

Social Network Mining and Analysis: Link Prediction

Peng Wang

Networked World



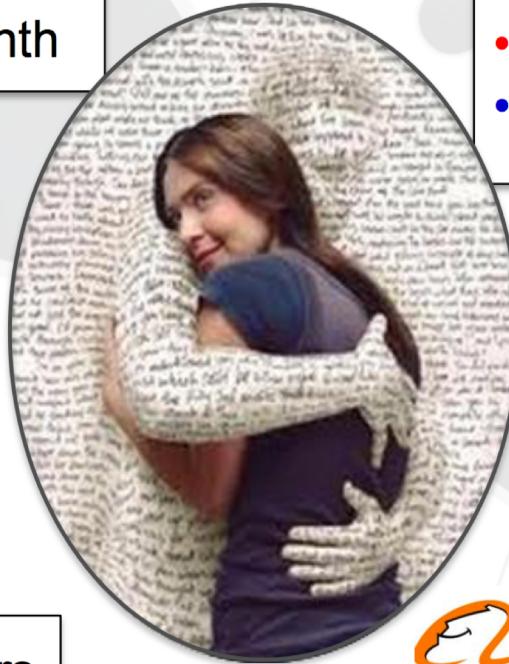
- **1.65 billion** MAU
- **2.5 trillion** minutes/month



- **320 million** MAU
- Peak: 143K tweets/s



- **304 million** active users
- **14 billion** items/year



- **QQ: 800 million** MAU
- **WeChat: 800 million** MAU



- **360 million** users
- **influencing** our daily life

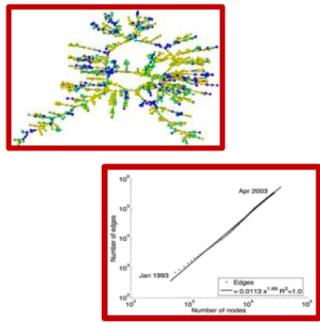
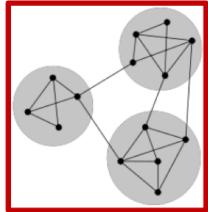


- ~**700 million** trans. (alipay)
- **120.7 billion** on 11/11

History of Social Network Analysis

Computational Social Science [Giles]

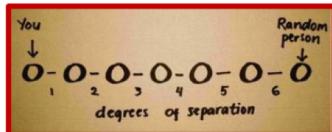
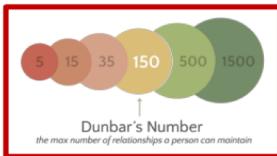
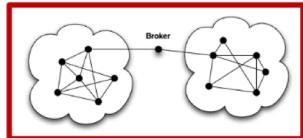
Computational Social Science [Lazer et al.]



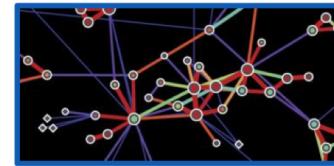
Scale Free [Barabási, Albert & Faloutsos et al.]

Small World [Watts, Strogatz]

HITS [Kleinberg] & PageRank [Brin&Page]



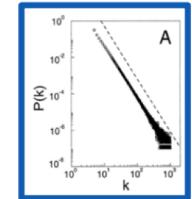
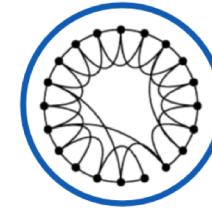
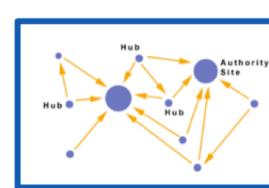
2012
2009



2009
2007
2005
2003
2002

Social Influence Analysis [Tang, Sun]
Spread of Obesity, Happiness [Christakis, Fowler]
Densification [Leskovec, Kleinberg, Faloutsos]
Link Prediction [Liben-Nowell, Kleinberg]
Influence Maximization [Kempe, Kleinberg, Tardos]
Community Detection [Girvan, Newman]

1999
1998
1997

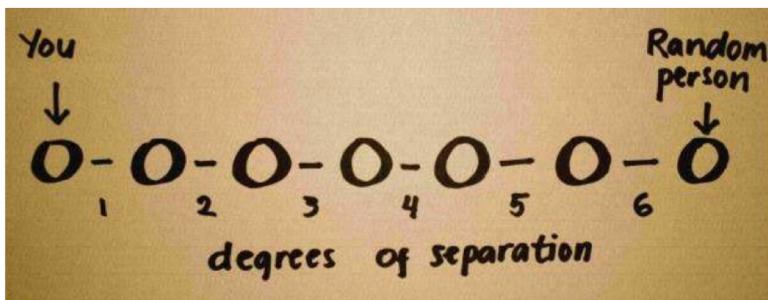


1995
1992
1973
1967

Structural Hole [Burt]
Dunbar's Number [Dunbar]
Weak Tie [Granovetter]
Six Degrees of Separation [Milgram]

1967: Six Degrees of Separation

- “Given two individuals selected randomly from the population, what is the probability that the minimum number of intermediaries required to link them is 0, 1, 2, ..., k ? ”
- Milgram “selected 296 volunteers and to distribute a mail to a stockholder living in Boston.”
- “The average number of intermediaries in the mailing chains was 5.2.”

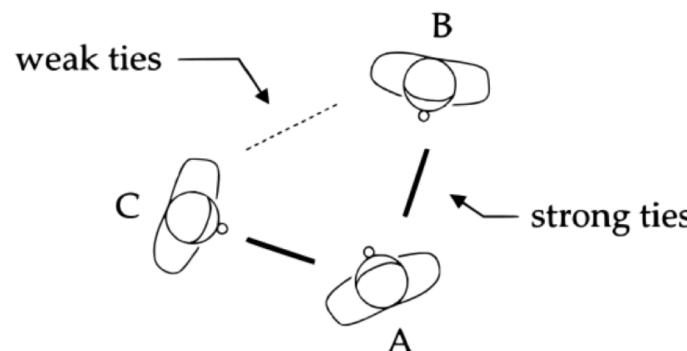


Stanley Milgram. The small world problem. *Psychology Today*, 2(1):60–67, 1967. Cited by 6967 (as of May 2016)

1995	○	Structural Hole [Burt]
1992	○	Dunbar's Number [Dunbar]
1973	○	Weak Tie [Granovetter]
1967	○	Six Degrees of Separation [Milgram]

1973: Weak Tie

- The **weak tie hypothesis** argues, if A is linked to both B and C, then there is a greater-than-chance probability that B and C are linked to each other.
- Essentially form “A friend’s friend is also my friend”

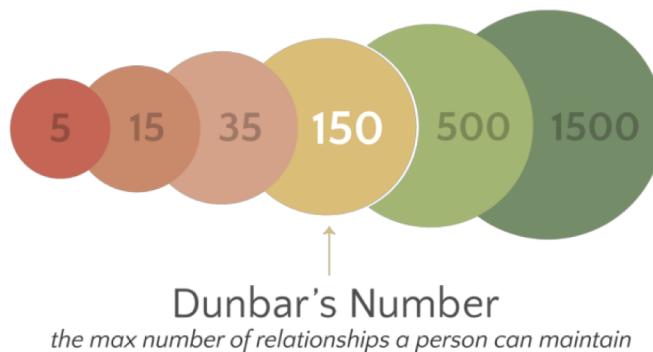


Granovetter, Mark S. The strength of weak ties. *American journal of sociology* 78.6 (1973): 1360-1380. Cited by 43186 (as of April 2017)

1995	○	Structural Hole [Burt]
1992	○	Dunbar's Number [Dunbar]
1973	○	Weak Tie [Granovetter]
1967	○	Six Degrees of Separation [Milgram]

1992: Dunbar's Number

- “**Dunbar’s number** is a suggested cognitive limit to the number of people with whom one can maintain stable social relationships.”
- Dunbar “proposed that humans can only comfortably maintain **150** stable relationships.”

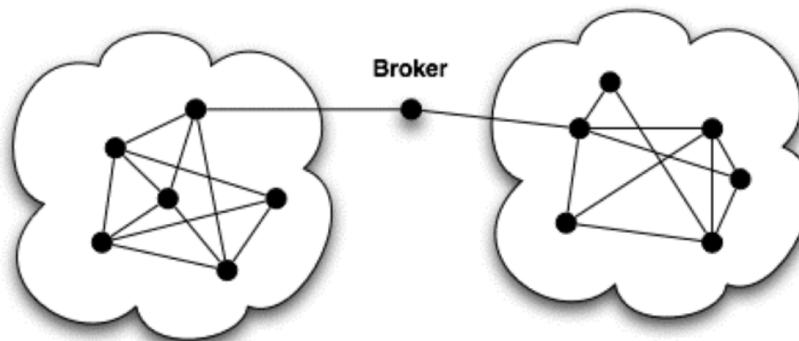


Robin I. M. Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution* 22 (6): 469–493.
Cited by 1393 (as of May 2016)

1995	○	Structural Hole [Burt]
1992	○	Dunbar's Number [Dunbar]
1973	○	Weak Tie [Granovetter]
1967	○	Six Degrees of Separation [Milgram]

1995: Structural Holes

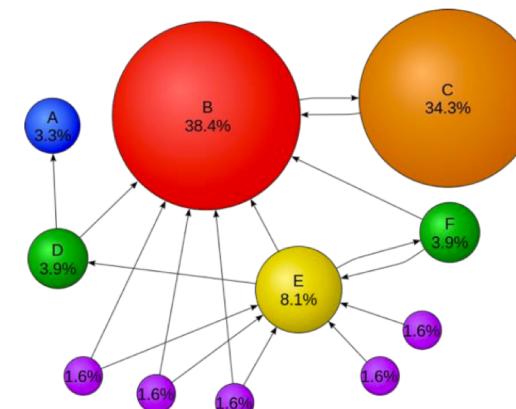
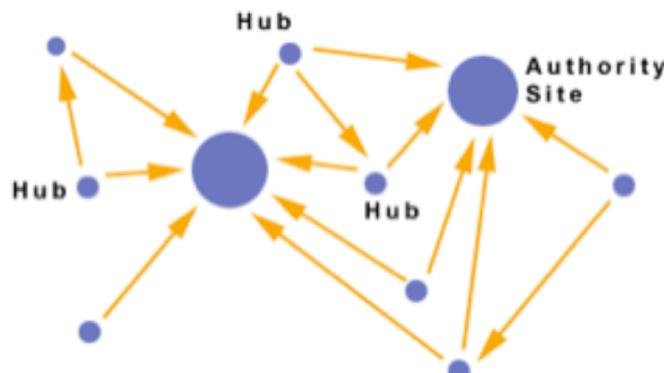
- “The position of a bridge between distinct groups allows him or her to transfer valuable information from one group to another.”
- “The individual can combine all the ideas he or she receives from different sources and come up with the most innovative idea among all.”



Ronald S. Burt. *Structural Holes: The Social Structure of Competition*. Cambridge: Harvard University Press. 1995.
Cited by 20207 (as of May 2016)

1995	○	Structural Hole [Burt]
1992	○	Dunbar's Number [Dunbar]
1973	○	Weak Tie [Granovetter]
1967	○	Six Degrees of Separation [Milgram]

1997-1998: HITS and PageRank



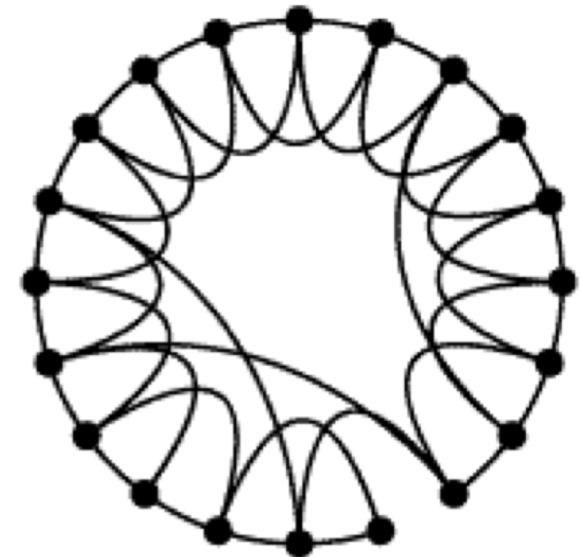
Scale Free [Barabási,Albert & Faloutsos et al.] 1999
Small World [Watts, Strogatz] 1998
HITS [Kleinberg]&PageRank [Brin&Page] 1997

1. Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In ACM SODA, 1998. Extended version in Journal of the ACM 46 (1999). Also at IBM Research Report RJ 10076, May 1997. Cited by 9896 (as of May 2016)
2. Sergey Brin, Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. In WWW'07, Pages 107-117,1998. Cited by 14618 (as of May 2016)

1998: Small World

Small-World Properties:

1. short average path lengths
2. high clustering coefficients



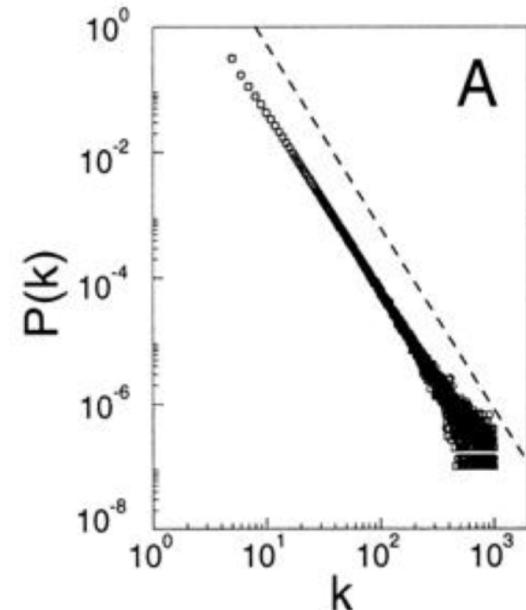
Duncan Watts, Steven Strogatz. Collective dynamics of ‘small-world’ networks. Nature 393 (6684): 440–442. Cited by 29044 (as of May 2016)

1999: Scale Free

- “A scale-free network is a network whose degree distribution follows a power law.”
- The fraction $P(k)$ of nodes having k connections to other nodes:

$$P(k) \sim k^{-\gamma}$$

where γ is a parameter whose value is typically in the range $2 < \gamma < 3$



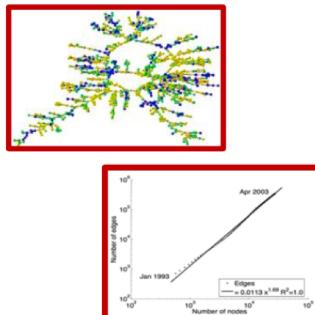
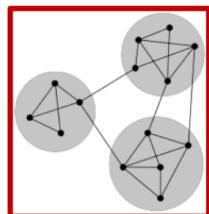
Albert-László Barabási, Réka Albert. Emergence of scaling in random networks, Science, 286:509–512, 1999. Cited by 25189 (as of May 2016)

Michalis Faloutsos, Petros Faloutsos, Christos Faloutsos. On power-law relationships of the internet topology. In ACM SIGCOMM 1999. Cited by 5631 (as of May 2016)

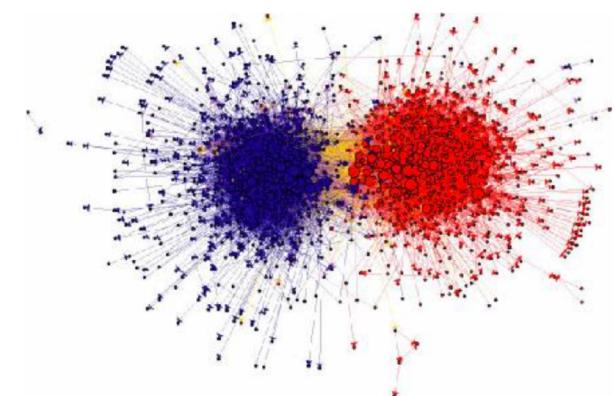
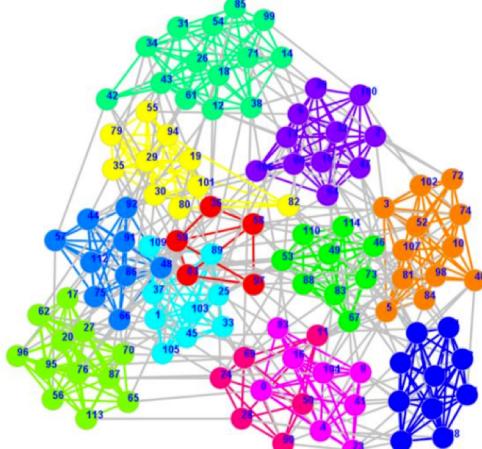
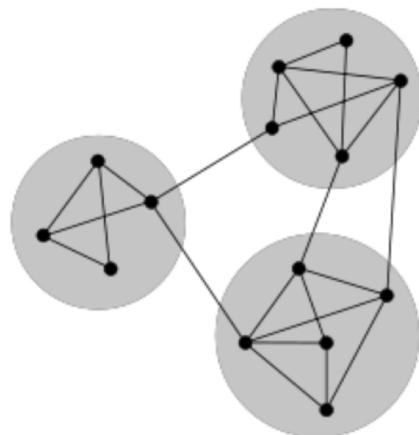
2002: Community Detection

- “The property of community structure, in which network nodes are joined together in tightly knit groups, between which there are only looser connections.”

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS* 99(12):7821-7826, 2002. Cited by 8088 (as of May 2016)



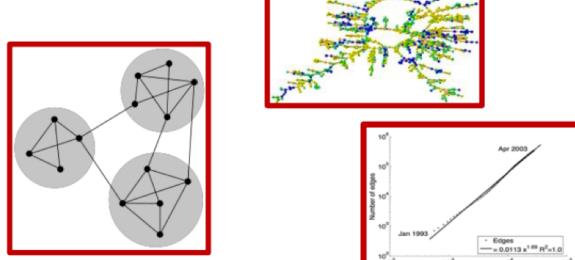
2009	○	Social Influence Analysis [Tang, Sun]
2007	○	Spread of Obesity, Happiness [Christakis, Fowler]
2005	○	Densification [Leskovec, Kleinberg, Faloutsos]
2003	○	Link Prediction [Liben-Nowell, Kleinberg] Influence Maximization [Kempe, Kleinberg, Tardos]
2002	○	Community Detection [Girvan, Newman]



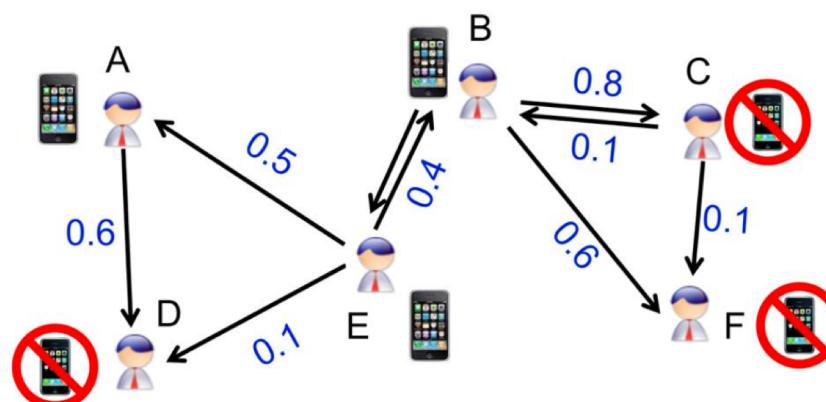
2003: Influence Maximization

- Minimize marketing cost and more generally to maximize profit, e.g., to get a small number of influential users to adopt a new product, and subsequently trigger a large cascade of further adoptions.

D. Kempe, J. Kleinberg, E. Tardos. Maximizing the Spread of Influence through a Social Network. In ACM KDD 2003.
Best Research Paper Award & Test of Time Paper Award. Cited by 3513 (as of May 2016)



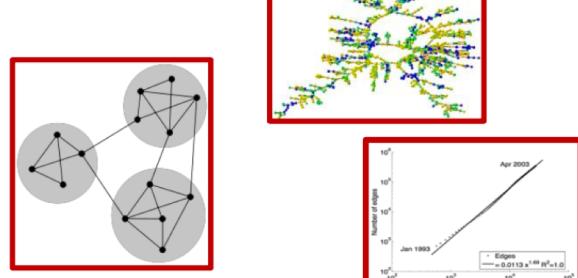
2009	Social Influence Analysis [Tang, Sun]
2007	Spread of Obesity, Happiness [Christakis, Fowler]
2005	Densification [Leskovec, Kleinberg, Faloutsos]
2003	Link Prediction [Liben-Nowell, Kleinberg] Influence Maximization [Kempe, Kleinberg, Tardos]
2002	Community Detection [Girvan, Newman]



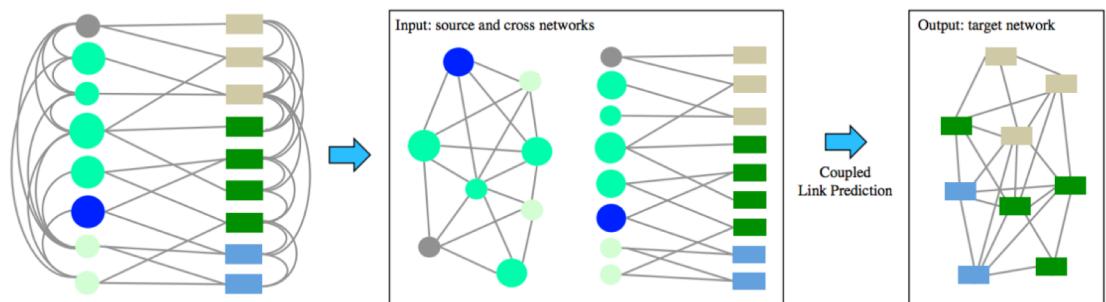
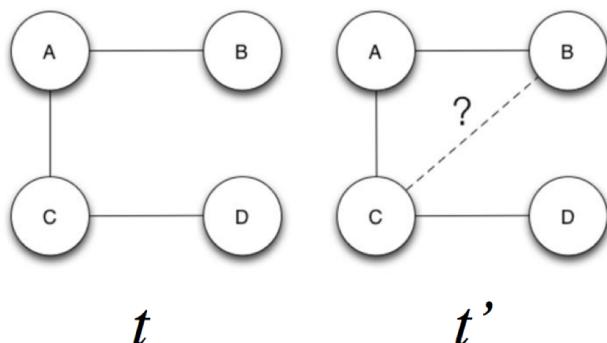
2003: Link Prediction

- “Given a snapshot of a social network at time t , we seek to accurately predict the edges that will be added to the network during the interval from time t to a given future time t' .”

D. Liben-Nowell, J. Kleinberg. The Link Prediction Problem for Social Networks. In ACM CIKM, 2003. Cited by 2616 (as of May 2016)



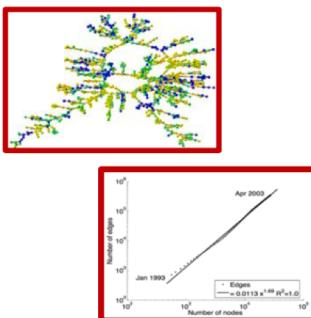
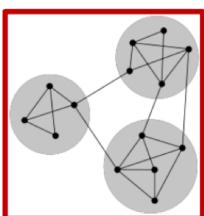
2009	Social Influence Analysis [Tang, Sun]
2007	Spread of Obesity, Happiness [Christakis, Fowler]
2005	Densification [Leskovec, Kleinberg, Faloutsos]
2003	Link Prediction [Liben-Nowell, Kleinberg] Influence Maximization [Kempe, Kleinberg, Tardos]
2002	Community Detection [Girvan, Newman]



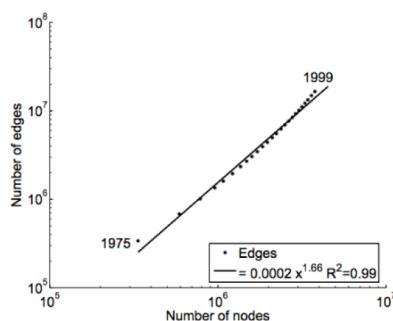
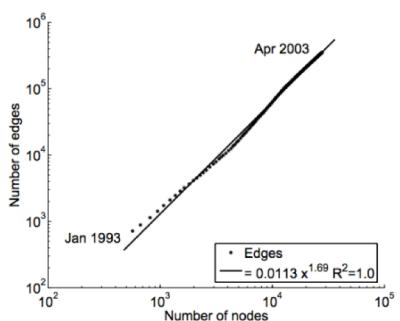
2005: Network Evolution

- “Most of graphs densify over time, with the number of edges growing superlinearly in the number of nodes.”
- “The average distance between nodes often shrinks over time.”

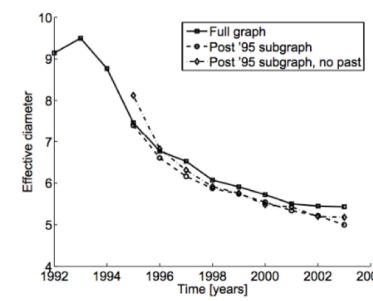
J. Leskovec, J. Kleinberg, C. Faloutsos. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In KDD, 2005. **Best Research Paper Award & Test of Time Award**. Cited by 1492 (as of May 2016)



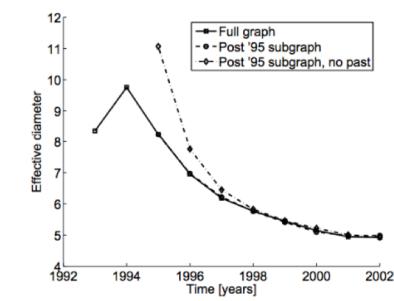
2009	Social Influence Analysis [Tang, Sun]
2007	Spread of Obesity, Happiness [Christakis, Fowler]
2005	Densification [Leskovec, Kleinberg, Faloutsos]
2003	Link Prediction [Liben-Nowell, Kleinberg] Influence Maximization [Kempe, Kleinberg, Tardos]
2002	Community Detection [Girvan, Newman]



Densification

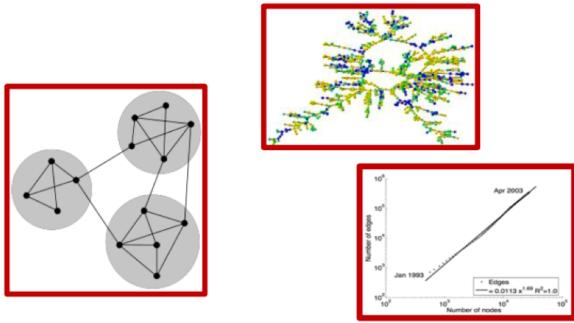


Shrinking Diameters

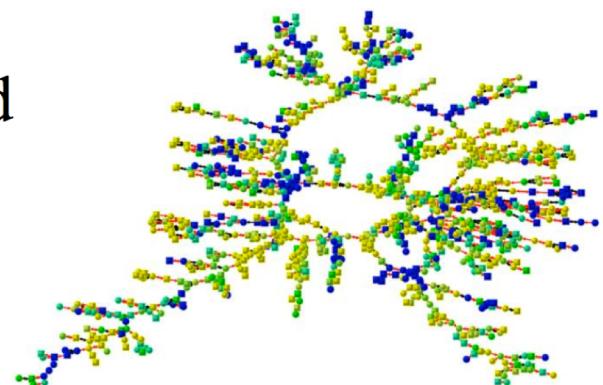


2007: Diffusion and Influence

1. Nicholas Christakis, James Fowler. The Spread of Obesity in a Large Social Network Over 32 Years. *The New England Journal of Medicine* 357 (4): 370–379, 2007. Cited by 3335 (as of May 2016)
2. Nicholas Christakis, James Fowler. The Collective Dynamics of Smoking in a Large Social Network. *The New England Journal of Medicine* 358 (21): 2249–2258, 2008. Cited by 1232 (as of May 2016)
3. James Fowler, Nicholas Christakis. The Dynamic Spread of Happiness in a Large Social Network. *British Medical Journal* 337 (768): a2338, 2009. Cited by 1178 (as of May 2016)



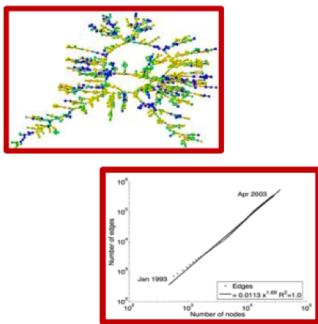
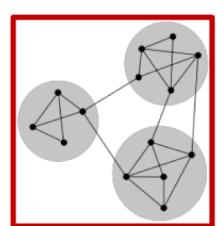
“If the husband became obese, the likelihood that his wife would become obese increased by **37%**.
—by tracking 10,000+ people for 32 years



2009: Social Influence Analysis

- “Social influence is a prevalent, complex and subtle force that governs the dynamics of all social networks.”

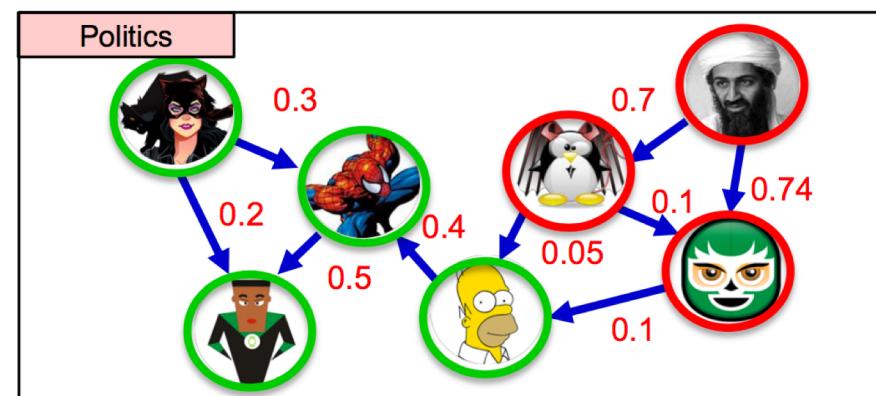
Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. In KDD'09, pages 807-816. **Top cited paper in KDD'09**. Cited by 624 (as of April 2017)



2009
2007
2005
2003
2002

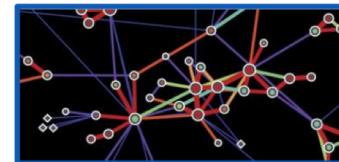
- Social Influence Analysis [Tang, Sun]**
Spread of Obesity, Happiness [Christakis, Fowler]
Densification [Leskovec, Kleinberg, Faloutsos]
Link Prediction [Liben-Nowell, Kleinberg]
Influence Maximization [Kempe, Kleinberg, Tardos]
Community Detection [Girvan, Newman]

“How to quantify the *social influences* from different topics”
—Topical Affinity Propagation (TAP),
an efficient model



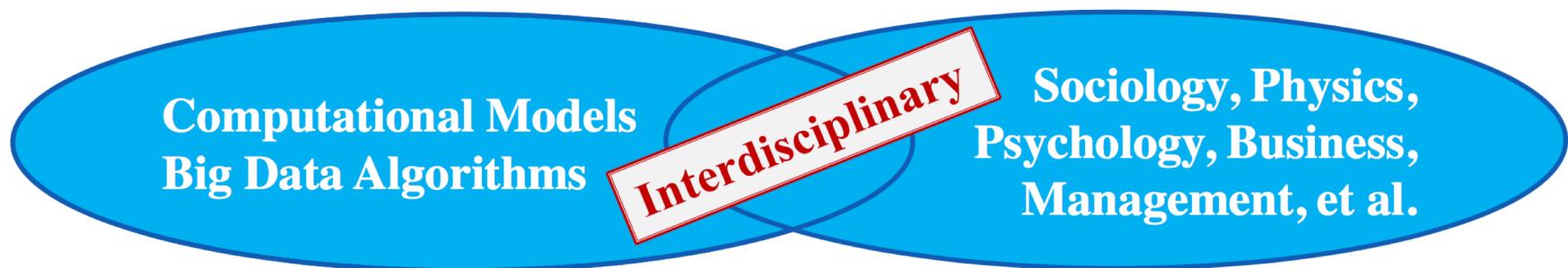
Computational Social Science

Computational Social Science [Giles]
Computational Social Science [Lazer et al.]



“A field is emerging that leverages the capacity to collect and analyze **data at a scale** that may reveal patterns of individual and group behaviors.”

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Alber-Laszlo Barabasi, et al. from Departments of Sociology, Computer Science, Physics, Business, Government, etc. at Harvard, MIT, Northeastern, Northwestern, Columbia, Cornell, etc.

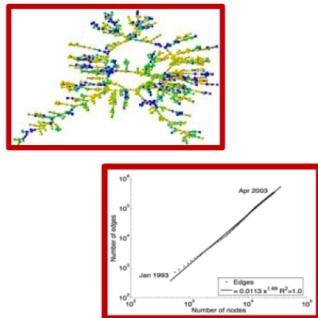
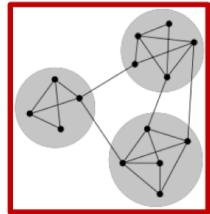


1. **David Lazer et al. Computational Social Science. *Science* 2009.**
2. James Giles. Computational Social Science: Making the Links. *Nature* 2012.

History of Social Network Analysis

Computational Social Science [Giles]

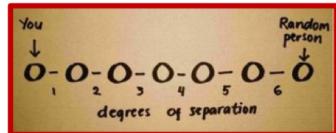
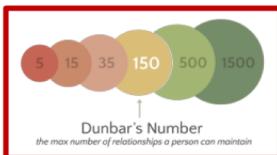
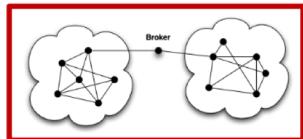
Computational Social Science [Lazer et al.]



Scale Free [Barabási, Albert & Faloutsos et al.]

Small World [Watts, Strogatz]

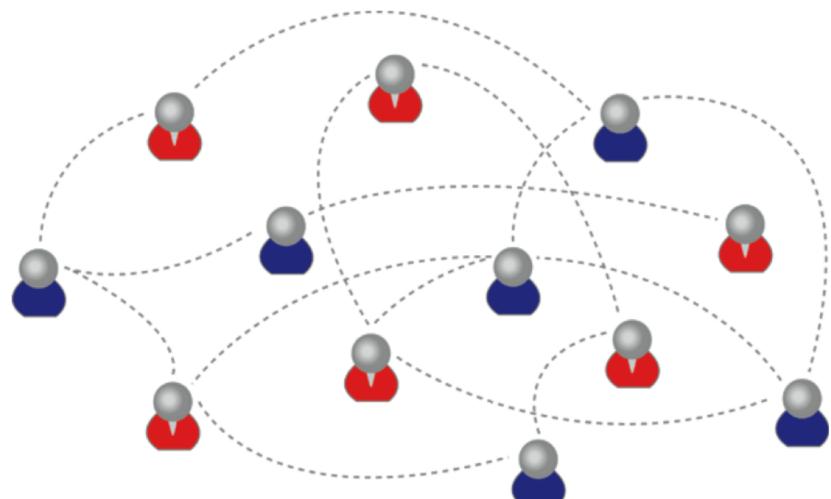
HITS [Kleinberg] & PageRank [Brin&Page]



		2012	
		2009	
2009	Social Influence Analysis [Tang, Sun]		
2007	Spread of Obesity, Happiness [Christakis, Fowler]		
2005	Densification [Leskovec, Kleinberg, Faloutsos]		
2003	Link Prediction [Liben-Nowell, Kleinberg] Influence Maximization [Kempe, Kleinberg, Tardos]		
2002	Community Detection [Girvan, Newman]		
1999			
1998			
1997	A		
1995	Structural Hole [Burt]		
1992	Dunbar's Number [Dunbar]		
1973	Weak Tie [Granovetter]		
1967	Six Degrees of Separation [Milgram]		

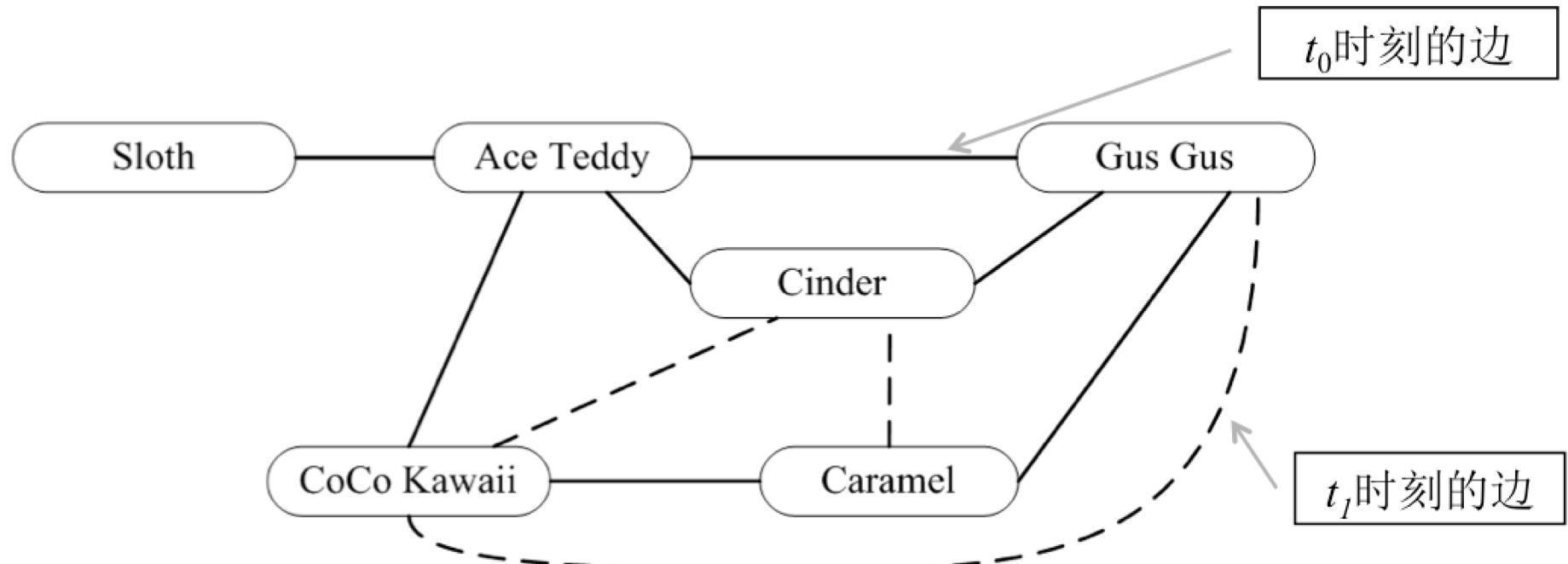
What is a social network?

- A **social network** is:
 - a **graph** made up of :
 - a set of **individuals**, called “nodes”, and
 - tied by one or more **interdependency**, such as friendship, called “edges”.



What is Link Prediction

Link Prediction: Given a network in t_0 , predict new links in future $t_1 (t_1 > t_0)$, or find missing and unseen links.



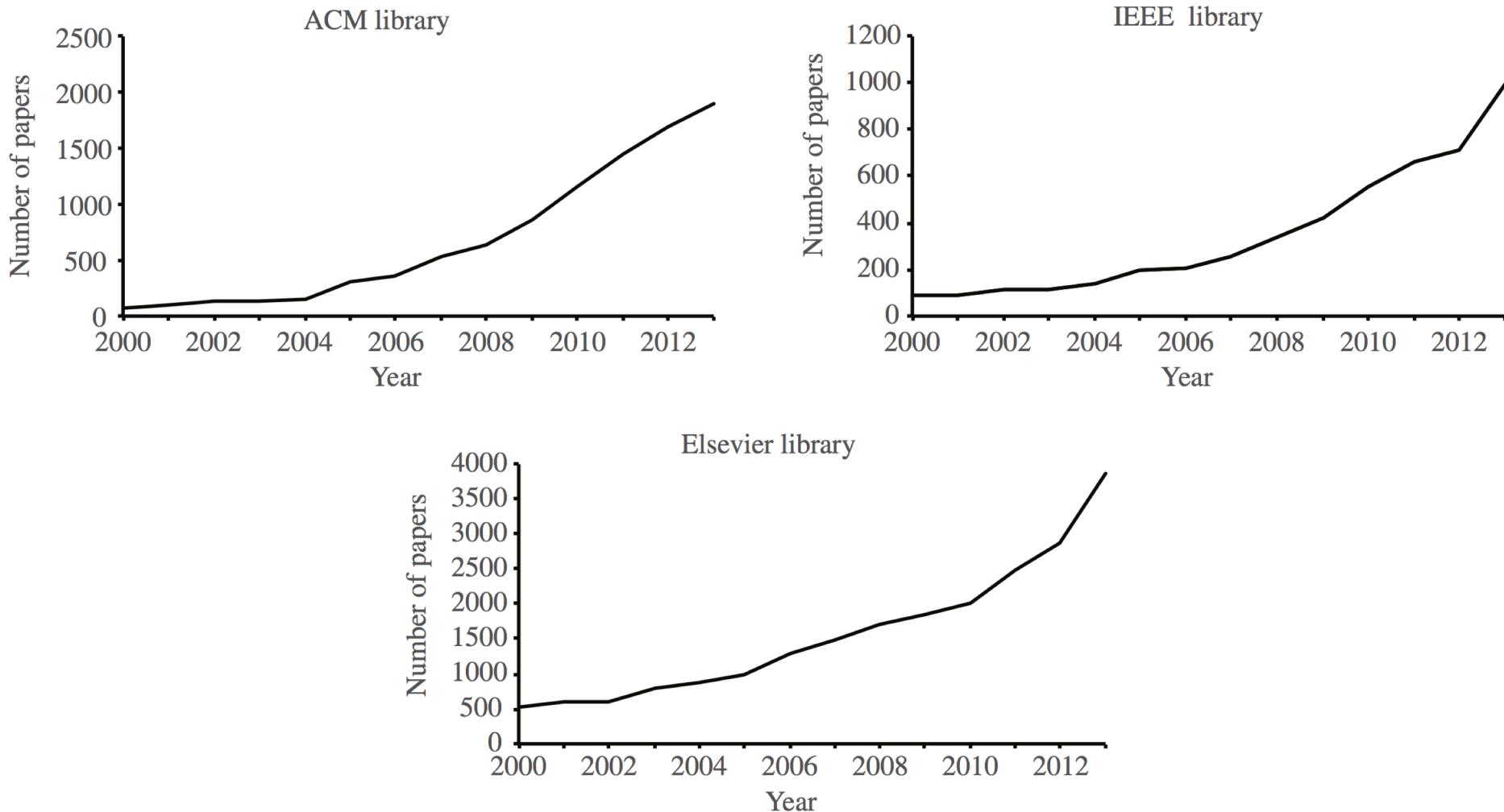
Wang P, et al. *Sci China Inf Sci* January 2015 Vol. 58 011101:2

Figure 1 Published papers related to link prediction problem on ACM, IEEE, and Elsevier libraries.

Link prediction framework

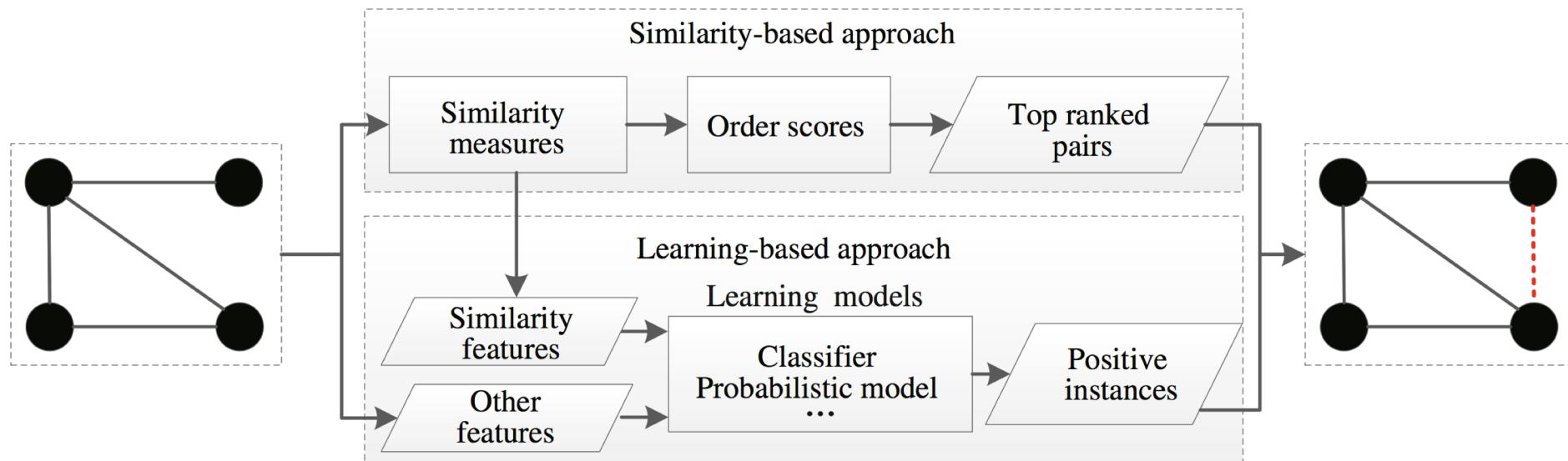
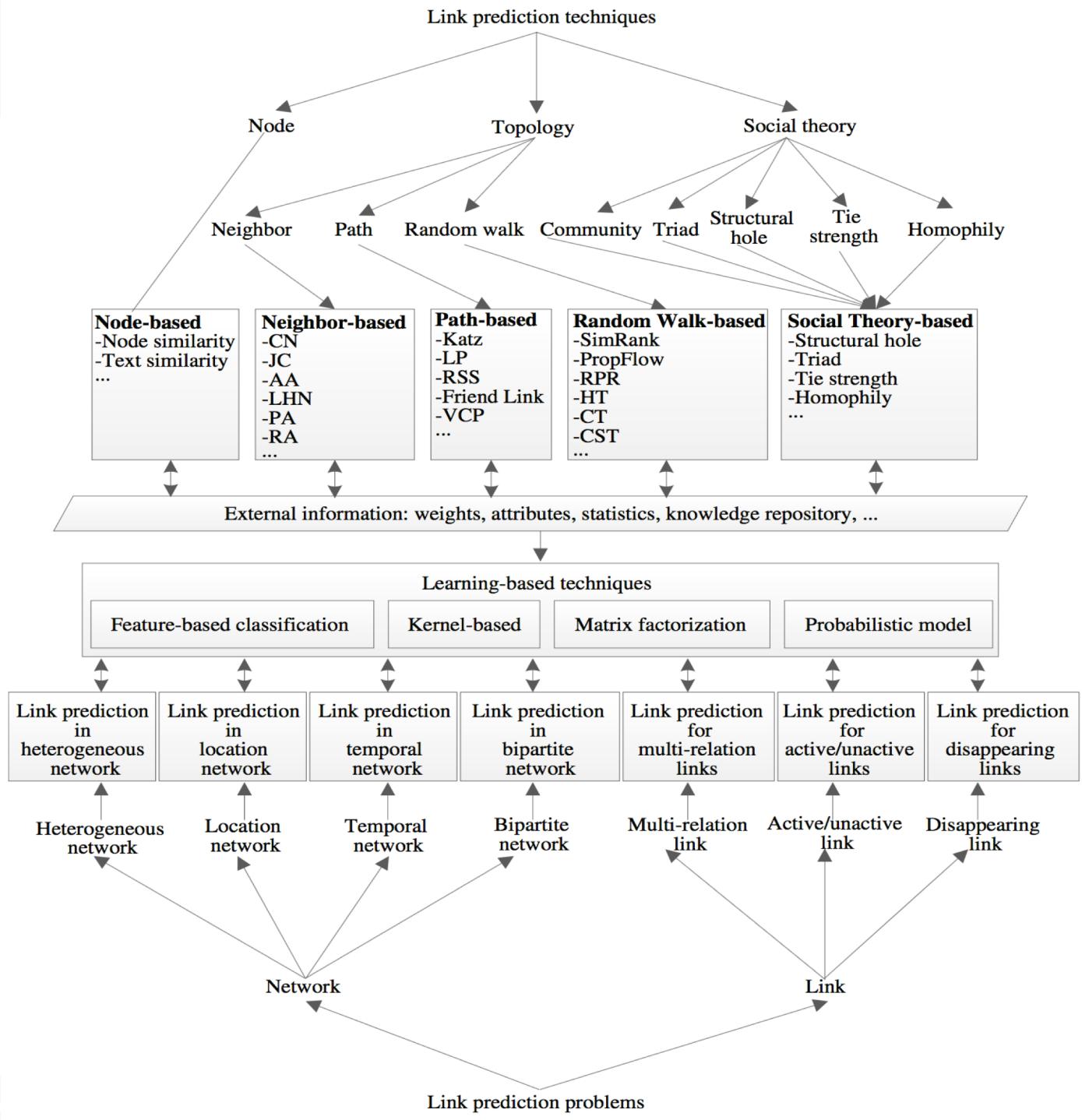


Figure 7 The generic link prediction framework.



Solutions

- **1. Based on node properties**

- ✓ Similarity between node properties

- **2. Based on topology**

- ✓ Based on common neighbors

CN,JC,AA,PA,RA,LHN,HP,HD,PD,SI,SC,...

- ✓ Based on paths

Katz, LP, RSS, VCP, ...

- ✓ Based on random walk

SimRank, Rooted PageRank, PropFlow,...

Solutions

- **3. Based on social theory**

- ✓ community, triadic closure, strong and weak ties, homophily, and structural balance

- **4. Based on machine learning**

- ✓ Feature-based classification
- ✓ Probabilistic graph model
- ✓ Matrix factorization
- ✓ Deep learning

Topology based methods

Common Neighbors (CN): The CN metric is one of the most widespread measurements used in link prediction problem mainly due to its simplicity [30]. For two nodes, x and y , the CN is defined as the number of nodes that both x and y have a direct interaction with. A bigger number of the common neighbors make it easier that a link between x and y will be created. This measure is defined as following formula.

$$\text{CN}(x, y) = |\Gamma(x) \cap \Gamma(y)|. \quad (1)$$

Since CN metric is not normalized, it usually reflects the relative similarities between node pairs. Therefore, some neighbor-based metrics consider how to normalize the CN metric reasonably.

Jaccard Coefficient (JC): Jaccard coefficient normalizes the size of common neighbors. It assumes higher values for pairs of nodes which share a higher proportion of common neighbors relative to total number of neighbors they have. This measure is defined as:

$$\text{JC}(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}. \quad (2)$$

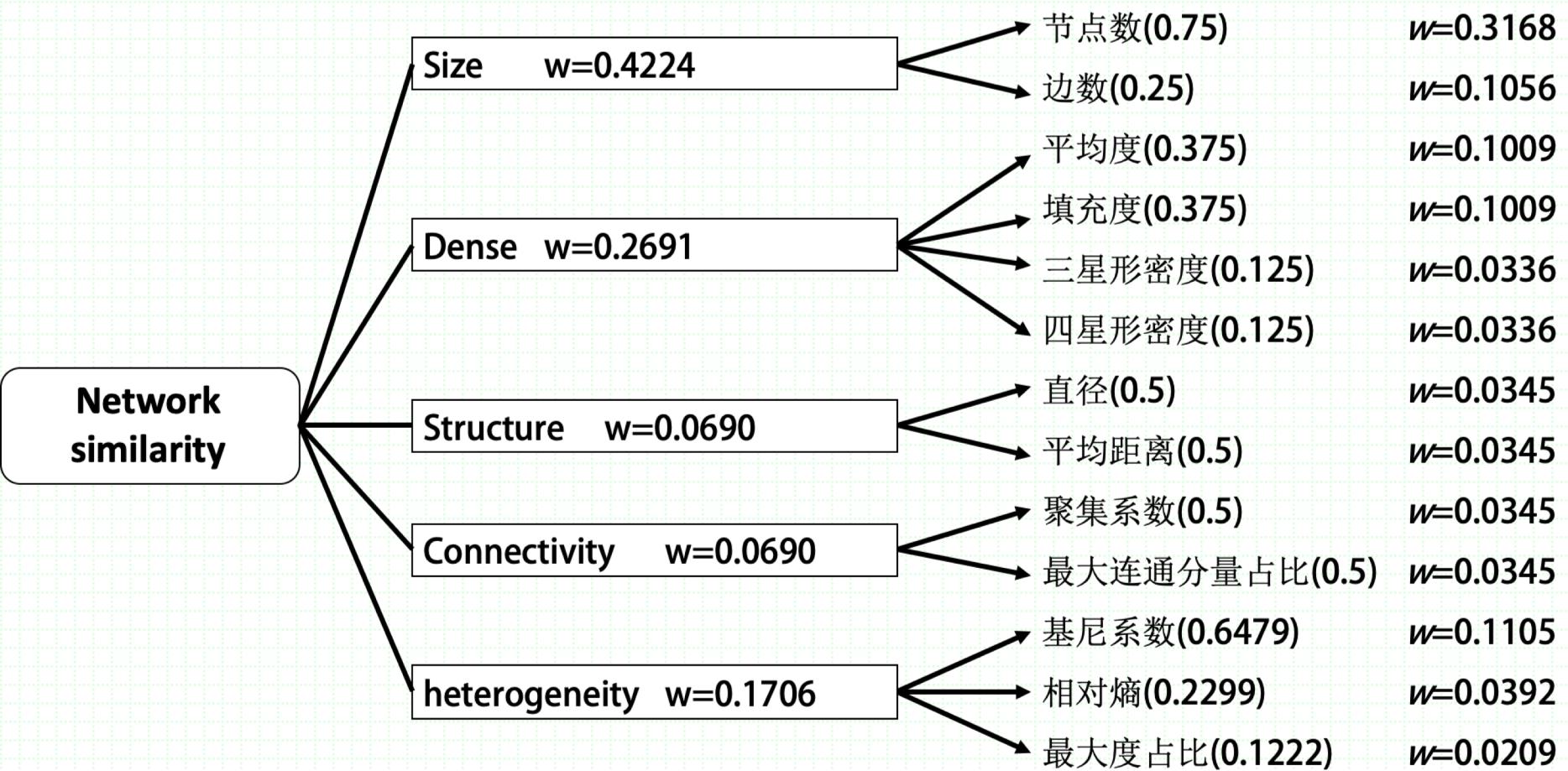
Sørensen Index (SI): This metric is defined as formula (3). Besides considering the size of the common neighbors, it also points out that lower degrees of nodes would have higher link likelihood.

$$\text{SI}(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x)| + |\Gamma(y)|}. \quad (3)$$

Salton Cosine Similarity (SC): SC is a common cosine metric for measuring the similarity between two nodes x and y . It is defined as:

$$\text{SC}(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{|\Gamma(x)| \cdot |\Gamma(y)|}}. \quad (4)$$

Link prediction evaluation: a benchmark methods



Link prediction evaluation: a benchmark methods

算法 1: 评测数据集筛选算法

Input: 各网络数据集网络拓扑结构属性矩阵, $M = (m_1, m_2, \dots, m_k)$, $m_i = (c_1, c_2, \dots, c_t)$, $i \in [1, k]$

Output: 评测数据集序号集合 $u = \{u_1, u_2, u_3, u_4\}$

1 队列 Q (初始只含一个成员 $\{1, 2, \dots, k\}$), 二叉树 B (初始只含根节点 $Root = \{1, 2, \dots, k\}$), 各数据集得分序列 $S = (s_1, s_2, \dots, s_k)$;

2 **while** 队列 Q 不空 **do**

3 Front = Q.front();

4 寻找两个集合 P_1 和 P_2 使得 $P_1 \cup P_2 = Front$ 且 P_1 和 P_2 的距离最近;

5 将 P_1 和 P_2 添加到二叉树 B 中;

6 Q.push(P_1);

7 Q.push(P_2);

8 **for** $i \in P_1$ **do**

9 $s_i = s_i + \frac{1}{Height(P_1)} \times \frac{|P_1|}{|Front|};$

10 **end**

11 **for** $j \in P_2$ **do**

12 $s_j = s_j + \frac{1}{Height(P_2)} \times \frac{|P_2|}{|Front|};$

13 **end**

14 **end**

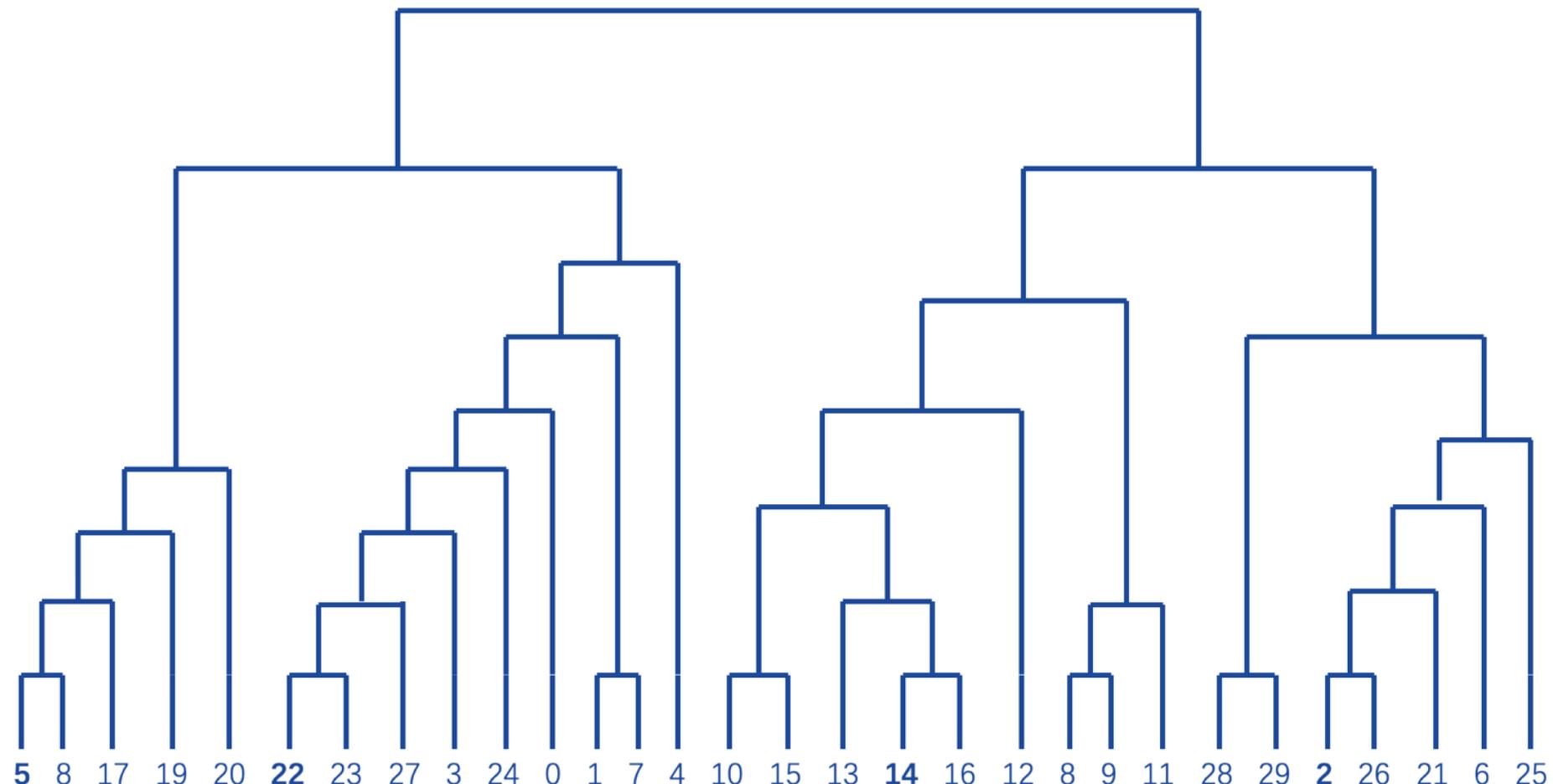
15 count = 0;

16 **for** $P \in B$ 且 $Height(P) = 2$ **do**

17 $u_{count} = S$ 中得分最高且对应下标属于 P 中的项对应的下标;

18 **end**

Link prediction evaluation: a benchmark methods

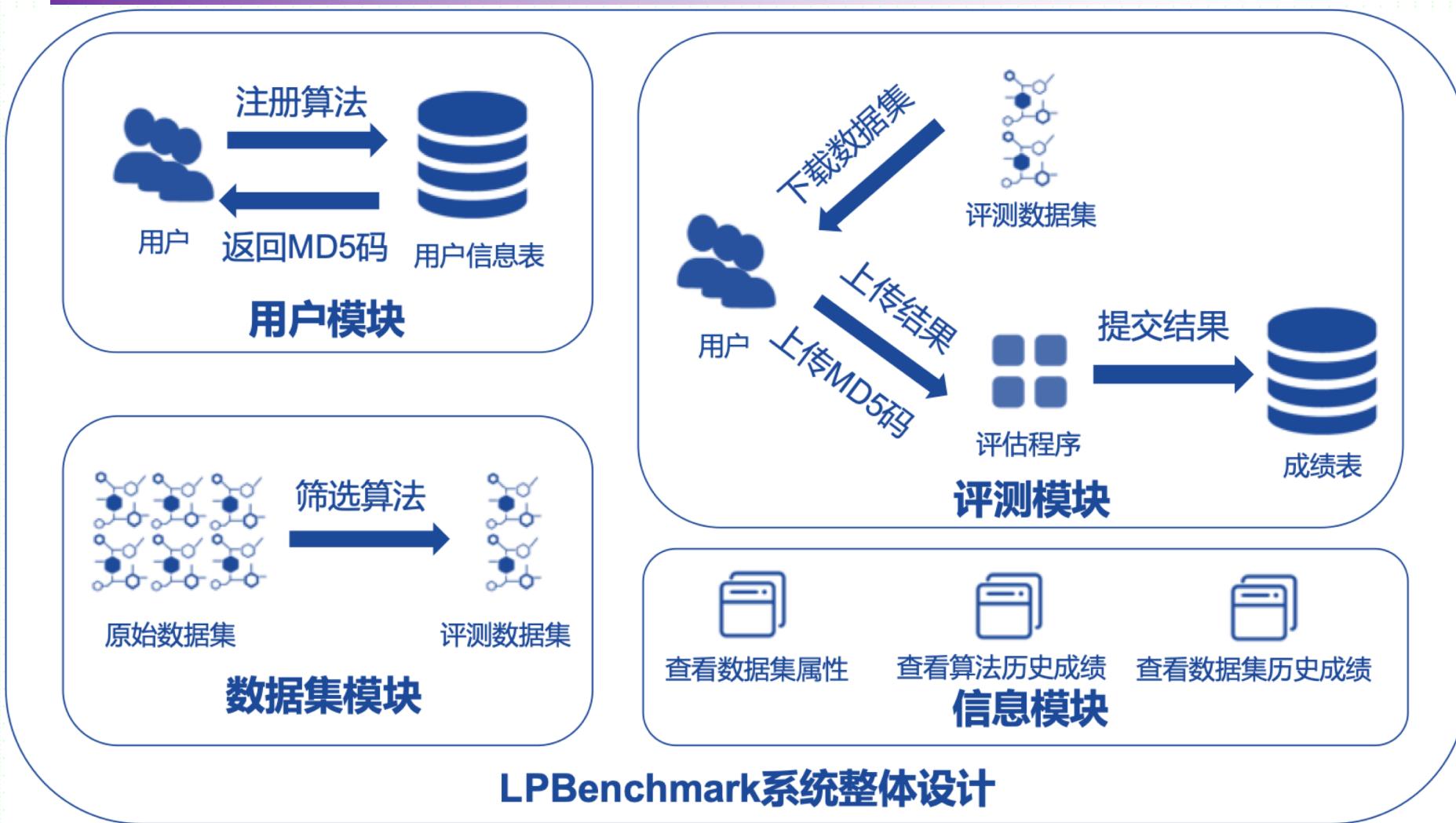


Link prediction evaluation: a benchmark methods

LPBenchmark Dataset

数据集名称	动态属性	节点数	边数	平均度	基尼系数
arXiv HepPh	动态网络	34546	421578	24.407	52.5%
Enron	动态网络	87273	1148072	26.31	90.8%
Manufacturing Emails	动态网络	167	82927	993.14	61.9%
Wikipedia talk,Portuguese	动态网络	541335	2424962	8.9589	76.3%
Berkeley/Stanford	静态网络	685230	7600595	22.184	65.9%
Blogs	静态网络	1224	19025	31.087	63.3%
CoraCitation	静态网络	23166	91500	7.8995	52.1%
Douban	静态网络	154908	327162	8.9589	69.4%

Link prediction evaluation: a benchmark methods



Link prediction evaluation: a benchmark methods

LPBenchmark Network Datasets User Performance About

Register Your Algorithm

Welcome to LPBenchmark!

LPBenchmark is an effective and fair evaluation tool for link prediction methods.

Propose a new method for link prediction and want to evaluate its effectiveness?

#	Network Type	Network Name	Performance Rank	Datasets List
1	Static Network	Berkeley/Stanford		
		Blogs		<button>Get Performance Rank</button>
		Cora Citation		<button>Get Datasets List</button>
		Douban		
		axXiv HepPh		

Link prediction evaluation: a benchmark methods

LPBenchmark Network Datasets User Performance About

Register Your Algorithm

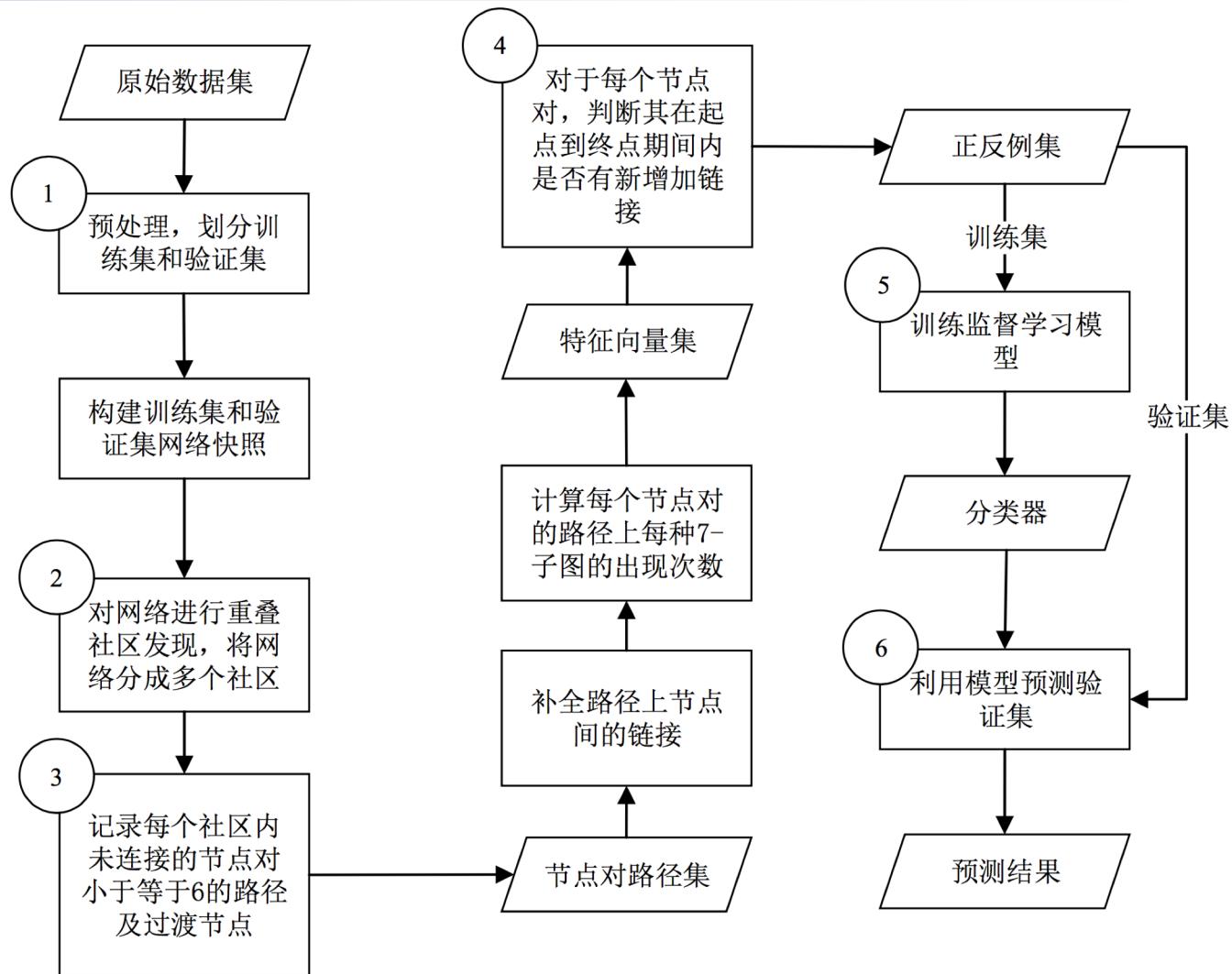
PropFlow

Southeast University

vincent shaw

Rank	Dataset Name	Dataset Type	Grade	AUC
1	Berkeley/Stanford	Temporal Network	0.9822	1.0412
2	Blogs	Temporal Network	0.8701	1.0359
3	Cora Citation	Temporal Network	0.8421	1.31
4	Douban	Temporal Network	0.9848	1.4458

Link prediction based on OVCP structure



Link prediction based on OVCP structure

● Representation of nodes

- ✓ Limited subgraphs in n -size graph

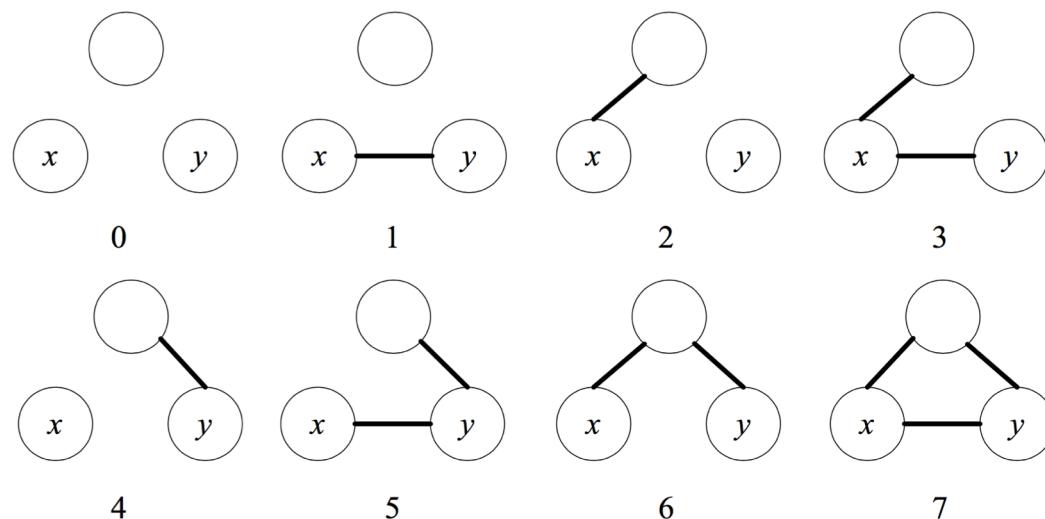
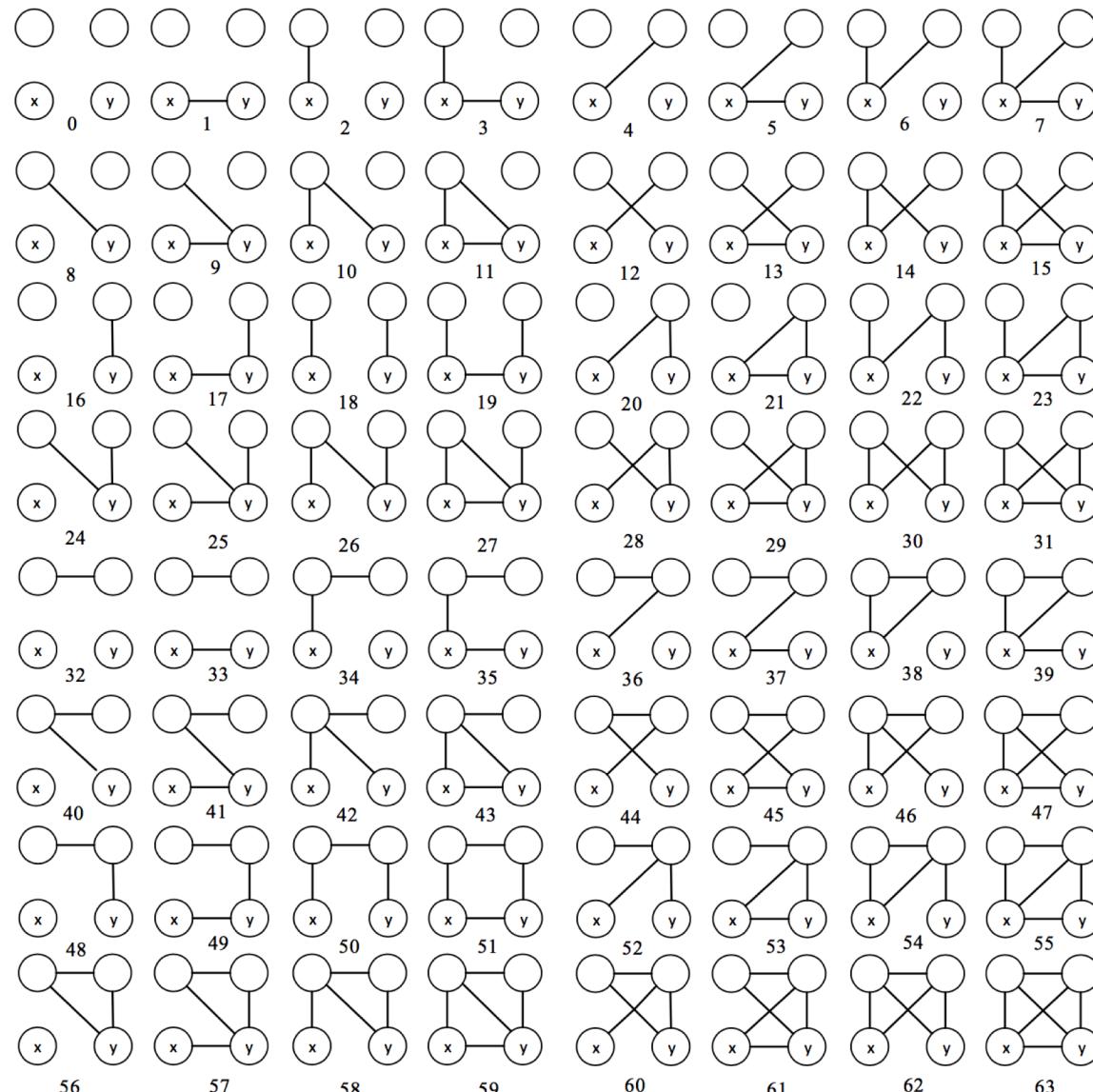


图 3.2 $n = 3$ 的所有可能子图

Link problem



0:{0}	7:{7}	12:{12,18}	25:{25}	32:{32}	39:{39}	44:{44,50}	57:{57}
1:{1}	8:{8,16}	13:{13,19}	26:{26,28}	33:{33}	40:{40,48}	45:{45,51}	58:{58,60}
2:{2,4}	9:{9,17}	14:{14,22}	27:{27,29}	34:{34,36}	41:{41,49}	46:{46,54}	59:{59,61}
3:{3,5}	10:{10,20}	15:{15,23}	30:{30}	35:{35,37}	42:{42,52}	47:{47,55}	62:{62}
6:{6}	11:{11,21}	24:{24}	31:{31}	38:{38}	43:{43,53}	56:{56}	63:{63}

Link prediction based on OVCP structure

● OVCP feature selection for nodes

- ✓ Do not consider long path: six degrees separation

We calculate that all subgraphs with the longest paths <6

Total: 78084 subgraphs

Link prediction based on OVCP structure

- **OVCP feature selection for nodes**
- ✓ Transfer (V_x, V_y) to vector

定义 3.11 (7-子图的位矩阵) 具有 7 个节点的无权无向图 G , 其位矩阵 V 是 7×7 阶矩阵, 设矩阵所有元素集为 N , 则 V 的定义如下

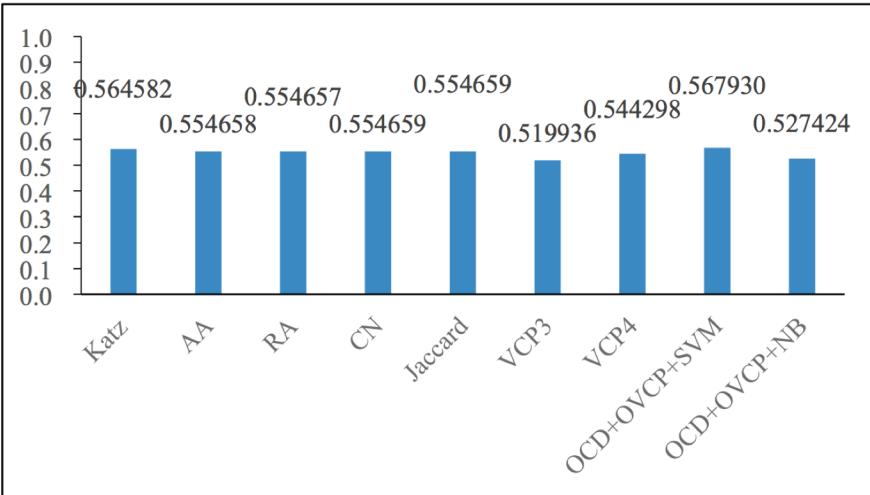
$$V = \left\{ \{v_{ij} \in N \mid v_{ij} = \begin{cases} 2^{(7(i-1)-\frac{1}{2}(i-1)(i+1)+j-2)} & \text{if } i < j \\ 0 & \text{if } i = j \\ 2^{(7(j-1)-\frac{1}{2}(j-1)(j+1)+i-2)} & \text{if } i > j \end{cases}\} \right\} \quad (3-7)$$

用另一种形式表示, $V = \begin{bmatrix} 0 & 1 & 2 & 4 & 8 & 16 & 32 \\ 1 & 0 & 64 & 128 & 256 & 512 & 1024 \\ 2 & 64 & 0 & 2^{11} & 2^{12} & 2^{13} & 2^{14} \\ 4 & 128 & 2^{11} & 0 & 2^{15} & 2^{16} & 2^{17} \\ 8 & 256 & 2^{12} & 2^{15} & 0 & 2^{18} & 2^{19} \\ 16 & 512 & 2^{13} & 2^{16} & 2^{18} & 0 & 2^{20} \\ 32 & 1024 & 2^{14} & 2^{17} & 2^{19} & 2^{20} & 0 \end{bmatrix}$

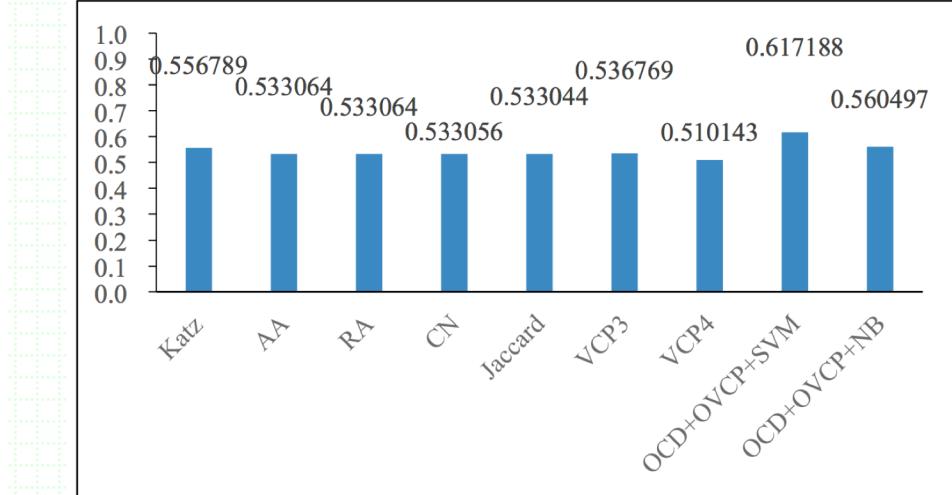
定义 3.12 (7-子图的地址) 7-子图 G 的地址可以用邻接矩阵与位矩阵做内积得到, 即

$$\Psi(G) = \sum_{i=1}^7 \sum_{j=1}^7 (v_{ij} \times d_{ij}) \quad (3-8)$$

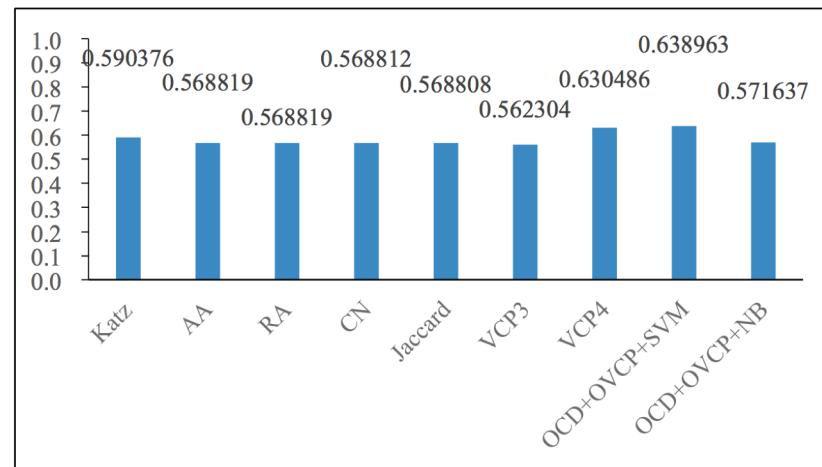
Link prediction based on OVCP structure



(a) DBLP 第一组数据集预测效果

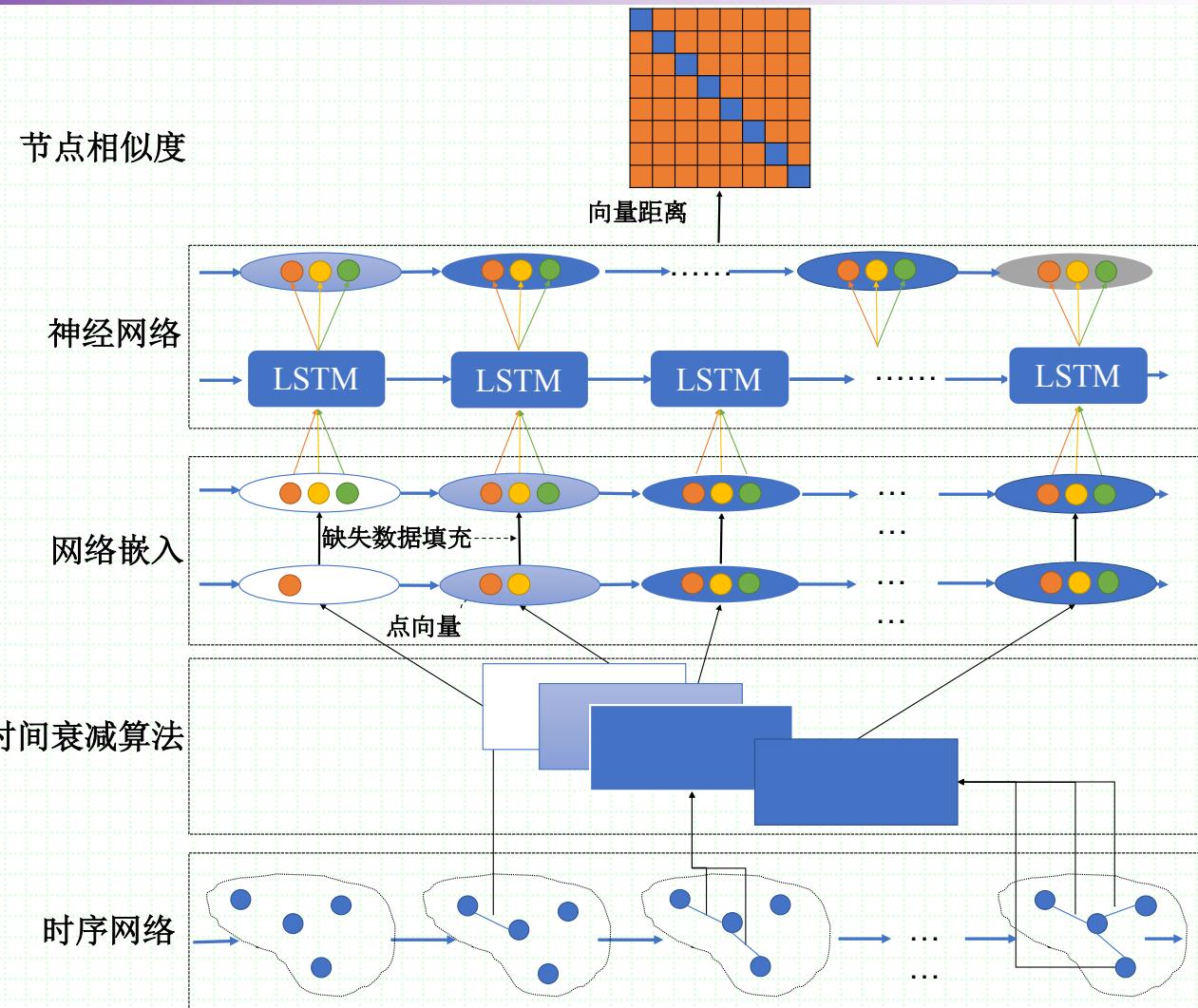


(c) Facebook Friendship 第一组数据集预测效果



(e) Facebook Wall 第一组数据集预测效果

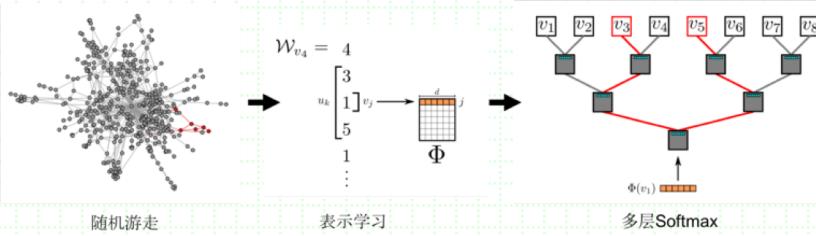
Temporal Link prediction based on DL



Temporal Link prediction based on DL

Network embedding

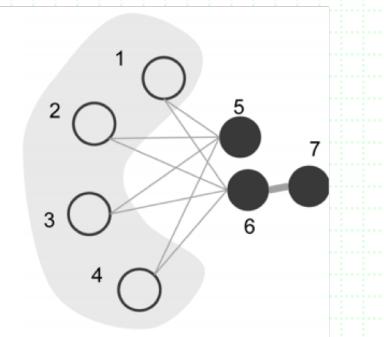
DeepWalk



$$Pr(u|v_j) = \frac{e^{\phi(v_j) \cdot \phi'(u)}}{\sum_{u \in W} e^{\phi(v_j) \cdot \phi'(u)}}$$

$$J = - \sum_{u \in C(v_j)} \log Pr(u|v_j)$$

LINE



$$p_1(v_i, v_j) = \frac{1}{1 + e^{-u_i^T \cdot u_j}}$$

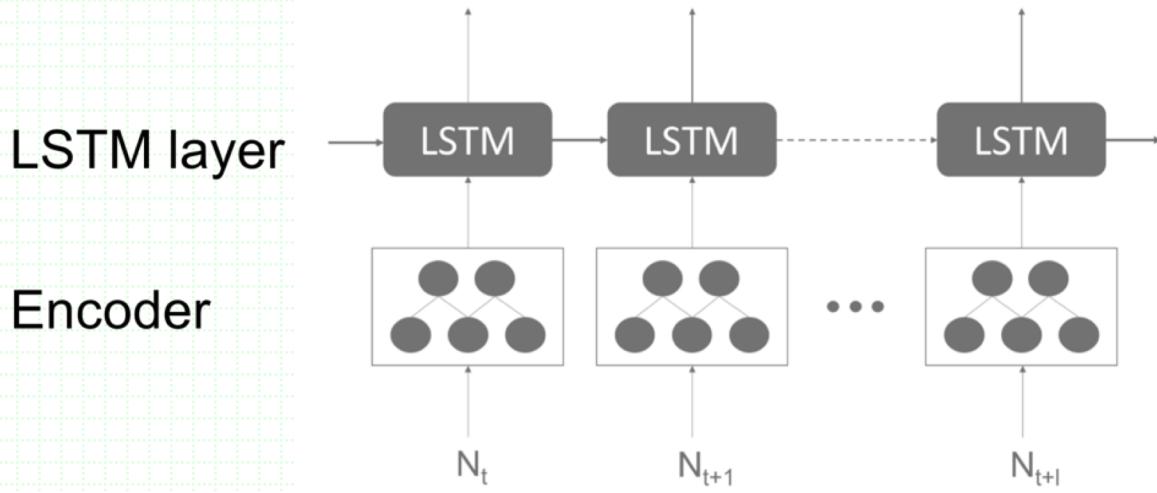
$$p_2(v_j|v_i) = \frac{e^{-u_j'^T \cdot u_i}}{\sum_{k=1}^{|V|} e^{-u_k'^T \cdot u_i}}$$

$$J = - \sum w_{ji} \log p_1(v_j, v_i) - \sum w_{ji} \log p_2(v_j|v_i)$$

只适用于无权网络

网络嵌入方法	上下文节点的来源	学习方法
DeepWalk	随机游走路径	带有分层 Softmax 的 Skip-gram
LINE	一阶相似度和二阶相似度	带有负采样的 Skip-gram

Temporal Link prediction based on DL



$$L_{t+l} = \|v'_{t+l+1} - v_{t+l+1}\|^2 = \|f(v_t, \dots, v_{t+l}) - v_{t+l+1}\|^2$$

当使用单层LSTM时，前向传播：

$$v'_{t+1} = o_t * \tanh(C_t) \quad o_t = \sigma_t(W_{LSTM}[v'_t, v_t] + b_o)$$

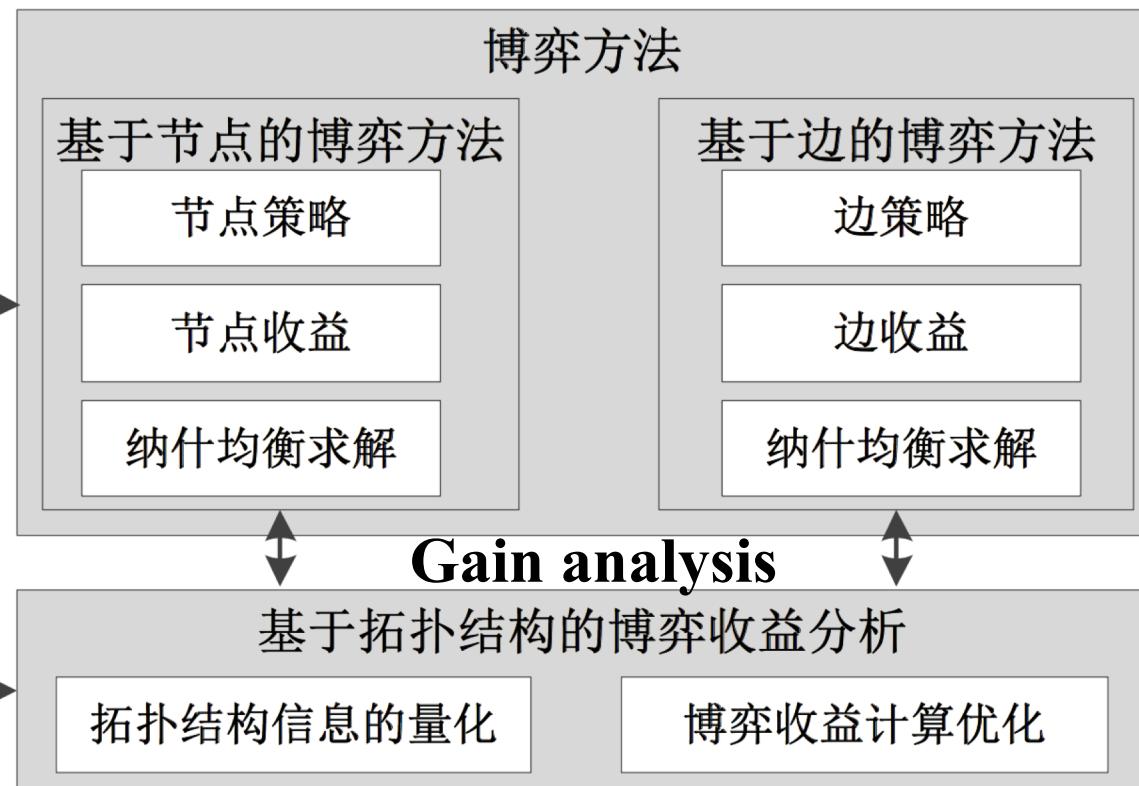
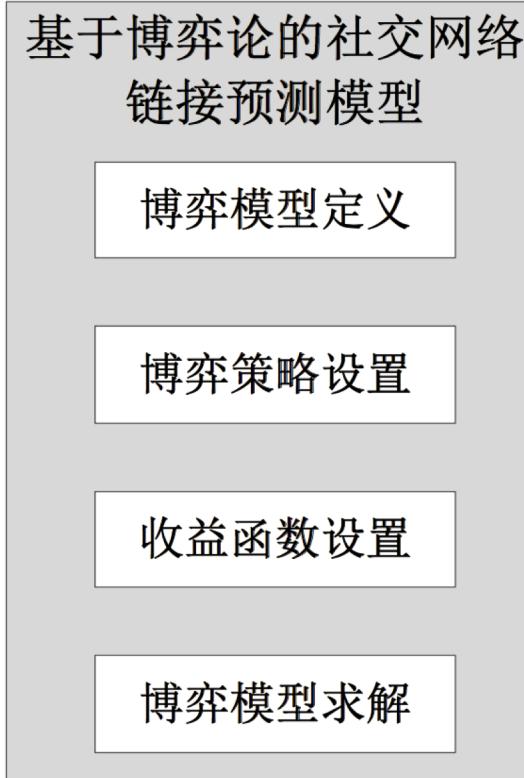
利用梯度下降法优化：

$$\frac{\partial L_t}{\partial W} = [2(v'_{t+1} - v_{t+1})] \cdot \left[\frac{\partial f(V * W + b)}{\partial W} \right]$$

Link Prediction based on Game Theory

Model

Two game methods



Link Prediction based on Game Theory

● Current issues

- ✓ 预测结果的可解释性不强 (interpretability)
 - 无法精确捕获社交行为
 - 无法有效捕获用户决策
 - ✓ 忽略了社交网络本身的“社会”特性 (social features)
 - 与社会理论成果结合不深
 - 没有考虑社交用户——人的理性和自私

Link Prediction based on Game Theory

● Our novel model

基于博弈的社交网络链接预测模型

✓ 博弈参与者 (Player) :

社交网络中的用户

✓ 博弈策略 (Strategy) :

如何构造博弈策略

✓ 博弈收益 (Playoff) :

如何使收益最大化

博弈求解：纳什均衡，时间复杂度NP-Hard

Link Prediction based on Game Theory

● Node based game: NGLP

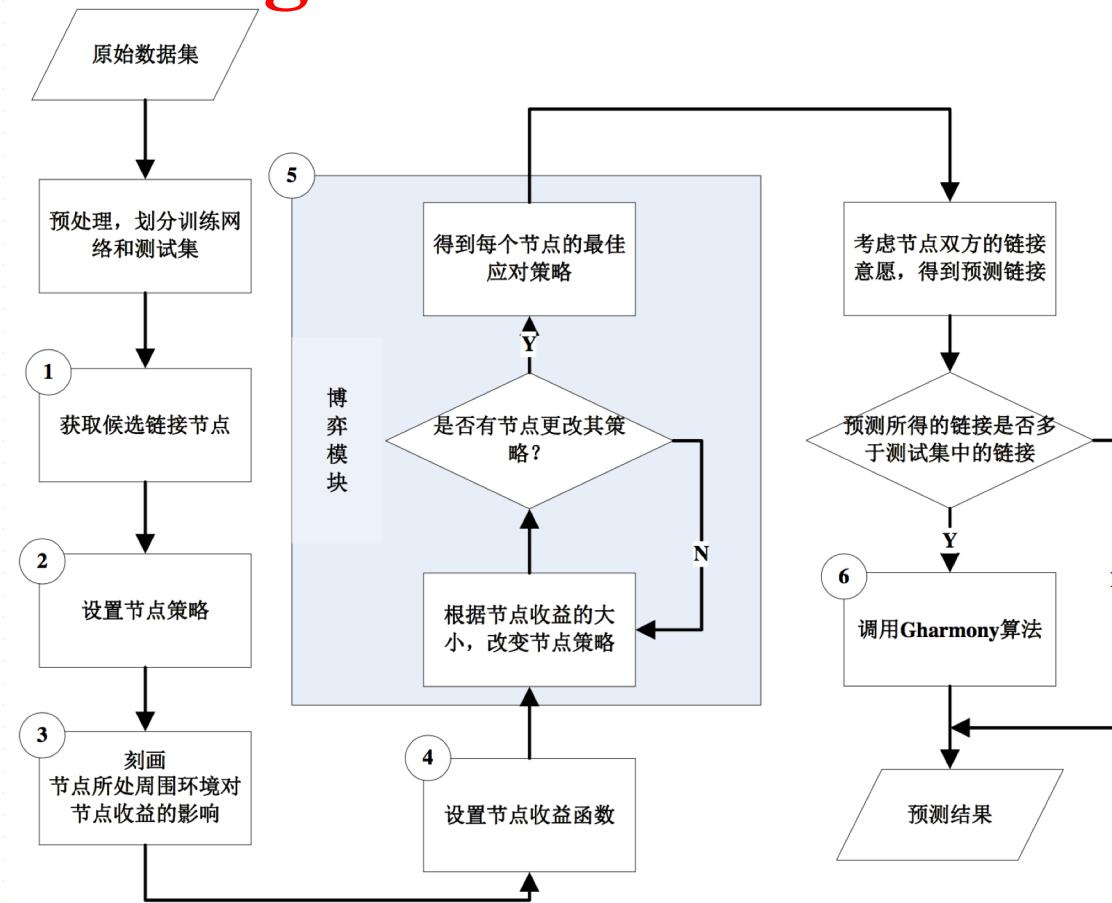


图 3-1NGLP 方法的流程图

Link Prediction based on Game Theory

● Node based game: NGLP

Most new links have small length

绝大部分新链接形成与距离2-3的节点之间

表3-1 社交网络中的路径距离对链接形成的影响

	$\delta = 2$	$\delta = 3$	$\delta = 4$	$\delta = 5$	$\delta = 6$	$\delta > 6$
Jazz	100%	0	0	0	0	0
Hamsterster	95.43%	2.52%	0.60%	0.06%	0.00%	1.39%
Email	71.43%	18.50%	5.86%	0.73%	0.00%	3.48%
Enron	90.96%	2.40%	0.20%	0.07%	0.01%	6.36%
Facebook	99.65%	0.22%	0.00%	0.01%	0.00%	0.01%
CA-GrQc	83.16%	2.48%	0.97%	0.00%	0.83%	11.25%
CA-HepTh	81.85%	3.76%	2.13%	1.20%	0.97%	10.09%
CA-HepPh	96.56%	1.23%	0.48%	0.22%	0.12%	1.39%
Twitter	99.08%	0.47%	0.00%	0.00%	0.00%	0.45%
Weibo	46.82%	16.28%	5.00%	0.00%	0.00%	31.90%

Link Prediction based on Game Theory

● Node based game: NGLP

Node strategy

节点策略：选择距离2-3的节点作为候选链接节点

定义 3-2(节点策略)：给定一个社交网络 $G(V,E)$ ，对于 $\forall i \in V$ ，节点 i 的策略 s_i 为节点 i 的链接候选节点的子集

$$s_i \subset \{t \mid t \in cNode(i)\} \quad (3-2)$$

此时节点策略方法的时间复杂度和空间复杂度均为 $O(2^{|cNode|})$ ，其中， $0 \leq |cNode| < N - \bar{d}$ ， $|cNode|$ 表示网络的平均链接候选节点个数。可以发现，当网络达到一定规模时，我们仍然很难在多项式时间和空间内给出所有的节点策略。

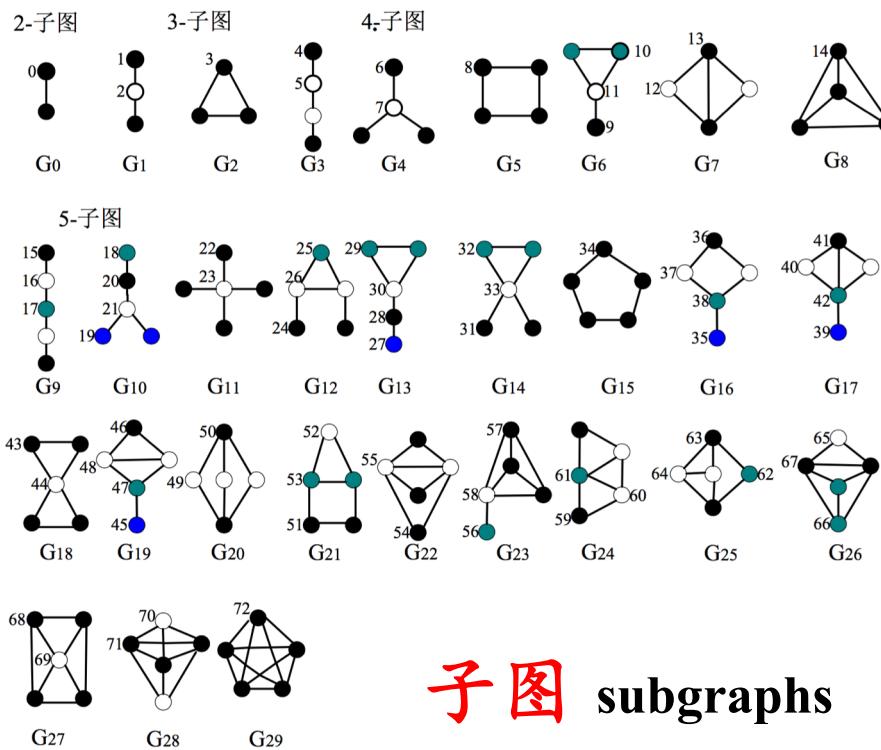
策略复杂度的解决：；去除重复策略，且每次只选择一个最佳策略

Link Prediction based on Game Theory

● Node based game: NGLP

Node gain
节点收益

节点收益：利用节点所处拓扑结构进行收益量化分析



子图 subgraphs

图 3-6 由 2,3,4,5 节点组成的所有子图(Graphlet)

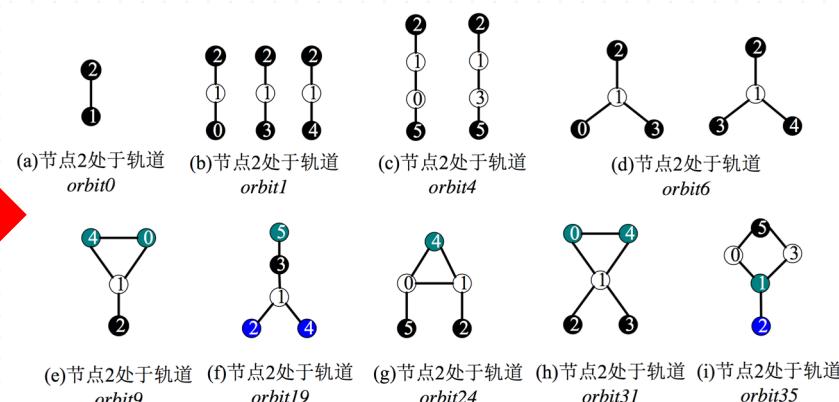
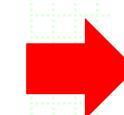


图 3-7 节点 2 与其他节点组成的所有子图

Orbit vector 轨道向量

Link Prediction based on Game Theory

● Node based game: NGLP

轨道向量：量化表示节点在特定拓扑结构下的收益情况

定义 3-12(权力向量)：给定社交网络 $G(V,E)$ ，以及任意轨道 $orbitx$ 的轨道占比 $ratio(x)$ 。我们称网络 $G(V,E)$ 的权力向量为各个轨道对节点收益的影响权重，

$$powerVector = (powerVector(0), powerVector(1), \dots, powerVector(k)) \quad (3-10)$$

其中 $powerVector(x)$ 为轨道 $orbitx$ 的权力参数，表明轨道 $orbitx$ 对节点收益的影响。 k 表示轨道的个数。

Link Prediction based on Game Theory

● Node based game: NGLP

收益函数 $Payoff_i(s) = GainNode_i(s) - LossNode_i(s)$

定义 3-14(福利函数—— $GainNode_i(s)$): 给定社交网络 $G(V,E)$ 以及网络的当前策略组合 $s=(s_i, s_{-i})$,
令节点 i 的福利为

特定拓扑结构中获得的福利

$$GainNode_i(s) = orbitVector(i) \times powerVector^T \quad (3-12)$$

维持链接需要付出成本 (例如脚踏两只船、一心二用)

定义 3-15(损失函数—— $LossNode_i(s)$): 给定社交网络 $G(V,E)$ 以及网络的当前策略组合 $s=(s_i, s_{-i})$,
节点 i 的损失开销为

$$LossNode_i(s) = d(i) \times c \quad (3-13)$$

Link Prediction based on Game Theory

● Node based game: NGLP

算法复杂度优化：

1、若节点 i 利用 Join 操作，往其最佳应对策略 s_i 中加入一个链接候选节点 j ，生成新策略 s'_i 。

那么，节点 i 选择新策略下所处的周围拓扑结构环境中将出现一条新边 $\langle i, j \rangle$

2、节点 i 利用 Leave 操作，从其最佳应对策略 s_i 中移除一个节点 j ，生成当前策略 s'_i 。那么，

节点 i 选择当前策略下所处的周围拓扑结构环境中的边 $\langle i, j \rangle$ 被移除。

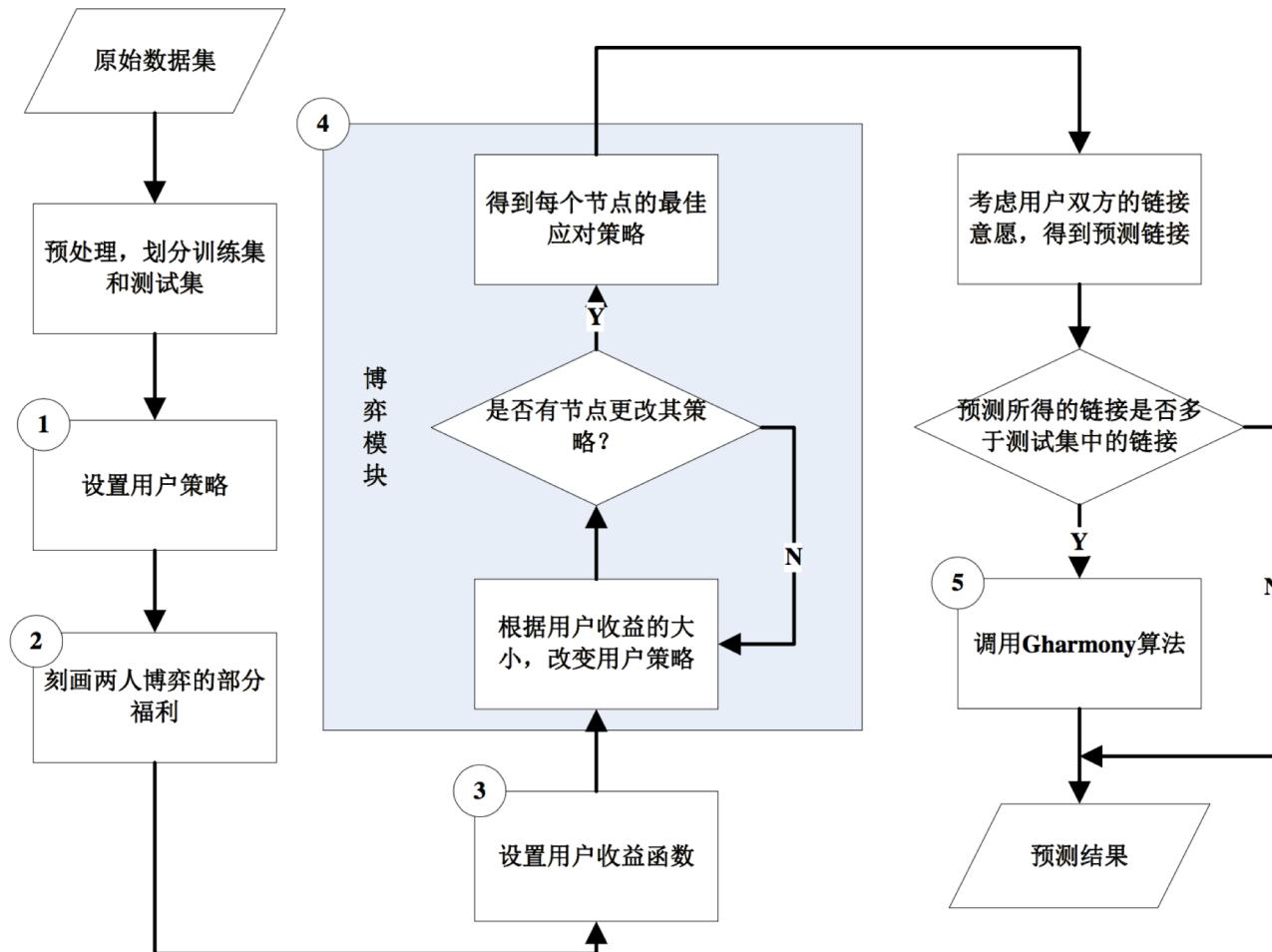
纳什均衡求解：

定理 3-1：在一定条件下，NLGP 博弈为潜能博弈，存在纯策略纳什均衡解。

博弈结果过滤：

Link Prediction based on Game Theory

Edge based game: NGLP



Link Prediction based on Game Theory

● 实验评估：数据集

表 5-1 各数据集的网络拓扑信息

数据集	节点数	边数	平均度	平均聚集系数	直径
Jazz	198	2742	27.7	0.6174	6.0
Hamsterster	2426	16631	13.71	0.5376	10
Email	1133	5451	9.62	0.2202	8
Enron	36692	183831	10.02	0.4970	11
CA-GrQc	5242	14496	5.53	0.5296	17
CA-HepTh	9877	25998	5.26	0.4714	17
CA-HepPh	12008	118521	19.74	0.6115	13
Facebook	4039	88234	43.69	0.6055	8
Twitter	81306	1768149	43.49	0.5653	7
Weibo	135623	276739	4.08	0.0958	4.0

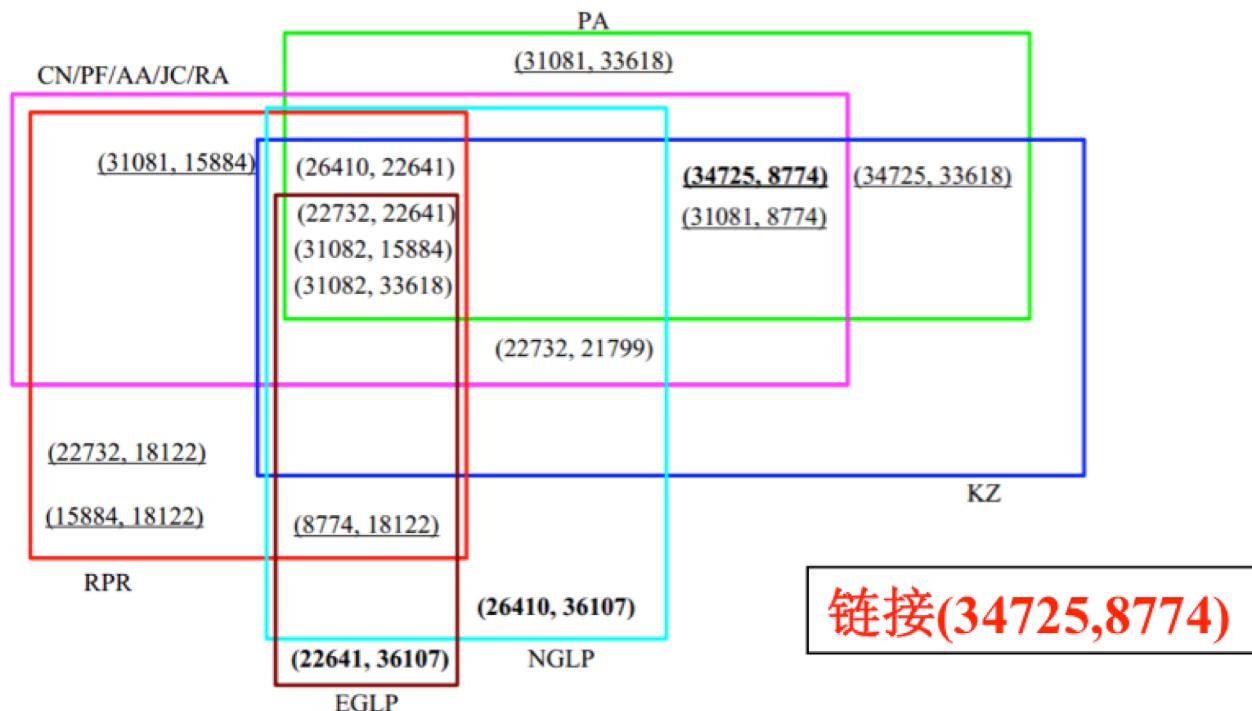
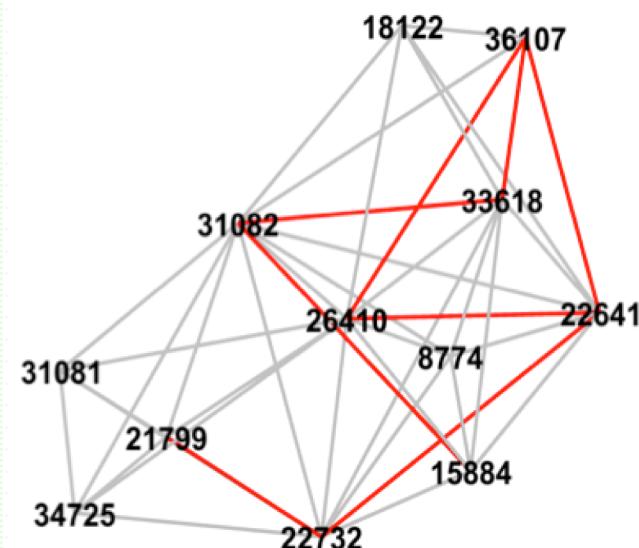
表 5-3 各个链接预测方法在多个数据集上的预测效果

		NGLP	EGLP	RA	PF	KZ	CN	AA	JC	PA	RPR	VCP
Jazz	F1	0.549	0.549	0.567	0.385	0.483	0.516	0.552	0.534	0.123	0.36	0.218
	AUC	0.59	0.591	0.549	0.279	0.471	0.498	0.552	0.509	0.082	0.269	0.175
Hamsterster	F1	0.332	0.325	0.343	0.297	0.198	0.202	0.257	0.232	0.046	0.243	0.127
	AUC	0.36	0.328	0.293	0.245	0.142	0.146	0.194	0.199	0.029	0.194	0.149
Email	F1	0.17	0.166	0.157	0.059	0.135	0.145	0.168	0.068	0.024	0.048	0.031
	AUC	0.239	0.161	0.101	0.03	0.106	0.114	0.124	0.033	0.012	0.026	0.048
Enron	F1	0.217	0.214	0.328	0.167	0.334	0.133	0.155	0.001	0.032	0.133	0.049
	AUC	0.22	0.231	0.271	0.139	0.897	0.096	0.114	0.011	0.019	0.106	0.057
CA-GrQc	F1	0.409	0.39	0.543	0.326	0.344	0.363	0.464	0.349	0.049	0.159	0.207
	AUC	0.474	0.436	0.515	0.252	0.315	0.327	0.449	0.26	0.037	0.125	0.331
CA-HepTh	F1	0.286	0.274	0.346	0.189	0.201	0.233	0.335	0.142	0.003	0.155	0.094
	AUC	0.306	0.291	0.3	0.137	0.164	0.184	0.284	0.099	0.002	0.127	0.172
CA-HepPh	F1	0.521	0.573	0.778	0.276	0.477	0.503	0.585	0.565	0.256	0.15	0.186
	AUC	0.608	0.657	0.805	0.216	0.465	0.492	0.577	0.533	0.233	0.099	0.197
Facebook	F1	0.352	0.364	0.452	0.186	0.315	0.317	0.334	0.323	0.055	0.17	0.141
	AUC	0.412	0.409	0.411	0.119	0.255	0.259	0.275	0.249	0.024	0.126	0.108
Twitter	F1	0.246	N/A	0.296	0.128	0.066	0.197	0.231	0.077	0.014	0.134	0.057
	AUC	0.282	N/A	0.233	0.136	0.044	0.14	0.169	0.032	0.008	0.09	0.046
Weibo	F1	0.043	N/A	0.027	0.012	0.05	0.049	0.04	0	0.015	0.012	0.001
	AUC	0.036	N/A	0.021	0.023	0.034	0.032	0.026	0	0.009	0.007	0.003

Link Prediction based on Game Theory

● Advantage 1:

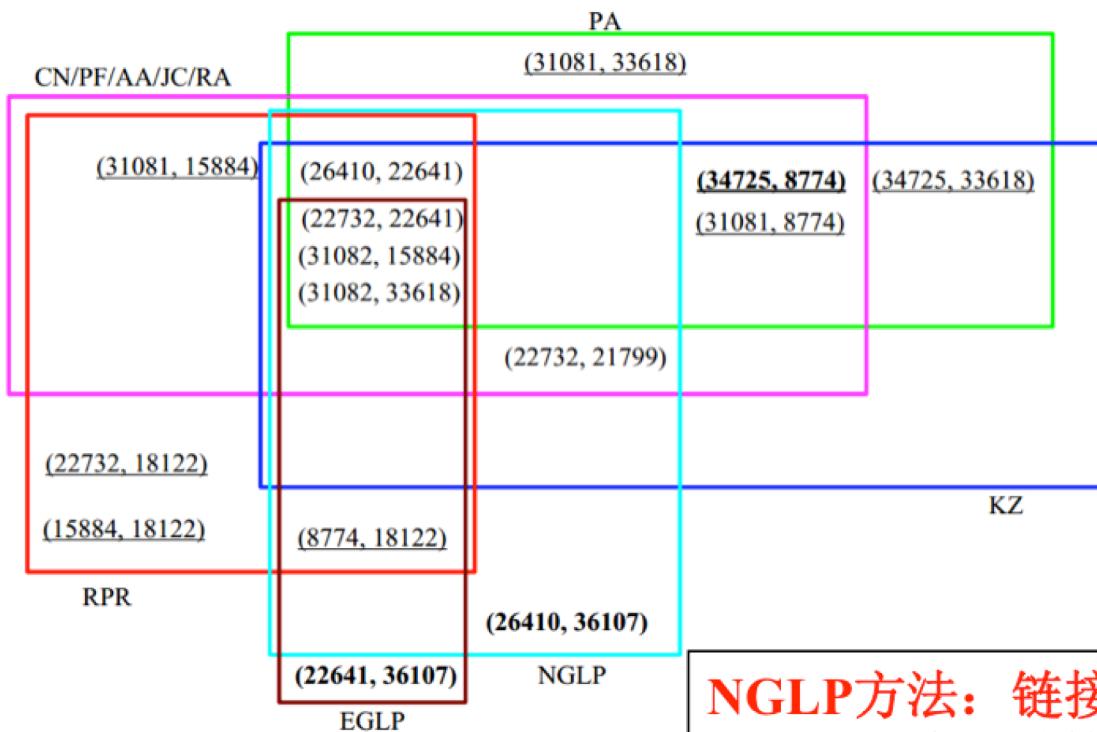
能过滤掉一些节点之间相似度较高的错误的链接



Link Prediction based on Game Theory

● Advantage 2:

能够预测到其他方法预测不到的链接。



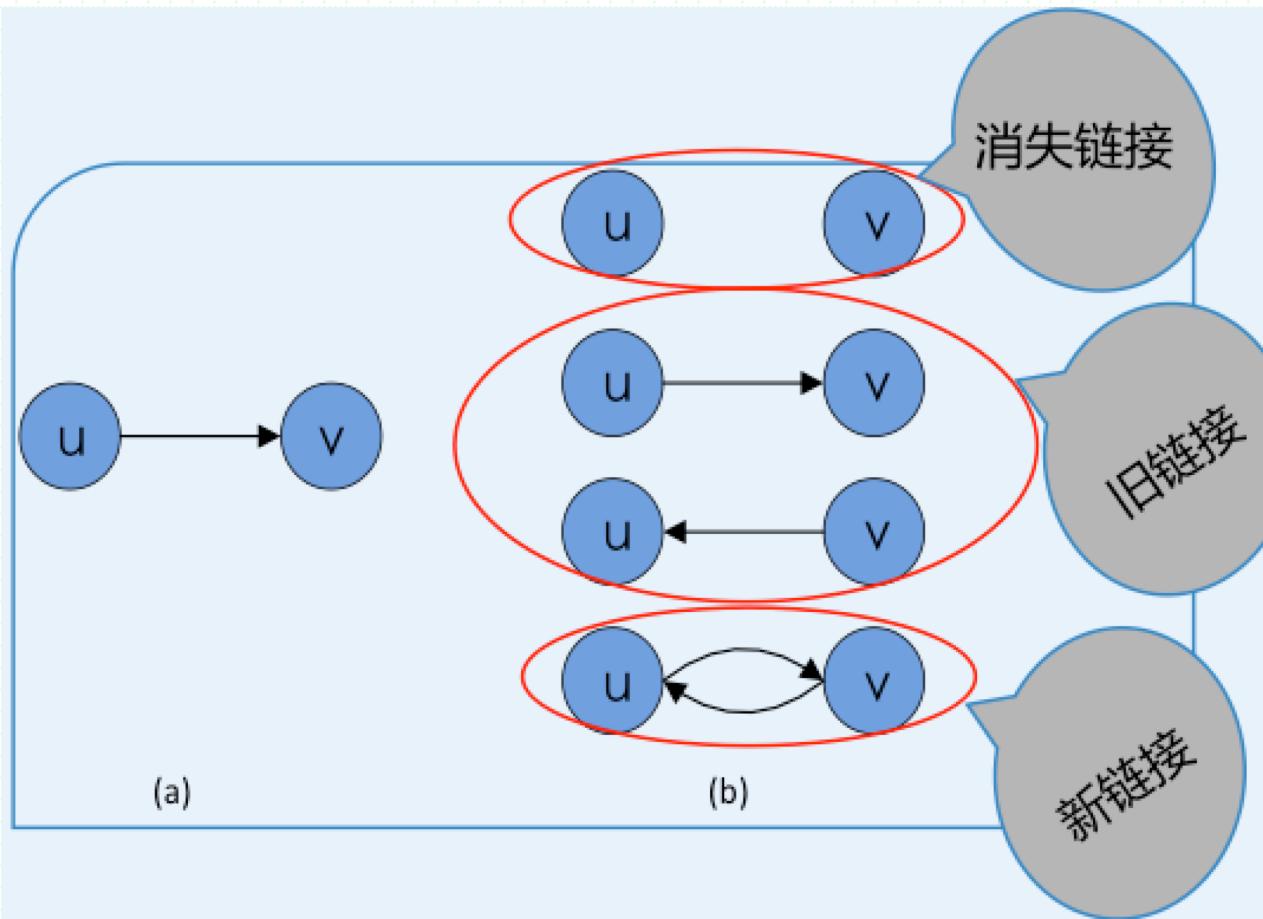
链接(26410,22641)

新链接的形成
对网络节点策
略选择的影响

NGLP方法：链接(26410,36107)
EGLP方法：链接(22641,36107)

The predictability of disappearing links

● Problem



The predictability of disappearing links

● 消失链接现象实证分析 empirical analysis

链接比例分布

各符号表示含义：

- E_1 和 E_2 ----子网络 G_1 和 G_2 共有节点链接集合
- S_{Dis} ----消失链接集合
- S_{new} ----新链接集合
- S_{Old} ----旧链接集合

消失链接比例 (*ratio_dis*)

$$ratio_dis = \frac{|S_{Dis}|}{|E_1|}$$

新链接比例 (*ratio_new*)

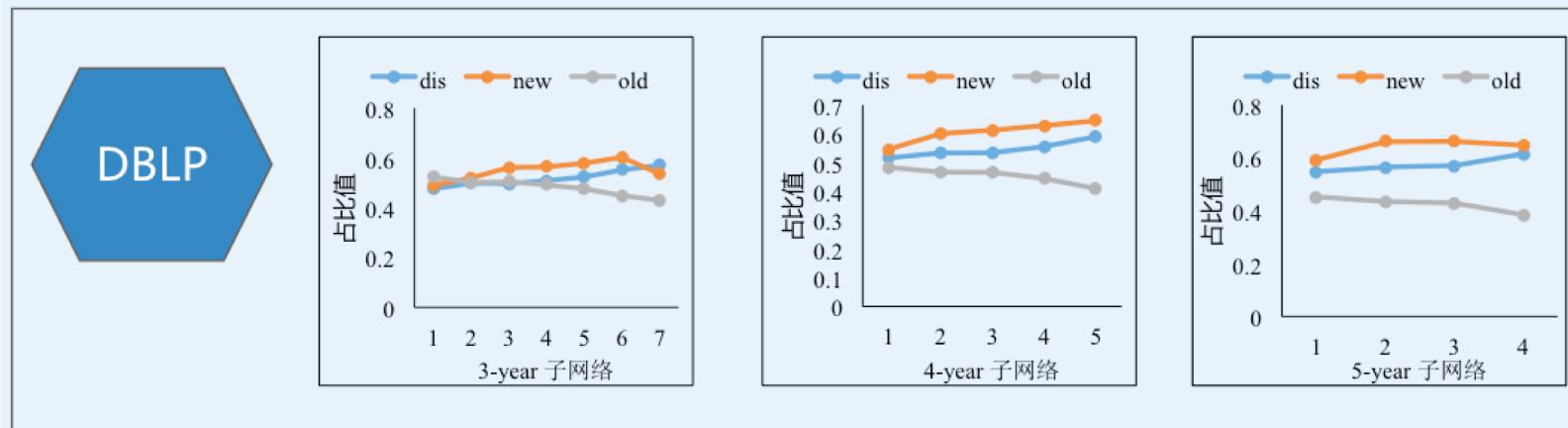
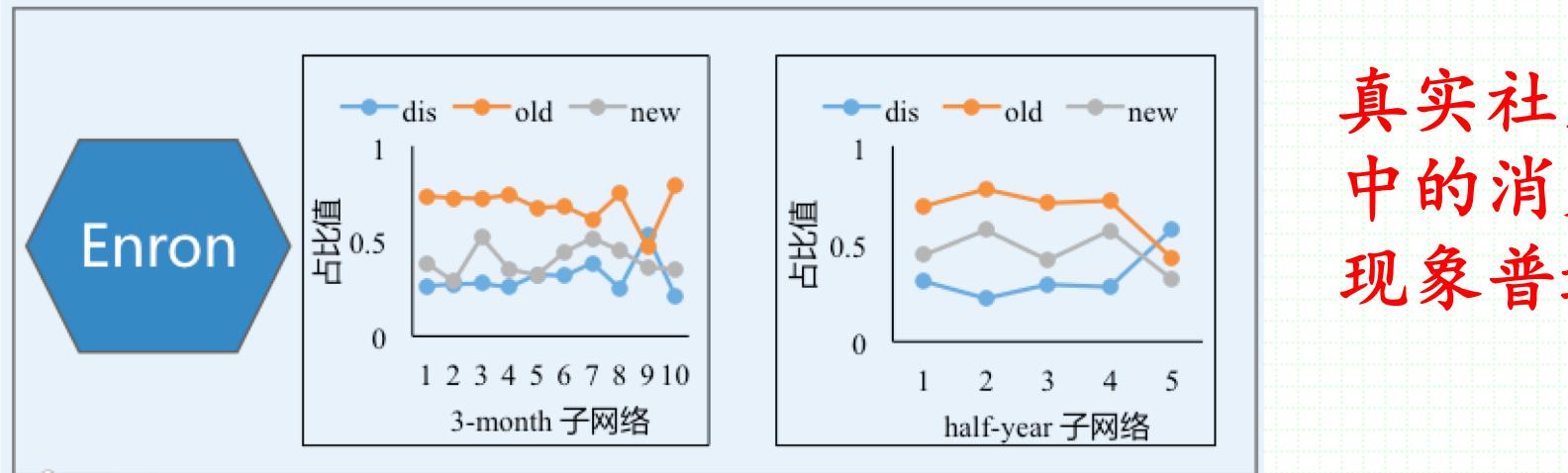
$$ratio_new = \frac{|S_{New}|}{|E_2|}$$

继续存在链接比例 (*ratio_old*)

$$ratio_old = \frac{|S_{Old}|}{|E_1|}$$

The predictability of disappearing links

● 消失链接现象实证分析 empirical analysis



The predictability of disappearing links

● 消失链接现象实证分析 empirical analysis

互惠性

嵌入性

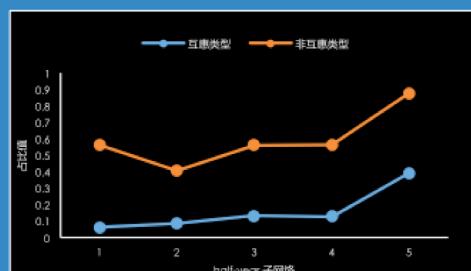
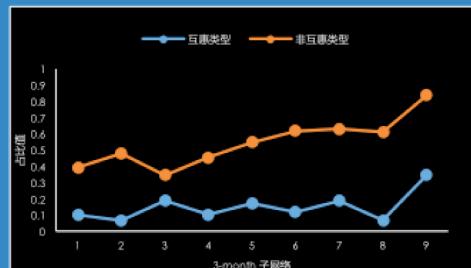
互动频率

时间属性

Enron实验结果：满足互
惠性质的链接消失的可能
性明显小于不满足互惠性
质的情况

互惠性：有向网络，具体表现为两节点之间链接的双向性。

Enron数据集



The predictability of disappearing links

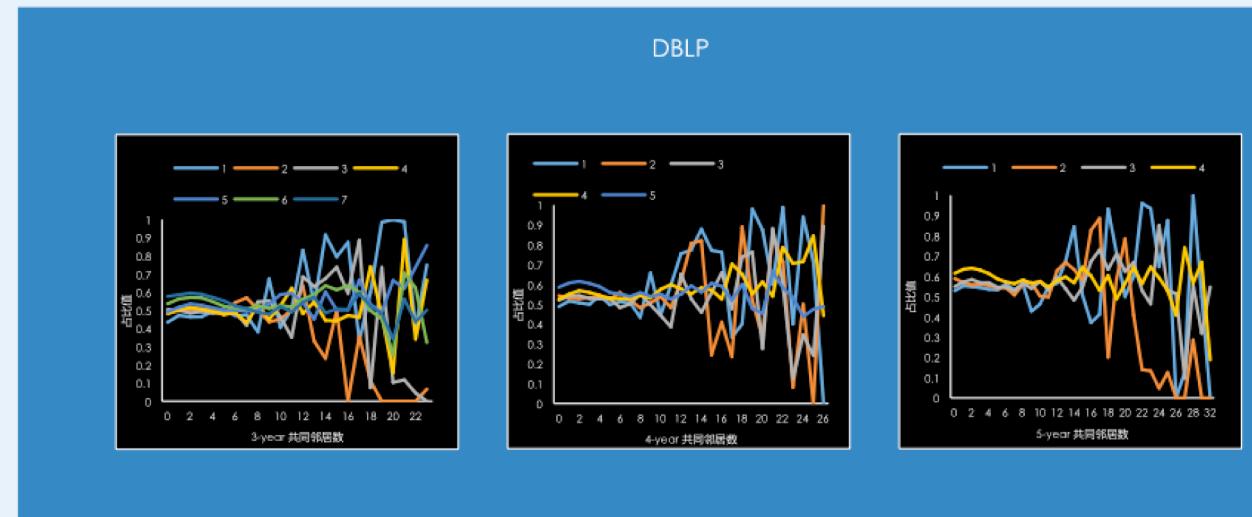
● 消失链接现象实证分析 empirical analysis

互惠性

嵌入性

互动频率

时间属性



嵌入性：定义为两节点的共同邻居数

DBLP: 嵌入性链接的消失预测作用不明显

The predictability of disappearing links

● 消失链接现象实证分析 empirical analysis

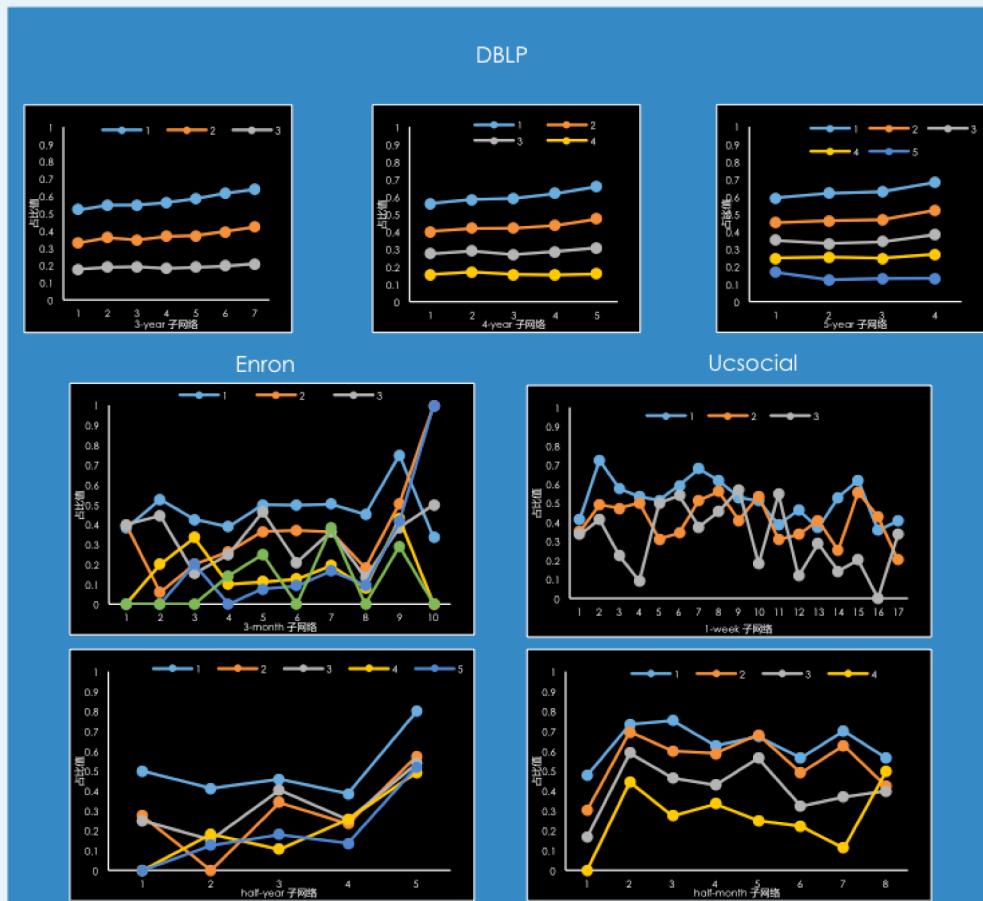
互惠性

嵌入性

互动频率

时间属性

互动频率：相互交流的次数衡量



实验结果：
链接互动频率的增加可以减小消失的概率。

The predictability of disappearing links

● 消失链接现象实证分析 empirical analysis

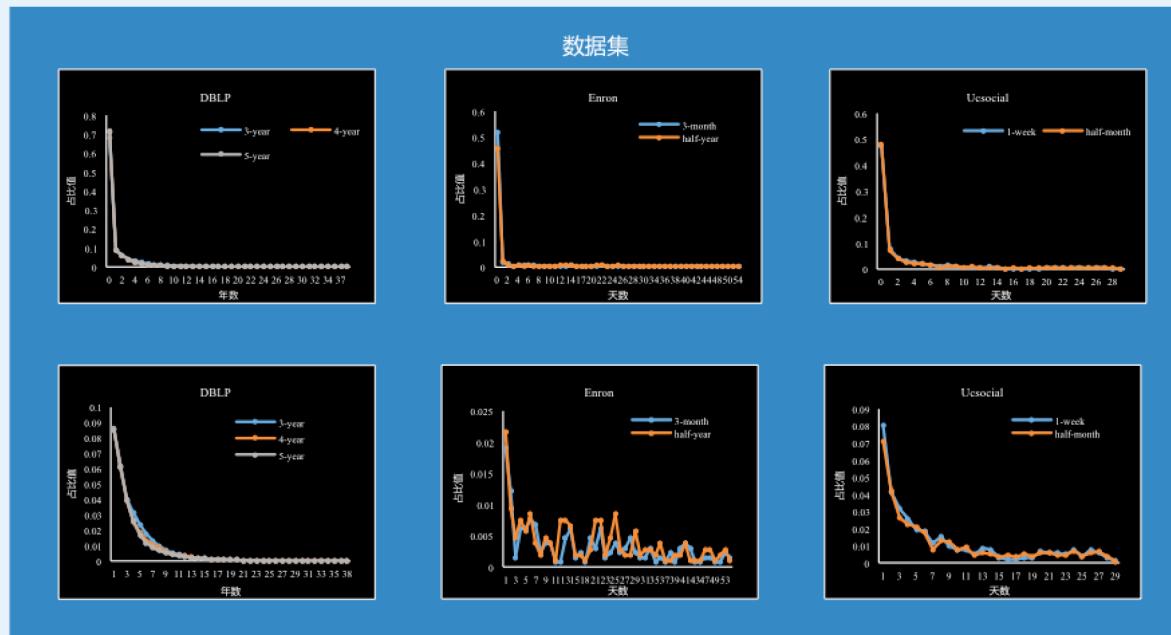
互惠性

嵌入性

互动频率

时间属性

时间属性：链接生存时间



实验结果表明：

网络中的时序信息对于实验结果有着显著的影响，特别是当链接的生存时间小于某一阈值时，其消失概率随着生存时间的增加显著减小；当其生存时间大于某一阈值时，链接的消失概率同生存时间不在具有显著的相关性

The predictability of disappearing links

● 消失链接预测方法 Prediction methods 无向无权网络

同质性

优先连接

三度影响力

社会比较理论

子图转移概率矩阵

链接联系强度

基于同质性理论的相似性指标

□ CN指标 : $s_{xy} = |\Gamma(x) \cap \Gamma(y)|$ □ Salton指标 : $s_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k_x k_y}}$ □ Jaccard指标 : $s_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$ □ Sorenson指标 : $s_{xy} = \frac{2 \times |\Gamma(x) \cap \Gamma(y)|}{k_x + k_y}$ □ HPI指标 : $s_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\min\{k_x, k_y\}}$ □ HDI指标 : $s_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\max\{k_x, k_y\}}$ □ LHN-I指标 : $s_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{k_x k_y}$ □ AA指标 : $s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z}$ □ RA指标 : $s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}$

各符号表示含义 :

□ $\Gamma(x)$ ---节点 v_x 邻居节点集合□ k_x ----节点 v_x 度数值二者关系 : $k_x = |\Gamma(x)|$

The predictability of disappearing links

● 消失链接预测方法 Prediction methods 无向有权网络

无向无权网络

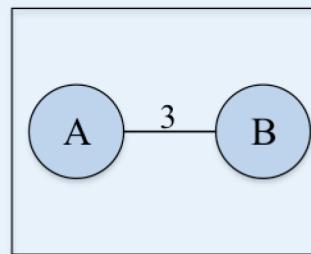
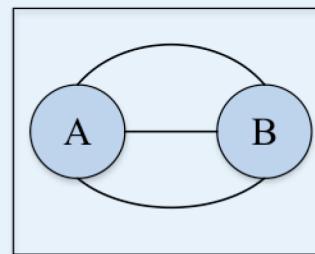
无向有权网络

时序无向网络

极大连通图

综合实验评估

同质性



权重示例

Weight指标 : $s_{xy} = w_{xy}$

WCN指标 : $s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w_{xz} + w_{yz}}{2}$

$s_{xy} = \frac{\sum_{z \in \Gamma(x) \cap \Gamma(y)} w_{xz} + w_{yz}}{\sum_{s \in \Gamma(x)} w_{xs} \sum_{t \in \Gamma(y)} w_{yt}}$

WJaccard指标 :

WAA指标 : $s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w_{xz} + w_{yz}}{2 \log(|\Gamma(z)|)}$

WRA指标 : $s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w_{xz} + w_{yz}}{2 * |\Gamma(z)|}$

$$s_{xy} = \frac{w_{xy}}{\sum_{s \in \Gamma(x)} w_{xs} \sum_{t \in \Gamma(y)} w_{yt}}$$

WRPA指标 :

链接联系强度

TS-B和TS-V预测方法

The predictability of disappearing links

● 消失链接预测方法 Prediction methods 无向时序网络

无向无权网络

无向有权网络

时序无向网络

极大连通图

综合实验评估

预测指标

时序示例

- **FirstLast (FL) 指标:** 主要衡量的是链接存在的时间长度；本文认为，该链接的存在时间越长，则该链接的稳定性越高；其中 $First(x,y)$ 表示的是链接 (x,y) 首次出现的时间， $Last(x,y)$ 表示的是链接 (x,y) 最新出现的时间。定义为
$$T_{xy} = Last(x,y) - First(x,y)$$
- **ActiveMulti (AM) 指标:** 主要通过链接两端节点的最近活跃时间去度量链接的活跃程度；本文认为该链接两端的节点越活跃，则该链接的稳定性越高；其中， T 表示当前时间， $last(x)$ 表示的是节点 v_x 的最近活跃时间。为了使得分母不为0，故加1处理。定义为
$$T_{xy} = \frac{1}{(T - last(x) + 1)(T - last(y) + 1)}$$
- **TimeList (TL) 指标:** 主要是通过引入影响因子去衡量不同的时序信息对整个链接的存在稳定性的作用；本文认为链接两端节点的合作时间越近，则对当前链接的稳定性影响程度越大；其中 $\alpha \in [0,1]$ 为影响因子， $[t_1, t_2, L]$ 为网络中的当前链接的时序信息列表。定义为
$$T_{xy} = \sum_{t \in [t_1, t_2, L]} (1 - \alpha)^{(T-t)}$$



The predictability of disappearing links

Factor analysis ● 消失链接预测因素的相关性分析

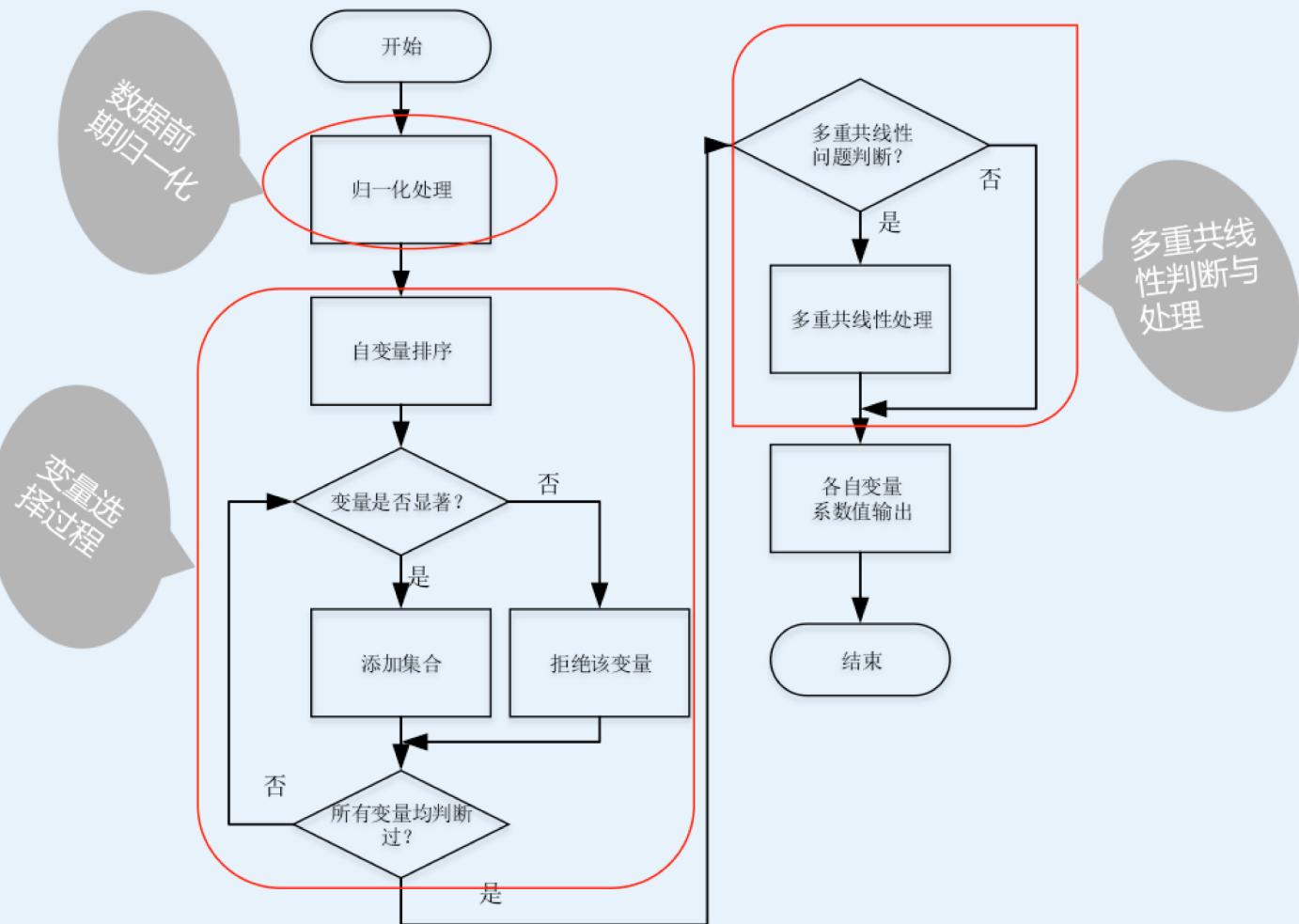
拓扑结构度量	符号/定义	含义
节点数	n	描述网络中节点总数
链接数	m	描述网络中链接总数
最大度	k_{\max}	描述所有节点中度数最大值
平均度	$\langle k \rangle$	描述网络中节点的平均链接数
密度	D	衡量网络稀疏程度
聚集系数	C	描述节点的平均聚集程度
网络效率	E	描述网络连通性能
楔形数目	s	描述2-星数目
爪形数目	z	描述3-星数目
三角数目	t	描述三元闭包的数目
幂律	λ	节点度分布幂律值
相关链接分布熵	H_{er}	网络中度分布的等价程度
同配系数	r	描述链接形成的偏好性

The predictability of disappearing links

Factor analysis

● 消失链接预测因素的相关性分析

多元线性相关性分析过程



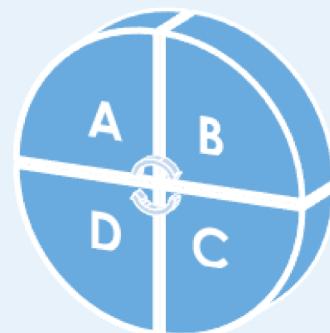
The predictability of disappearing links

● 消失链接预测因素的相关性分析

Factor analysis

主要工作一：无向无权网络

- 基于无向网络中各拓扑结构和预测指标效果值之间的多元相关性分析结果，经过数据归一化、变量选择和多重共线性处理，得到最后结果，分析并得出结论。



主要工作四：综合结果

- 综观所有结果，选择显著影响的拓扑特征

主要工作二：最大连通图

- 基于最大连通图中各拓扑结构和预测指标效果值之间的多元相关性分析结果，经过数据归一化、变量选择和多重共线性处理，得到最后结果，分析并得出结论。

主要工作三：有向网络

- 基于有向网络中各拓扑结构和预测指标效果值之间的多元相关性分析结果，经过数据归一化、变量选择和多重共线性处理，得到最后结果，分析并得出结论。

The predictability of disappearing links

● 消失链接预测因素的相关性分析

Factor analysis

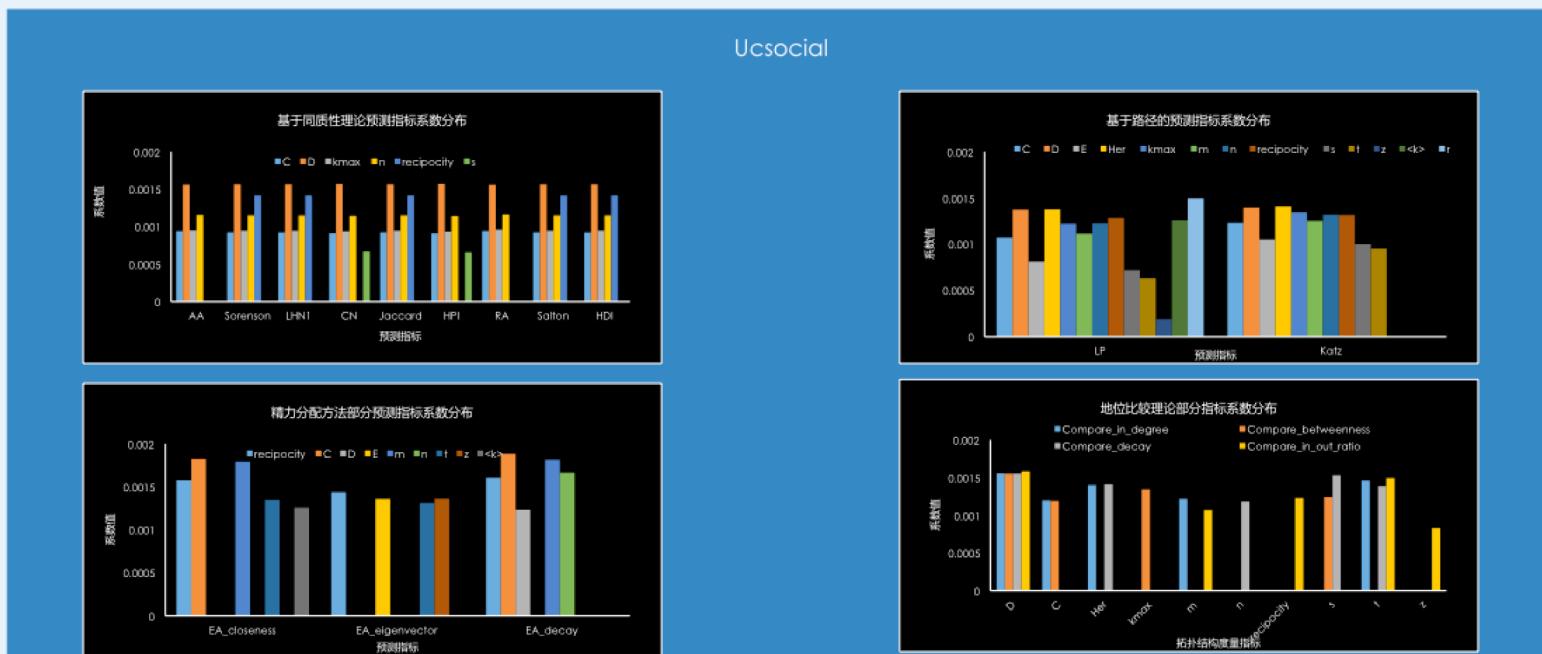
消失链接现象

链接预测方法

相关性分析

网络演化模型

实验结果表明



- 有向无权Ucsocial网络中经岭回归方法处理后可知C和D度量方式对各指标综合影响最深。

Evolution model for disappearing links

描述

前提：初始状态为 n_0 个节点构成的完全网络 G_0 ，网络 G_t 在下个时间步 $t+1$ ，将以不同的概率发生下列的三种事件，且每个时间步有且仅有一种事件发生。

1) 新节点添加：

以概率 p_a 发生该事件。事件描述为：将一个新节点添加到网络中，以偏好概率

$$P_{t+1}[u] = \frac{d_t(u)}{\sum_{w \in V_t} d_t(w)} = \frac{d_t(u)}{2m_t}$$

同网络中已有节点之间建立 C_1 条链接。

Evolution model for disappearing links

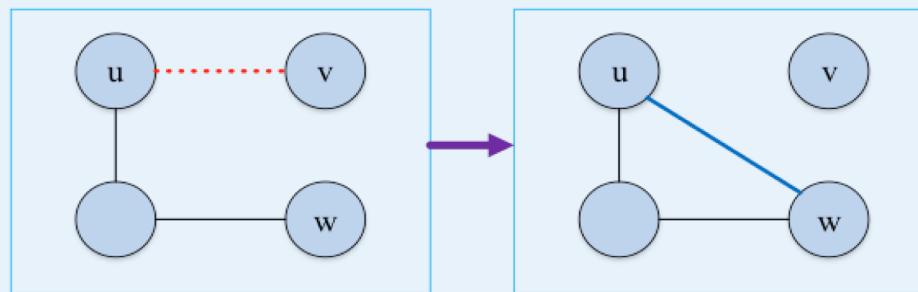
描述

2) 链接重连:

以概率 p_b 发生该事件。事件描述为：以反偏好概率

$$P_{t+1}[u] = \frac{n_t - d_t(u)}{n_t^2 - 2m_t}$$

选择节点，同时在节点的邻居节点中以反偏好概率选择节点，删除二者之间的连线；再以偏好概率在网络中选择节点同该节点建立链接。此事件重复发生 c_2 次。



3) 旧节点消失:

以概率 $p_c = 1 - p_a - p_b$ 发生该事件。事件描述为：以反偏好概率选择节点删除，同时将该节点相关链接删除。

Evolution model for disappearing links

证明目标：LWBA模型中节点数、链接数、平均节点度等特征如何变化？LWBA模型中度分布是否符合幂律分布？

证明
思路



节点数

- 同新节点添加事件；
旧节点消失事件有关

- 表示为：
$$n_t = (p_a - p_c)t$$



链接数

- 同新节点添加事件；
旧节点消失事件有关

- 表示为：
$$E[m_t] = \frac{c_1 p_a}{1 + \left(\frac{2p_c}{p_a - p_c} \right)} t$$



平均节点度

- 同节点数和链接数相
关，表达式：

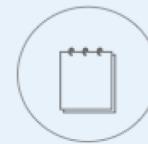
$$\bar{d}_t = \frac{2m_t}{N_t} = \frac{\left[\frac{2c_1 p_a}{1 + \left(\frac{2p_c}{p_a - p_c} \right)} t \right]}{(p_a - p_c)t} = \frac{2c_1 p_a}{p_a + p_c}$$

Evolution model for disappearing links

LWBA模型中度分布证明思路： $P[k] = \lim_{t \rightarrow +\infty} \frac{N_{k,t}}{N_t}$



新节点添加



链接重连



旧节点消失

- 新节点的影响：分为 $k=c/1$ 和 $k \neq c/1$
- 已有节点的影响：节点度数为 k 和 $k-1$ 的情况

- 链接消失相关节点影响：节点度数为 k 和 $k+1$ 的情况
- 链接新生相关节点影响：节点度数为 k 和 $k-1$ 的情况

- 消失节点的影响：节点度数为 k
- 消失节点相关邻居节点：节点度数为 k 和 $k+1$ 的情况

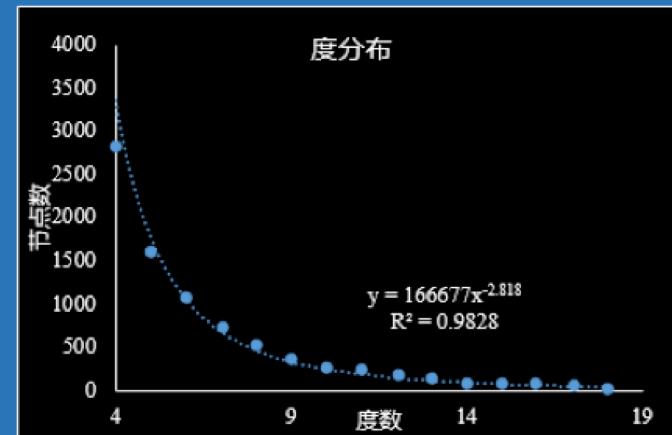
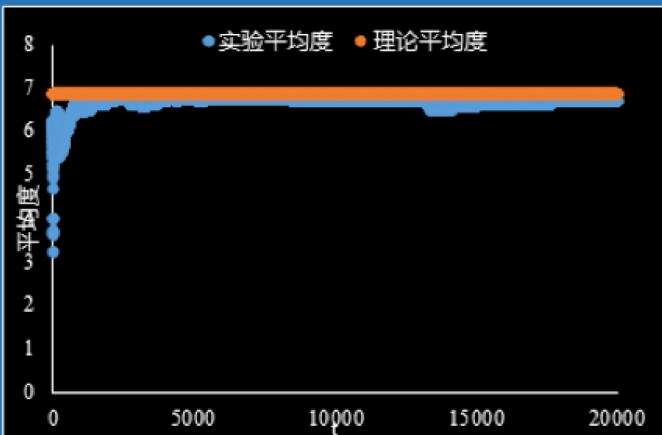
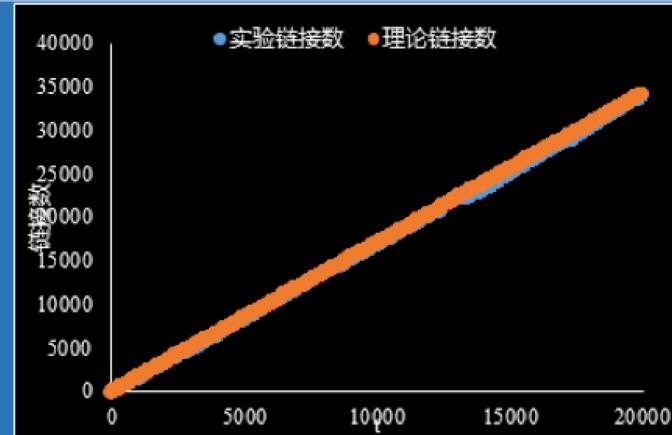
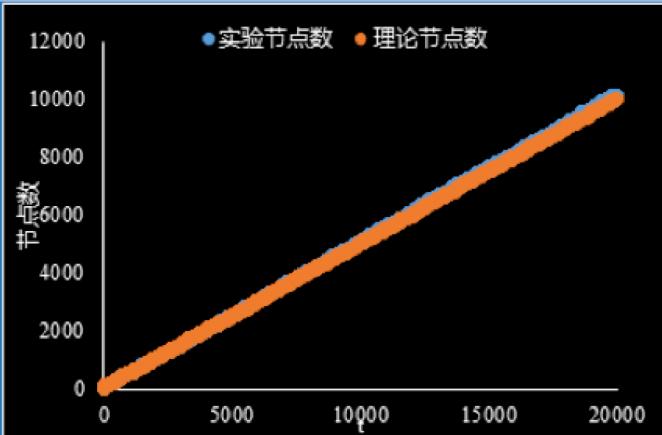
$k-1$ 的情况

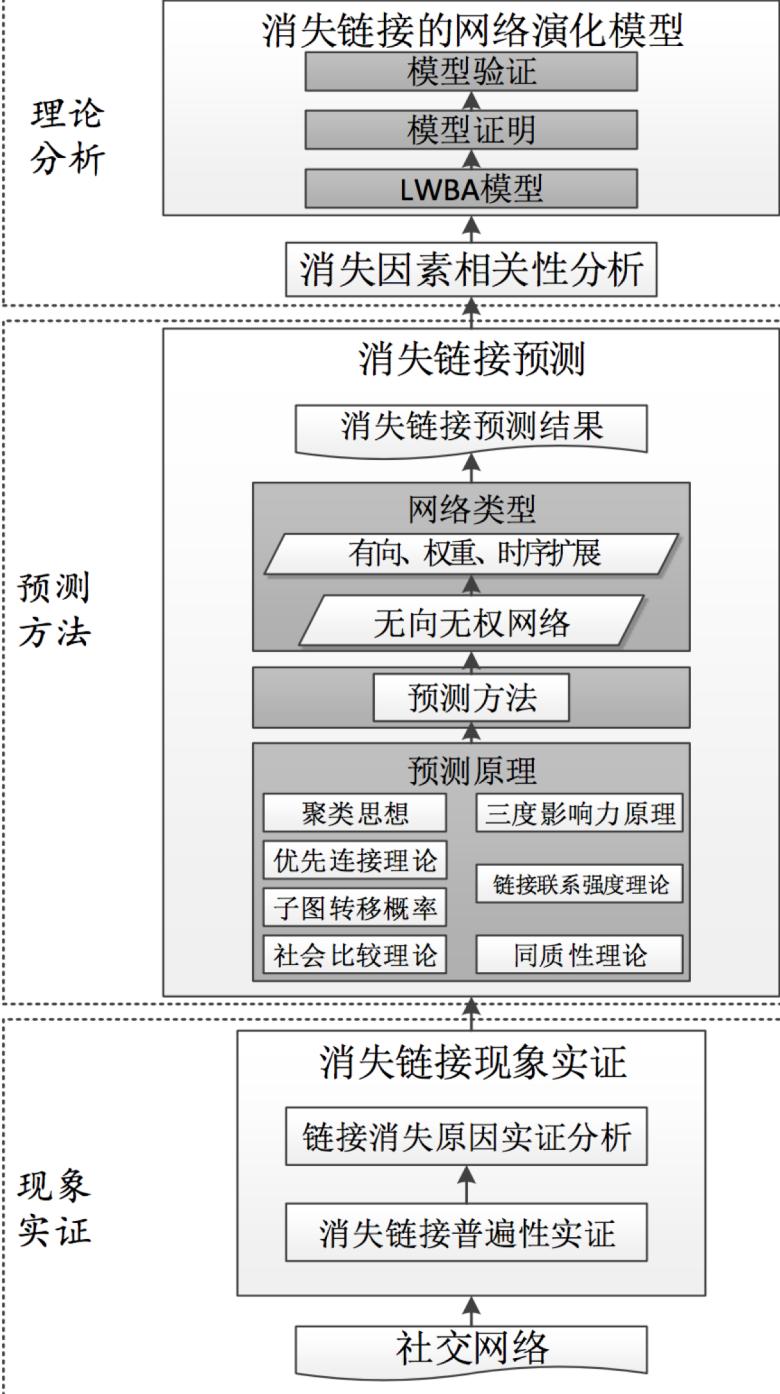
度分布结果为： $p[k]: k^{-1 - \left[\frac{-2c_1 p_a p_a}{2c_1 p_a p_c - (p_a c_1 + p_b c_2)(p_a + p_c)} \right]}$

需要注意的是当 $p_a = 1, p_b = p_c = 0$ 时，整个模型退化为BA模型

Evolution model for disappearing links

➤ 正常情况 case2:Pa=0.6,pb=0.3,pc=0.1: 度分布-2.78





Some KGCode members in the above work



धन्यवाद

Hindi

多謝

Traditional Chinese

ขอบคุณ

Thai

Спасибо

Russian

شُكْرًا

Arabic

Thank You

English

Gracias

Spanish

Obrigado

Brazilian Portuguese

Danke

German

Merci

French

நன்றி

Tamil

ありがとうございました

Japanese

감사합니다

Korean

Email: pwang@seu.edu.cn