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

Yu Shi, Po-Wei Chan, Honglei Zhuang, Huan Gui, and Jiawei Han
University of Illinois at Urbana-Champaign (UIUC)
In real world applications, objects of different types can have different relations, which form heterogeneous information networks (HINs).

- **Typed** nodes: objects
- **Typed** edges: relations

![Diagram of relationships between people and universities]
In real world applications, objects of different types can have different relations, which form heterogeneous information networks (HINs).

- **Typed** nodes: objects
- **Typed** edges: relations

![Diagram with nodes and edges representing various entities and their relationships.]

- Wong
- Kaecilius
- Christine
- Stephen
- Mordo
- Christine
- Stephen
- Mordo
- Wong

- UC Berkeley
- Urbana, IL
- Stanford University
- Computer Science

- Lives in
- Attends
- Majors in
Heterogeneous information networks (HINs) are ubiquitous.

IMDb Network

Bibliographical Network

Biomedical Network

Economic Graph

Social Network

Facebook Open Graph

A fundamental problem in network mining:

**defining relevance measure**

a.k.a., similarity, proximity.

A good relevance measure can benefit downstream applications.
In the context of HIN, a relevance measure should be able to answer:

- How relevant are Mordo (person) and Stephen (person)?
- How relevant are UC Berkeley (university) and Berkeley, CA (location)?
Many existing HIN relevance measures are defined upon **meta-path**.

A **meta-path** (type of paths):

\[
\text{[location]} \xrightarrow{\text{lives-in}^{-1}} \text{[person]} \xrightarrow{\text{majors-in}} \text{[discipline]}
\]

A concrete **path instance** under this **meta-path**:

Berkeley, CA $\rightarrow$ Mordo $\rightarrow$ ComputerScience

Another **example**:

\[
\text{[person]} \xrightarrow{\text{attends}} \text{[university]} \xrightarrow{\text{attends}^{-1}} \text{[person]}
\]

Wong $\rightarrow$ UC Berkeley $\rightarrow$ Mordo
For given meta-path $t$ and a pair of node $s = (u, v)$

- $P_{st}$ or $P_{(uv)t}$: the path count between $s = (u, v)$ under meta-path $t$.

Examples:

$$P_{\langle \text{Won. Mor.} \rangle_t} = 1$$

$$P_{\langle \text{Mor. Mor.} \rangle_t} = 2$$
Widely-used HIN relevance measures:

- **PathCount** [1]: simply the path count between $u$ and $v$

\[
\text{PathCount}^{(t)}(u, v) := \text{P}_{\langle uv \rangle t}
\]

$\text{PathCount}^{(t)}(\text{Won.}, \text{Mor.}) = 1$

\[
\mathcal{M}_t \quad (\longrightarrow) : \text{[person]} \xrightarrow{\text{attends}} \text{[university]} \xrightarrow{\text{attends}^{-1}} \text{[person]}
\]
Widely-used HIN relevance measures:

- **PathCount** [1]: simply the path count between $u$ and $v$
  \[
  \text{PathCount}^{(t)}(u, v) := P_{(uv)t}^{(t)}
  \]
  \[
  \text{PathCount}^{(t)}(\text{Won.}, \text{Mor.}) = 1
  \]

- **PathSim** [1]: further penalizes nodes with more “cycles”
  \[
  \text{PathSim}^{(t)}(u, v) := \frac{2 \cdot P_{(uv)t}^{(t)}}{P_{(uu)t}^{(t)} + P_{(vvt)}^{(t)}}
  \]
  \[
  \text{PathSim}^{(t)}(\text{Won.}, \text{Mor.}) = \frac{2 \cdot 1}{1 + 2} \approx 0.67
  \]

Mordo attends multiple universities. It is hence less significant for Wong and Mordo to be schoolmates.
Widely-used HIN relevance measures:

- **PathCount [1]**: simply the path count between \( u \) and \( v \)
  \[
  \text{PathCount}^t(u, v) := P_{(uv)}^t
  \]
  \[
  \text{PathCount}^t(\text{Won}, \text{Mor.}) = 1
  \]

- **PathSim [1]**: further penalizes nodes with more “cycles”
  \[
  \text{PathSim}^t(u, v) := \frac{2 \cdot P_{(uv)}^t}{P_{(uu)}^t + P_{(vv)}^t}
  \]
  \[
  \text{PathSim}^t(\text{Won}, \text{Mor.}) = \frac{2 \cdot 1}{1 + 2} \approx 0.67
  \]

- **JoinSim [2]**: another way to penalize
  \[
  \text{JoinSim}^t(u, v) := \frac{P_{(uv)}^t}{\sqrt{P_{(uu)}^t \cdot P_{(vv)}^t}}
  \]
  \[
  \text{JoinSim}^t(\text{Won}, \text{Mor.}) = \frac{1}{\sqrt{1 \cdot 2}} \approx 0.71
  \]
Linear combination is usually used to combine multiple meta-paths.

- Let $\mathbf{w} = \{w_1, \ldots, w_T\}$, where $w_t$ is the weight for meta-path $t$

$$\text{PathCount}_\mathbf{w}(u, v) := \sum_{t=1}^{T} w_t \cdot \text{PathCount}^{(t)}(u, v)$$

$\text{PathCount}_\mathbf{w}(\text{Won., Mor.}) = w_1 \cdot 1 + w_2 \cdot 1 = w_1 + w_2$

$$\text{PathSim}_\mathbf{w}(u, v) := \sum_{t=1}^{T} w_t \cdot \text{PathSim}^{(t)}(u, v)$$

$$\text{JoinSim}_\mathbf{w}(u, v) := \sum_{t=1}^{T} w_t \cdot \text{JoinSim}^{(t)}(u, v)$$
Why can these heuristic measures reflect relevance?
• Most node pairs are not connected by path instances. It is a significant event to observe (many) path instance(s) between a pair of nodes as measured by \textbf{PathCount}.

\[
\text{PathCount}^t(u, v) := P_{(uv)}^t
\]

• \textbf{PathSim} penalizes nodes with more “cycles”, because it is a less significant event to have path instances with these nodes.

\[
\text{PathSim}^t(u, v) := \frac{2 \cdot P_{(uv)}^t}{P_{(uu)}^t + P_{(vv)}^t}
\]

Can we establish probabilistic interpretation to quantify such significance?

• Yes.
By assuming the generating process of path instances via exponential distribution.

\[ P_{st} \sim \text{Exp} (\lambda) \]

\( P_{st} \): the path count between \( s = (u, v) \) under meta-path \( t \).

The negative log-likelihood of observing such path instances:

\[
-LL^{(t)}(s) = -\log(\lambda e^{-\lambda P_{st}}) = \lambda P_{st} - \log \lambda
\]

\[ \propto P_{st} + \text{const} = \text{PathCount}^{(t)}(s) + \text{const}. \]

Existing relevance measure
If we assume path instances are generated with a **meta-path-specific rate** $w_t$,

\[ P_{st} \sim \text{Exp}(w_t) \]

Likelihood

\[ -LL(s) = -\log(\prod_t w_t e^{-w_t P_{st}}) = \sum_t w_t P_{st} - \sum_t \log w_t \]

\[ = \sum_t w_t P_{st} + \text{const} = \text{PathCount}_w(s) + \text{const}. \]

Existing relevance measure

Similar results are derived for PathSim and JoinSim by adding a node-pair-specific component $\kappa_s$.

\[ P_{st} \sim \text{Exp}(w_t / \kappa_s) \]

Likelihood $\leftrightarrow$ Relevance
Beyond the probabilistic interpretation, we identify three characteristics important for path-based HIN relevance.

1. Node visibility

\[ M_1 (\rightarrow) : [\text{person}] \xrightarrow{\text{attends}} [\text{university}] \xrightarrow{\text{attends}^{-1}} [\text{person}] \]
\[ M_2 (\leftarrow) : [\text{person}] \xrightarrow{\text{livesIn}} [\text{location}] \xrightarrow{\text{livesIn}^{-1}} [\text{person}] \]
\[ M_3 (\cdots) : [\text{person}] \xrightarrow{\text{majorsIn}} [\text{discipline}] \xrightarrow{\text{majorsIn}^{-1}} [\text{person}] \]

Modeled by PathSim and JoinSim.
Beyond the probabilistic interpretation, we identify three characteristics important for path-based HIN relevance.

1. Node visibility
2. Path selectivity

Modeled by weights of meta-paths in linear combination.
Beyond the probabilistic interpretation, we identify three characteristics important for path-based HIN relevance.

1. **Node visibility**

2. **Path selectivity**

3. **Cross-meta-path synergy**

   - It is less likely to observe the co-occurrence of path instances under multiple uncorrelated meta-paths, and observing it implies high relevance.

\[
\mathcal{M}_1 (\rightarrow): [\text{person}] \xrightarrow{\text{attends}} [\text{university}] \xrightarrow{\text{attends}^{-1}} [\text{person}]
\]

\[
\mathcal{M}_2 (\Leftarrow): [\text{person}] \xrightarrow{\text{livesIn}} [\text{location}] \xrightarrow{\text{livesIn}^{-1}} [\text{person}]
\]

\[
\mathcal{M}_3 (\cdots): [\text{person}] \xrightarrow{\text{majorsIn}} [\text{discipline}] \xrightarrow{\text{majorsIn}^{-1}} [\text{person}]
\]
Beyond the probabilistic interpretation, we identify three characteristics important for path-based HIN relevance.

1. **Node visibility**

2. **Path selectivity**

3. **Cross-meta-path synergy**

   - It is less likely to observe the co-occurrence of path instances under multiple uncorrelated meta-paths, and observing it implies high relevance.

   ![Diagram showing node visibility and path selectivity](image)

   ![Graph showing one meta-path and multiple meta-paths](image)

Not modeled by existing measures, and observed in real-world data.
By

• generalizing the probabilistic interpretation,
• with intention to model the three characteristics,

we propose a novel Path-based Relevance from Probabilistic perspective:

PReP
1. Models the generating process of path instances under each meta-path.

2. Estimates model parameters by fitting the given HIN.
   • To find what scenario is most likely in this dataset.
   • PReP is therefore a relevance measure tailored for each dataset.

3. Computes relevance score for each node pair with negative log-likelihood.
   • PReP is a generalization of PathCount, PathSim, and JoinSim.
$\eta_t$ models the path selectivity of paths under meta-path $t$

$s = (u, v)$ denotes a node pair; $t$ denotes a meta-path

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

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

$\rho_u$ and $\rho_v$ model the node visibility of $u$ and $v$, respectively.

Each node further regularized by a gamma prior:

$\rho_z \sim \text{Gamma}(\alpha, 1)$

$\psi_{st}$ governs the distribution of meta-paths between $s$.

It is given by a mixture of $K$ generating patterns with $\phi_{sk}$ from the $k$-th, and the $k$-th contains a portion $\theta_{kt}$ of path instances under meta-path $t$:

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

where $\sum_{k=1}^{K} \phi_{sk} = 1$ and $\sum_{t=1}^{T} \theta_{kt} = 1$.

Each node pair adopts a few generating patterns to model cross-meta-path synergy:

$\phi_s \sim \text{Dir}_K(\beta)$

After model inference, the relevance between $u$ and $v$ is derived from negative log-likelihood:

$r(u, v) = \sum_{t=1}^{T} \frac{P_{st}}{\rho_u \rho_v \eta_t} \frac{1}{\sum_{k=1}^{K} \phi_{sk} \theta_{kt}} + (1 - \beta) \sum_{k=1}^{K} \log \phi_{sk}$
We find the maximum a posteriori (MAP) estimate for model parameters.

- The proposed algorithm iteratively update model parameters: $\eta$, $\rho$, $\Phi$, and $\Theta$.

**Update $\eta$**

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

Closed-form

**Update $\rho$** by solving

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

Closed-form for each node $u$

**Update $\Phi$** using projected gradient descent (PGD) in parallel

$$
\frac{\partial O}{\partial \Phi} = \left[ \frac{1}{\Phi \Theta} - \frac{P}{(\tau (\eta^{\theta-1})^T \circ (\Phi \Theta)^{\theta_2})} \right] \Theta^T - \frac{\beta^{-1}}{\Phi}
$$

s.t. $\sum_{k=1}^{K} \phi_{sk} = 1$ and $\phi_{st} \geq 0$

Rows of $\Phi$ are independent and can be updated in parallel

**Update $\Theta$** using PGD

$$
\frac{\partial O}{\partial \Theta} = \Phi^T \left[ \frac{1}{\Phi \Theta} - \frac{P}{(\tau (\eta^{\theta-1})^T \circ (\Phi \Theta)^{\theta_2})} \right]
$$

s.t. $\sum_{t=1}^{T} \theta_{kt} = 1$ and $\theta_{kt} \geq 0$

size of $\Theta \ll$ size of $\Phi$
Datasets and evaluation tasks

• **Facebook**: to infer whether two users are friends.

  Meta-paths [user]--[X]--[user] are used, where X is one of 10 node types in this HIN. The area under the receiver operating characteristic curve (ROC-AUC) and the area under precision-recall curve (AUPRC) are used as evaluation metrics.

• **DBLP**: to resolve duplicates of author node.

  Meta-paths [author]--[paper]--[X]--[paper]--[author] are used, where X is one of the 14 computer sciences research areas papers are published in. Each author node queried in this task is designed to have exactly one duplicate. The mean reciprocal rank (MRR) is used as the evaluation metric.
Experiments

Baselines

- (i) PathCount, (ii) PathSim, (iii) JoinSim, and (iv) SimRank are used as baselines to compute relevance scores for a single meta-path.

- Without any supervision, we use 2 heuristics to determine the weights \( w = \{w_1, ..., w_T\} \) for linear combination: Mean and SD (standard deviation).

Variants of PReP

- 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: (i) No node visibility (No-NV); (ii) No path selectivity (No-PS); (iii) No cross-meta-path synergy (No-CS).
Experiments

- **PReP** outperformed all baselines, which demonstrates the effectiveness of the proposed PReP model.
- **PReP** generally outperformed all variants (partial models), which suggests each model component has a positive effect on the performance of the full model.
- Heuristic methods cannot yield robust relevance measures, while **PReP** is tailored for each dataset.
  - E.g., with different heuristics on node visibility, PathSim and JoinSim cannot consistently outperform the other.
- Please check out our paper for more results and observations.

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>PathCount</th>
<th>PathSim</th>
<th>JoinSim</th>
<th>SimRank</th>
<th>PReP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Facebook</td>
<td>ROC-AUC</td>
<td>uni.</td>
<td>0.8056</td>
<td>0.8598</td>
<td><strong>0.8367</strong></td>
<td><strong>0.8556</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>rel.</td>
<td>0.8612</td>
<td>0.8879</td>
<td>0.8578</td>
<td>0.8888</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tot.</td>
<td>0.8558</td>
<td>0.8849</td>
<td>0.8577</td>
<td>0.8866</td>
</tr>
<tr>
<td></td>
<td>AUPRC</td>
<td>uni.</td>
<td>0.2456</td>
<td>0.2832</td>
<td>0.2370</td>
<td>0.2845</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rel.</td>
<td>0.2496</td>
<td>0.3048</td>
<td>0.2142</td>
<td>0.2873</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tot.</td>
<td>0.2107</td>
<td>0.2542</td>
<td>0.1841</td>
<td>0.2460</td>
</tr>
<tr>
<td>DBLP</td>
<td>MRR</td>
<td>uni./rel.</td>
<td>0.8091</td>
<td>0.8130</td>
<td>0.6922</td>
<td>0.7003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tot.</td>
<td>0.7839</td>
<td>0.7871</td>
<td>0.6612</td>
<td>0.6731</td>
</tr>
</tbody>
</table>
Future Work

   • Without supervision, current model assumes uninformative prior on $\eta_t$.
   • The best weights on meta-paths for different task can differ significantly.

2. Instead of MAP estimate on parameters of the proposed model, treating all model parameters as hidden variables and define the relevance as the marginal likelihood of the observed path instances.

3. Further add-on designs to adapt the proposed model to a supervised setting.
Summary

1. We establish the **probabilistic interpretation** for path-based HIN relevance measures.

2. We identify node visibility, path selectivity, and **cross-meta-path synergy** as three important characteristic in path-based HIN relevance, where cross-meta-path synergy is not modeled by existing methods.

3. We propose an novel relevance measure (**PReP**) based on a generative model, which is tailored for each HIN.

4. Experiments on two real-world HINs corroborated the effectiveness of our proposed model and relevance measure.