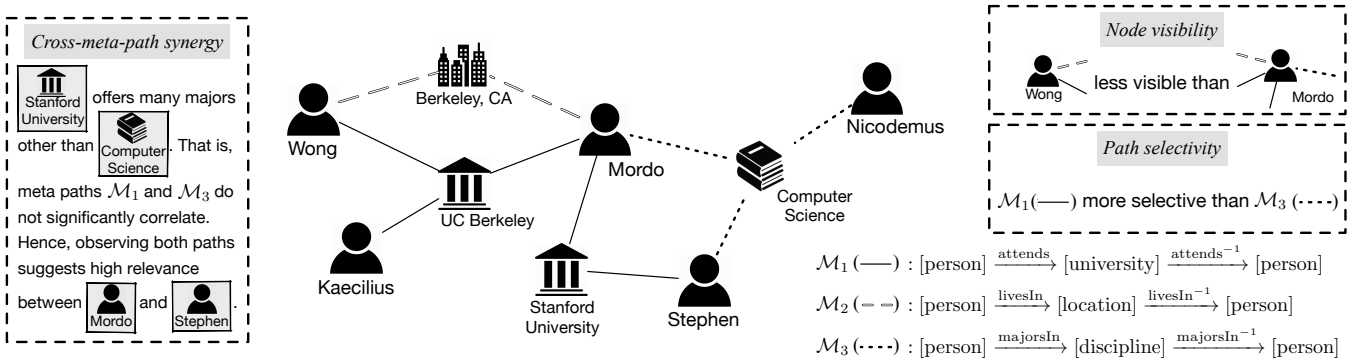


Figure 2: Toy example for one part of an HIN, consisting of four node types: person, university, location, and discipline.

(u, v) under meta-path t is generated from an exponential distribution with rate λ_{st}, *i.e.*, $P_{st} \sim \text{Exp}(\lambda_{st})$. To capture *node visibility*, *path selectivity*, and *cross-meta-path synergy*, we must design λ_{st} in a way that can model these three characteristics.

According to the property of exponential distribution, if a random variable X is generated from $\text{Exp}\left(\tilde{\lambda}\right)$, then the expectation of X will be $1/\tilde{\lambda}$. Bearing this in mind, we introduce three components to model the three characteristics as follows.

- Both the *node visibility* of u and that of v affect the generation of path instances. We consider the visibility of this pair of node as *node pair visibility*, τ_s, which is positively correlated with the expectation of P_{st}.
- We let path instances under the same meta-path share the same *path selectivity*. Denote η_t the *path selectivity* for meta-path t. η_t is negatively correlated with the expectation of P_{st}.
- Each node pair with paths in between can be linked by path instances under a different set of meta-paths. We assume an underlying *meta-path distribution* $\boldsymbol{\psi}_s = [\psi_{s1}, \ldots, \psi_{sT}]$ for node pair s, where $\sum_{t=1}^{T} \psi_{st} = 1$ and $\psi_{st} \geq 0$. As a distribution over meta-paths, $\boldsymbol{\psi}_s$ models the semantics of the relevance between this node pair, because each meta-path carries its own semantic meaning. With further design to be introduced, $\boldsymbol{\psi}_s$ also serves as the basis to capture *cross-meta-path synergy*. ψ_{st} is positively correlated with the expectation of P_{st}.

Putting the above three components together considering their correlation with the expectation of P_{st}, we find path count generating process as

$$P_{st} \sim \text{Exp}\left(\frac{\eta_t}{\tau_s \psi_{st}}\right), \tag{1}$$

where the detailed illustration and design of the three components are to be further discussed in this section. Note that while we only discuss unweighted HINs in this paper, the use of exponential distribution in Eq. (1) enables the model to handle weighted HINs, where paths are associated with real-valued path strengths, and P_{st} may not be integers to reflect the path strengths.

Since node pairs with no paths under any predefined meta-path should trivially receive the lowest possible relevance score, we only model the generation of path counts between node pairs with paths in between – henceforth referred to as nontrivial node pairs – and we denote $\mathcal{S}$ the set of all nontrivial node pairs.

Illustrative example. To better illustrate how each component design affects the path generation process, we present a toy example in Fig. 2, which shows a part of an HIN with four node types: person, university, location, and discipline. We concern three meta-paths in this network: $\mathcal{M}_1$: [person] $\xrightarrow{\text{attends}}$ [university] $\xrightarrow{\text{attends}^{-1}}$ [person], $\mathcal{M}_2$: [person] $\xrightarrow{\text{livesIn}}$ [location] $\xrightarrow{\text{livesIn}^{-1}}$ [person], $\mathcal{M}_3$: [person] $\xrightarrow{\text{majorsIn}}$ [discipline] $\xrightarrow{\text{majorsIn}^{-1}}$ [person].

Decoupling node pair visibility. To model *node visibility*, we decouple node pair visibility τ_s in Eq. (1) into two parts as in PathSim and JoinSim discussed in Sec. 3. The two parts correspond to the *node visibility* ρ_u and ρ_v, respectively, where $s = (u, v)$, and $\rho_z > 0$ for all $z \in \mathcal{V}$. In our design, we let

$$\tau_{(u,v)} = \rho_u \rho_v \tag{2}$$

as in JoinSim because decoupling by multiplication eases model inference, which will be made clear in the next paragraph.

Since a trivial rescaling – multiplying all ρ_z by a constant and multiplying all η_t by the square of the same constant – leads to exactly the same model (Eq. (1)), we further regularize ρ_z by a gamma prior with a constant rate parameter

$$\rho_z \sim \Gamma(\alpha, 1). \tag{3}$$

Note that we arbitrarily set the rate parameter to be 1 since the shape of the distribution is solely determined by the shape parameter α. We choose gamma distribution as the prior for ρ_z because it is the conjugate prior for the exponential distribution, and this fact will largely facilitate the inference algorithm as we will show in Sec. 5.2. To determine the shape parameter α, we fit the gamma distribution to the total path count each node has, $\{\sum_{t=1}^{T} \sum_{\tilde{z} \in \mathcal{V}} P_{\langle z\tilde{z}\rangle t}\}_{z \in \mathcal{V}}$, in the HIN as a rough prior information.

Path selectivity at meta-path level. We assume path instances under meta-path t share the same *path selectivity* η_t. In the scope of this paper, where supervision is not available, we assume uninformative prior on η_t. In future work where supervision is provided, we can further learn η_t by minimizing the difference between supervision and model output to derive a task-specific relevance measure.

Cross-meta-path synergy and generating patterns. As discussed in Sec. 1, we have observed the existence of *cross-meta-path*

Measure	Node Pair	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	Composite	Truth
PathCount	*Mordo & Wong*	1	1	0	$w_1 + w_2$	−
	Mordo & Stephen	1	0	1	$w_1 + w_3$	+
PathSim	*Mordo & Wong*	0.67	1	0	$0.67w_1 + w_2$	−
	Mordo & Stephen	0.67	0	1	$0.67w_1 + w_3$	+
RWR ($C = 0.9$)	*Mordo & Wong*	0.29	0.47	0	$0.29w_1 + 0.47w_2$	−
	Mordo & Stephen	0.25	0	0.31	$0.25w_1 + 0.31w_3$	+
PReP	*Mordo & Wong*	1 generating pattern				−
	Mordo & Stephen	2 generating patterns				+

Table 2: Existing measures cannot yield desired relevance, unless we assert $\mathcal{M}_3$ (discipline) is always more selective than $\mathcal{M}_2$ (location), while PReP can achieve this by recognizing the co-occurrence of multiple generating patterns.

synergy in real-world HIN, and this characteristic has not been modeled by existing HIN relevance measures. In case meta-paths do not correlate, we may simply add a Dirichlet prior, with concentration parameter smaller than 1, over meta-path distribution $\boldsymbol{\psi}_s$ for all node pair s. This use of Dirichlet prior resembles latent Dirichlet allocation (LDA) [1], where the Dirichlet prior prefers sparse distributions, *i.e.*, most entries of $\boldsymbol{\psi}_s$ tend to be 0. Therefore, the co-occurrence of paths under different meta-paths gets a lower likelihood from this prior, and attains a higher relevance score under our relevance–likelihood connection.

However, in reality, it would not be surprising to see two people attending UC Berkeley also both live in the City of Berkeley. This implies *cross-meta-path synergy* does not necessarily exist between all pairs of meta-paths, *e.g.*, it may not exist between meta-path $\mathcal{M}_1$ and meta-path $\mathcal{M}_2$ in the toy example of Fig. 2. To address this situation, we introduce a new component – *generating patterns*. Each of a total of K generating patterns is a distribution over the T meta-paths, where meta-paths that often co-occur between node pairs will also be included in a common generating pattern, and when a node pair s generates a path instance in between, it would first choose generating pattern k with probability ϕ_{sk}, and then choose meta-path t from this generating pattern with probability θ_{kt}. Formally, we describe this process as

$$\psi_{st} = \sum_{k=1}^{K} \phi_{sk}\theta_{kt}, \tag{4}$$

where $\boldsymbol{\phi}_s = [\phi_{s1}, \ldots, \phi_{sK}]$ is node pair s's choices of generating patterns, such that $\sum_{k=1}^{K} \phi_{sk} = 1, \phi_{sk} \geq 0$; and $\boldsymbol{\theta}_k = [\theta_{k1}, \ldots, \theta_{kT}]$ is generating pattern k's distribution over meta-paths, such that $\sum_{t=1}^{T} \theta_{kt} = 1, \theta_{kt} \geq 0$.

A symmetric Dirichlet prior is then enforced on $\boldsymbol{\phi}_s$, so that synergy will be recognized between and only between meta-paths from different generating patterns

$$\boldsymbol{\phi}_s \sim \mathrm{Dir}_K(\beta), \tag{5}$$

where $\beta \in (0, 1)$ is the concentration hyperparameter.

With this design, our model gives a lower likelihood and higher relevance score to *Mordo* and *Stephen* (same university, same major) than *Mordo* and *Wong* (attending UC Berkeley and living in the City of Berkeley) in the toy example of Fig. 2 by learning a generating pattern that includes both $\mathcal{M}_1$ and $\mathcal{M}_2$. Whereas, other relevance measures cannot achieve this desired relationship as presented in Tab. 2, unless we set the weights $w_2 > w_3$, or equivalently assert $\mathcal{M}_2$ (location) is always less selective than $\mathcal{M}_3$ (discipline).

The unified model. For notation convenience, we use the bold italic form to represent the corresponding matrix or vector of each symbol with subscripts. For instance, the (s, t) element of $\boldsymbol{P}$ is P_{st} and the t-th element of $\boldsymbol{\eta}$ is η_t. Under this notation, combining Eq. (1), (3), and (5), with Eq. (2) and (4) substituted into Eq. (1), yields the total likelihood of the full PReP model

$$\begin{aligned}\mathcal{L} &= p(\boldsymbol{P}, \boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\Phi}, \boldsymbol{\Theta} \mid \alpha, \beta) \\ &= \left\{\prod_{u \in \mathcal{V}} \Gamma(\rho_u \,;\, (\alpha, 1))\right\} \cdot \left\{\prod_{s \in \mathcal{S}} \mathrm{Dir}_K(\boldsymbol{\phi}_s \,;\, \beta)\right\} \\ &\quad \cdot \left\{\prod_{\substack{s \in \mathcal{S} \\ (u,v)=s}} \prod_{t=1}^{T} \mathrm{Exp}\left(P_{st} \,;\, \frac{\eta_t}{\rho_u \rho_v \sum_{k=1}^{K} \phi_{sk}\theta_{kt}}\right)\right\}\end{aligned} \tag{6}$$

4.2 The PReP Relevance Measure

Given the unified model (Eq. (6)), we have two options to derive relevance measure using likelihood: (i) find the maximum a posteriori estimate for all parameters and compute the total likelihood of the observed data, and (ii) consider all model parameters as hidden variables and define the relevance as the marginal likelihood of the observed data. However, the marginal likelihood does not have a closed-form representation in our case, nor can we approximate it with regular Markov chain Monte Carlo algorithms due to the large number of hidden variables. Therefore, we adopt the first option and defer the other to future work.

Once the model parameters $\{\boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\Phi}, \boldsymbol{\Theta}\}$ are estimated, we define the PReP relevance for a node pair $s = (u, v)$ as the negative log-likelihood involving this node pair, $-\log p(\boldsymbol{P}_{s,:}, \boldsymbol{\phi}_s \mid \boldsymbol{\Theta}, \boldsymbol{\rho}, \boldsymbol{\eta}, \alpha, \beta)$, without the log term as in the derivation of PathSim in Sec. 3

$$r(s) = \sum_{t=1}^{T} \frac{P_{st}}{\rho_u \rho_v \eta_t \sum_{k=1}^{K} \phi_{sk}\theta_{kt}} + (1 - \beta) \sum_{k=1}^{K} \log \phi_{sk}. \tag{7}$$

Note that PathCount, PathSim, and JoinSim discussed in Sec. 3 are special cases of this PReP relevance measure, when $\{\boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\Phi}, \boldsymbol{\Theta}\}$ are heuristically specified accordingly.

5 MODEL INFERENCE

In this section, we introduce the inference algorithm for the PReP model (Eq. (6)) proposed in Sec. 4.

5.1 The Optimization Problem

We find the maximum a posteriori (MAP) estimate for model parameters by minimizing the negative log-likelihood of the proposed model (Eq. 6), which, with an offset of a constant, is given by

Algorithm 1: Inference algorithm for the PReP model

Input : the observed path counts $\boldsymbol{P}$ and the hyperparameters
Output: the model parameters $\boldsymbol{\eta}$, $\boldsymbol{\rho}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Theta}$
begin
  Initialize $\boldsymbol{\rho}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Theta}$
  while *not converged* **do**
    Update $\boldsymbol{\eta}$ by the closed-form Eq. (10)
    while *not converged* **do**
      for $u \in \mathcal{V}$ **do**
        Update ρ_u by the closed-form solution to Eq. (11)
    Update $\boldsymbol{\Phi}$ via parallelized PGD with gradient in Eq. (13)
    Update $\boldsymbol{\Theta}$ via PGD with gradient in Eq. (12)

$$\begin{aligned} O = &\sum_{u\in\mathcal{V}} (\rho_u - (\alpha-1)\log\rho_u) - (\beta-1)\sum_{s\in\mathcal{S}}\sum_{k=1}^{K}\log\phi_{sk} \\ &+ T\sum_{(u,v)\in\mathcal{S}} (\log\rho_u + \log\rho_v) - |\mathcal{S}|\sum_{t=1}^{T}\log\eta_t \\ &+ \sum_{\substack{s\in\mathcal{S}\\(u,v)=s}}\sum_{t=1}^{T}\left[\log\sum_{k=1}^{K}\phi_{sk}\theta_{kt} + \frac{\eta_t P_{st}}{\rho_u\rho_v\sum_{k=1}^{K}\phi_{sk}\theta_{kt}}\right], \end{aligned} \tag{8}$$

and the optimization problem is therefore

$$\min_{\boldsymbol{\eta},\boldsymbol{\rho},\boldsymbol{\Phi},\boldsymbol{\Theta}} O(\boldsymbol{\eta},\boldsymbol{\rho},\boldsymbol{\Phi},\boldsymbol{\Theta}). \tag{9}$$

We solve the above minimization problem with an iterative algorithm to be detailed in the following Sec. 5.2.

5.2 The Inference Algorithm

We iteratively update one of $\boldsymbol{\eta}$, $\boldsymbol{\rho}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Theta}$ when the others are fixed. The inference algorithm is summarized in Algorithm 1.

Update $\boldsymbol{\eta}$ given $\{\boldsymbol{\rho},\boldsymbol{\Phi},\boldsymbol{\Theta}\}$. Once given $\boldsymbol{\rho}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Theta}$, the optimal $\boldsymbol{\eta}$ that minimizes O in Eq. (8) has a closed-form solution. One can derive this closed-form update formula by looking back to the total likelihood Eq. (6), since

$$\begin{aligned} \mathcal{L} &\propto \prod_{s\in\mathcal{S}}\prod_{t=1}^{T}\mathrm{Exp}\left(P_{st}\ ;\ \frac{\eta_t}{\tau_s\sum_{k=1}^{K}\phi_{sk}\theta_{kt}}\right) \\ &= \prod_{t=1}^{T}\left[\mathrm{Exp}\left(\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\frac{P_{st}}{\tau_s\sum_{k=1}^{K}\phi_{sk}\theta_{kt}}\ ;\ \eta_t\right)\right]^{|\mathcal{S}|}, \end{aligned}$$

where $\tau_s = \rho_u\rho_v$ for node pair $s = (u,v)$. Using the property of exponential distributions, we find the $\boldsymbol{\eta}$ that maximizes $\mathcal{L}$, and hence minimizes O, can be computed by

$$\eta_t = \left(\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\frac{P_{st}}{\tau_s\sum_{k=1}^{K}\phi_{sk}\theta_{kt}}\right)^{-1}. \tag{10}$$

Update $\boldsymbol{\rho}$ given $\{\boldsymbol{\eta},\boldsymbol{\Phi},\boldsymbol{\Theta}\}$. Unlike $\boldsymbol{\eta}$, closed-form formula for updating $\boldsymbol{\rho}$ does not exist because (i) $\boldsymbol{\rho}$ has an informative prior, and (ii) the generating process for paths between node pair (u,v) involves the coupling of ρ_u and ρ_v. Fortunately, the gamma distribution is the conjugate prior to the exponential distribution. Therefore, for each u, when the rest $\{\rho_v\}_{v\neq u}$ are fixed, the closed-form update formula for ρ_u can be derived as follows. Denote $\{\xi_s\}_{s\in\mathcal{S}}$ the following quantities that are fixed during the $\boldsymbol{\rho}$ update phase

$$\xi_s := \sum_{t=1}^{T}\frac{\eta_t P_{st}}{\sum_{k=1}^{K}\phi_{sk}\theta_{kt}},$$

and we have $\frac{\partial O}{\partial\rho_u} = \sum_{\substack{v\in\mathcal{V}\setminus\{u\}\\ s=(u,v)}}\left[\sum_{t=1}^{T}\frac{1}{\rho_u} - \frac{\xi_s}{\rho_u^2\rho_v}\right] - \frac{\alpha-1}{\rho_u} + 1$. Setting this partial derivative to 0 leads to

$$\rho_u^2 + [(|\mathcal{V}|-1)\cdot T - (\alpha-1)]\,\rho_u - \sum_{\substack{v\in\mathcal{V}\setminus\{u\}\\ s=(u,v)}}\frac{\xi_s}{\rho_v} = 0. \tag{11}$$

Note that Eq. (11) is a single-variable quadratic equation with one positive and one negative roots. Furthermore, O is convex *w.r.t.* ρ_u on the positive half-axis, and the positive root is a minimum of O. Therefore, the optimal ρ_u that minimizes O is given by the positive root of the quadratic equation (Eq. (11)), which has closed-form solution. Holistically, we update $\boldsymbol{\rho}$ by iterating through $u \in \mathcal{V}$ to update ρ_u with the aforementioned closed-form solution to Eq. (11).

Update $\boldsymbol{\Theta}$ given $\{\boldsymbol{\eta},\boldsymbol{\rho},\boldsymbol{\Phi}\}$. To update $\boldsymbol{\Theta}$, we use the projected gradient descent (PGD) algorithm [12]. The gradient is given by

$$\frac{\partial O}{\partial\boldsymbol{\Theta}} = \boldsymbol{\Phi}^\top\left[\frac{1}{\boldsymbol{\Phi\Theta}} - \frac{\boldsymbol{P}}{(\boldsymbol{\tau}(\boldsymbol{\eta}^{\circ-1})^\top)\circ(\boldsymbol{\Phi\Theta})^{\circ 2}}\right], \tag{12}$$

where $[\cdot]\circ[\cdot]$, $\frac{[\cdot]}{[\cdot]}$, and $[\cdot]^{\circ[\cdot]}$ are element-wise multiplication, division, and power. Additional constraint fed into PGD is that each row of $\boldsymbol{\Theta}$ lies in the standard $(T-1)$-simplex, *i.e.*, $\sum_{t=1}^{T}\theta_{kt} = 1$ for all $k \in \{1,\ldots,K\}$ and $\theta_{kt} \geq 0$ for all $(k,t) \in \{1,\ldots,K\}\times\{1,\ldots,T\}$. Projection onto the standard simplex or the direct product of multiple standard simplices can be achieved efficiently using the method introduced in [3].

Update $\boldsymbol{\Phi}$ given $\{\boldsymbol{\eta},\boldsymbol{\rho},\boldsymbol{\Theta}\}$. Similarly, we use PGD to update $\boldsymbol{\Phi}$, where the gradient is given by

$$\frac{\partial O}{\partial\boldsymbol{\Phi}} = \left[\frac{1}{\boldsymbol{\Phi\Theta}} - \frac{\boldsymbol{P}}{(\boldsymbol{\tau}(\boldsymbol{\eta}^{\circ-1})^\top)\circ(\boldsymbol{\Phi\Theta})^{\circ 2}}\right]\boldsymbol{\Theta}^\top - \frac{\beta-1}{\boldsymbol{\Phi}}. \tag{13}$$

However, directly updating the entire $\boldsymbol{\Phi}$ using PGD can be problematic, because the row number of $\boldsymbol{\Phi}$ is the same as the number of nontrivial node pairs, $|\mathcal{S}|$, which can be significantly larger than that of $\boldsymbol{\Theta}$.

Fortunately, we can decompose the update scheme for $\boldsymbol{\Phi}$ by rows, because each row is independent from the others. Specifically, we update each row s using PGD in parallel, with gradient $\frac{\partial O}{\partial\boldsymbol{\Phi}_{s,:}} = \left[\frac{1}{\boldsymbol{\Phi}_{s,:}\boldsymbol{\Theta}} - \frac{\boldsymbol{P}_{s,:}}{(\tau_s(\boldsymbol{\eta}^{\circ-1})^\top)\circ(\boldsymbol{\Phi}_{s,:}\boldsymbol{\Theta})^{\circ 2}}\right]\boldsymbol{\Theta}^\top - \frac{\beta-1}{\boldsymbol{\Phi}_{s,:}}$, and constraints $\sum_{k=1}^{K}\phi_{sk} = 1$ for all $s \in \mathcal{S}$ and $\phi_{st} \geq 0$ for all $(s,k) \in \mathcal{S}\times\{1,\ldots,K\}$.

5.3 Implementation Details

For program reproducibility, we provide details in parameter initialization and computational singularity handling.

Since the inference algorithm starts with updating $\boldsymbol{\eta}$, no initialization for $\boldsymbol{\eta}$ is needed. $\boldsymbol{\rho}$ is initialized by drawing random samples from its prior distribution, $\Gamma(\alpha, 1)$, where α is estimated from data as discussed in Sec. 4. $\boldsymbol{\Phi}$ is initialized uniformly at random within the row-wise simplex constraint. For $\boldsymbol{\Theta}$, the first T rows of this $K \times T$ matrix are initialized to be an identity matrix, because many node pairs with paths in between involve only one meta-path, and we initialize the rest $K - T$ rows uniformly at random within the row-wise simplex constraint. This choice is out of the consideration that the PReP model is not convex over all parameters.

Dirichlet distribution is defined over open sets with unbounded probability density function. As a result, when using MAP, certain components of $\boldsymbol{\Phi}$ can be inferred to approach the singularities along the boundary. Therefore, in practice, we let $\boldsymbol{\Phi}$ to be bounded away from the boundary with an infinitesimal quantity δ, *i.e.*, each of its entries must not only be positive, but also be greater or equal to δ. In this way, we keep the capability of Dirichlet distribution in modeling *cross-meta-path synergy*, while ensuring the model is computationally meaningful. In our experiment, we set $\delta = 10^{-50}$. With this constraint, the domain of definition for $\boldsymbol{\Phi}$ is no longer a standard simplex as discussed in [3]. For this reason, we provide the algorithm for efficient projection onto this shrunken simplex, $\{\mathbf{x} \in \mathbb{R}^K | x_i \geq \delta, \sum_{i=1}^{K} x_i = 1\}$ in the Appendix, which is required by the inference algorithm. Note that if one wishes to evade the point estimation of parameters in the PReP model, Eq. (6), and thereby avoid computational singularity, they can treat all model parameters as hidden variables and derive relevance from the marginal likelihood of the observed data as discussed in Sec. 4.2. The exploration of this direction requires novel method, such as a sampling algorithm design for our model, to efficiently calculate marginal likelihood, and we defer this to future work.

6 EXPERIMENTS

In this section, we quantitatively evaluate the proposed model on two publicly available real-world HINs: Facebook and DBLP. We first describe the datasets and the unsupervised tasks used for evaluation. Baselines and model variations for comparison are then introduced. Afterward, we present experiment results together with discussions, which demonstrate the advantage of using probability as the backbone of relevance.

6.1 Data Description and Evaluation Tasks

In this section, we introduce the two publicly available real-world datasets and the evaluation tasks.

The Facebook dataset. This dataset [11] contains nodes of 11 types, including user, major, degree, school, hometown, surname, location, employer, work-location, work-project, and other. It consists of 5,621 nodes and 98,023 edges, among which 4,167 nodes are of the user type. We aim to determine the relevance between users, using 10 meta-paths, each of the form [user]–[X]–[user], where X is any of the above 11 node types except for other.

To derive ground truth label between user pairs for evaluation, we use being friends on Facebook as a proxy for being relevant. This dataset is collected by recruiting participants to label their own Facebook friends It consists of 10 distinct ego networks, where an ego network consists of one ego user and all her friends together with edges attached to these users. We hence perform one sub-task for each ego network, where the compared measures are used to calculate the relevance between all pairs of non-ego users in this ego network.

We use two evaluation metrics widely adopted for tasks with multiple relevant instances: the area under the receiver operating characteristic curve (ROC-AUC) and the area under precision-recall curve (AUPRC). The receiver operating characteristic curve (ROC) is created by plotting true positive rate against false positive rate as the threshold varies, while the precision-recall curve (PRC) is drawn by plotting precision against recall as the threshold varies. Higher values are more preferred for both ROC-AUC and AUPRC. We further average each of the above metrics across ego networks with the following methods – uni.: averaging over all ego networks uniformly; rel.: weighting by the number of relevant pairs in each ego network; tot.: weighting by the total number of pairs in each ego network.

The DBLP dataset. This dataset is derived from the DBLP dataset processed by Tang et al. [18] containing computer science research papers together with author names and publication venue associated to each paper. It consists of 13,697 nodes and 19,665 edges, among which 1,546 nodes are of the author type. Notably, in this dataset, the same author name associated with two papers may not necessarily be the same person. Based on this fact, we design an entity resolution task as follows. First, we use the labels made available by Tang et al. [18] to group all author name mentions corresponding to one person to define an author node. In this way, an author node is linked to multiple papers written by her. Then, for each author name, we split the author node with the most author name mentions into two nodes, and we define two nodes to be relevant if and only if they actually refer to the same person. Finally, we perform one sub-task for each author name, where the compared measures are used to calculate the relevance between all pairs of nodes with the same author name.

We use 14 meta-paths in this task, each of the form [author]–[paper]–[venue domain]–[paper]–[author], where a node of the venue domain type corresponds to one of the 14 computer science research areas. The definition of the 14 areas is derived from the Wikipedia page: List of computer science conferences[1]. Since only one relevant pair exists in each sub-task, the mean reciprocal rank (MRR) is used as the evaluation metric, where, for each sub-task, the reciprocal rank is the reciprocal of the rank of the relevant pair. Higher values indicate better results for MRR. We also average the above metrics across different sub-tasks using three methods: uni., rel., and tot. Note that uni. and rel. are equivalent in this entity resolution task because each sub-task has exactly one relevant pair.

6.2 Baselines and Variations

In this section, we describe the meta-path-based baseline methods and variations of the PReP model, which are used to compare with our proposed full PReP model. Existing meta-path-based unsupervised HIN measures define relevance computation method on each meta-path and then use linear combination to find the composite score. Therefore, each baseline consists of two parts: (i) the base

[1] https://en.wikipedia.org/wiki/List_of_computer_science_conferences

Dataset	Metric		PathCount		PathSim		JoinSim		SimRank		PReP			
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	No-NV	No-PS	No-CS	(full)
Facebook	ROC-AUC	uni.	0.8056	0.8598	0.8367	0.8586	0.8326	0.8547	0.7977	0.8303	0.8310	0.6702	0.8689	**0.8850**
		rel.	0.8612	0.8879	0.8578	0.8888	0.8556	0.8872	0.8076	0.8596	0.8556	0.6713	0.8880	**0.9133**
		tot.	0.8558	0.8849	0.8577	0.8866	0.8557	0.8851	0.8096	0.8594	0.8547	0.6773	0.8893	**0.9139**
	AUPRC	uni.	0.2456	0.2832	0.2370	0.2845	0.2340	0.2803	0.2055	0.2435	0.2183	0.1650	**0.3273**	**0.3269**
		rel.	0.2496	0.3048	0.2142	0.2873	0.2117	0.2837	0.1764	0.2408	0.2067	0.1283	0.3354	**0.3486**
		tot.	0.2107	0.2542	0.1841	0.2460	0.1821	0.2432	0.1523	0.2071	0.1760	0.1089	0.3010	**0.3080**
DBLP	MRR	uni./rel.	0.8091	0.8130	0.6922	0.7003	0.7454	0.7538	0.6636	0.6738	0.8223	0.8494	0.8365	**0.8517**
		tot.	0.7839	0.7871	0.6612	0.6731	0.7128	0.7244	0.6302	0.6357	0.8234	**0.8407**	0.8264	0.8391

Table 3: Quantitative evaluation results on two real-world datasets using the proposed measure, PReP, and other measures.

measure that calculates the relevance score on one meta-path, and (ii) the weights assigned to different meta-paths used in the linear combination. The 4 base measures we used are:

- **PathCount** [17]. $PathCount_{\mathbf{w}}(s) \coloneqq \sum_t w_t P_{st}$.
- **PathSim** [17]. $PathSim_{\mathbf{w}}(s) \coloneqq \sum_t w_t \cdot \frac{2 \cdot P_{\langle uv \rangle t}}{P_{\langle uu \rangle t} + P_{\langle vv \rangle t}}$.
- **JoinSim** [20]. $JoinSim_{\mathbf{w}}(s) \coloneqq \sum_t w_t \cdot \frac{P_{\langle uv \rangle t}}{\sqrt{P_{\langle uu \rangle t} \cdot P_{\langle vv \rangle t}}}$.
- **SimRank**. We adopt SimRank [8] with meta-path constraints. Let $\boldsymbol{A}$ be a matrix, where A_{uv} is the number of paths under this meta-path between node pair (u, v) after column normalization. The SimRank score is then given by S_{uv}, where $\boldsymbol{S}$ is the solution to $\boldsymbol{S} = \max\{C \cdot (\boldsymbol{A}^\top \boldsymbol{S} \boldsymbol{A}), \boldsymbol{I}\}$, and C is the decay factor to be specified. Note that we use SimRank instead of random walk with restart because SimRank is a symmetric relevance measure.

Without any supervision available, we use 2 heuristics to determine the weights $\mathbf{w}$ for linear combination.

- **Mean**. Let w_t be the reciprocal of the mean of all scores computed using the corresponding base measure on meta-path t.
- **SD**. Let w_t be the reciprocal of the standard deviation of all scores computed using the corresponding base measure on meta-path t. Note that this heuristic normalized the original score in the way similar to z-score.

Combining the aforementioned 4 base measures and 2 heuristic for setting weights, we have 8 baselines in total.

Additionally, we also experiment with three variations of PReP, which are partial models with one of the three components knocked out from the full PReP model.

- No *node visibility* (No-NV): Set $\boldsymbol{\rho} = \mathbf{1}_{|\mathcal{V}|}$, and do not update $\boldsymbol{\rho}$ during model inference.
- No *path selectivity* (No-PS): Set $\boldsymbol{\eta} = \mathbf{1}_T$, and do not update $\boldsymbol{\eta}$ during model inference.
- No *cross-meta-path synergy* (No-CS): Set $\boldsymbol{\Phi} = \mathbf{1}_{|\mathcal{V}|\times K}/K$, $\boldsymbol{\Theta} = \mathbf{1}_{|\mathcal{V}|\times T}/T$, and do not update $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ during model inference.

Note that $\mathbf{1}_M$ stands for all one column vector of size M and $\mathbf{1}_{M\times N}$ denotes all one matrix of size $M \times N$.

6.3 Effectiveness and Discussion

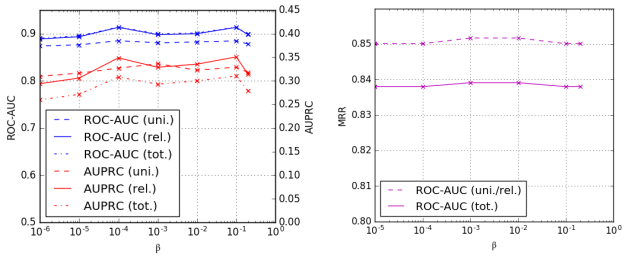
In this section, we present the quantitative evaluation results on both the Facebook and the DBLP datasets. We tune the decay factor C in the baseline measure, SimRank, to have the best performance with $C = 0.5$ for both SimRank-Mean and SimRank-SD on Facebook, and $C = 0.8$ for SimRank-Mean, $C = 0.7$ for SimRank-SD on DBLP. We set hyperparameters of PReP as $K = 15$ and $\beta = 10^{-4}$ for Facebook and $K = 14$ and $\beta = 10^{-2}$ for DBLP. The choice of hyperparameters will be further discussed in this section.

As presented in Tab. 3, PReP outperformed all 8 baselines under various metrics. Moreover, PReP outperformed its 3 variations under most metrics, suggesting each component of the model generally has a positive effect on the performance of the full PReP model. Note that under MRR (tot.), PReP performed slightly worse than PReP-No-PS, the partial model without η_t for *path selectivity*. This happened because, as discussed in Sec. 4, we cannot enforce task-specific design on *path selectivity* η_t due to the lack of supervision, and we expect *path selectivity* η_t to play a more important role in future work where relevance labels are provided as supervision.

Additionally, we have made the following observations.

Heuristic methods cannot yield robust relevance measures. Compared with PathCount, both PathSim and JoinSim further model *node visibility*, which penalizes the relevance with nodes that are highly visible. However, as Tab. 3 presents, PathSim and JoinSim cannot always outperform PathCount. Moreover, JoinSim performs better than PathSim on DBLP, while PathSim is slightly better than JoinSim on Facebook. We interpret these results as, PathSim and JoinSim model *node visibility* in a deterministic heuristic way. Unlike our generative-model-based measure that derives relevance measure based on parameters inferred from each HIN, the heuristic approaches adopted by PathSim and JoinSim have varying performance on different HINs. This suggests being data-driven is a favorable property of PReP.

Non-one-hot generating patterns help only when meta-paths correlate. In our experiment, we set $K = 14 = T$ for DBLP. Recall that we initialized the first T rows of $\boldsymbol{\Theta}$, the matrix representing the K generating patterns, to be T one-hot vectors corresponding to T meta-paths. We observed in the DBLP experiment that after model fitting, $\boldsymbol{\Theta}$ was still the same as its initialization, meaning each inferred generating pattern only generated path instances under exactly one meta-path. Moreover, by increasing the value of K, we did not see improvement in performance. This observation is inline with the situation that it is not frequently seen that two authors both publish papers in two distinct research areas, where the 14 areas on the Wikipedia page have been defined to be distinct areas including theory, software, parallel computing, *etc.* In this case, it is preferred to model synergy across every pair of meta-paths, and not to employ any non-one-hot generating patterns.



(a) Facebook **(b) DBLP**

Figure 3: Performance with varying β.

On the other hand, we used $K = 15 > T$ for Facebook, and we did observe non-one-hot generating patterns after model fitting. The most popular non-one-hot generating pattern consisted of three meta-paths: [user]–[hometown]–[user], [user]–[school]–[user], and [user]–[user]–[user], where we define popularity of a generating pattern as the fraction of node pairs adopting this pattern, *i.e.*, $pop(k) = \sum_{s \in \mathcal{S}} \phi_{sk}$. This generating pattern corresponds to two users sharing the same hometown, the same school, and having common friends. This scenario is common for two people sharing similar friend group back in the hometown school.

Sensitivity of β in modeling *cross-meta-path synergy*. In the PReP model (Eq. (6)) and relevance measure (Eq. (7)), the concentration parameter β of the Dirichlet prior controls the extent to which we boost *cross-meta-path synergy*. Experiment results in Fig 3 shows performance of PReP do not significantly change around the values we have set for β, *i.e.*, 10^{-4} for Facebook and 10^{-2} for DBLP.

7 RELATED WORK

In this section, we review the study on HIN relevance. The problem of deriving relevance between node pairs has been extensively studied for homogeneous information networks. Relevance measures of this type include the random walk based *Personalized PageRank* and *SimRank* [8], the neighbor-based *common neighbors* and *Jaccard's coefficient*, the path-based *Katz* [5], *etc.* To generalize relevance from the homogeneous networks to the typed heterogeneous case, researchers have been exploring from multiple perspectives. One perspective, as in *PathCount* and *PathSim* from [17] and *Path-Constrained Random Walk* from [10], is to first compute relevance score along each meta-path, and then glue scores from all types together via linear combination to establish the composite measure. A great many applications [9, 14, 16, 22, 24] based on this meta-path paradigm with linear combination have been proposed. Our proposed method follows this meta-path paradigm, but goes beyond linear combination to model *cross-meta-path synergy* that we have observed from real-world HINs. Another perspective is to go beyond meta-path and derive relevance based on the more complex graph structures [4, 7]. While these approaches can yield good performance, they differ from our proposed methods for further entailing label information or expertise in designing graph structure. Also, they do not carry probabilistic interpretations. Besides, people have explored the idea of incorporating richer information [6, 21] to define more effective relevance scoring functions, or adding supervision to derive task-specific relevance measures [2, 19, 23]. While being valuable, these works are out of the scope of the problem we study in this paper, where we address the basic, unsupervised case with no additional information as our starting point of studying HIN relevance from the probabilistic perspective.

8 CONCLUSION AND FUTURE WORK

Inspired by the probabilistic interpretation of existing path-based relevance measures, we studied HIN relevance from a probabilistic perspective. We identified *cross-meta-path synergy* as one of the three characteristics that we deem important for HIN relevance. A generative model was proposed to derive a novel path-based relevance measure, PReP, which could capture the three important characteristics. An inference algorithm was also developed to find the maximum a posteriori (MAP) estimate of the model parameters, which entailed non-trivial tricks. Experiments on real-world HINs demonstrated the effectiveness of our relevance measure, which is data-driven and tailored for each HIN.

Future work includes the exploration of defining relevance from the proposed PReP model with marginal likelihood as discussed in Sec. 4.2. Further add-on designs to adapt the proposed model to a supervised setting are also worth exploring to unleash the potential of our model.

Acknowledgments. We thank our colleagues and friends for the enlightening discussions: Jason Jian Ge, Jiasen Yang, Carl Ji Yang, and many members of the Data Mining Group at UIUC. We also thank the anonymous reviewers for their insightful comments. This work was sponsored in part by the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1320617 and IIS 16-18481, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies of the U.S. Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

APPENDIX

We provide the algorithm for efficient projection onto the standard simplex shrunk by δ, $\{\mathbf{x} \in \mathbb{R}^K | x_i \geq \delta, \sum_{i=1}^K x_i = 1\}$, in Algorithm 2.

Algorithm 2: Efficient projection onto shrunk simplex

Input : the original vector $\mathbf{z} \in \mathbb{R}^K$ and the shrinking factor δ
Output: the projection $\mathbf{x} \in \mathbb{R}^K$
begin
  Sort $\mathbf{z}$ into $\mathbf{u}$: $u_1 \geq u_2 \geq \ldots \geq u_K$
  $\rho \leftarrow \max\{1 \leq j \leq K | u_j + \frac{1}{j}(1 - \delta K - \sum_{i=1}^{j} u_i) > 0\}$
  $\lambda \leftarrow \frac{1}{\rho}(1 - \delta K - \sum_{i=1}^{\rho} u_i)$
  $x_i \leftarrow \max\{z_i + \lambda, 0\} + \delta$

The validity of this algorithm can be established in a way similar to the proof of the algorithm for standard simplex [3].

REFERENCES

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

[2] Ting Chen and Yizhou Sun. 2017. Task-Guided and Path-Augmented Heterogeneous Network Embedding for Author Identification. In *WSDM*. ACM.

[3] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. 2008. Efficient projections onto the l 1-ball for learning in high dimensions. In *ICML*. ACM.

[4] Yuan Fang, Wenqing Lin, Vincent W Zheng, Min Wu, Kevin Chen-Chuan Chang, and Xiao-Li Li. 2016. Semantic Proximity Search on Graphs with Metagraph-based Learning. In *ICDE*. IEEE.

[5] Jiawei Han, Jian Pei, and Micheline Kamber. 2011. *Data mining: concepts and techniques*. Elsevier.

[6] Jiazhen He, James Bailey, and Rui Zhang. 2014. Exploiting transitive similarity and temporal dynamics for similarity search in heterogeneous information networks. In *DASFAA*. Springer.

[7] Zhipeng Huang, Yudian Zheng, Reynold Cheng, Yizhou Sun, Nikos Mamoulis, and Xiang Li. 2016. Meta Structure: Computing Relevance in Large Heterogeneous Information Networks. In *KDD*. ACM.

[8] Glen Jeh and Jennifer Widom. 2002. SimRank: a measure of structural-context similarity. In *KDD*. ACM.

[9] Jonathan Kuck, Honglei Zhuang, Xifeng Yan, Hasan Cam, and Jiawei Han. 2015. Query-based outlier detection in heterogeneous information networks. In *ICDE*. IEEE.

[10] Ni Lao and William W Cohen. 2010. Relational retrieval using a combination of path-constrained random walks. *Machine learning* 81, 1 (2010), 53–67.

[11] Julian J McAuley and Jure Leskovec. 2012. Learning to Discover Social Circles in Ego Networks. In *NIPS*. NIPS Foundation.

[12] Yurii Nesterov. 2013. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media.

[13] Chuan Shi, Xiangnan Kong, Yue Huang, S Yu Philip, and Bin Wu. 2014. Hetesim: A general framework for relevance measure in heterogeneous networks. *TKDE* 26, 10 (2014), 2479–2492.

[14] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S Yu Philip. 2017. A survey of heterogeneous information network analysis. *TKDE* 29, 1 (2017), 17–37.

[15] Yu Shi, Myunghwan Kim, Shaunak Chatterjee, Mitul Tiwari, Souvik Ghosh, and Rómer Rosales. 2016. Dynamics of Large Multi-View Social Networks: Synergy, Cannibalization and Cross-View Interplay. In *KDD*. ACM.

[16] Yizhou Sun and Jiawei Han. 2013. Mining heterogeneous information networks: a structural analysis approach. *SIGKDD Explorations* 14, 2 (2013), 20–28.

[17] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In *Proceedings of VLDB*. VLDB Endowment.

[18] Jie Tang, Alvis CM Fong, Bo Wang, and Jing Zhang. 2012. A unified probabilistic framework for name disambiguation in digital library. *TKDE* 24, 6 (2012), 975–987.

[19] Chi Wang, Rajat Raina, David Fong, Ding Zhou, Jiawei Han, and Greg Badros. 2011. Learning relevance from heterogeneous social network and its application in online targeting. In *SIGIR*. ACM.

[20] Yun Xiong, Yangyong Zhu, and S Yu Philip. 2015. Top-k similarity join in heterogeneous information networks. *TKDE* 27, 6 (2015), 1710–1723.

[21] Kun Yao, Hoi Fong Mak, and others. 2014. PathSimExt: Revisiting pathsim in heterogeneous information networks. In *WAIM*. Springer.

[22] Xiao Yu, Xiang Ren, Yizhou Sun, Quanquan Gu, Bradley Sturt, Urvashi Khandelwal, Brandon Norick, and Jiawei Han. 2014. Personalized entity recommendation: A heterogeneous information network approach. In *Proceedings of WSDM*. ACM.

[23] Xiao Yu, Yizhou Sun, Brandon Norick, Tiancheng Mao, and Jiawei Han. 2012. User guided entity similarity search using meta-path selection in heterogeneous information networks. In *CIKM*. ACM.

[24] Honglei Zhuang, Jing Zhang, George Brova, Jie Tang, Hasan Cam, Xifeng Yan, and Jiawei Han. 2014. Mining query-based subnetwork outliers in heterogeneous information networks. In *ICDM*. IEEE.