Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks

Yizhou Sun (UIUC), Rick Barber (UIUC), Manish Gupta (UIUC), Charu C. Aggarwal (IBM), Jiawei Han (UIUC)

ASONAM 2011

Acknowledgements: NSF, ARL, NASA, AFOSR (MURI), IBM, DHS

August 13, 2011

Content

- Background and Motivation
- Problem Definition
- *PathPredict*: Meta Path-Based Relationship Prediction Model
  - Meta path-based topological features
  - The supervised learning framework and model
- Experiments
- Conclusions

Background

- Homogeneous networks
  - One type of objects
  - One type of links
  - Ex: Friendship network in Facebook

- Link prediction in homogeneous networks
  - Predict whether a link between two objects will appear in the future, according to:
    - Topological feature of the network
    - Attribute feature of the objects (usually cannot be fully obtained)

Motivation

- In reality, heterogeneous networks are ubiquitous
  - Multiple types of objects, multiple types of links
    - Ex: bibliographic network, movie network
- From link prediction to relationship prediction
  - A relationship between two objects could be a composition of two or more links
    - Ex: Co-author relationship iff they have co-written a paper
  - Re-design topological features in heterog. info. networks
- Our goal: Study the topological features in heterogeneous networks in predicting the co-author relationship building

Content

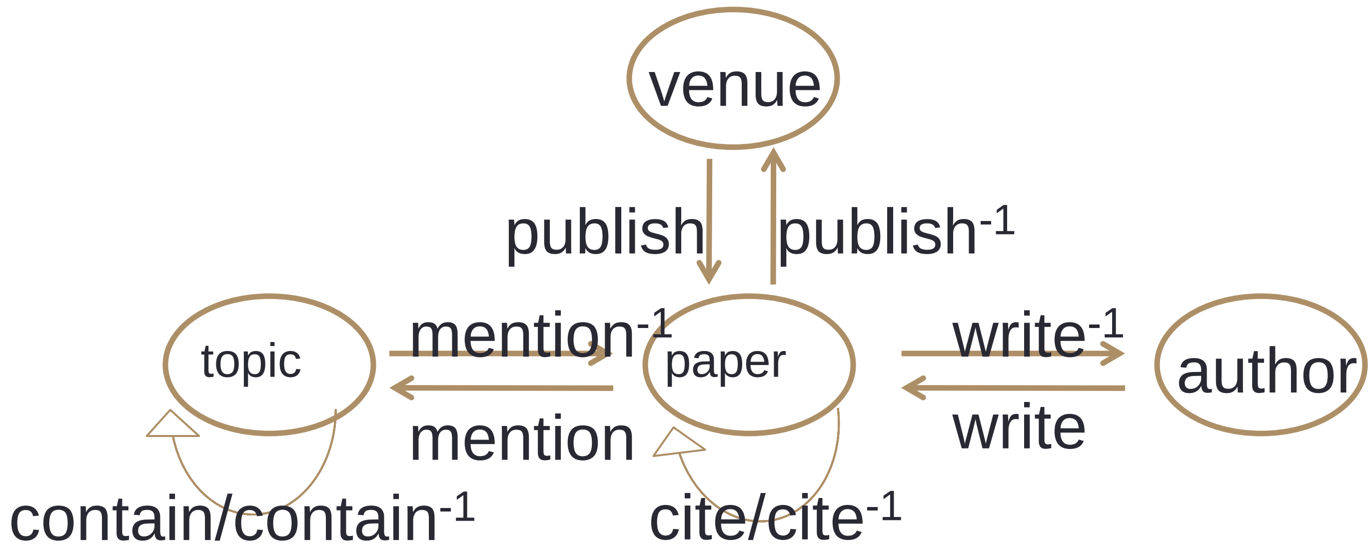
- Background and Motivation
- Problem Definition
- *PathPredict*: Meta Path-Based Relationship Prediction Model
    - Meta path-based topological features
    - The supervised learning framework and model
- Experiments
- Conclusions

Heterogeneous Information Networks

- Heterogeneous information networks:
  - A directed network containing multiple types of objects and links
- Network schema:
  - A meta graph structure that summarizes a heterogeneous information network
    - Node: types of objects
    - Edge: relations between types of objects (types of links)
- Meta path
  - A path defined over network schema
  - Denotes a composition relation
  - Example: co-author relation
    - $A \xrightarrow{write} P \xrightarrow{write^{-1}} A$ (short for $A - P - A$)

Guidance: Meta Path in Bibliographic Network

- Relationship prediction: meta path-guided prediction
- Meta path relationships among similar typed links share similar semantics and are comparable and inferable



Co-author Relationship Prediction

- Target relation and relationship
  - A target relation is the relation to be predicted
  - A relationship following a target relation is an instance of the target relation
- Co-author relationship prediction:
  - Co-author relation is encoded by the meta path:

$$A \xrightarrow{write} P \xrightarrow{write^{-1}} A$$

  - Predict whether two existing authors will build a relationship in the future following co-author relation
    - Namely, for two authors a_i and a_j, $\exists p, a_i - p - a_j$
- Topological features:
  - Relations between the same types of objects as the target relation, also encoded by meta paths
    - E.g., citation relations between authors: $A \xrightarrow{write} P \xrightarrow{cite} P \xrightarrow{write^{-1}} A$

Content

PathPredict: A Path-Guided Model

- *PathPredict*: Meta path-based relationship prediction model
  - Propose meta path-based topological features in heterog. info. network
    - Topological features used in homogeneous networks cannot be directly used
  - Using logistic regression-based supervised learning methods to learn the coefficients associated with each feature

Selection Among Competitive Measures

- Path Count: #path instances between authors following *R*

$$PC_R(a_i, a_j)$$

- Normalized Path Count: Normalize path count following *R* by the "degree" of authors

$$NPC_R(a_i, a_j) = \frac{PC_R(a_i, a_j) + PC_{R^{-1}}(a_j, a_i)}{PC_R(a_i, \cdot) + PC_R(\cdot, a_j)}$$

- Random Walk: Consider one way random walk following *R*

$$RW_R(a_i, a_j) = \frac{PC_R(a_i, a_j)}{PC_R(a_i, \cdot)}$$

- Symmetric Random Walk: Consider random walk in both directions

$$SRW_R(a_i, a_j) = RW_R(a_i, a_j) + RW_{R^{-1}}(a_j, a_i)$$

Meta Path-based Topological Features

- From a space of
  - $\{Meta\ Path \times Measure\}$
- Meta path
  - Specify the topology structure
  - Denote a new composite relation
  - Different meta paths represent different semantic meanings
- Measure
  - Quantify the meta path
  - Different measures focus on different aspects, e.g.:
    - Count: the strength of the connectivity;
    - PathSim: find similar peers
    - …

Meta Paths for Co-authorship Prediction in DBLP

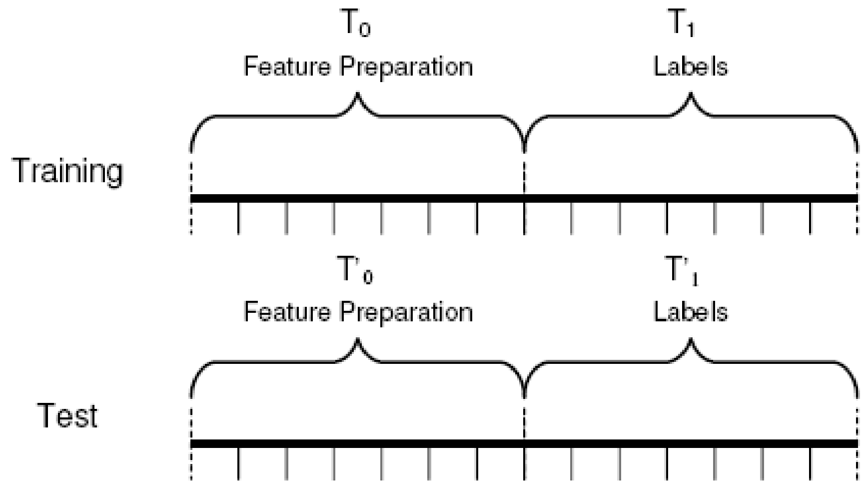
- List of all the meta paths between authors under length 4

Table II
META PATHS UNDER LENGTH 4 BETWEEN AUTHORS IN THE DBLP NETWORK

Meta Path	Semantic Meaning of the Relation
$A - P - A$	a_i and a_j are coauthors (the target relation)
$A - P \rightarrow P - A$	a_i cites a_j
$A - P \leftarrow P - A$	a_i is cited by a_j
$A - P - V - P - A$	a_i and a_j publish in the same venues
$A - P - A - P - A$	a_i and a_j are co-authors of the same authors
$A - P - T - P - A$	a_i and a_j write the same topics
$A - P \rightarrow P \rightarrow P - A$	a_i cites papers that cite a_j
$A - P \leftarrow P \leftarrow P - A$	a_i is cited by papers that are cited by a_j
$A - P \rightarrow P \leftarrow P - A$	a_i and a_j cite the same papers
$A - P \leftarrow P \rightarrow P - A$	a_i and a_j are cited by the same papers

Supervised Learning Framework

- Training:
  - T_0-T_1 time framework
    - T_0: feature collection (x)
    - T_1: label of relationship collection (y)
- Testing:
  - T_0'-T_1' time framework, which may have a shift of time compared with training stage



Prediction Model

- Training and test pair: <$\mathbf{x}_i$, y_i> = <history feature list, future relationship label>
- Logistic Regression Model
  - Model the probability for each relationship as
    - $$p_i = \frac{e^{\mathbf{x}_i \boldsymbol{\beta}}}{e^{\mathbf{x}_i \boldsymbol{\beta}} + 1}$$
    - β is the coefficients for each feature (including a constant 1)
  - MLE estimation
    - Maximize the likelihood of observing all the relationships in the training
    - $$L = \prod_i p_i^{y_i} (1 - p_i)^{(1 - y_i)}$$

Content

- Background and Motivation
- Problem Definition
- *PathPredict*: Meta Path-Based Relationship Prediction Model
  - Meta path-based topological features
  - The supervised learning framework and model
- Experiments
- Conclusions

Experiment Setting

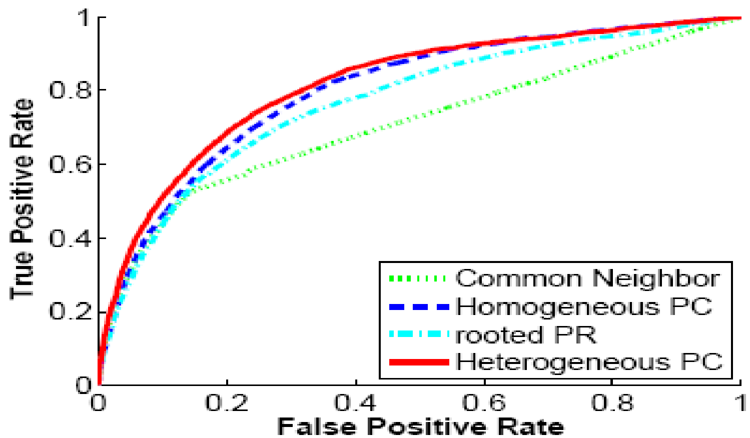
- Datasets:
  - DBLP bibliographic network
  - Time window
    - T_0: 1990-1996
    - T_1: 1997-2003
    - T_2: 2004-2010
  - Data property summarization for four datasets under time framework $T_0 - T_1$

Table III
FOUR TRAINING DATASETS IN TIME FRAMEWORK $T_0 - T_1$ SUMMARIZATION

Source author type	Constraint	# Source authors	# Source author with new relationships	# New relationships	# Avg. target authors
highly productive	2-hop✓	2538	1548 (64.91%)	4986 (19.43%)	159.01
	3-hop✓	2538	1860 (77.99%)	9215 (35.91%)	930.65
	no	2538	2385 (100%)	25661 (100%)	119246
less productive	2-hop✓	13075	3367 (36.58%)	6189 (12.51%)	47.97
	3-hop✓	13075	4333 (47.08%)	10710 (21.64%)	271.06
	no	13075	9204 (100%)	49483 (100%)	119246

Homogeneous Measures vs. Heterogeneous Measures

- Homogeneous measures:
  - Only consider co-author sub-network : common neighbor; rooted PageRank
  - Consider the whole network and mix all types together: total path count
- Heterogeneous measure: Heterogeneous path count



(b) $HP3hop$

Heterogeneous Path Count produces the best accuracy!

Over Four Datasets

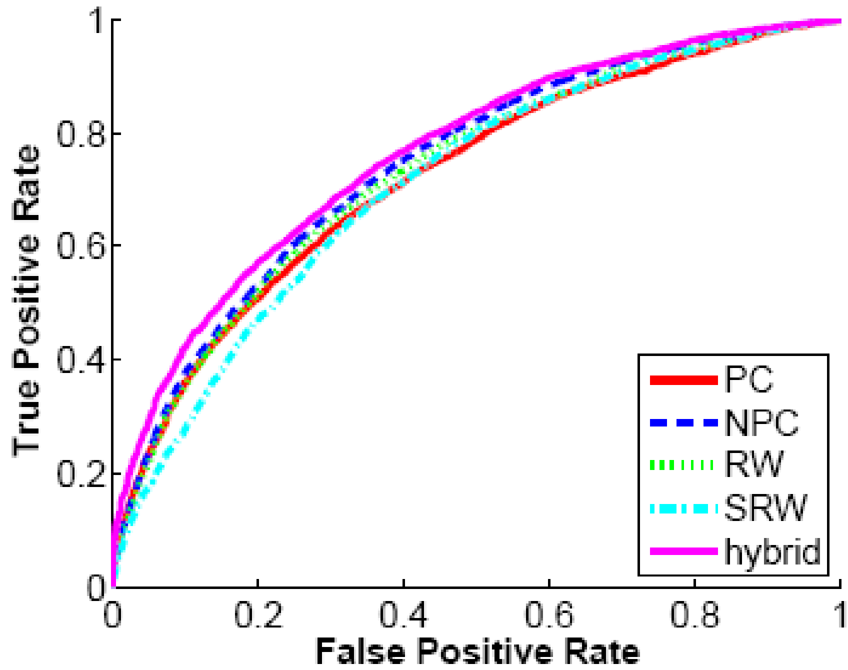
Table IV

HOMOGENEOUS TOPOLOGICAL FEATURES VS. HETEROGENEOUS ONES

Dataset	Topological features	Accuracy	AUC
$HP2hop$	common neighbor	0.6053	0.6537
	homogeneous PC	0.6433	0.7098
	heterogeneous PC	**0.6545**	**0.7230**
$HP3hop$	common neighbor	0.6589	0.7078
	homogeneous PC	0.6990	0.7998
	rooted PageRank	0.6433	0.7098
	heterogeneous PC	**0.7173**	**0.8158**
$LP2hop$	common neighbor	0.5995	0.6415
	homogeneous PC	0.6154	0.6868
	heterogeneous PC	**0.6300**	**0.6935**
$LP3hop$	common neighbor	0.6804	0.7195
	homogeneous PC	0.6901	0.7883
	heterogeneous PC	**0.7147**	**0.8046**

Compare among Different Heterogeneous Measures

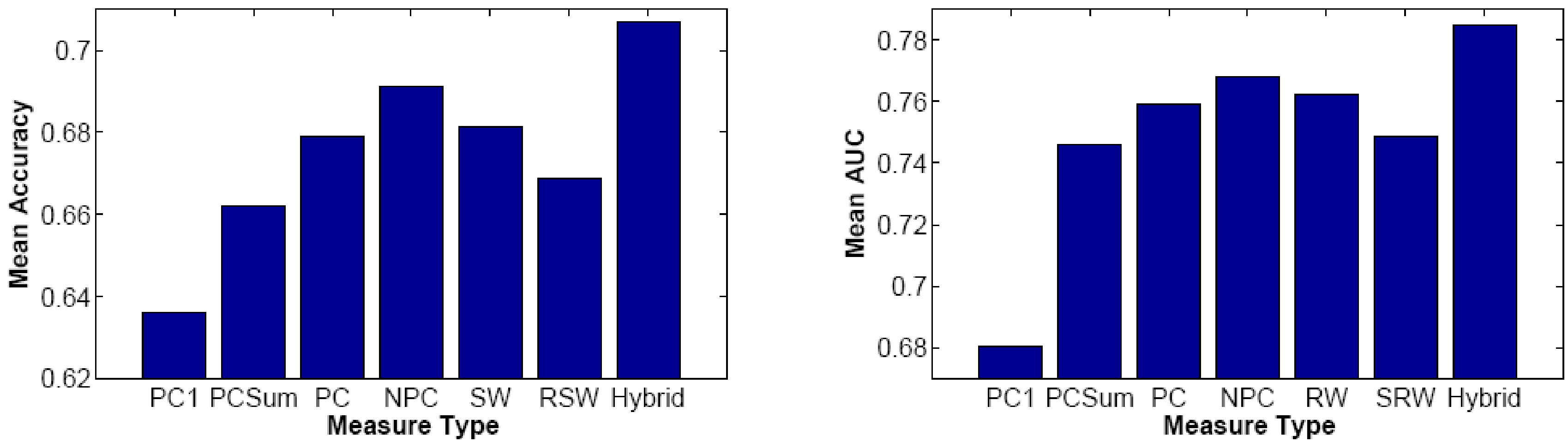
- Normalized path count is slightly better and the hybrid measure that combines all measures is the best



(a) $HP2hop$

Over Four datasets

- Four datasets: DB, DM, IR, AI

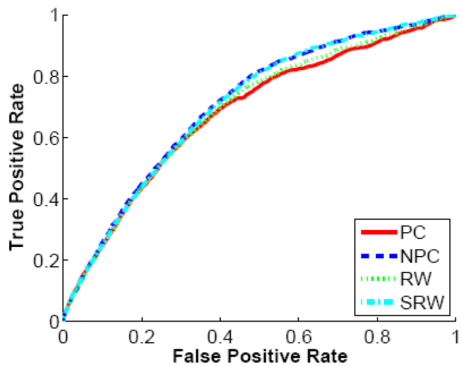


(a) Mean accuracy (b) Mean AUC

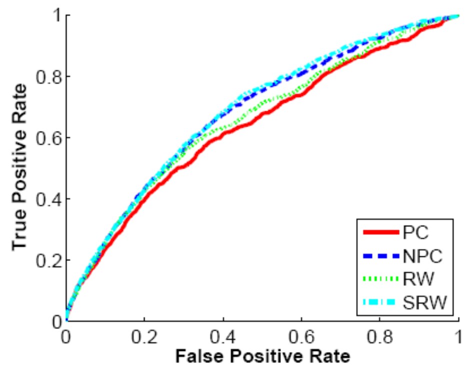
Figure 6. Average Accuracy over 4 Datasets for Different Features

Impacts of Collaboration Frequency on Different Measures

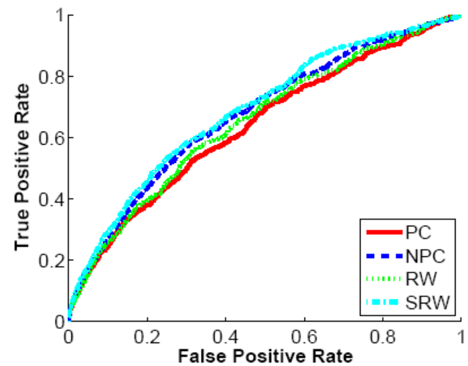
- Symmetric random walk is better for predicting frequent co-author relationship



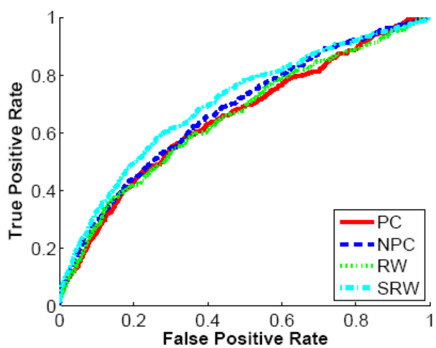
(a) $freq > 1$



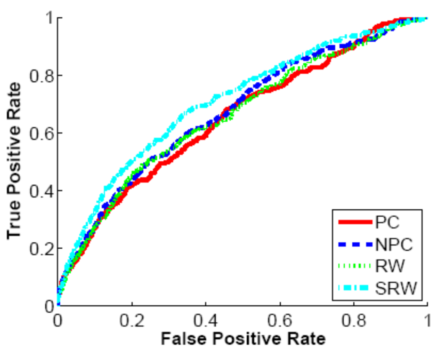
(b) $freq > 2$



(c) $freq > 3$



(d) $freq > 4$



(e) $freq > 5$

Model Generalization Over Time

- Historical training can help predict future relationship building

Table V

MODEL GENERALIZATION TEST OVER TIME EVOLVING

Training framework	Test framework	Prediction Accuracy	
		Accuracy	AUC
$T_0 - T_1$	$T_0 - T_1$	0.7368	0.8211
$T_0 - T_1$	$T_1 - T_2$	0.7123	0.8325
$T_1 - T_2$	$T_1 - T_2$	0.7442	0.8313

Learned Significance for Each Topological Feature

- The co-attending venues and the shared co-authors are very critical in determining two authors' future collaboration

Table VI

SIGNIFICANCE OF META PATHS WITH NORMALIZED PATH COUNT MEASURE FOR $HP3hop$ DATASET

Meta Path	p-value	significance level[1]
$A - P \rightarrow P - A$	0.0378	**
$A - P \leftarrow P - A$	0.0077	***
$A - P - V - P - A$	1.2974e-174	****
$A - P - A - P - A$	1.1484e-126	****
$A - P - T - P - A$	3.4867e-51	****
$A - P \rightarrow P \rightarrow P - A$	0.7459	
$A - P \leftarrow P \leftarrow P - A$	0.0647	*
$A - P \rightarrow P \leftarrow P - A$	9.7641e-11	****
$A - P \leftarrow P \rightarrow P - A$	0.0966	*

[1] *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$, ****: $p < 0.001$

Case Studies for Queries

QUERY AUTHOR SUMMARIZATION

Query author	# Candidates	# True relationships
Jiawei Han	11934	36
Christos Faloutsos	12945	45
Charu Aggarwal	5166	12
Jian Pei	4809	42
Xifeng Yan	1617	8

TOP-5 PREDICTED CO-AUTHORS FOR JIAN PEI IN 2003-2009

Rank	Hybrid heterogeneous features	# Shared authors
1	**Philip S. Yu**	**Philip S. Yu**
2	**Raymond T. Ng**	Ming-Syan Chen
3	Osmar R. Zaïane	Divesh Srivastava
4	**Ling Feng**	Kotagiri Ramamohanarao
5	**David Wai-Lok Cheung**	**Jeffrey Xu Yu**

* Authors in bold format are the true new co-authors of Jian in the time period 2003-2009.

TOP-10 PREDICTED CO-AUTHORS FOR JIAWEI HAN

Rank	Hybrid features	# Shared authors
1	**Hans-Peter Kriegel**	Elisa Bertino
2	Christos Faloutsos	Sushil Jajodia
3	Divesh Srivastava	Hector Garcia-Molina
4	H. V. Jagadish	**Hans-Peter Kriegel**
5	Bing Liu[1]	Christos Faloutsos
6	Johannes Gehrke	Divyakant Agrawal
7	George Karypis	Elke A. Rundensteiner
8	**Charu C. Aggarwal**	Amr El Abbadi
9	Mohammed Javeed Zaki	Krithi Ramamritham
10	Wynne Hsu	Stefano Ceri

[1] Although not included in the time interval T_2, Bing Liu co-authored with Jiawei in Year 2010.

Recall@50 COMPARISON

Query author	Hybrid Features	Random	# Shared authors
Jiawei Han	0.1111	0.0042	0.0833
Christos Faloutsos	0.0889	0.0039	0.1111
Charu Aggarwal	0.4167	0.0097	0.3333
Jian Pei	0.2619	0.0104	0.2619
Xifeng Yan	0.875	0.0309	0.5
Avg.	**0.3507**	0.0118	0.2579

Content

- Background and Motivation
- Problem Definition
- *PathPredict*: Meta Path-Based Relationship Prediction Model
  - Meta path-based topological features
  - The supervised learning framework and model
- Experiments
- Conclusions

Conclusions

- Problem:
  - Extend link prediction problem in homogeneous networks into relationship prediction in HIN, using co-authorship prediction as a case study
- Solution
  - Propose meta path-based topological features and measures in HIN
  - Using logistic regression-based supervised learning methods to learn the coefficients associated with each feature
- Results
  - Hetero. measures beats homo. measures
  - Hybrid measure beats single measures

- Thank you!

Q & A

Existing Topological Measure in Homogeneous Networks

- Common Neighbors
  - $|\Gamma(a_i) \cap \Gamma(a_j)|$
- Jaccard Coefficient
  - $\frac{|\Gamma(a_i) \cap \Gamma(a_j)|}{|\Gamma(a_i) \cup \Gamma(a_j)|}$
- $Karz_\beta$
  - $\sum_{l=1}^{\infty} \beta^l |path_{a_i,a_j}^{\langle l \rangle}|$
- Propflow
  - Random walk over a path with fixed length
- Rooted PageRank
  - Random walk with restart

- However, in heterogeneous networks, neighbor sets and paths are with different semantic meanings
  - These measures cannot be directly used!

Four Meta Path-based Measures

- Given a meta path encoded relation *R*

1. Path Count: $PC_R(a_i, a_j)$
   - Number of path instances between authors following *R*

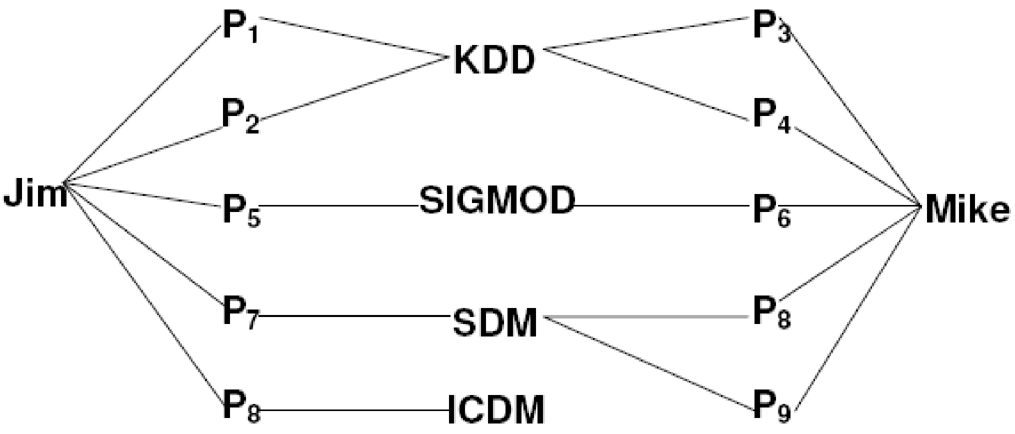
2. Normalized Path Count: $NPC_R(a_i, a_j)$
   - Normalized by the "degree" of authors

3. Random Walk: $RW_R(a_i, a_j)$
   - Consider one way random walk following *R*

4. Symmetric Random Walk: $SRW_R(a_i, a_j)$
   - Consider random walk in both directions

Example

- A meta-path: A-P-V-P-A



- PC(J,M) = 7
- NPC(J,M) = (7+7)/(7+9)
- RW(J,M) = ½
- SRW(J,M) = ½ + 1/16