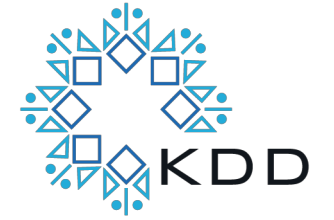


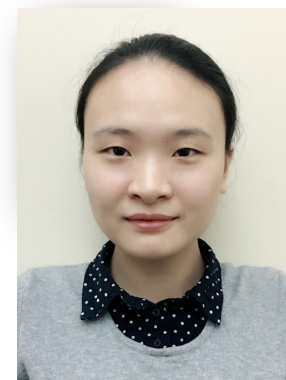


ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN



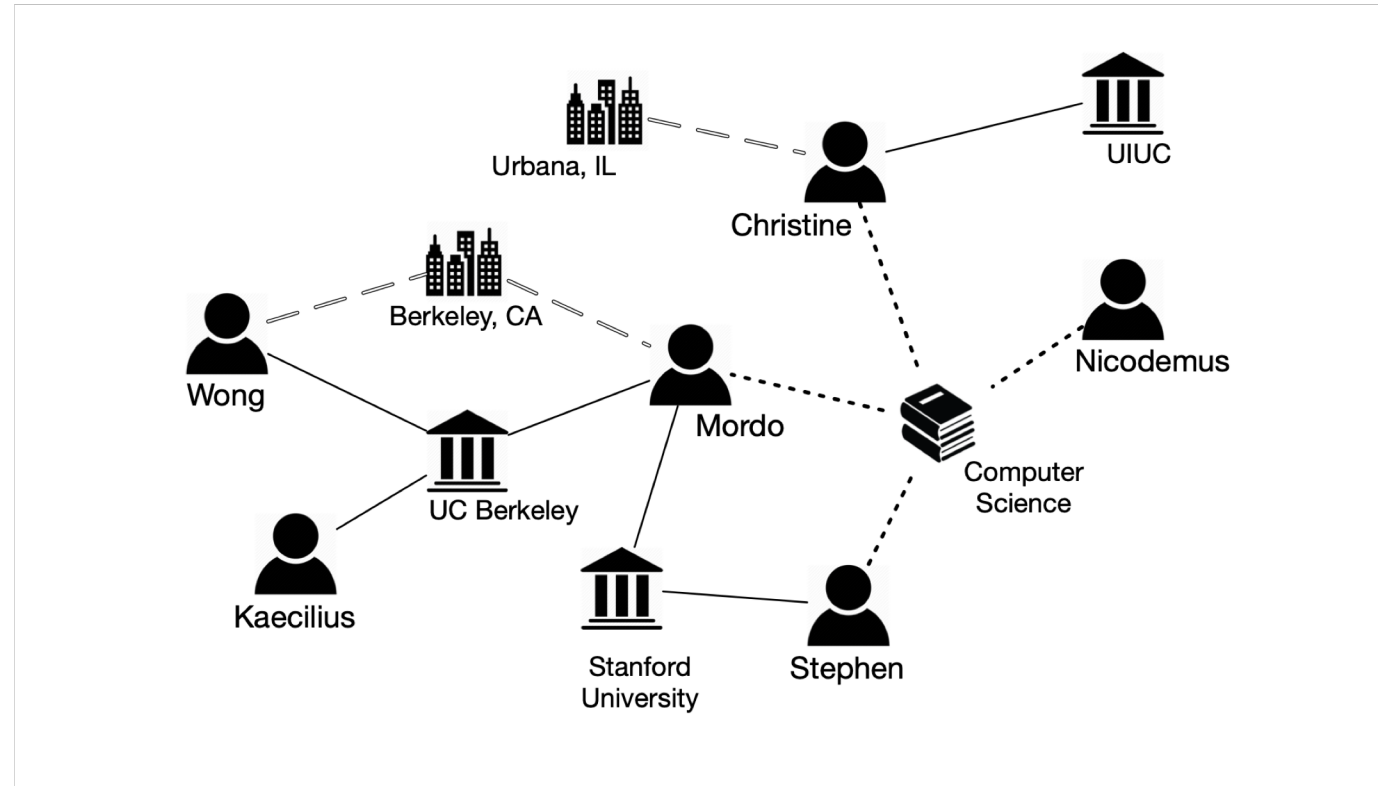
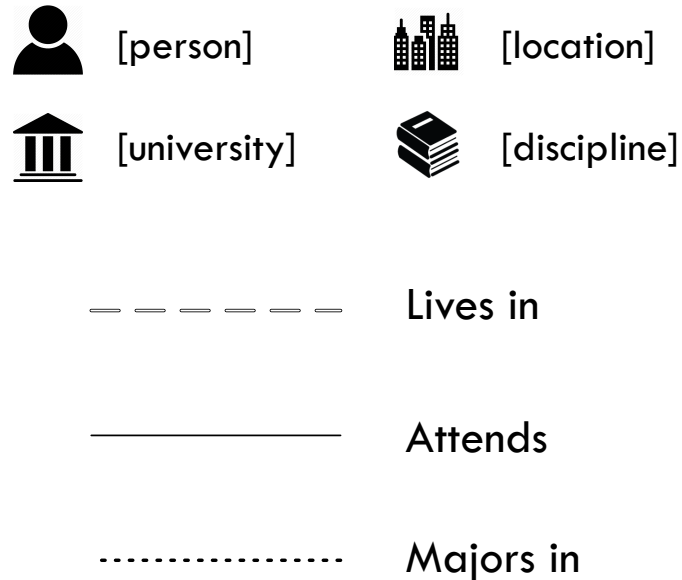
PReP: Path-Based Relevance from a Probabilistic Perspective in Heterogeneous Information Networks

Yu Shi, Po-Wei Chan, Honglei Zhuang, Huan Gui, and Jiawei Han
University of Illinois at Urbana-Champaign (UIUC)



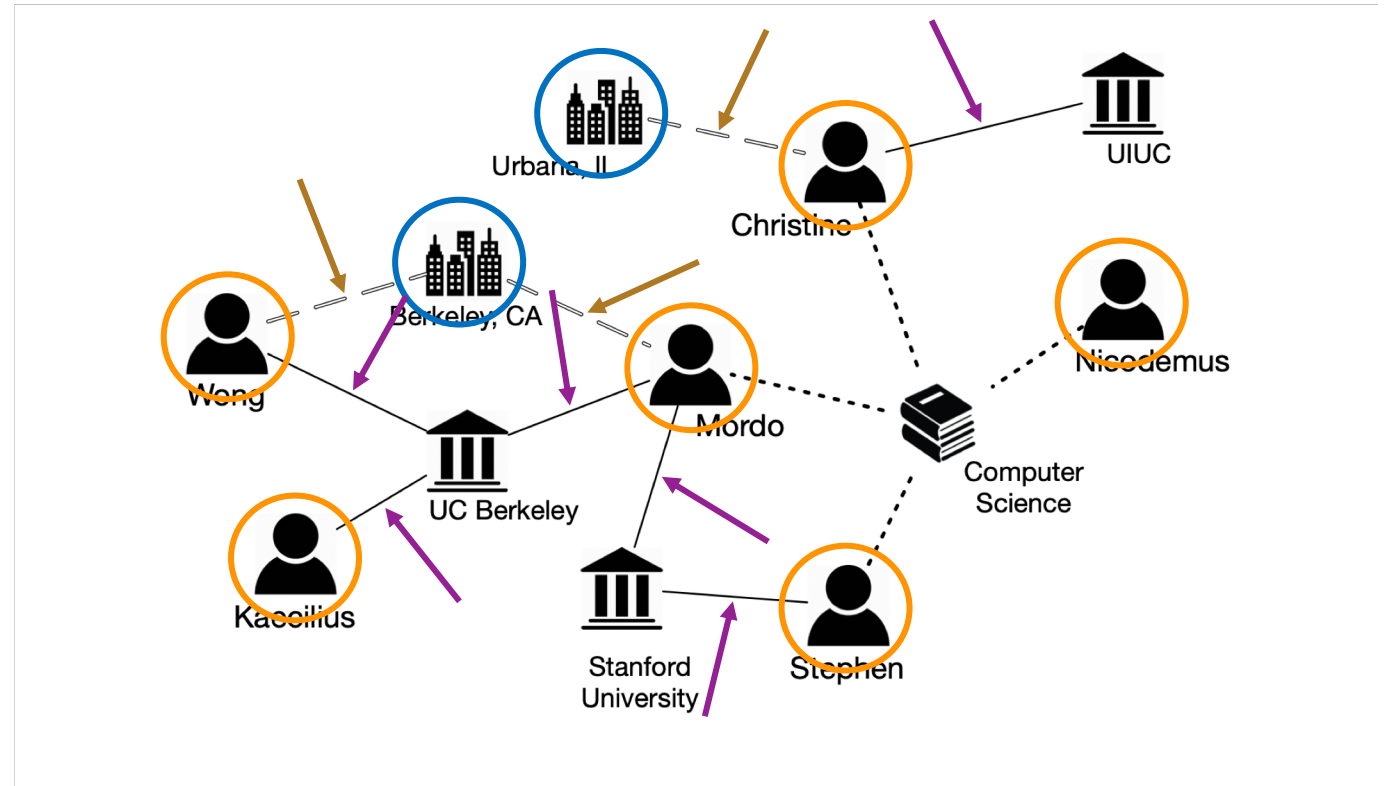
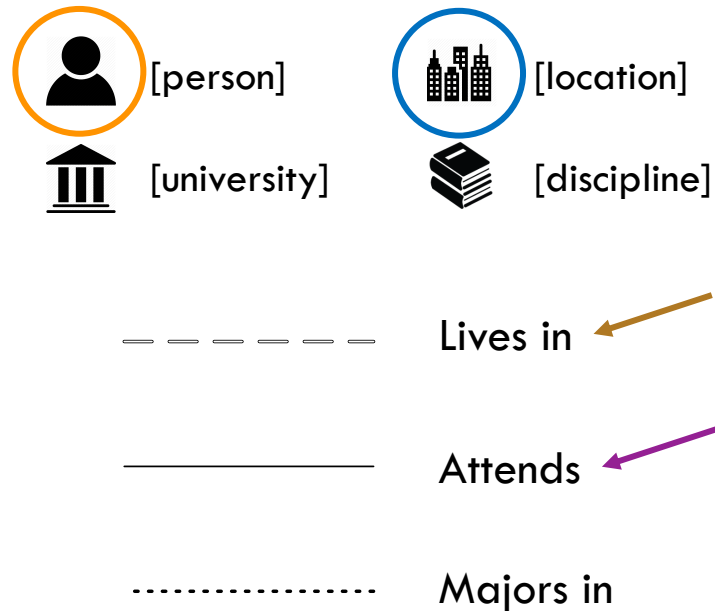
In real world applications, objects of different types can have different relations, which form **heterogeneous information networks (HINs)**.

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

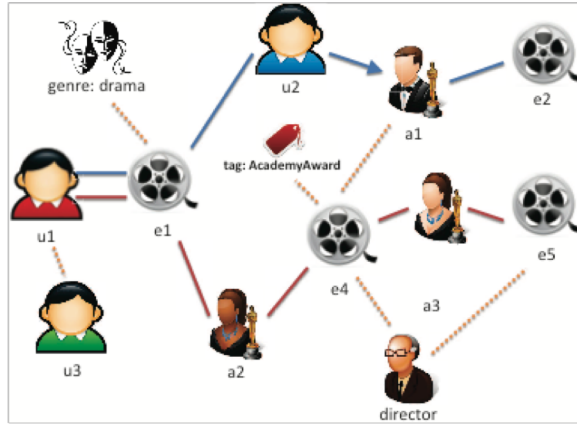


In real world applications, objects of different types can have different relations, which form **heterogeneous information networks (HINs)**.

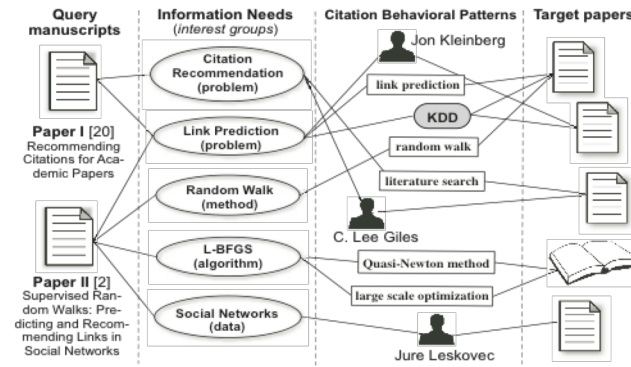
- **Typed nodes: objects**
- **Typed edges: relations**



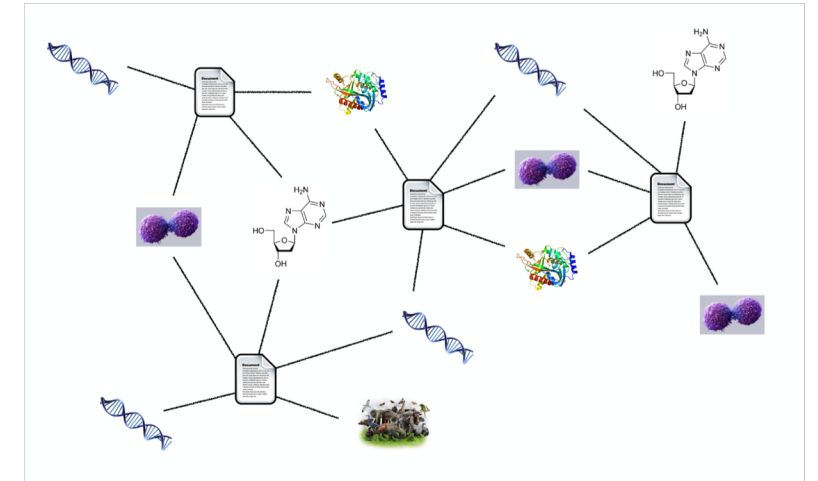
Heterogeneous information networks (HINs) are ubiquitous.



IMDb Network



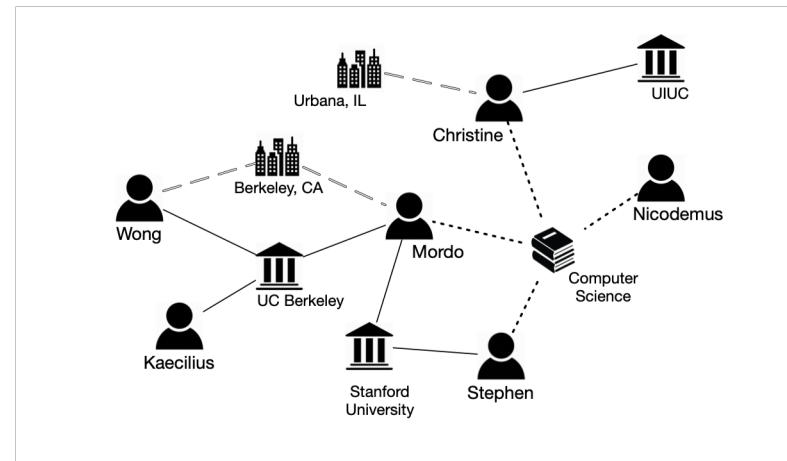
Bibliographical Network



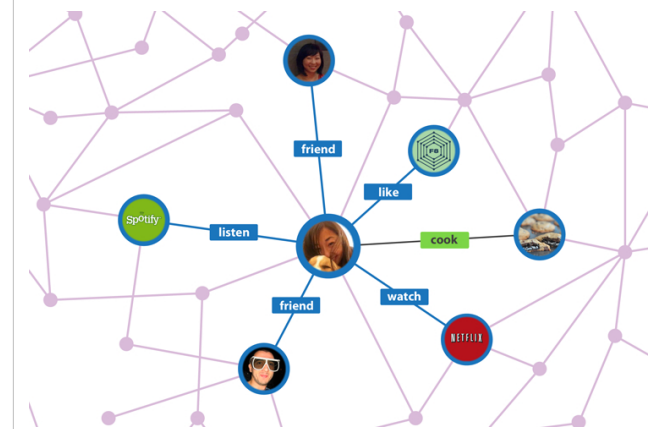
Biomedical Network



Economic Graph



Social Network



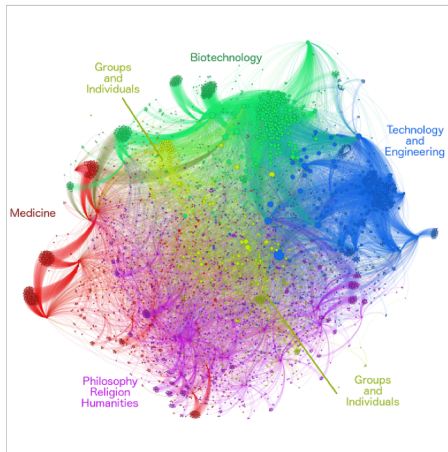
Facebook Open Graph

A fundamental problem in network mining:

defining **relevance measure**

a.k.a., **similarity, proximity.**

A good relevance measure can benefit downstream applications.



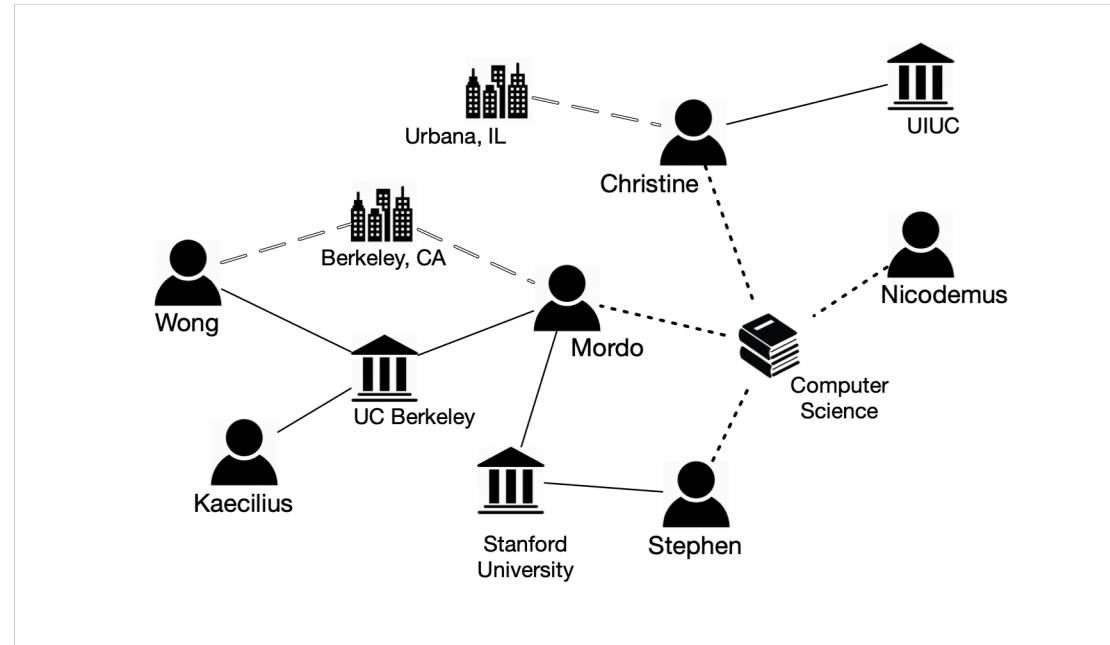
Community detection







Link prediction



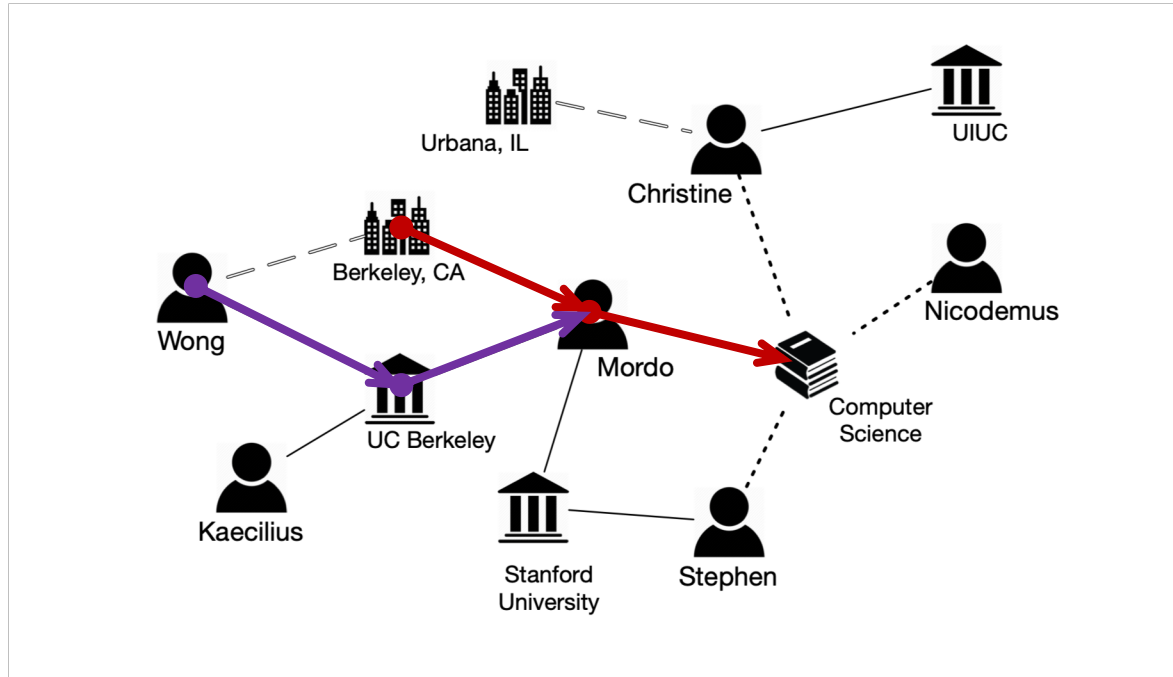
Recommendation



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**:

$$\text{Berkeley, CA} \rightarrow \text{Mordo} \rightarrow \text{Computer Science}$$

Another **example**:

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

$$\text{Wong} \rightarrow \text{UC Berkeley} \rightarrow \text{Mordo}$$

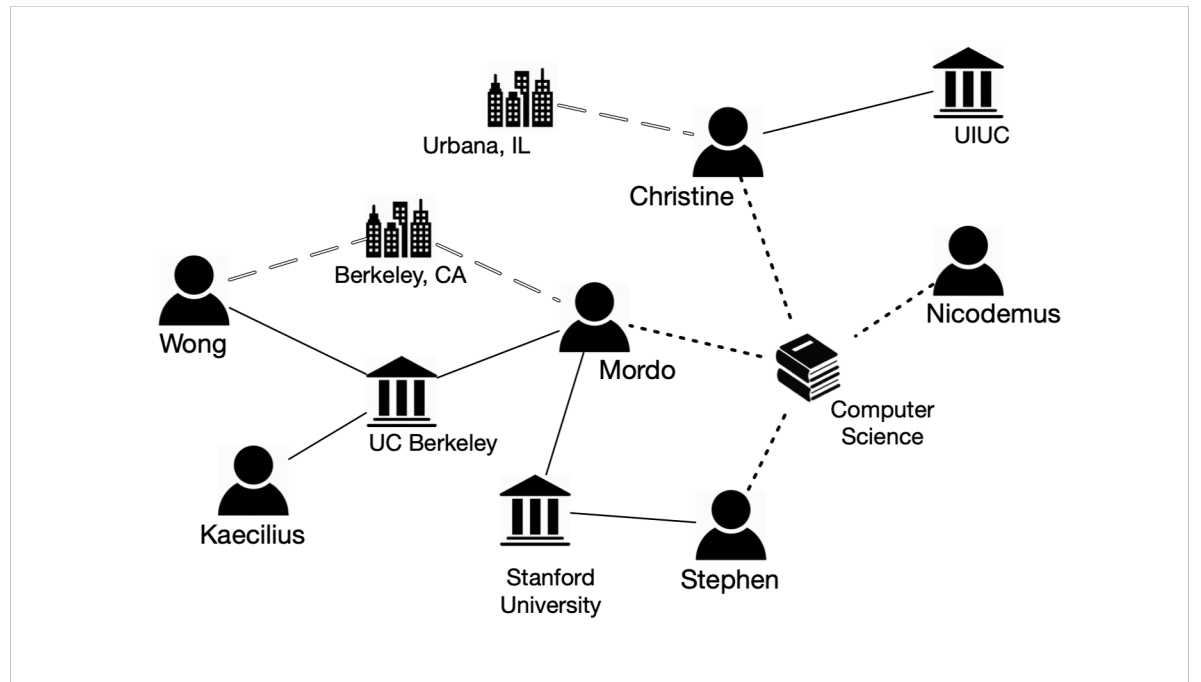
For given **meta-path** t and a **pair of node** $s = (u, v)$

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

Examples:

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

$$P_{\langle \text{Mor. Mor.} \rangle t} = 2$$



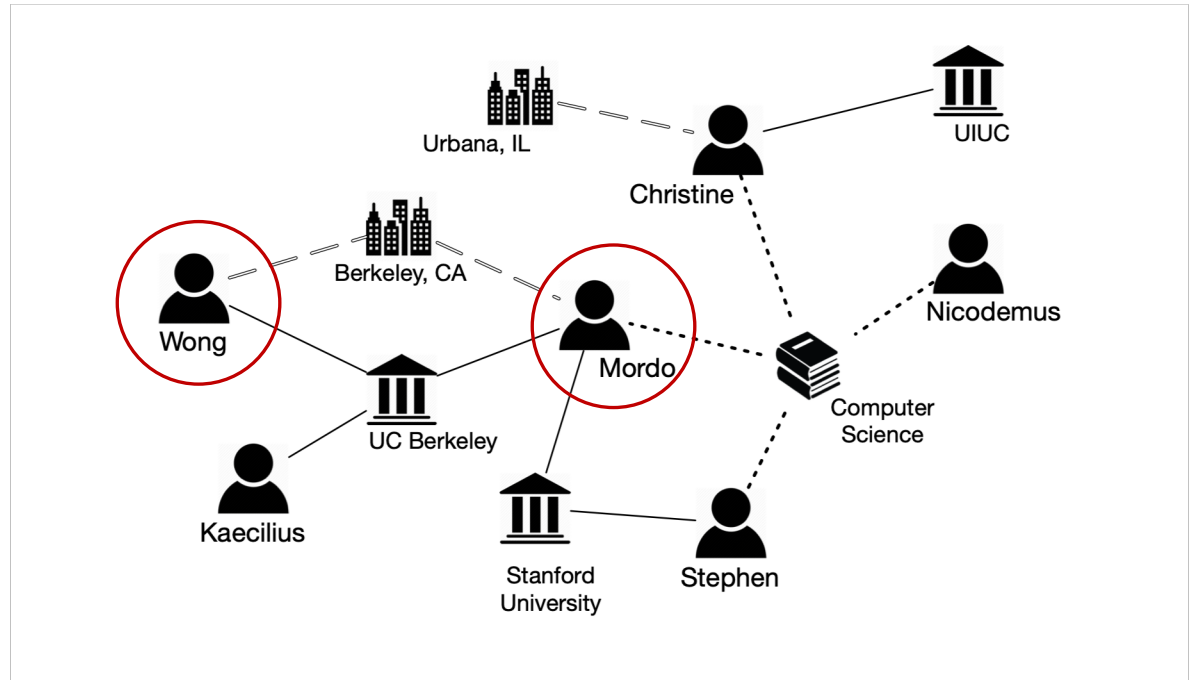
$$\mathcal{M}_t (\text{---}) : [\text{person}] \xrightarrow{\text{attends}} [\text{university}] \xrightarrow{\text{attends}^{-1}} [\text{person}]$$

Widely-used HIN relevance measures:

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

$$\text{PathCount}^{(t)}(u, v) := P_{\langle uv \rangle t}$$

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



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

Widely-used HIN relevance measures:

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

$$\text{PathCount}^{(t)}(u, v) := P_{\langle uv \rangle t}$$

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

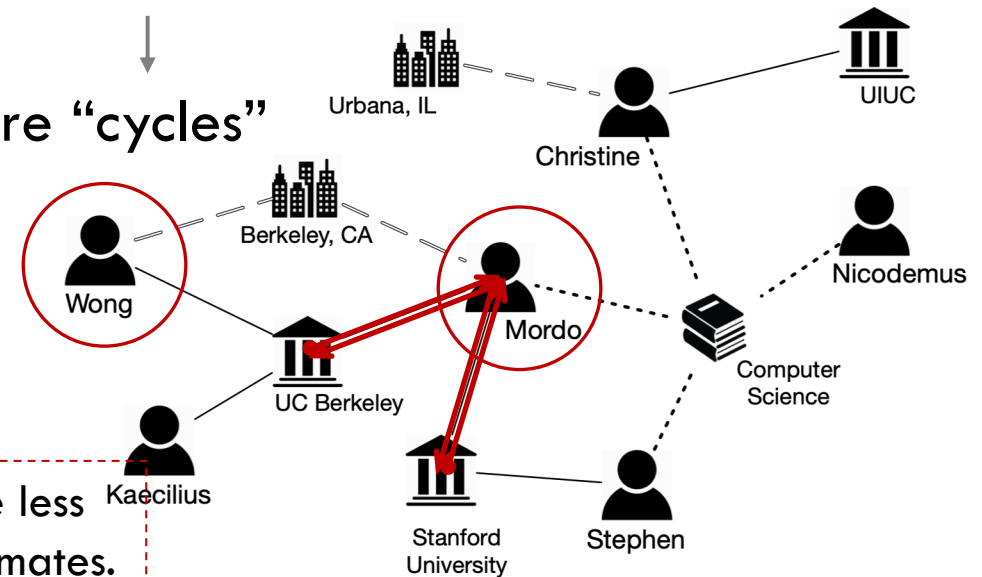
- **PathSim** [1]: further **penalizes** nodes with more “cycles”

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

$$\text{PathSim}^{(t)}(\text{Won.}, \text{Mor.}) = \frac{2 \cdot 1}{1+2} \approx 0.67$$

Mordo attends multiple universities. It is hence less significant for Wong and Mordo to be schoolmates.

Path instance from a node back to itself.



\mathcal{M}_t (—) : [person] $\xrightarrow{\text{attends}}$ [university] $\xrightarrow{\text{attends}^{-1}}$ [person]

Widely-used HIN relevance measures:

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

$$\text{PathCount}^{(t)}(u, v) := P_{\langle uv \rangle t}$$

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

- **PathSim** [1]: further **penalizes** nodes with more “cycles”

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

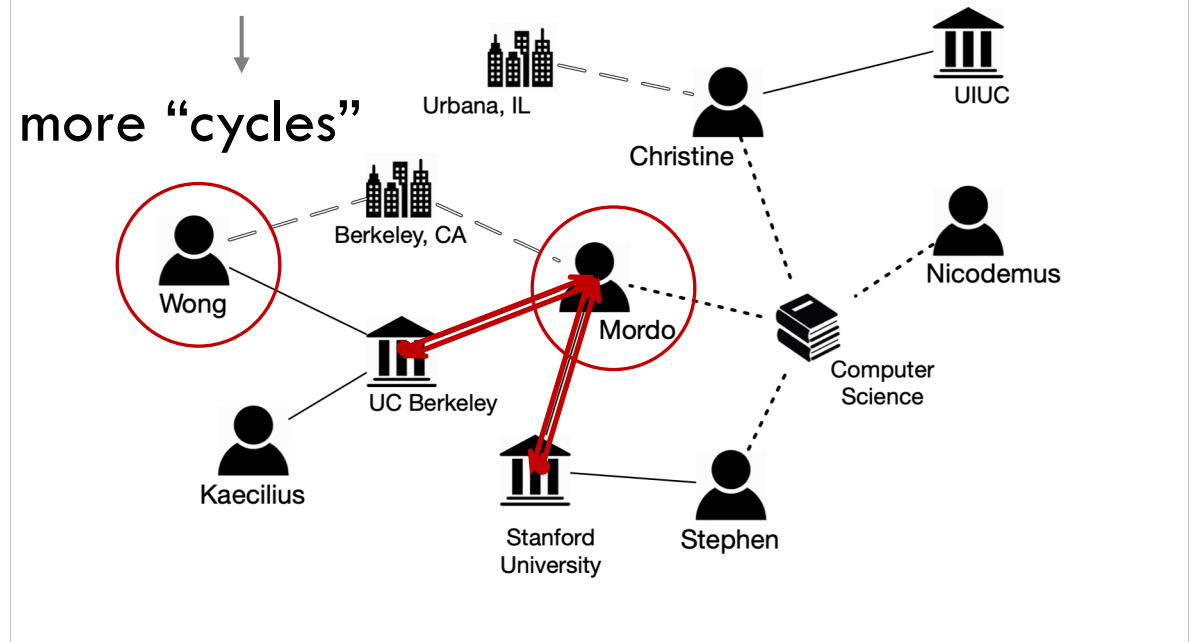
$$\text{PathSim}^{(t)}(\text{Won.}, \text{Mor.}) = \frac{2 \cdot 1}{1+2} \approx 0.67$$

- **JoinSim** [2]: another way to penalize

$$\text{JoinSim}^{(t)}(u, v) := \frac{P_{\langle uv \rangle t}}{\sqrt{P_{\langle uu \rangle t} \cdot P_{\langle vv \rangle t}}}$$

$$\text{JoinSim}^{(t)}(\text{Won.}, \text{Mor.}) = \frac{1}{\sqrt{1 \cdot 2}} \approx 0.71$$

Path instance from a node back to itself.



\mathcal{M}_t (—) : [person] $\xrightarrow{\text{attends}}$ [university] $\xrightarrow{\text{attends}^{-1}}$ [person]

Linear combination is usually used to combine multiple meta-paths.

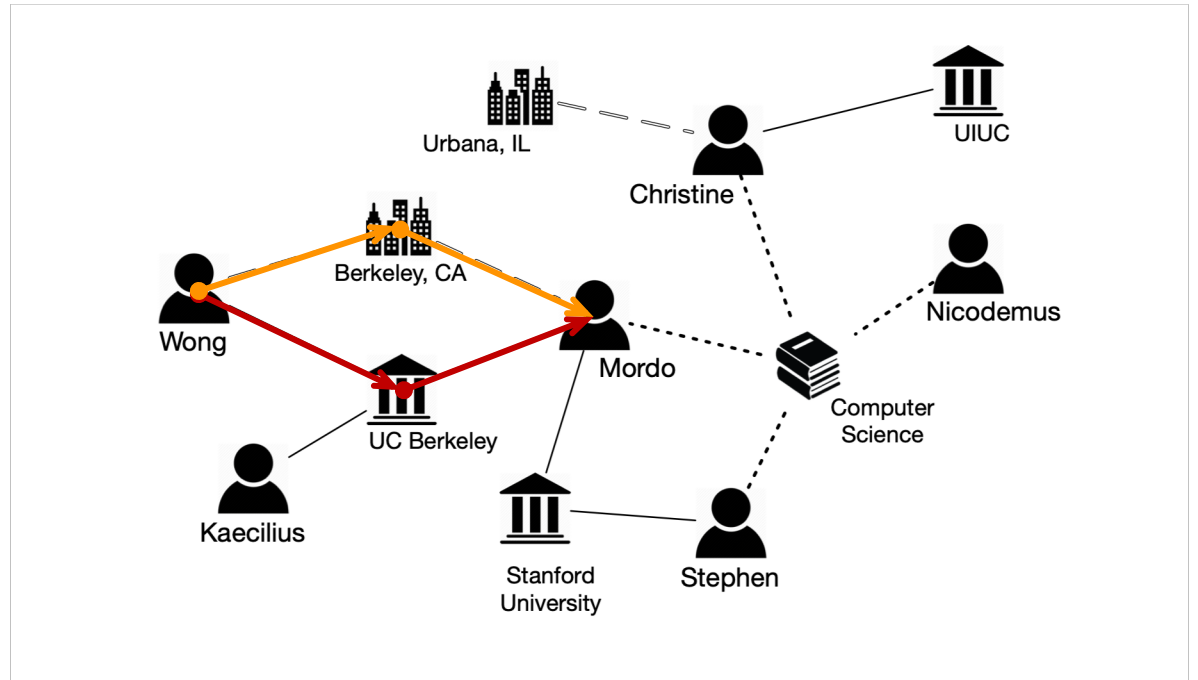
- Let $\mathbf{w} = \{w_1, \dots, w_T\}$, where w_t is the weight for meta-path t

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

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

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

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



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

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

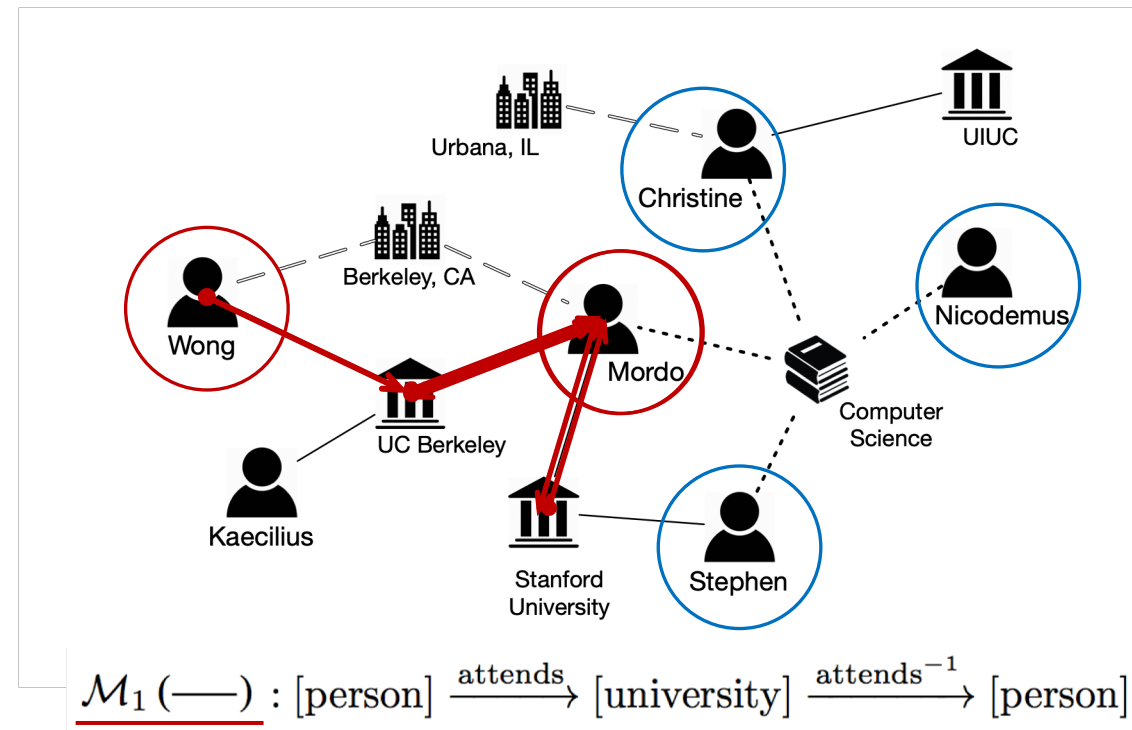
Why can these heuristic measures reflect relevance?

- 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 **PathCount**.

$$\text{PathCount}^{(t)}(u, v) := P_{\langle uv \rangle t}$$

- PathSim** penalizes nodes with more “cycles”, because it is a **less significant event** to have path instances with these nodes.

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



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:

Likelihood

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

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

$$\begin{aligned} -LL(s) &= -\log\left(\prod_t w_t e^{-w_t P_{st}}\right) = \sum_t w_t P_{st} - \sum_t \log w_t \\ &= \sum_t w_t P_{st} + \text{const} = \text{PathCount}_{\mathbf{w}}(s) + \text{const}. \end{aligned}$$

Existing relevance measure

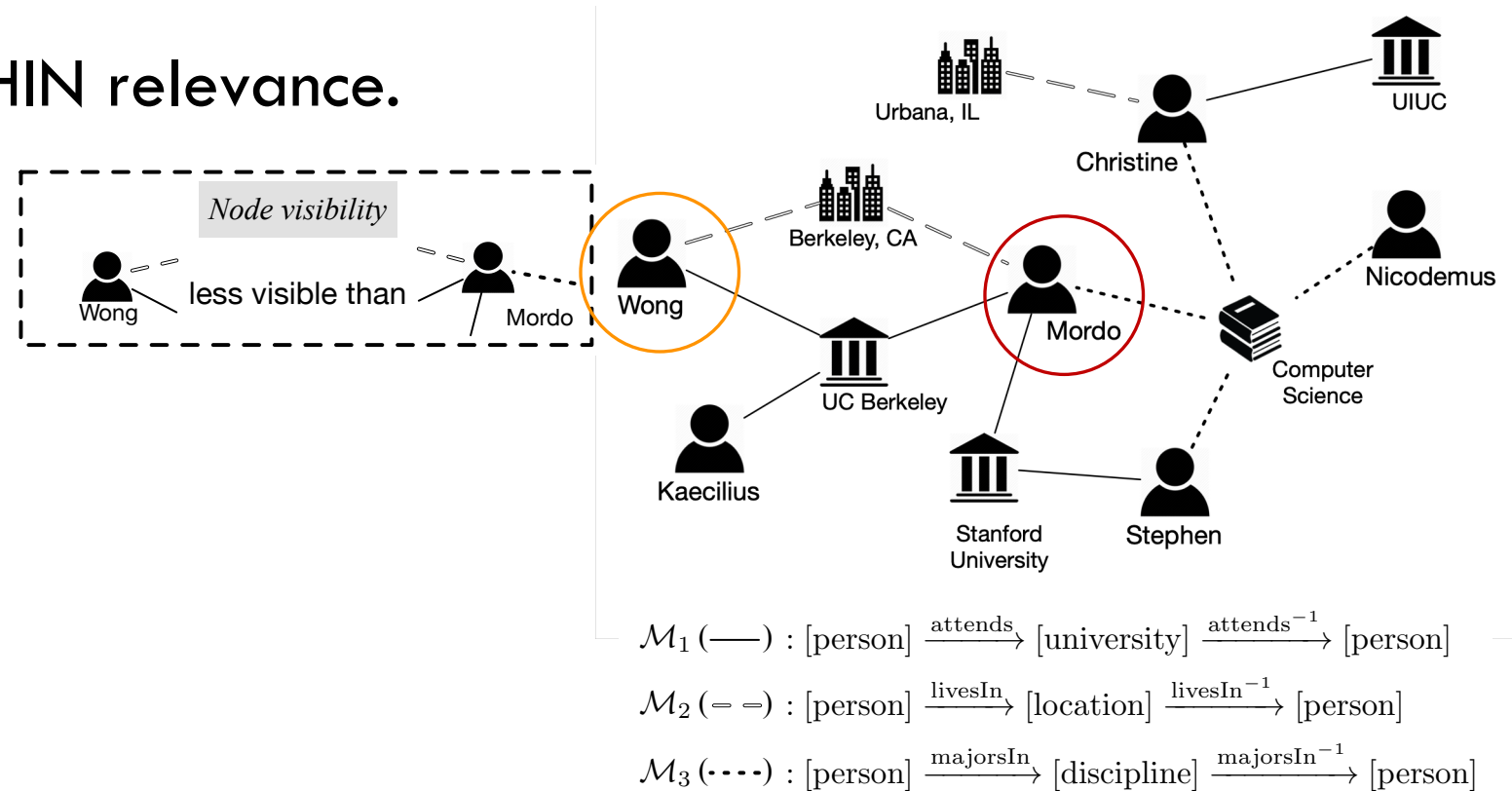
Similar results are derived for PathSim and JoinSim by adding a node-pair-specific component κ_s .

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

Likelihood \longleftrightarrow Relevance

Beyond the probabilistic interpretation, we identify **three characteristics** important for path-based HIN relevance.

1. Node visibility

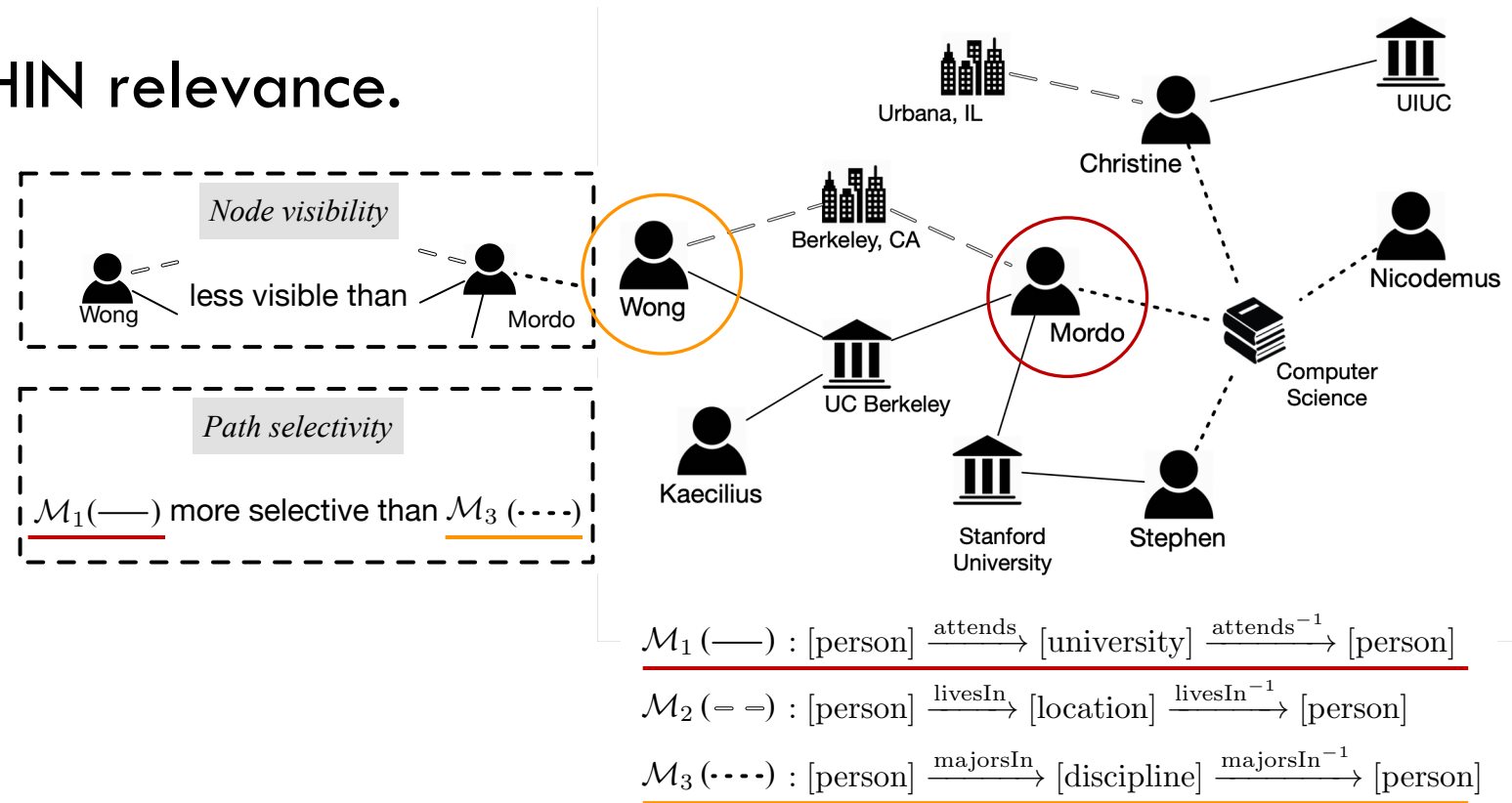


Modeled by PathSim and JoinSim.

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**

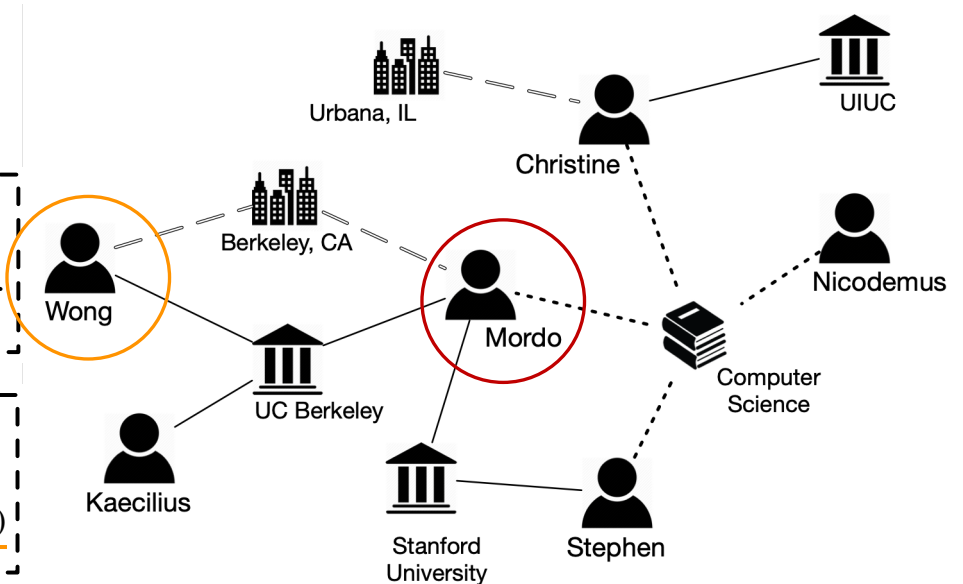
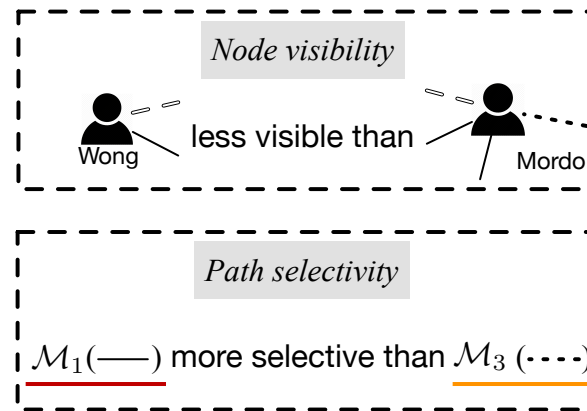
important for path-based HIN relevance.

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.



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

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

$$\mathcal{M}_3(\text{⋯⋯}) : [\text{person}] \xrightarrow{\text{majorsIn}} [\text{discipline}] \xrightarrow{\text{majorsIn}^{-1}} [\text{person}]$$

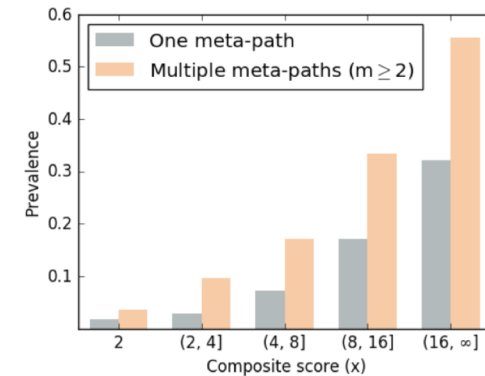
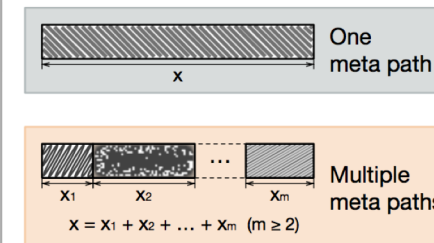
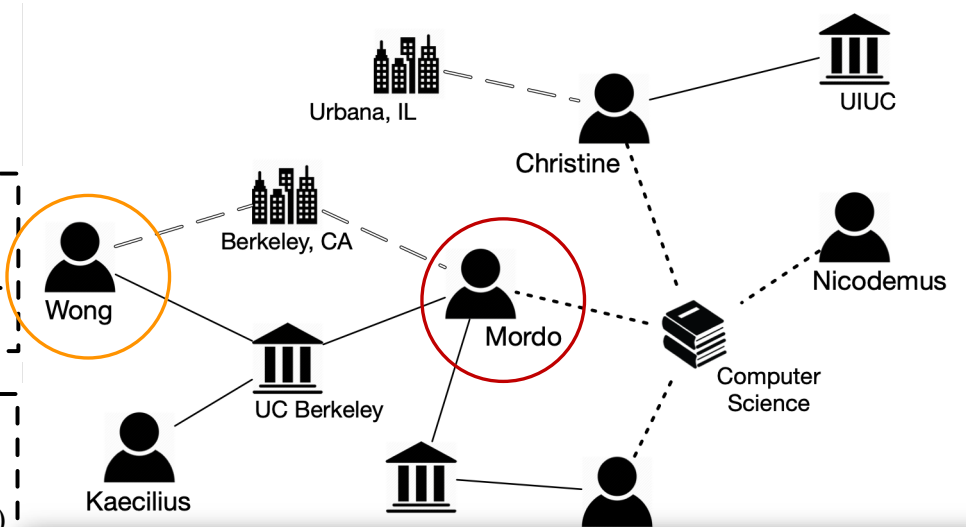
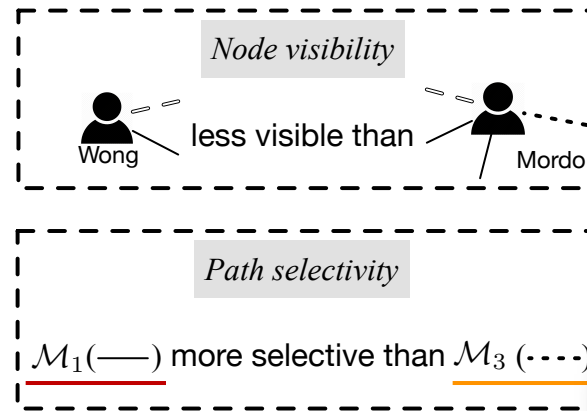
Beyond the probabilistic interpretation, we identify **three characteristics** important for path-based HIN relevance.

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



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 **P**ath-based **R**elevance from **P**robabilistic perspective:

PReP

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.

η_t models the **path selectivity** of paths under meta-path t

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

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

ψ_{st} governs the distribution of meta-paths between s .

It is given by a mixture of K **generating patterns** with ϕ_{sk} from the k -th, and the k -th contains a portion θ_{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 $\sum_{t=1}^T \theta_{kt} = 1$.

$\tau_{(u,v)} = \rho_u \rho_v$
 ρ_u and ρ_v model the **node visibility** of u and v , respectively.

Each node further regularized by a gamma prior:

$$\rho_z \sim \Gamma(\alpha, 1)$$

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**:

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

We find the maximum a posteriori (**MAP**) estimate for model parameters.

- The proposed algorithm iteratively update model parameters: η , ρ , Φ , and Θ .

Update η

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

Closed-form

Update ρ by solving

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

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

Closed-form for each node u

Update Φ using projected gradient descent (PGD) in parallel

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

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

Rows of Φ are **independent**
and can be updated **in parallel**

Update Θ using PGD

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

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

size of $\Theta \ll$ size of Φ

Experiments

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 $\mathbf{w} = \{w_1, \dots, w_T\}$ 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

Baselines

Partial models

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.

- **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.

Future Work

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

1. Better modeling of **path selectivity**.
 - Without supervision, current model assumes uninformative prior on η_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 characteristics in path-based HIN relevance, where cross-meta-path synergy is not modeled by existing methods.
3. We propose a 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.