On Link Prediction

Linyuan Lü

E-mail: babyann519@hotmail.com

linyuan.lv@gmail.com

Outline

- What is *Link Prediction* problem?
- Why it is significant?
- How it relates to personalized recommendation?
- Representative Algorithms on Link prediction
- Similarity-based link prediction algorithms
- Our current works
- Outlook

What is Link Prediction?

- Estimating the likelihood of the existence of a link between two nodes, based on the observed topology.
- Prediction of existed yet unknown links for sampling networks, such as *food webs*, *protein-protein interaction networks* and *metabolic networks*.
- Prediction of future links for evolving networks, like *on-line friendship networks*.

Why it is significant?

- Theoretical insights

  Accurate prediction indeed gives evidence to some underlying mechanism that drives the network evolution.

- Practical value
  - □ Reduce experimental costs

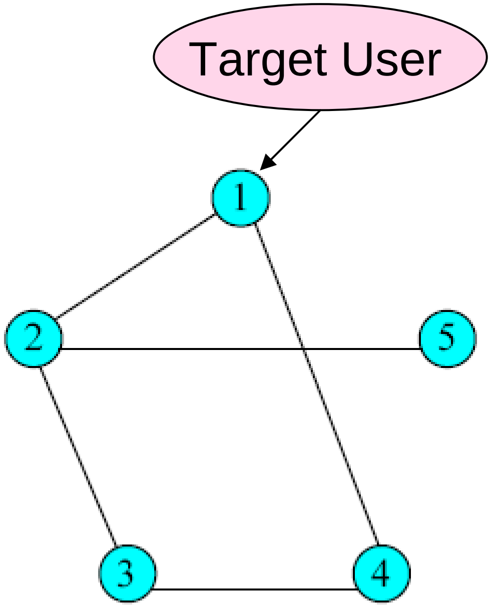
    Discovery of links/interactions of biological networks costs much. Instead of blindly checking all possible links, to predict in advance and focus on the most likely existing links can sharply reduce costs if the predictions are accurate enough.
  - □ Classification in labeled networks

    Predict the label of unlabeled node.
  - □ On-line recommendations

    Recommend new friends to the users in web society.

How it relates to personalized recommendation?

- Link prediction algorithm sorts all the nonexistent links, the top ones are most likely to exist.
- Personalized recommendation can be made by picking up the relevant links of the target user.



Rank of LP

1	U2-U4
2	U1-U3
3	U4-U5
4	U1-U5
5	U3-U5

PR for U1 is U3

PR can be considered as a sub-problem of LP!

Representative Algorithms on Link prediction

- Markov Chains
  - *R. R. Sarukkai, Computer Networks, 33, 377 (2000)*
  - *J. Zhu, J. Hong and J.-G. Hughes, Proceedings of the thirteenth ACM conference on Hypertext and hypermedia (2002)*
- Machine Learning
  - *A. Popescul and L. Ungar, in Workshop on Learning Statistical Models from Relational Data, ACM Press, New York, 2003.*
  - *K. Yu, W. Chu, S. Yu, V. Tresp and Z. Xu, Stochastic Relational Models for Discriminative Link Prediction, in Advance in Neural Information Processing Systems 19, MIT Press, Cambridge, MA, 2007.*
- Collaborative Filtering
  - *Z. Huang, X. Li, H. Chen, Link prediction approach to collaborative filtering, In Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, ACM Press, New York, 2005.*
- To Predict Based on Node Similarities
  - *D. Liben-Nowell and J. Kleinberg, J. Am. Soc. Inform. Sci. &. Technol. 58, 1019, 2007.*
- To Predict Based on Prior Knowledge
  - *A. Clauset, C. Moore and M. E. J. Newman, Nature 453, 98 (2008)*
  - *S. Redner, Nature 453, 47-48 (1 May 2008)*

Similarity-based link prediction algorithms

- Similarity Indices
  - Attributes
  - Network structure based
    - Node-dependent *vs.* Path-dependent
    - Local information *vs.* Global information
    - Parameter free *vs.* Parameter-dependent

- Method
  - Consider an unweighted undirected network G(V, E), V is the set of nodes, E is the set of links.
  - Calculate similarities corresponding to all nonexistent links.
  - Sort all the nonexistent links in descending order according to their similarities, the top ones are most likely to exist.
- Data
  - Training set as known information
  - Probe set for testing

Metric

- AUC (area under the receiver operating characteristic curve)
  - □ Probability that a randomly chosen link in the probe set (PL) has higher similarity than a randomly chosen nonexistent link (NL).
  - □ Independently testing *n* times, and *n1* PL>NL, *n2* PL=NL, then

$$Accuracy = \frac{n_1 + 0.5 n_2}{n}$$

Note: for pure chance Accuracy=0.5.

- Precision
  - □ The ratio of relevant items selected to the number of items selected.
  - □ Rank all the nonexistent links in decreasing order according to their score. How many links are predicted right among the top *L*.
- Leave one out
  - □ Select one link as probe link
  - □ Ranking Score: if it is of rank *r* among all nonexistent links(*M*), RankS=*r/M.*
  - □ Small size networks

Node-dependent Indices

- Common Neighbors

  F. Lorrain, H. C. White, J. Math. Sociol. 1, 49 (1971).

$$s_{xy} = |\Gamma(x) \cap \Gamma(y)|$$

- Salton Index

  G. Salton, Introduction to modern information retrieval (MuGraw-Hill, Auckland, 1983).

$$s_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k(x) \times k(y)}}$$

- Jaccard Index

  P. Jaccard, Bulletin de la societe Vaudoise des Science Naturelles 37, 547 (1901).

$$s_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

- Sorensen Index

  T. Sorensen, Biologiske Skrifter 5, 1 (1948).

$$s_{xy} = \frac{2 \times |\Gamma(x) \cap \Gamma(y)|}{k(x) + k(y)}$$

- Hub Promoted Index

  E. Ravasz, et al.,Science 297, 1553 (2002).

$$s_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\min\{k(x), k(y)\}}$$

- Hub Depressed Index

$$s_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\max\{k(x), k(y)\}}$$

- Leicht-Holme-Newman Index-I

  E. A. Leicht, et al., Phys. Rev. E 73, 026120 (2006).

$$s_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{k(x) \times k(y)}$$

- Preferencial Attachment

  A.-L. Barabási, R. Albert, Science 286, 509 (1999).

$$s_{xy} = k(x) \times k(y)$$

- Adamic-Adar Index

  L. A. Adamic, E. Adar, Social Networks, 25 211 (2003).

$$s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k(z)}$$

- Resource Allocation (RA)

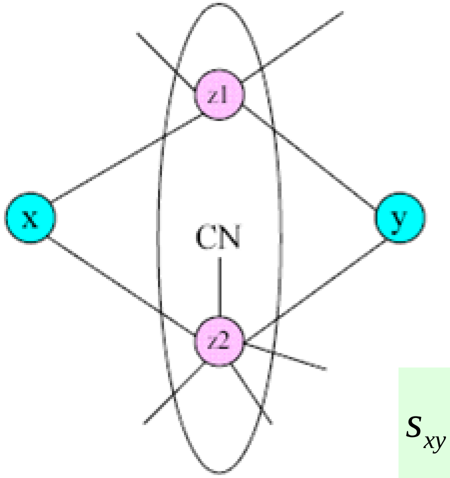
Q. Ou, et al., Phys. Rev. E, 75, 021102 (2007).

T. Zhou, et al., Phys. Rev. E, 76, 046115 (2007).

T. Zhou, L. Lü, Y.-C. Zhang, Eur. Phys. J. B 71, 623-630 (2009).

The node x can send some resource to y with their common neighbors playing the role of transmitters. Assume that each transmitter has a unit of resource, and will equally distribute it between all its neighbors. Then S(x,y) is defined as the amount of resource y received from x.

$$s_{xy} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{k(z)}$$



$$s_{xy} = \frac{1}{4} + \frac{1}{6} = \frac{5}{12}$$

Path-dependent Indices

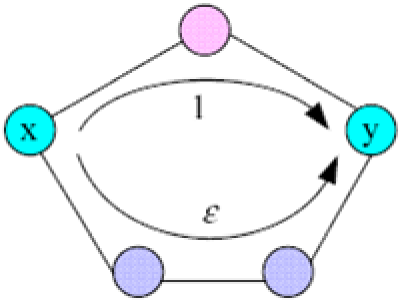
$$S = I + \phi A + \phi^2 A^2 + \phi^3 A^3 + \cdots$$

- Local Path (LP)

  T. Zhou, L. Lü, Y.-C. Zhang, Eur. Phys. J. B 71, 623-630 (2009).
  L. Lü, C.-H. Jin, T. Zhou, Phys. Rev. E 80, 046122 (2009).

  $$S = A^2 + \varepsilon A^3$$

  - □ where A is the adjacent matrix
  - □ ε is a free parameter



- Katz Index
  - □ Based on the ensemble of all paths, which sum over the collection of paths and exponentially damped by length to give short paths more weights. *(L. Katz, Psychmetrika 18(1) (1953) 39-43)*

  $$s_{xy} = \sum_{l=1}^{\infty} \beta^l \cdot \left| paths_{xy}^{\langle l \rangle} \right| \longrightarrow S = (I - \beta A)^{-1} - I$$

  - □ Where $paths_{xy}^{\langle l \rangle}$ is the set of all paths with length *l* connecting x and y
  - □ β is a free parameter

- Leicht-Holme-Newman Index-II

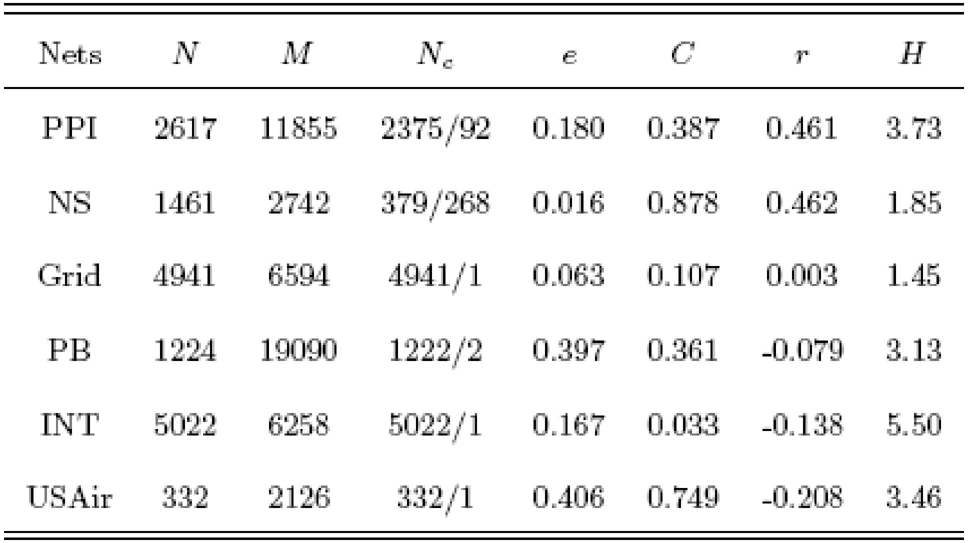
  (E. A. Leicht, Petter Holme, M. E. J. Newman, Phys. Rev. E 73, 026120, 2006)

  $$S = 2m\lambda D^{-1}\left(I - \frac{\alpha}{\lambda}A\right)^{-1} D^{-1}$$

  - m is the number of links of the network
  - λ is the largest eigenvalue of A.
  - a<1 is a free parameter

Local indices on six real networks

- **PPI**—A protein-protein interaction network.
- **NS**—A network of co-authorships between scientist.
- **Grid**—An electrical power grid of western US.
- **PB**—A network of the US political blogs
- **INT**—The router-level topology of the Internet
- **USAir**—The network of Us air transportation system



Nets	N	M	N_c	e	C	r	H
PPI	2617	11855	2375/92	0.180	0.387	0.461	3.73
NS	1461	2742	379/268	0.016	0.878	0.462	1.85
Grid	4941	6594	4941/1	0.063	0.107	0.003	1.45
PB	1224	19090	1222/2	0.397	0.361	-0.079	3.13
INT	5022	6258	5022/1	0.167	0.033	-0.138	5.50
USAir	332	2126	332/1	0.406	0.749	-0.208	3.46

N -node *M* -edge

Nc -giant component

e -efficiency

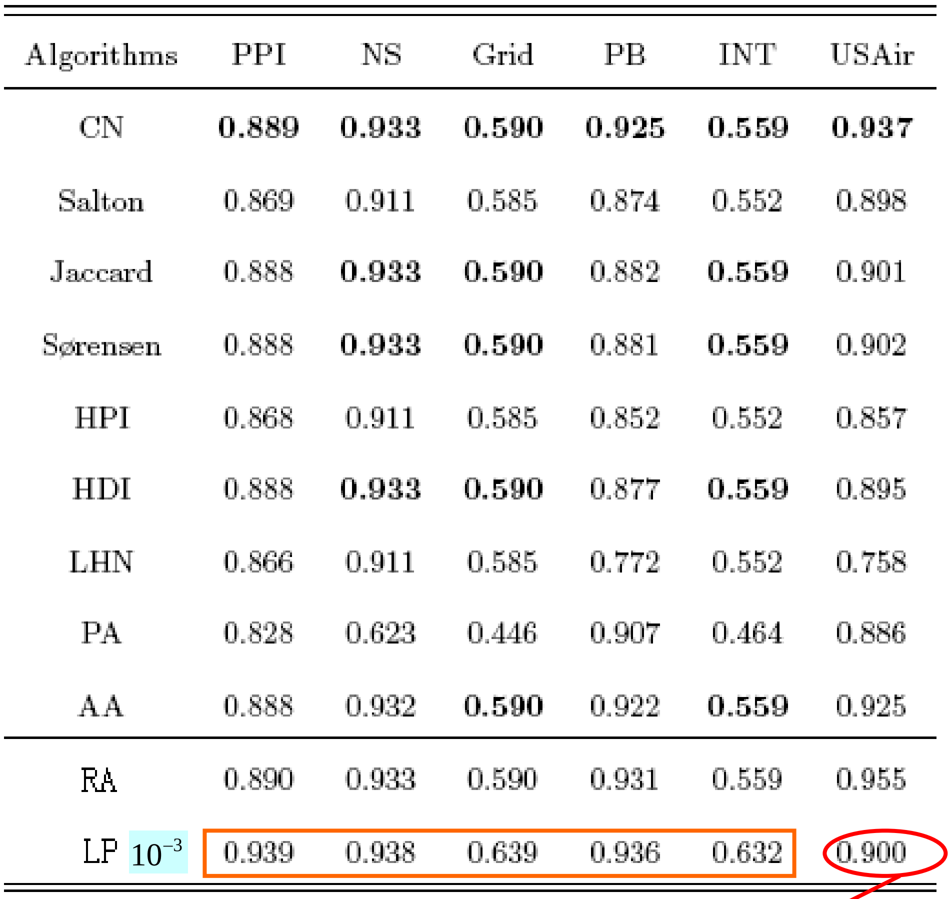
C -clustering coefficient

r -assortative coefficient

H -degree heterogeneity

(for detail see *Eur. Phys. J. B 71, 623-630 (2009).*)

Empirical analysis



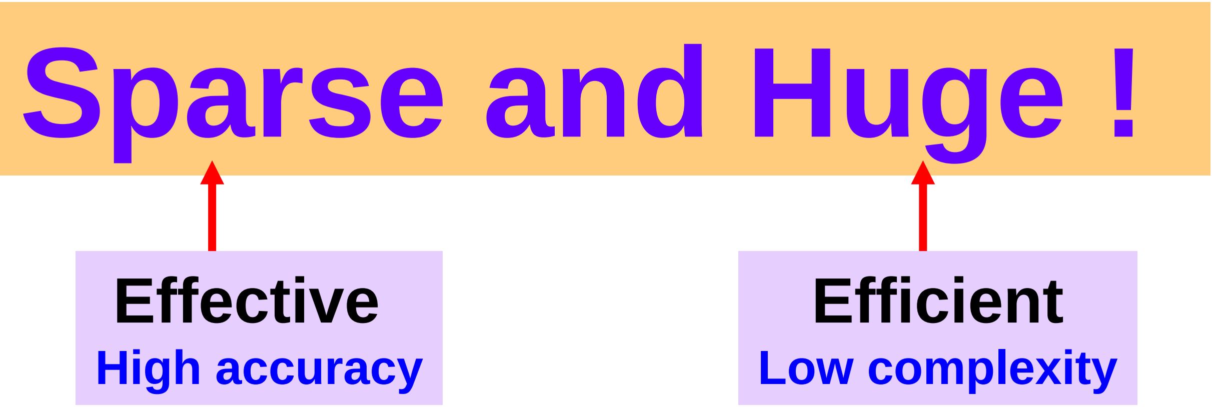
Algorithms	PPI	NS	Grid	PB	INT	USAir
CN	**0.889**	**0.933**	**0.590**	**0.925**	**0.559**	**0.937**
Salton	0.869	0.911	0.585	0.874	0.552	0.898
Jaccard	0.888	**0.933**	**0.590**	0.882	**0.559**	0.901
Sørensen	0.888	**0.933**	**0.590**	0.881	**0.559**	0.902
HPI	0.868	0.911	0.585	0.852	0.552	0.857
HDI	0.888	**0.933**	**0.590**	0.877	**0.559**	0.895
LHN	0.866	0.911	0.585	0.772	0.552	0.758
PA	0.828	0.623	0.446	0.907	0.464	0.886
AA	0.888	0.932	**0.590**	0.922	**0.559**	0.925
RA	0.890	0.933	0.590	0.931	0.559	0.955
LP 10^{-3}	0.939	0.938	0.639	0.936	0.632	0.900

0.945

-10^{-3}

1) CN performs the best among first nine indices, followed by AA.
2) RA outperforms all above indices.
3) LP requires further more information, and it performs the best except in USAir network.
4) LP is not sensitive to the parameter, which means a small positive ε can distinctly enhance the accuracy.
5) Negative parameter enhances the accuracy for USAir (for detail see *Eur. Phys. J. B 71, 623-630 (2009).*)

Great Challenges



Sparse and Huge !

Effective
High accuracy

Efficient
Low complexity

Effective and Efficient

- Comparison of three similarity indices
  - □ Common Neighbors
  - □ Local Path Index
  - □ Katz Index

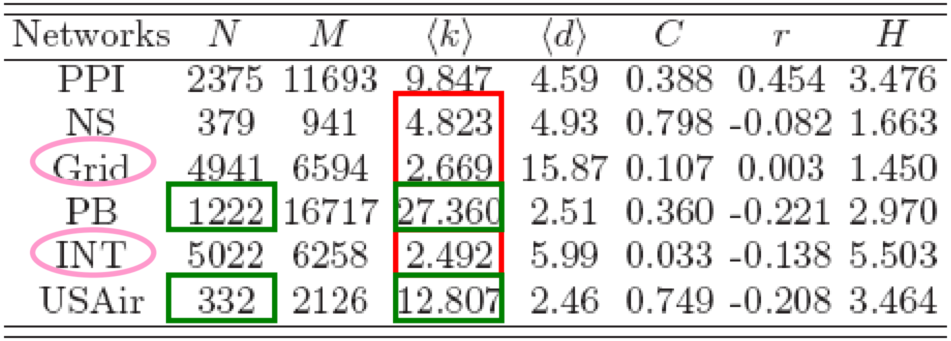
$$s_{xy} = \sum_{l=1}^{\infty} \beta^{l} \cdot \left| paths_{xy}^{\langle l \rangle} \right|$$

→

l=2 for CN
l=2,3 for LP
l=2,3,…, ∞ for Katz

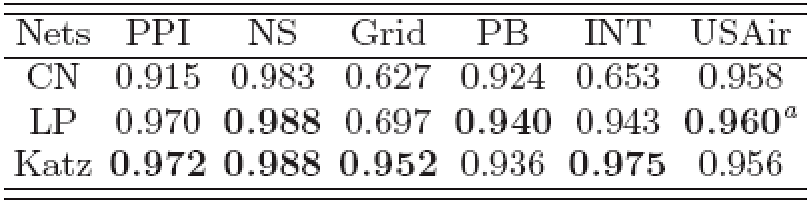
Empirical results

- Topological features
  - □ Giant component
  - □ <d> average shortest distance



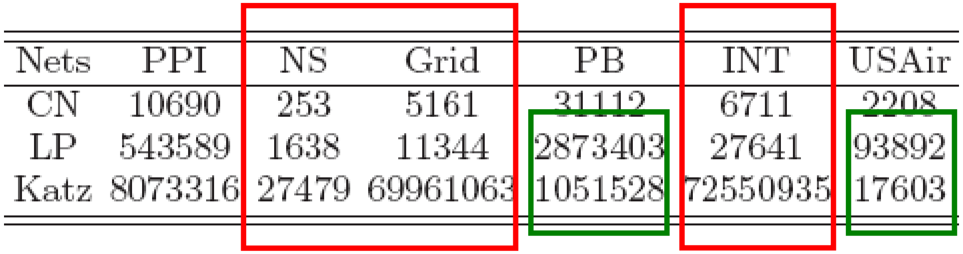
Networks	N	M	$\langle k \rangle$	$\langle d \rangle$	C	r	H
PPI	2375	11693	9.847	4.59	0.388	0.454	3.476
NS	379	941	4.823	4.93	0.798	-0.082	1.663
Grid	4941	6594	2.669	15.87	0.107	0.003	1.450
PB	1222	16717	27.360	2.51	0.360	-0.221	2.970
INT	5022	6258	2.492	5.99	0.033	-0.138	5.503
USAir	332	2126	12.807	2.46	0.749	-0.208	3.464

- Accuracy
  - □ Optimal parameter



Nets	PPI	NS	Grid	PB	INT	USAir
CN	0.915	0.983	0.627	0.924	0.653	0.958
LP	0.970	**0.988**	0.697	**0.940**	0.943	**0.960**[a]
Katz	**0.972**	**0.988**	**0.952**	0.936	**0.975**	0.956

- Time complexity
  - □ In microsecond



Nets	PPI	NS	Grid	PB	INT	USAir
CN	10690	253	5161	31112	6711	2208
LP	543589	1638	11344	2873403	27641	93892
Katz	8073316	27479	69961063	1051528	72550935	17603

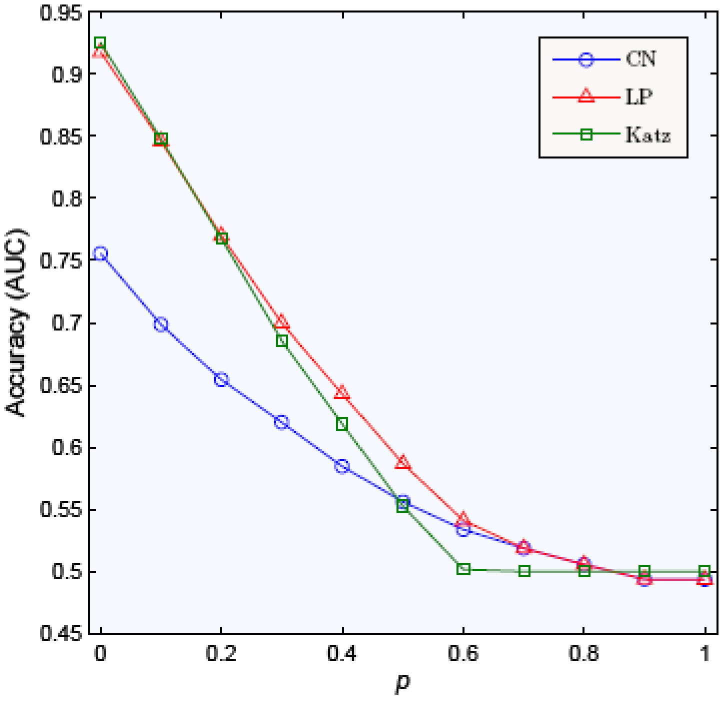
Model

- N nodes with identical degree *k (density)*
- Each node is characterized by a 10-dementional vector $\vec{f}$ with each element random selected in (-1,1)
- Intrinsic similarity of two nodes

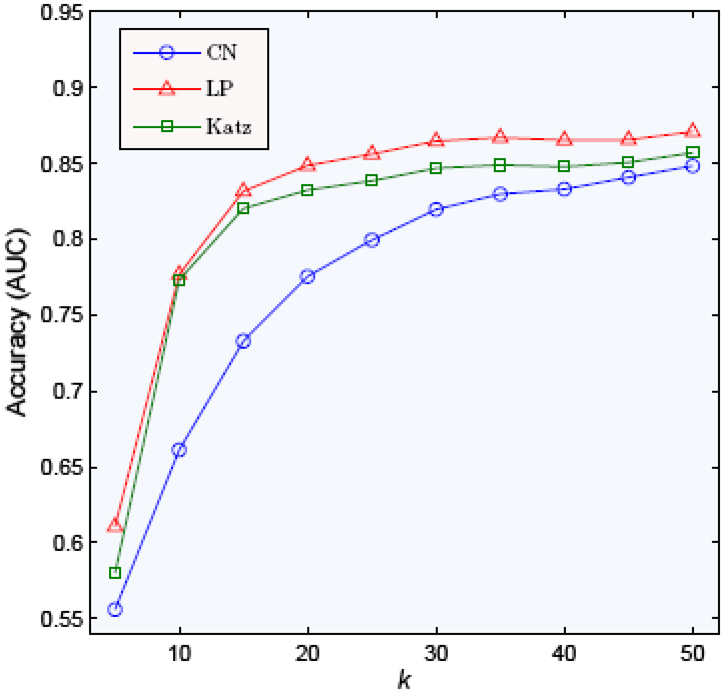
$$s_{xy}^{I} = \overrightarrow{f_x} \cdot \overrightarrow{f_y} = s_{yx}^{I}$$

- Process
  - □ A node with smallest degree is randomly select
  - □ *p (randomness* ∈*[0，1])* random choosing one node among all other nodes whose degree small than k
  - □ *1-p* choose the most similar node
  - □ *p* represents the strength of randomness in generating links, which can be understood as noise or irrationality in real system.

Accuracy (effective)

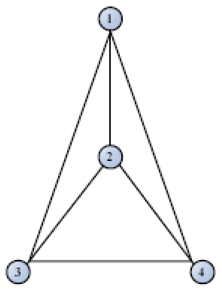


Prediction accuracy vs. the strength of randomness. *N=1000, k=10*.

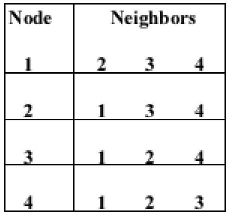


Prediction accuracy vs. network density. *N=1000, p=0.2*.

Computational Complexity (efficient)

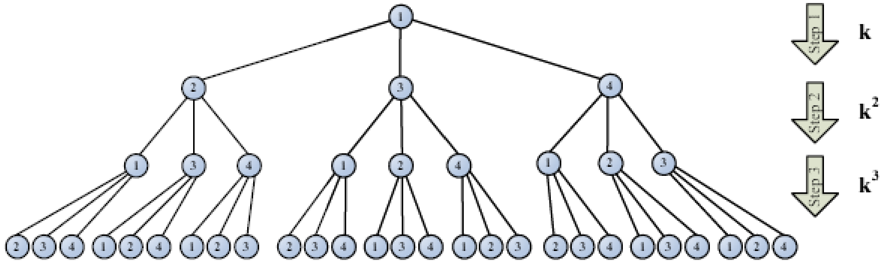


(a)



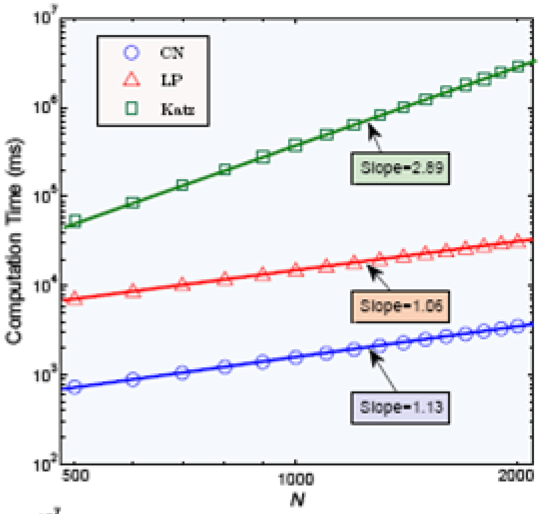
Node	Neighbors		
1	2	3	4
2	1	3	4
3	1	2	4
4	1	2	3

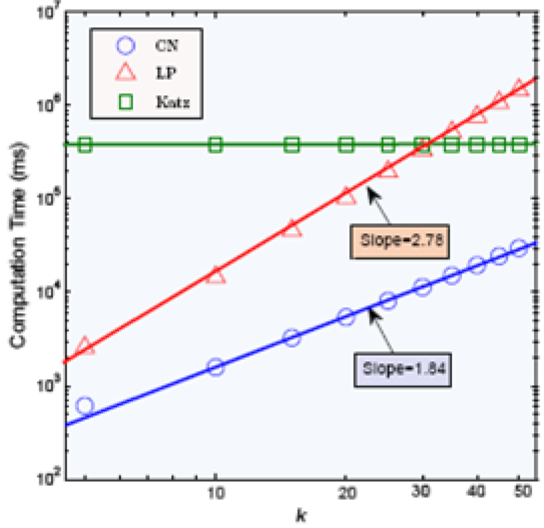
(b)



(c)

	Time complexity	Memory
CN	$\Theta(Nk^2)$	$\Theta(Nk)$
LP	$\Theta(Nk^3)$	$\Theta(Nk)$
Katz	$\Theta(N^3)$	$\Theta(N^2)$





Random-walk-based indices

- Average commute time (ACT)

  (D. J. Klein, M. Randic, J. Math. Chemistry, 12 (1993) 81-95)
  (F. Fouss, A. Pirotte, J.-M. Renders, M. Saerens, IEEE Trans. Knowl. Data. Eng. 19 (2007) 355)

  $$s_{xy} = V(l_{ii}^{+} + l_{jj}^{+} - 2l_{ij}^{+})$$ where *V* is the total degree

  - □ L^{+} is pseudoinverse of the Laplacian matrix. (L=D-A)
- Cosine based on the Pseudoinverse of the Laplacian matrix

  $$s_{xy} = \cos^{+}(x,y) = l_{ij}^{+} / \sqrt{l_{ii}^{+} \cdot l_{jj}^{+}}$$
- SimRank

  $$s_{xy} = C\frac{\sum_{a\in\Gamma(x)}\sum_{b\in\Gamma(y)} s_{ab}}{|\Gamma(x)|\cdot|\Gamma(y)|}$$ where C is decay factor
- Random walk with restart (RWR)
  - □ A random walker starting from node *i*, will iteratively moves to a random neighbor with probability *c* and return to node *i* with probability *1-c*.

$$\vec{s}_i = cP^{T}\vec{s}_i + (1-c)\vec{e}_i \longrightarrow \vec{s}_i = (1-c)(I - cP^{T})^{-1}\vec{e}_i$$

- Local random walk (LRW)
  W.-P. Liu, L. Lü, Link Prediction based on Local Random walk, EPL (submitted).
  - □ A random walker starts from *i*, denote by π_{ij} the probability that after *n* steps this walker happen to arrived at *j*.
  - □ The n-step LRW is defined as

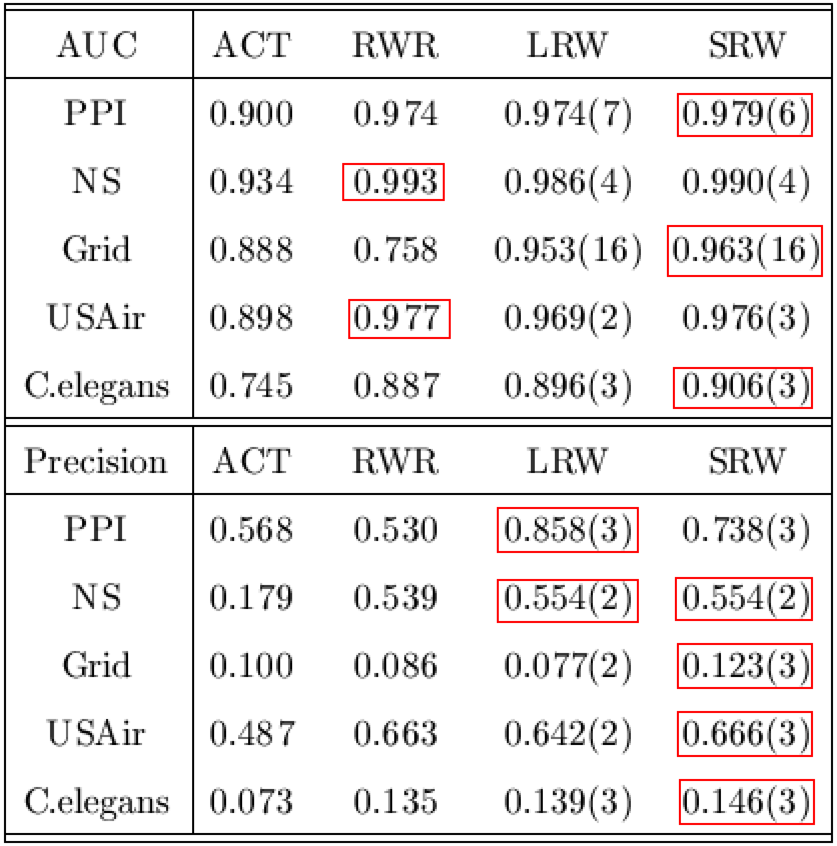
$$s_{ij}(n) = \pi_{ij}(n) \cdot k_i + \pi_{ji}(n) \cdot k_j$$

- Superposed random walk (SRW)
  - □ Reset the initial resource to node *i* at each time step
  - □ Each resetting can be considered as an independent random walk process
  - □ Then n-step SRW is defined as

$$s'_{ij}(n) = \sum_{l=1}^{n}[s_{ij}(l)] = k_i \sum_{l=1}^{n}[\pi_{ij}(l)] + k_j \sum_{l=1}^{n}[\pi_{ji}(l)]$$

Link prediction based on LRW

- **PPI**—A protein-protein interaction network.
- **NS**—A network of co-authorships between scientist.
- **Grid**—An electrical power grid of western US.
- **USAir**—The network of Us air transportation system
- **C.elegans**—The neural network of the nematode worm C.elegans, in which an edge joins two neurons if they are connected by either a synapse or a gap junction.



AUC	ACT	RWR	LRW	SRW
PPI	0.900	0.974	0.974(7)	**0.979(6)**
NS	0.934	**0.993**	0.986(4)	0.990(4)
Grid	0.888	0.758	0.953(16)	**0.963(16)**
USAir	0.898	**0.977**	0.969(2)	0.976(3)
C.elegans	0.745	0.887	0.896(3)	**0.906(3)**

Precision	ACT	RWR	LRW	SRW
PPI	0.568	0.530	**0.858(3)**	0.738(3)
NS	0.179	0.539	**0.554(2)**	**0.554(2)**
Grid	0.100	0.086	0.077(2)	**0.123(3)**
USAir	0.487	0.663	0.642(2)	**0.666(3)**
C.elegans	0.073	0.135	0.139(3)	**0.146(3)**

Computation Complexity

- LRW and SRW $O(N\langle k\rangle^{n})$
- ACT and RWR $O(N^{3})$

Other similarity indices

- Matrix Forest Theorem (Graph Theory)

  P. Chebotarev, E. Shamis, Automation and Remote Control 59 (1997) 1505-1514; 59 (1998) 1443-1459.

  $S=(I+\alpha L)^{-1}$ where $L=D-A$ and $D_{ij}=k_i\delta_{ij}$

  - □ S(i,j) indicates the ratio of the number of spanning rooted forests such that nodes i and j belong to the same tree rooted at i among all spanning rooted forests.

- Transferring similarity

  D. Sun, et al., Phys. Rev. E 80, 017101 (2009)

  - □ S denotes a similarity matrix
  - □ Denoting ε a decay factor of similarity transferred by a medi-user, a self consistent definition of *transferring similarity* can be written as:

  $$t_{ij}=\varepsilon\sum_{v}s_{iv}t_{vj}+s_{ij}$$

  - □ Using the matrix form, $T=(I-\varepsilon S)^{-1}S$ is the transferring similarity, where I is the identity matrix.

Link prediction in weighted networks

- Weighted similarity indices

  Replace adjacency matrix *A* with weighted matrix *Aw.*

  - □ Common Neighbors

$$s_{xy} = \left|\Gamma(x) \cap \Gamma(y)\right| \quad \Longrightarrow \quad s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,y)}{2}$$

  - □ Adamic-Adar Index

$$s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k(z)} \quad \Longrightarrow \quad s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,y)}{\log(1 + s(z))}$$

  - □ Resource Allocation

$$s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)} \quad \Longrightarrow \quad s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,y)}{s(z)}$$

Empirical results

a<0 weak ties effect

- WCN

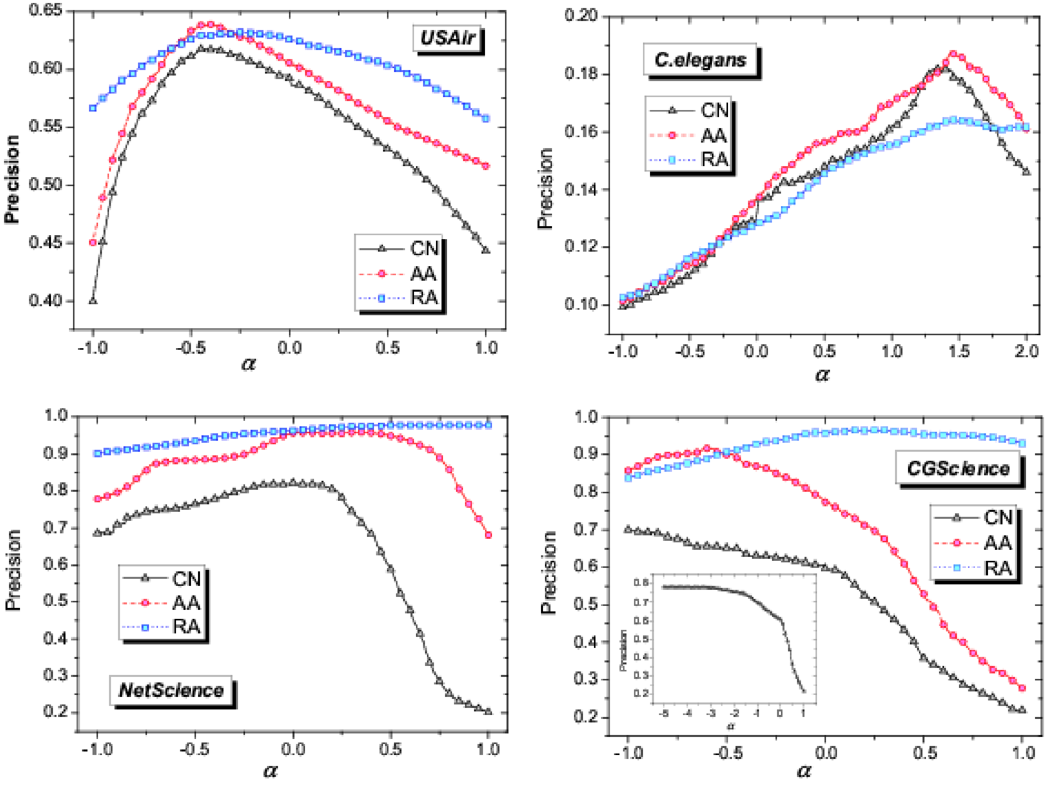
$$s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z)^{\alpha} + w(z,y)^{\alpha}}{2}$$

- WAA

$$s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z)^{\alpha} + w(z,y)^{\alpha}}{\log(1+s(z))}$$

- WRA

$$s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z)^{\alpha} + w(z,y)^{\alpha}}{s(z)}$$



a<0 weak>strong ; a=0 unweighted; a>0 strong>weak

0<a<1, weak links become stronger

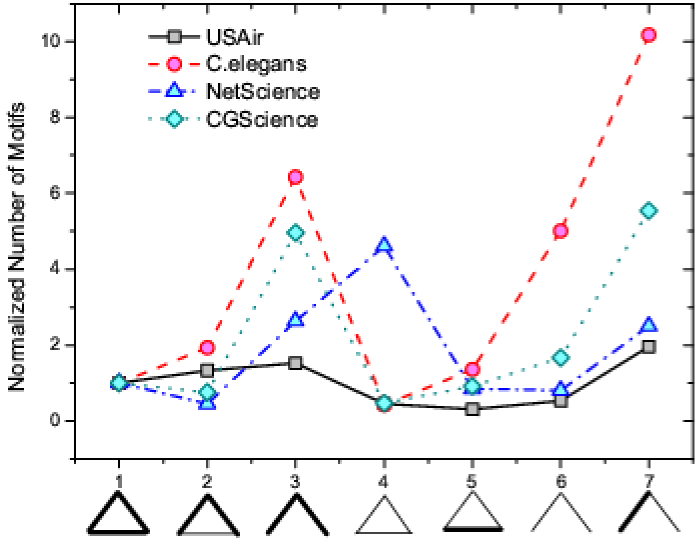
a>1, strong links become stronger

Motif analysis

	CN	WCN	WCN*	AA	WAA	WAA*	RA	WRA	WRA*
USAir	0.592	0.443	0.617(-0.41)	0.606	0.517	0.639(-0.40)	0.626	0.558	0.633(-0.24)
CE	0.132	0.162	0.182(1.41)	0.136	0.170	0.188(1.44)	0.128	0.155	0.164(1.56)
NetScience	0.822	0.202	0.822(0.00)	0.957	0.681	0.959(0.36)	0.962	0.978	0.978(0.80)
CGScience	0.625	0.299	0.782(-4.15)	0.780	0.292	0.917(-0.60)	0.963	0.938	0.969(0.13)

$$p_s = \frac{3N_1 + N_2}{3N_1 + N_2 + N_3} \qquad p_w = \frac{3N_4 + N_5}{3N_4 + N_5 + N_6}$$

	USAir	C.elegans	NetScience	CGScience
p_s	0.7393	0.4345	0.5667	0.4315
p_w	0.7572	0.3442	0.9479	0.5819



Thank you

Acknowledgement : Prof. Yi-Cheng Zhang, Prof. M.-S. Shang, Dr. M. Medo, Dr. Ci-hang Jin, Mr. Tao Zhou, Mr. Zi-ke Zhang, Mr. W.-P. Liu.

Main research interests

Link prediction:

T. Zhou, L. Lü, Y.-C. Zhang, Eur. Phys. J. B 71, 623-630 (2009).
L. Lü, C.-H. Jin, T. Zhou, Phys. Rev. E 80, 046122 (2009).
L. Lü, T. Zhou, Link Prediction in Weighted Networks: The Role of Weak Ties, EPL (accepted).
W.-P. Liu, L. Lü, Link Prediction based on Local Random walk, EPL (submitted).
L. Lü, T. Zhou, Link Prediction in Complex Networks: A Mini-Review (manuscript in progress).

Recommender Systems:

M.-S. Shang, L. Lü, Y.-C. Zhang, T. Zhou, Empirical analysis of web-based user-object bipartite networks, EPL (submitted).
M.-S. Shang, L. Lü, W. Zeng, Y.-C. Zhang, Relevance is More Significant than Correlation: Information Filtering on Sparse Data, EPL (accepted).

Semiotic dynamics:

Z.-K. Zhang, L. Lü, J. G. Liu, T. Zhou, Empirical analysis on a keyword-based semantic system, Eur. Phys. J. B, 66(2008) 557-561.
L. Lü, Z.-K. Zhang, T. Zhou, Zipf's law result in Heap's law, (manuscript in progress) .
L. Lü, Z.-K. Zhang, T. Zhou, Effects of vocabulary size: a deviation of Heaps' Law. (manuscript in progress)
L. Lü, Z.-K. Zhang, T. Zhou, Scaling Laws in Boundary Linguistic Systems, (manuscript in progress).

Informational economics:

L. Lü, M. Medo, Y.-C. Zhang, D. Challet, Emergence of product differentiation from consumer heterogeneity and asymmetric information, Eur. Phys. J. B, 64 (2008) 293.
L. Lü, M. Medo, Y. -C. Zhang, The role of matchmaker in a vender-buyer interaction market, Eur. Phys. J. B 71, 565-571 (2009).