Deep Learning Based Link Prediction with Social Pattern and External Attribute Knowledge in Bibliographic Networks

Chuanting Zhang, Haixia Zhang, Dongfeng Yuan and Minggao Zhang

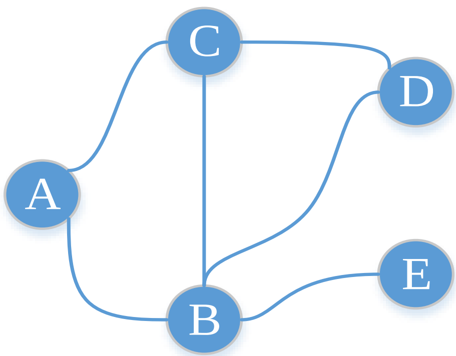
Shandong University

chuanting.zhang@gmail.com

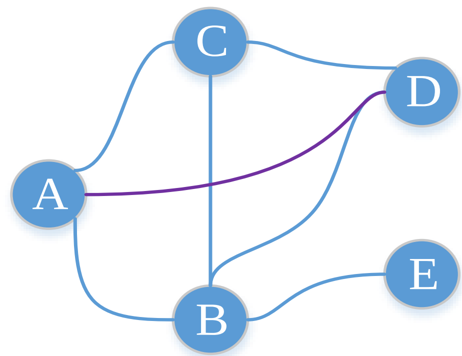
Outline

- Background & Motivation
- The SPEAK features and classification model
  - Dataset description
  - Social Pattern & External Attribute Knowledge
  - Deep neural networks
- Results and analysis
  - Experiment setup
  - AUC performance
  - Parameter sensitivity
- Conclusion

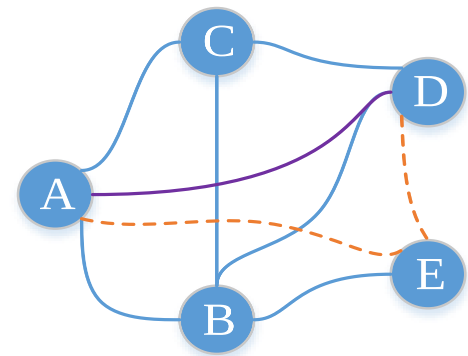
The link prediction problem



t_1

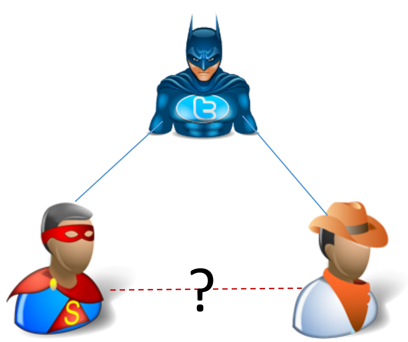


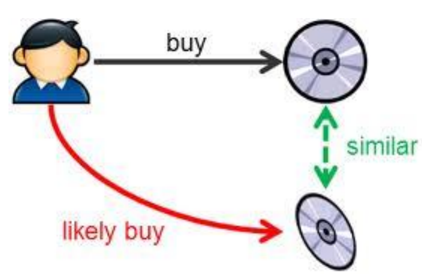
t_2

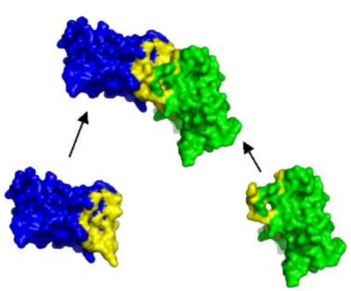


t_3

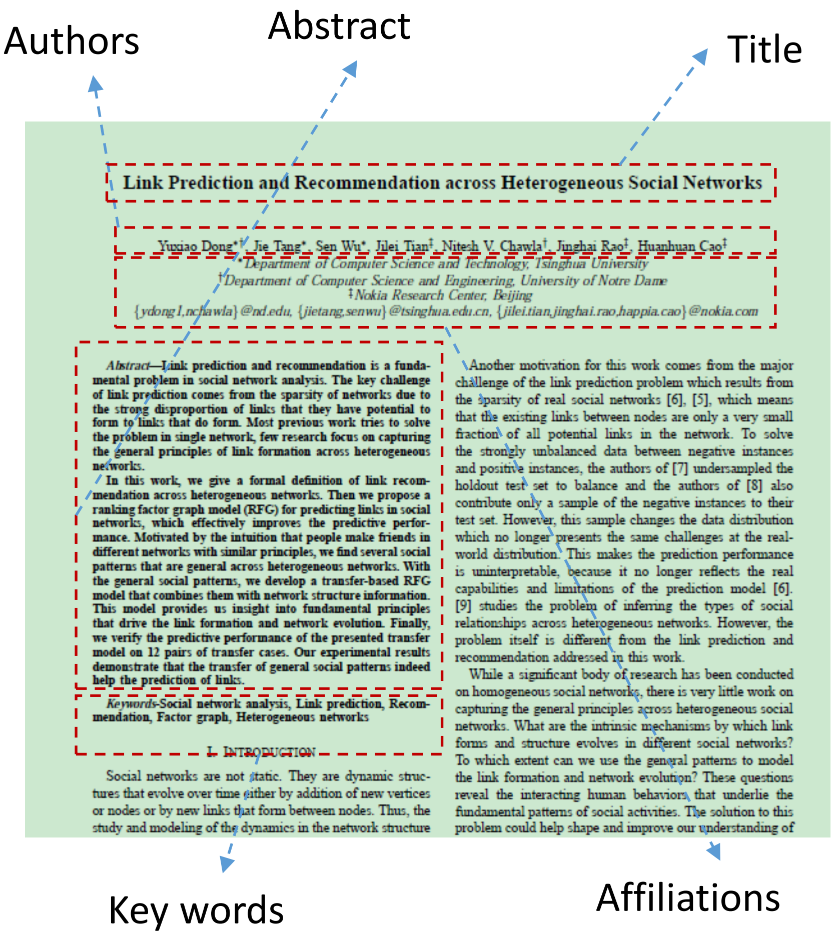
Given a series of snapshots of a network, we want to predict which link it will form in the next phase. That is, what will the network look like tomorrow?







LP in Bibliographic networks



Link Prediction and Recommendation across Heterogeneous Social Networks

Yuxiao Dong*†, Jie Tang*, Sen Wu*, Jilei Tian‡, Nitesh V. Chawla†, Jinghai Rao‡, Huanhuan Cao‡
*Department of Computer Science and Technology, Tsinghua University
†Department of Computer Science and Engineering, University of Notre Dame
‡Nokia Research Center, Beijing
{ydong1,nchawla}@nd.edu, {jietang,senwu}@tsinghua.edu.cn, {jilei.tian,jinghai.rao,happia.cao}@nokia.com

***Abstract*—Link prediction and recommendation is a fundamental problem in social network analysis. The key challenge of link prediction comes from the sparsity of networks due to the strong disproportion of links that they have potential to form to links that do form. Most previous work tries to solve the problem in single network, few research focus on capturing the general principles of link formation across heterogeneous networks.**

In this work, we give a formal definition of link recommendation across heterogeneous networks. Then we propose a ranking factor graph model (RFG) for predicting links in social networks, which effectively improves the predictive performance. Motivated by the intuition that people make friends in different networks with similar principles, we find several social patterns that are general across heterogeneous networks. With the general social patterns, we develop a transfer-based RFG model that combines them with network structure information. This model provides us insight into fundamental principles that drive the link formation and network evolution. Finally, we verify the predictive performance of the presented transfer model on 12 pairs of transfer cases. Our experimental results demonstrate that the transfer of general social patterns indeed help the prediction of links.

***Keywords*-Social network analysis, Link prediction, Recommendation, Factor graph, Heterogeneous networks**

I. INTRODUCTION

Social networks are not static. They are dynamic structures that evolve over time either by addition of new vertices or nodes or by new links that form between nodes. Thus, the study and modeling of the dynamics in the network structure

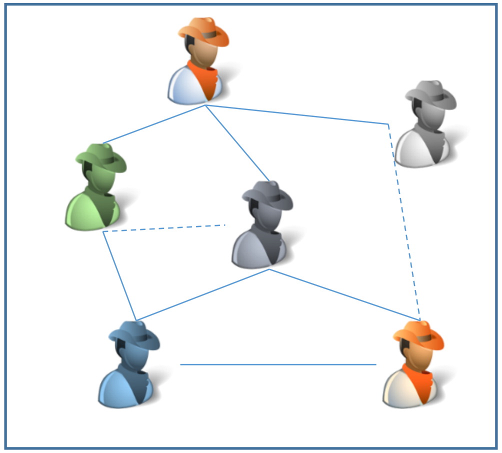
Another motivation for this work comes from the major challenge of the link prediction problem which results from the sparsity of real social networks [6], [5], which means that the existing links between nodes are only a very small fraction of all potential links in the network. To solve the strongly unbalanced data between negative instances and positive instances, the authors of [7] undersampled the holdout test set to balance and the authors of [8] also contribute only a sample of the negative instances to their test set. However, this sample changes the data distribution which no longer presents the same challenges at the real-world distribution. This makes the prediction performance is uninterpretable, because it no longer reflects the real capabilities and limitations of the prediction model [6]. [9] studies the problem of inferring the types of social relationships across heterogeneous networks. However, the problem itself is different from the link prediction and recommendation addressed in this work.

While a significant body of research has been conducted on homogeneous social networks, there is very little work on capturing the general principles across heterogeneous social networks. What are the intrinsic mechanisms by which link forms and structure evolves in different social networks? To which extent can we use the general patterns to model the link formation and network evolution? These questions reveal the interacting human behaviors that underlie the fundamental patterns of social activities. The solution to this problem could help shape and improve our understanding of

Meta data of a paper

Extract

Co-author relationship prediction



How?

- Unsupervised methods
  - Common neighbors $s_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)|$
  - Katz $s_{xy}^{Katz} = \sum_{l=1}^{\infty} \beta^l \cdot |\text{path}_{xy}^{<l>}|$
  - Resource allocation $s_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}$
- **Supervised methods**
  - **Classification**

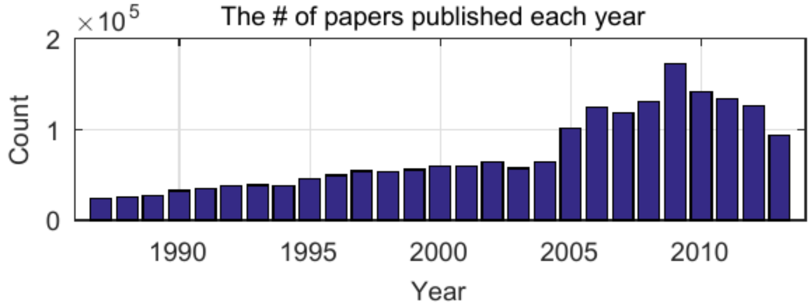
Dataset description

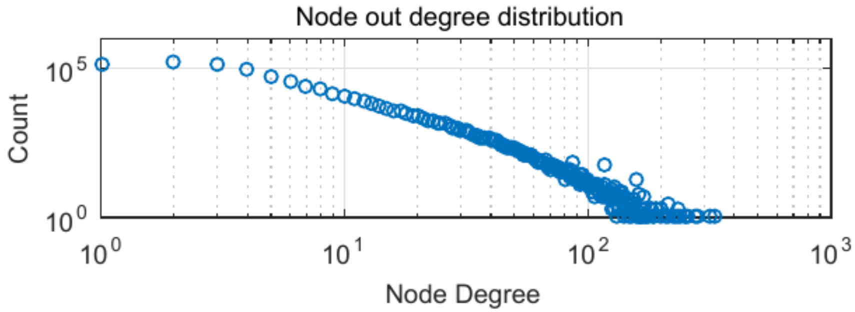
Overview

This dataset is designed for research purpose only.

The content of this data includes paper information, paper citation, author information and author collaboration. **2,092,356** papers and **8,024,869** citations between them are saved in the file AMiner-Paper.rar; **1,712,433** authors are saved in the file AMiner-Author.zip and **4,258,615** collaboration relationships are saved in the file AMiner-Coauthor.zip.

FileName	Node	***Number***	***Size***
AMiner-Paper.rar [download from mirror site]	Paper Citation	2,092,356 8,024,869	509 MB
AMiner-Author.zip [download from mirror site]	Author	1,712,433	167 MB
AMiner-Coauthor.zip [download from mirror site]	Collaboration	4,258,615	31.5 MB

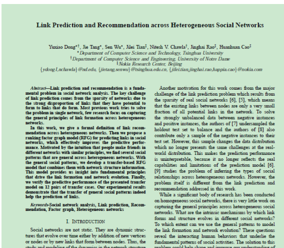


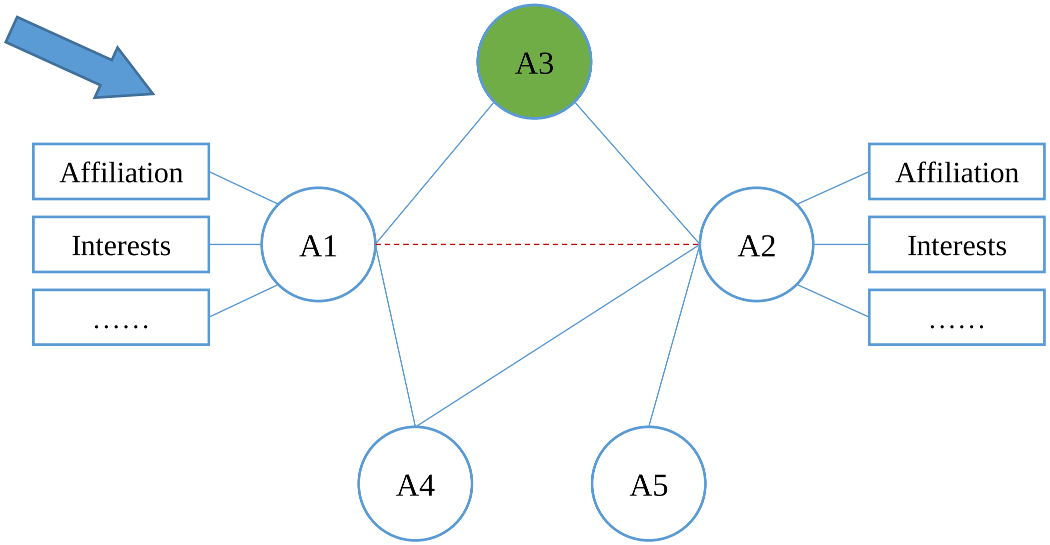


- AMiner Open Science Platform.
- 2.1 million papers, more than 1.7 million authors, 4.25 million collaboration links among authors.
- The content of the data includes the meta data of published papers such as title, abstract, author's name, affiliation and research topics.

https://cn.aminer.org/data

Features for Supervised Link Prediction


Link Prediction and Recommendation across Heterogeneous Social Networks

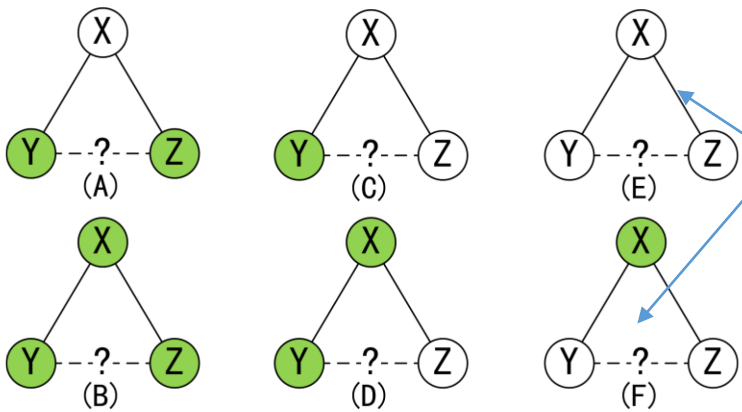


- Most works on link prediction only consider the topology features
  - Node degree, common neighbors, shortest paths.
- There are some other information can be used for link prediction
  - Social pattern (triadic relationships[1])
  - Node attributes (binary similarity[2])

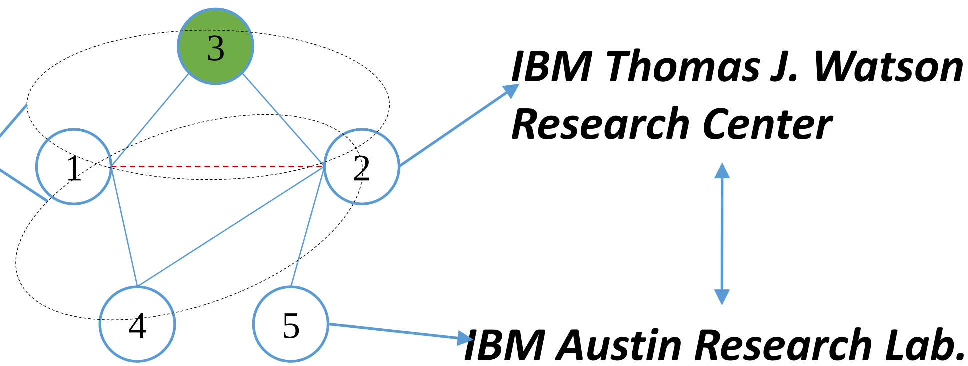
[1]Dong Y, Tang J, Wu S, et al. Link prediction and recommendation across heterogeneous social networks, IEEE ICDM.2012: 181-190.
[2]Al Hasan M, Chaoji V, Salem S, et al. Link prediction using supervised learning[C]//SDM06: workshop on link analysis, counter-terrorism and security. 2006.

Features for Supervised Link Prediction

Social pattern



Two different affiliations, but have something in common

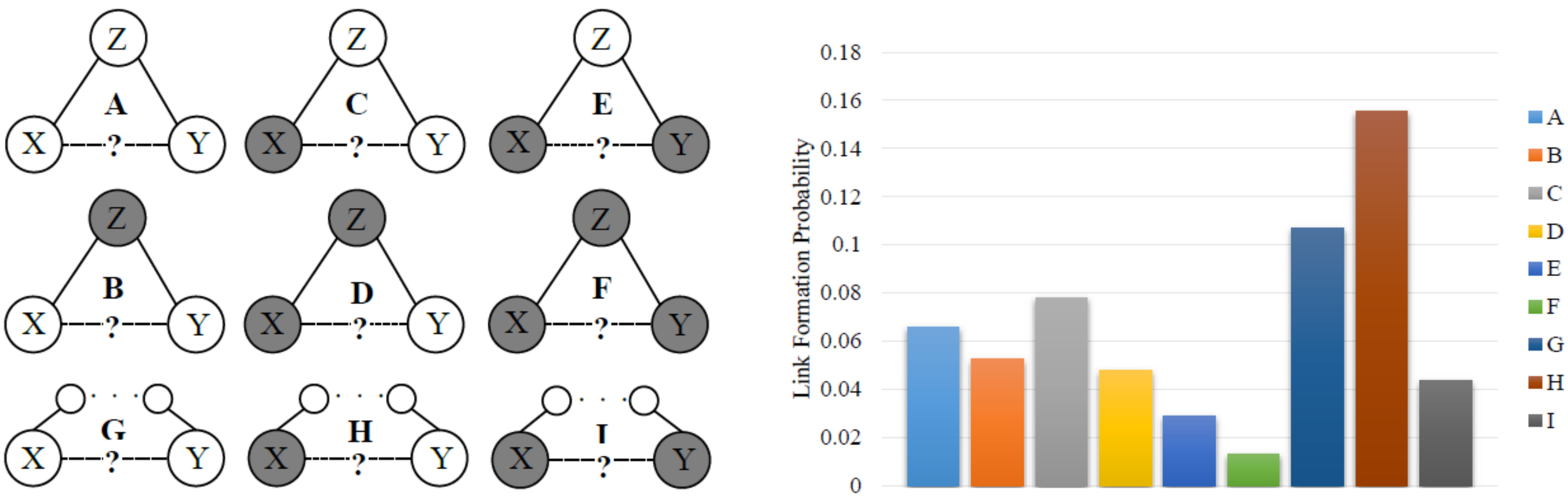


- Triadic relationship is only available when distance=2;
- Type inconsistent problem, for node 1 and node 2, (132) belongs to pattern F and (142) belongs to pattern E;
- Some potentially useful information may be lost by binary similarity in measuring author similarity .

Find a way to model the social patterns among authors and use an effective metric to measure author similarity.

[1]Dong Y, Tang J, Wu S, et al. Link prediction and recommendation across heterogeneous social networks, IEEE ICDM.2012: 181-190.
[2]Al Hasan M, Chaoji V, Salem S, et al. Link prediction using supervised learning[C]//SDM06: workshop on link analysis, counter-terrorism and security. 2006.

Social Pattern Feature



(a) Enumeration of author relationships

(b) Probability distribution of nine kinds of relationships

Figure 3: Different kinds of social patterns and their impact on link formation probability.

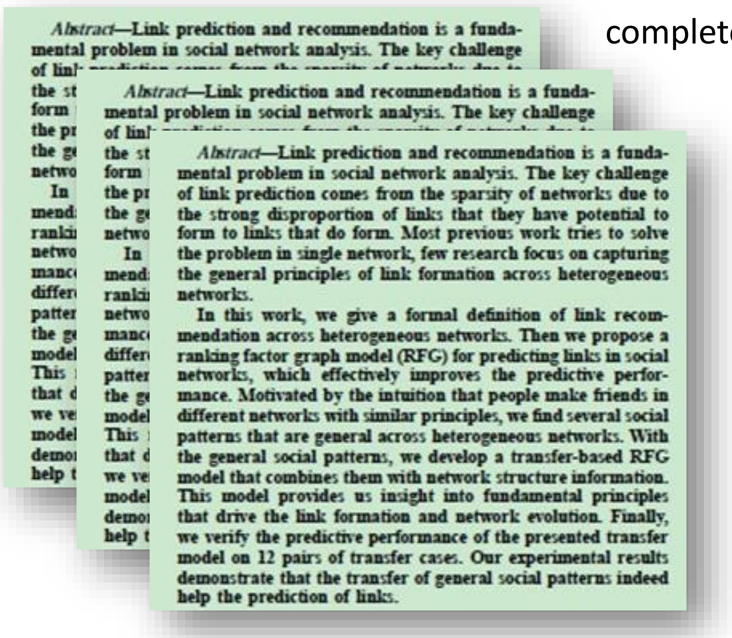
- Elite users: top 50% on the PageRank value list, others are treated as ordinary users
- Triadic relationship: CN=1 and 2-hops away from each other
- Dyadic relationship: CN>1 or 3-hops away from each other

Findings

- Ordinary users play a more important role in bridging two unconnected users than elite ones
- Positive correlation between the number of CN and link formation probability.

External Attribute Knowledge

For example, author u's affiliation ***IBM Thomas J. Watson Research Center*** and author v's affiliation ***IBM Austin Research Lab.*** will be treated as two completely different organizations. $K_{sim} = 0.408$



Abstract—Link prediction and recommendation is a fundamental problem in social network analysis. The key challenge of link prediction comes from the sparsity of networks due to the strong disproportion of links that they have potential to form to links that do form. Most previous work tries to solve the problem in single network, few research focus on capturing the general principles of link formation across heterogeneous networks.

In this work, we give a formal definition of link recommendation across heterogeneous networks. Then we propose a ranking factor graph model (RFG) for predicting links in social networks, which effectively improves the predictive performance. Motivated by the intuition that people make friends in different networks with similar principles, we find several social patterns that are general across heterogeneous networks. With the general social patterns, we develop a transfer-based RFG model that combines them with network structure information. This model provides us insight into fundamental principles that drive the link formation and network evolution. Finally, we verify the predictive performance of the presented transfer model on 12 pairs of transfer cases. Our experimental results demonstrate that the transfer of general social patterns indeed help the prediction of links.

TF-IDF

Cosine Similarity

Node Similarity

Tag1: ***data mining, information network***
Tag2: ***data analysis, social network***

Data	Mining	Information	Network	Analysis	Social
1	1	1	1	0	0
1	0	0	1	1	1

External Attribute Knowledge

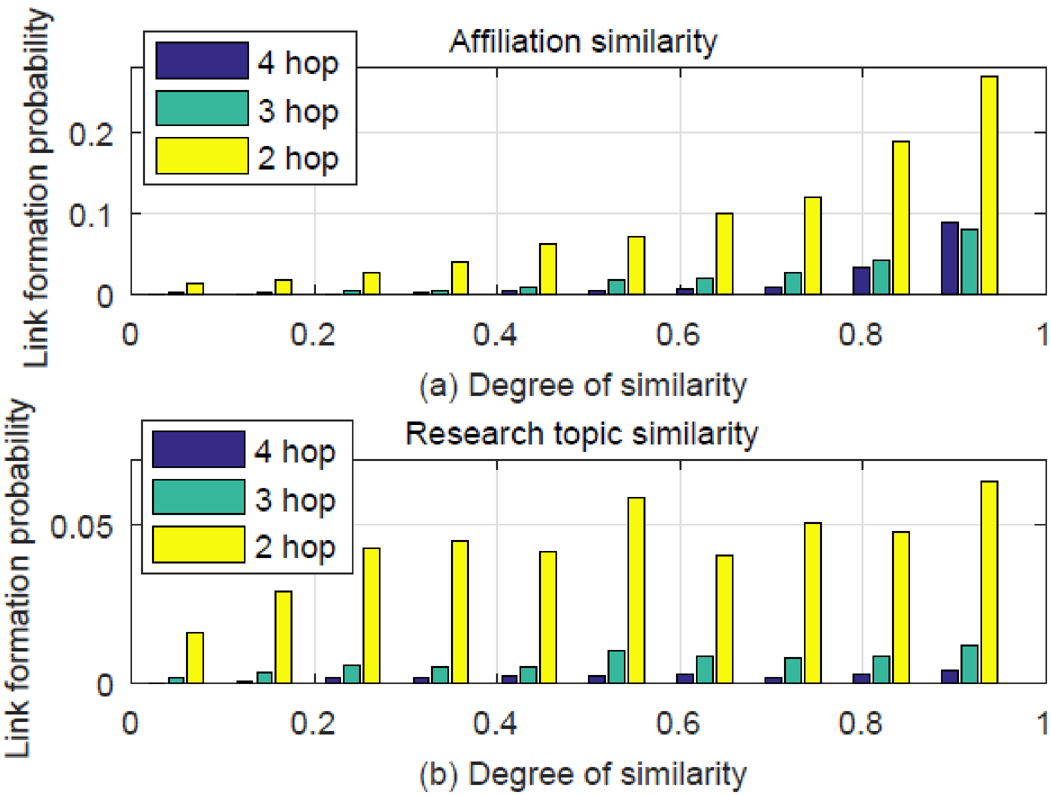


Figure 4: Link formation probability by external attribute knowledge.

Findings

- Authors with similar affiliation/research interests have a high probability to link to each other.
- The impact is sensitive to geodesic distance.
- Affiliation similarity plays a more important role than research interests in forming links.

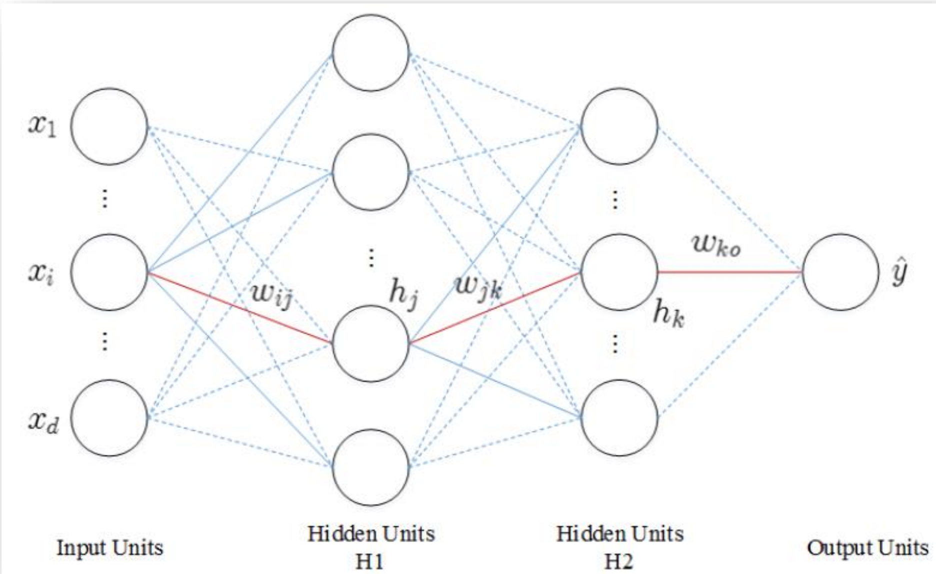
Feature list

TABLE 1: Full Feature Set.

Category	**Feature Name**	**Note**
Topology	Degree(u), Degree(v)	–
	Common Neighbors(u,v)	–
	Jaccard Coef.(u,v)	–
	Admic/Adar(u,v)	–
	Pref. attachment(u,v)	–
	Maximum flow(u,v)	–
	Shortest paths(u,v)	$l = 5$
	PropFlow(u,v)	$l = 5$
	Katz(u,v)	$l = 5, \beta = 0.005$
Social Pattern	Social pattern(u,v)	–
Attributes	Topic similarity(u,v)	–
	Aff. similarity(u,v)	–

We use Stanford SNAP as the network library to extract all these features.
http://snap.stanford.edu/

Deep Neural Networks for Classification



A diagram of DNNs with two hidden layers.

- The activation function is ReLU

$$h_j = f(a_j) + b_j = max(0, a_j) + b_j$$

- The output layer is mapped by a logistic function

$$\hat{y} = p_o = \text{logistic}(a_o) = \frac{1}{1 + e^{-a_o}}$$

- Cross-entropy loss function

$$C = -\frac{1}{m}\sum_D [y\ln\hat{y} + (1-y)\ln(1-\hat{y})]$$

Then we use Gradient Descent to update all the parameters. Such as w_{ko} can be updated as:

$$w_{ko} = w_{ko} - \eta\frac{\partial C}{\partial w_{ko}} \qquad \frac{\partial C}{\partial w_{ko}} = \frac{\partial C}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial a_o}\frac{\partial a_o}{\partial w_{ko}}$$

Experiment Setup

- D1, [1999, 2004], D2, [2005, 2010]
- The data of the first five years are used to extract features, the data of the sixth year is used to extract labels;
- We try to make predictions only for active users (K>=5);
- The author pair is treated separately according to their distance. We also sample an equal sized set of negative pairs.
- Model evaluation: 80%/20%, repeat 20 times
- Performance metrics: ROC and AUC

TABLE 2: Data Description.

Data	Nodes	Edges	Distance	Samples	New Edges
D_1	337,636	699,800	2 hop	99,008	2,520
			3 hop	427,573	1,203
			4 hop	1,481,272	974
D_2	797,511	1,993,100	2 hop	800,210	13,090
			3 hop	5,869,830	6,985
			4 hop	30,160,999	5,703

AUC Performance

TABLE 3: AUC Performance.

Dataset		PropFlow	HPLP	RBM	Node2vec	HPLP+SPEAK	RBM+SPEAK	DNN
	2 hop	0.711	0.728(0.016)	0.718(0.018)	0.744(0.010)	0.768(0.015)	0.780(0.010)	**0.799**(0.008)
D_1	3 hop	0.669	0.683(0.013)	0.667(0.022)	0.740(0.018)	0.783(0.012)	0.780(0.016)	**0.804**(0.020)
	4 hop	0.676	0.654(0.024)	0.625(0.018)	0.734(0.017)	0.792(0.024)	0.792(0.019)	**0.812**(0.015)
	2 hop	0.747	0.769(0.006)	0.768(0.005)	0.742(0.005)	0.787(0.004)	0.789(0.005)	**0.812**(0.006)
D_2	3 hop	0.736	0.723(0.005)	0.702(0.008)	0.728(0.007)	0.749(0.007)	0.775(0.006)	**0.797**(0.007)
	4 hop	0.739	0.737(0.007)	0.643(0.010)	0.719(0.010)	0.794(0.008)	0.836(0.007)	**0.865**(0.007)

- With the increase of geodesic distance, the performance of topology-based methods gradually decline due to the loss of available structure information;
- Methods with SPEAK features perform consistently well especially when geodesic distance is greater than two;
- This implies the SPEAK features can be served as compensation when limited topology information is used.

ROC Performance

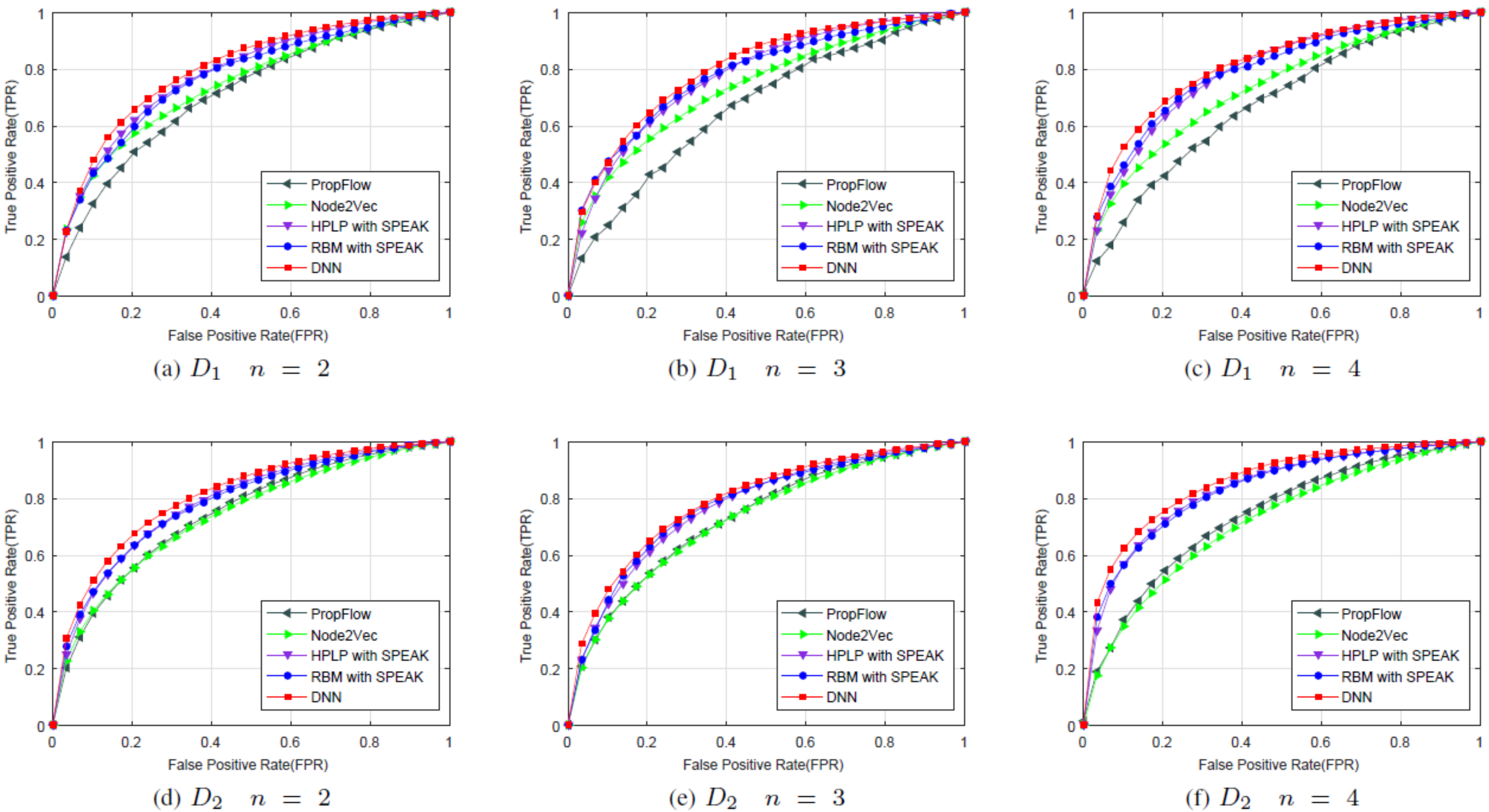


Figure 5: ROC curve for D_1 and D_2.

Parameter Sensitivity

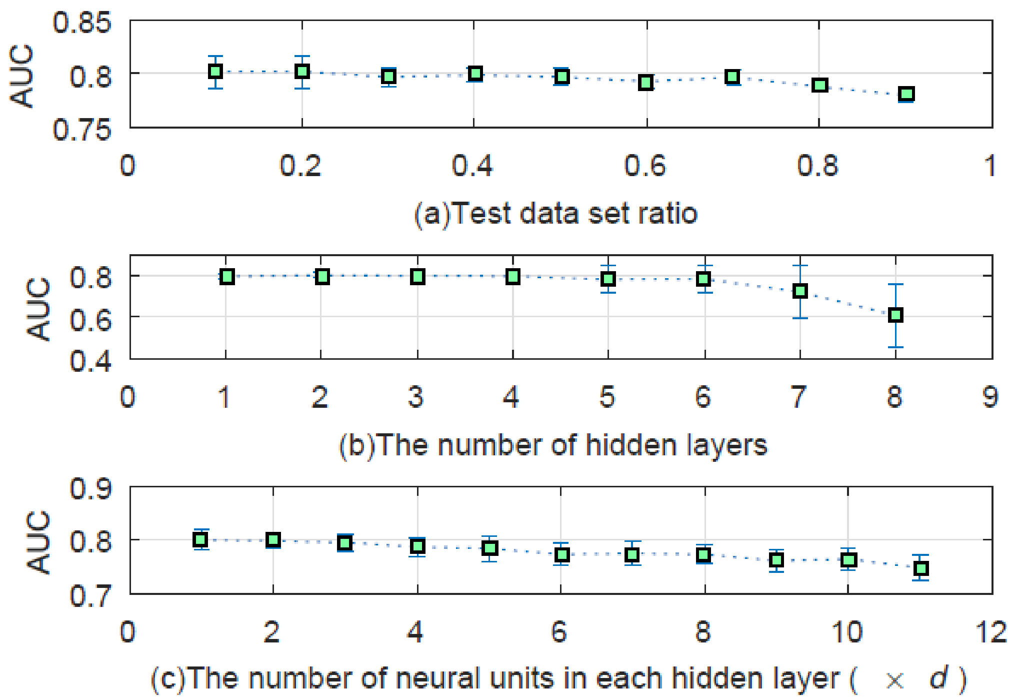


Figure 6: AUC values with different parameters

- Test data set ratio has slight influence on the AUC value;
- Too complex model (with many hidden layers or many nerual units) will overfit the data and reduce the performance;
- The number of hidden layers should be chosen according to the dataset ;
- The number of neural units should be one to four times of the feature dimension.

Conclusion

- Two kinds of novel features were introduced to capture node similarity based on social pattern and external attribute knowledge (SPEAK), respectively. The SPEAK features can boost the performance of link prediction.
- A deep learning approach using DNN was proposed to incorporate both topological features and the SPEAK features.
- We have released all the source code along with part of the dataset for the readers to reproduce our work.

https://github.com/zctzzy/speak_lp

Thanks for your attention!

Email: Chuanting.zhang@gmail.com
GitHub: https://github.com/zctzzy/