Statistical Mechanics of Information Systems

Information
Filtering on Complex Networks

Tao Zhou

Department of Physics, University of Fribourg, Switzerland

Email Address: zhutou@ustc.edu; zhutouster@gmail.com
Homepage: http://sites.google.com/site/zhutouster/

Content

- Motivation and Background
- Ranking
- Predicting
- **Recommending (Main Part)**
- Social filtering
- Looking into the future
- Summary of Scientific Achievement

Interdisciplinary Research of Statistical Physics

- **Biophysics**: using the methods of physics and physical chemistry to study biological systems.
- **Econophysics**: applying theories and methods originally developed by physicists in order to solve problems in economics, usually those including uncertainty or stochastic processes and nonlinear dynamics.
- **Sociophysics**: aiming at a statistical physics modeling of large scale social phenomena, like opinion formation, cultural dissemination, the origin and evolution of language, crowd behavior, social contagion.
- **Infophysics**???

Main Topics in Infophysics

- Organization and evolution of information systems (**Structures**)
- Physical dynamics taking place on information systems (**Functions**)
- Statistical analysis on internet-relevant activities (**Internet-Web-Object-User**)
- Information filtering: ranking, predicting and recommending
- ……

Motivation & Background

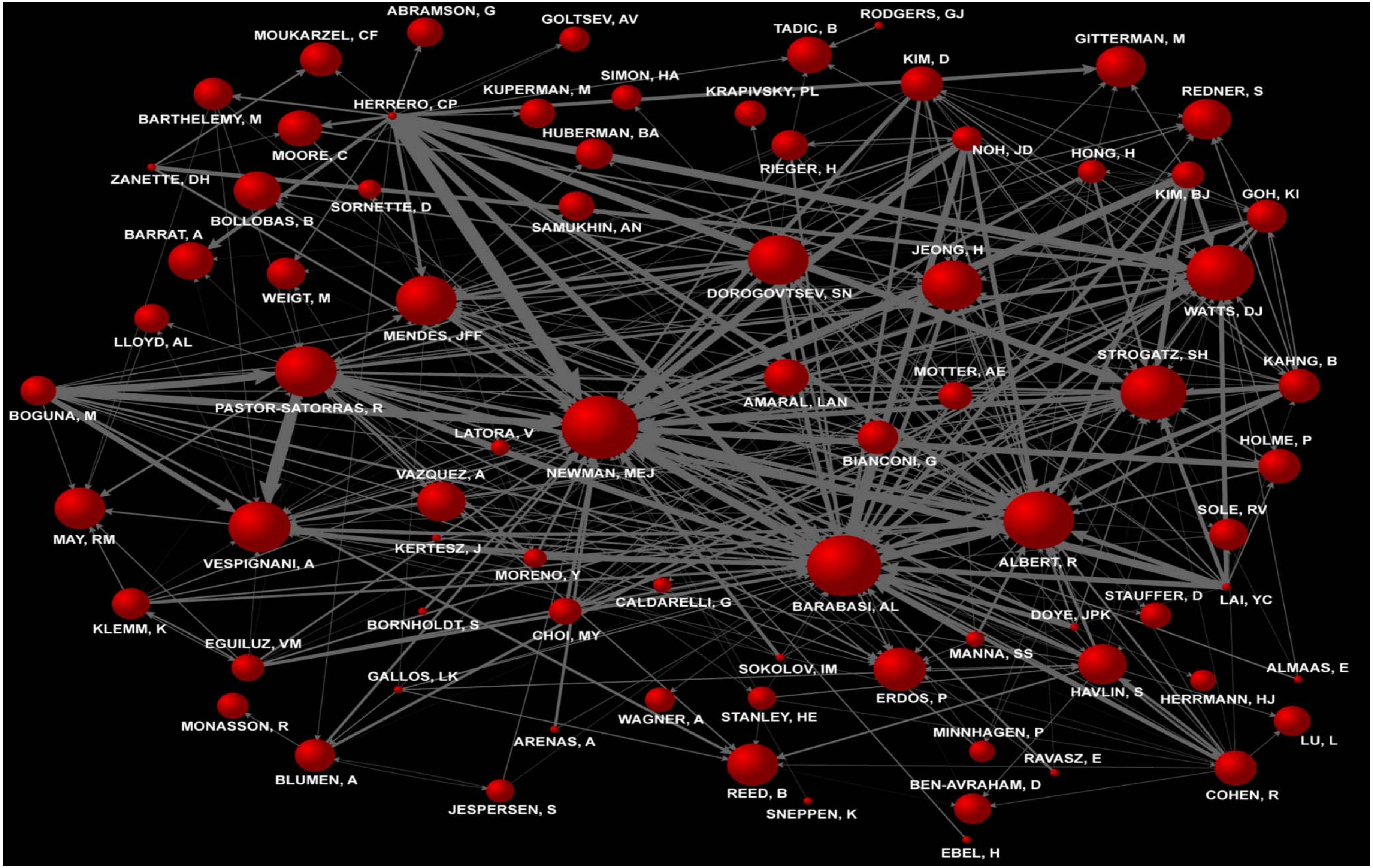
- Explosive growth of information asks for advanced theories and techniques to automatically find out the objects we like
- Internet and WWW are typical many-body systems that are favored by statistical physicists
- Physical concepts and approaches have already been successfully applied in uncovering important regularities and solving challenges in information science
- Information filtering shifts from finding out what you want to what you like, from centralized to decentralized, from population-based to personlized.

Ranking

Objects	Methods / Metrics
Simple Graph	Centralities
Directed Graph	PageRank, Random Walk with Restart
Nodes with Mixing Roles	HITS
Bipartite Rating Systems	**Reputation Systems**
Collaboration Graph with Citations	CiteRank, SARA
……	……

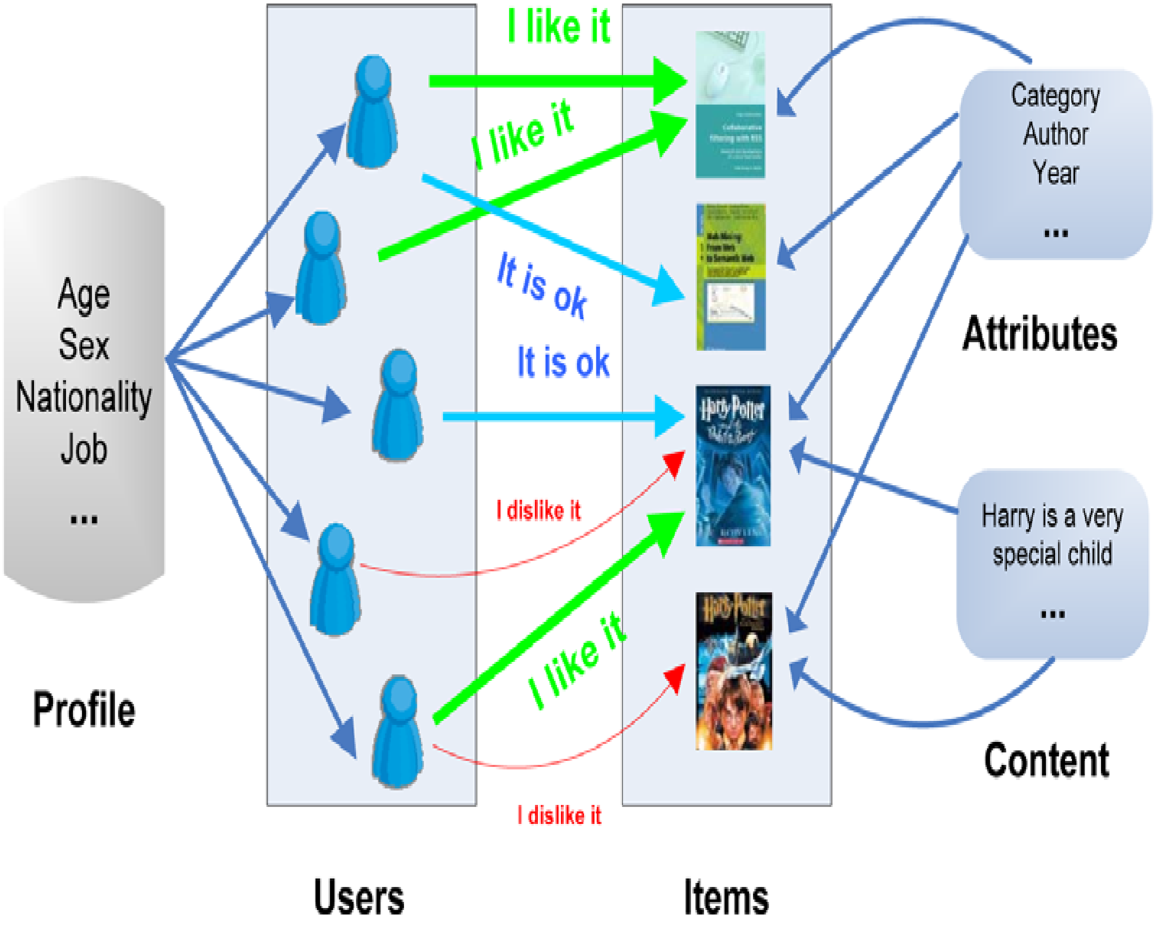
Core Method: Iterative Refinement !!

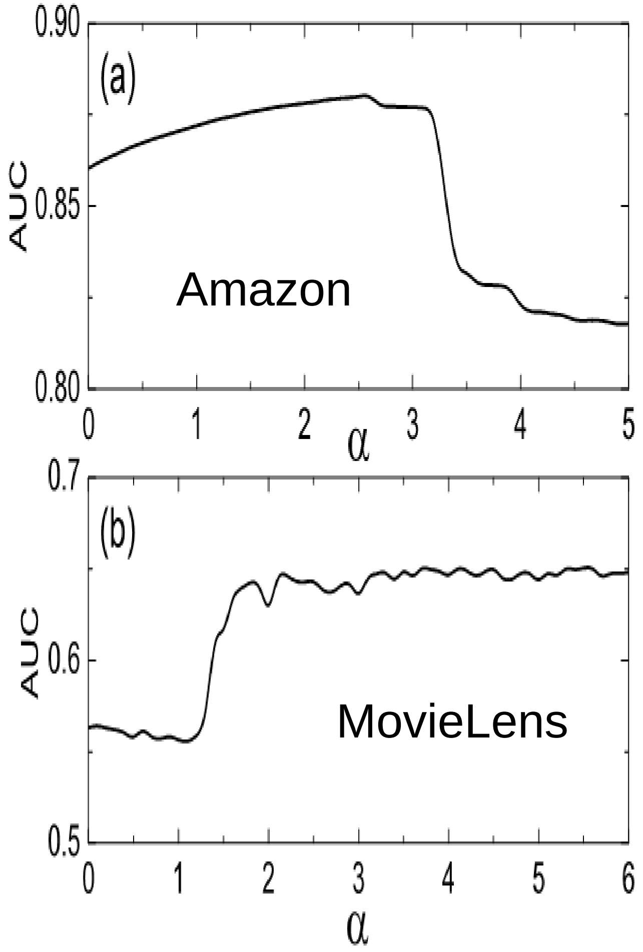
Ranking Based on the Diffusion of Scientific Credits



F. Radicchi *et al.*, PRE 80 (2009) 056103

Building Reputation Systems for Better Ranking





L.-L. Jiang, M. Medo, J. R. Wakeling, Y.-C. Zhang, T. Zhou, arXiv: 1001.2186

Link Prediction

- It aims at estimating the likelihood of the existence of a link between two nodes.
- It can help in understanding the factors underlying network evolution.
- It can help in evaluating various measurements of node similarity.
- For biological networks, it may reduce the experimental costs.
- For online social networks, it can generate good recommendations.
- It can be applied in solving the link classification problem in partially labeled networks

A. Clauset, C. Moore, M. E. J. Newman, Nature 453 (2008) 98
S. Render, Nature 453 (2008) 47
R. Guimera, M. Sales-Pardo, PNAS 106 (2009) 22073

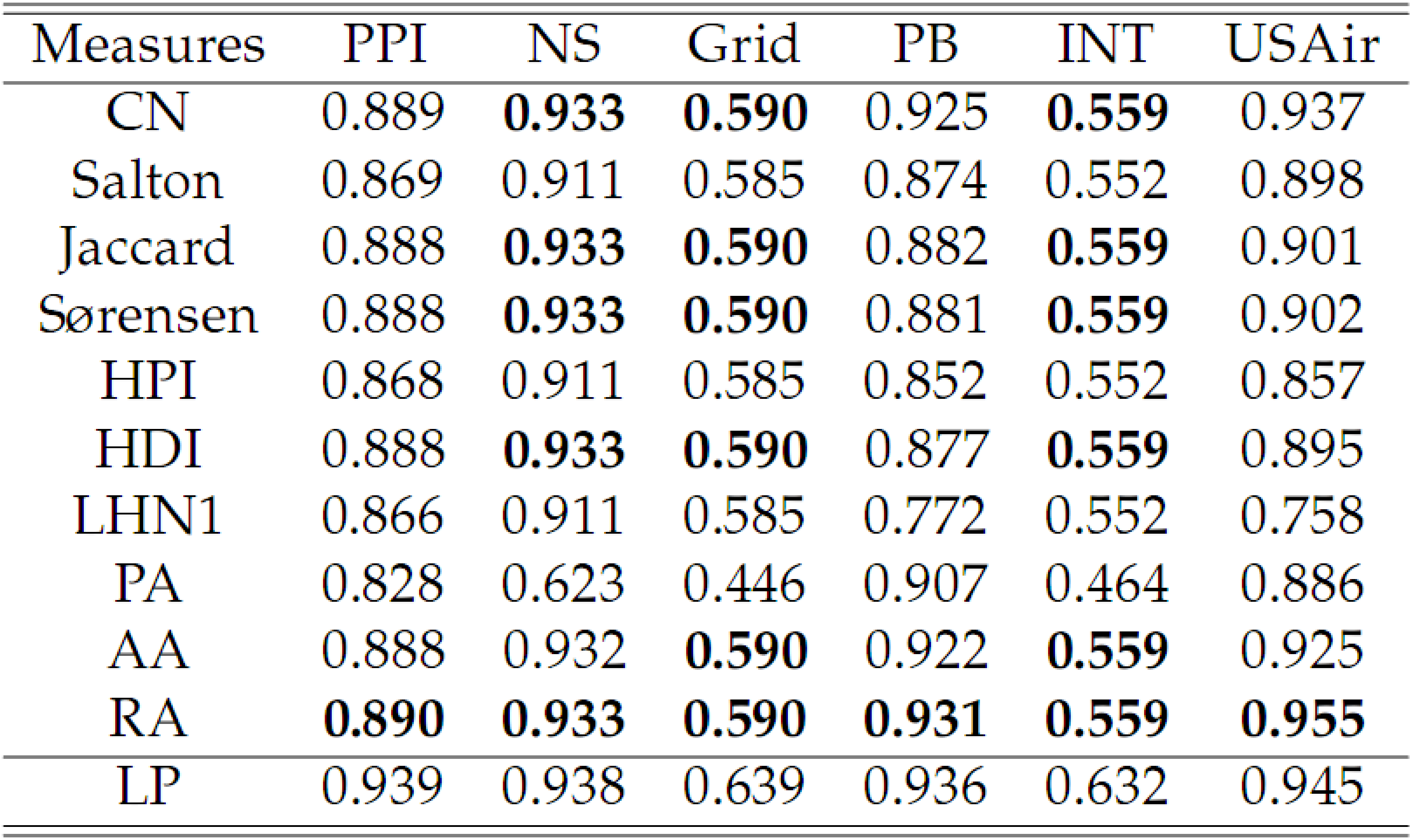
Main Methods

- Attribute-Aware and Content-Based Algorithms
- Relational Models
- Markov Chain
- Hierarchical Model
- Maximum Likelihood Methods
- Similarity-Based Methods

* Straightforward Comparison

* Collaboration Filtering

Local Similarity Indices

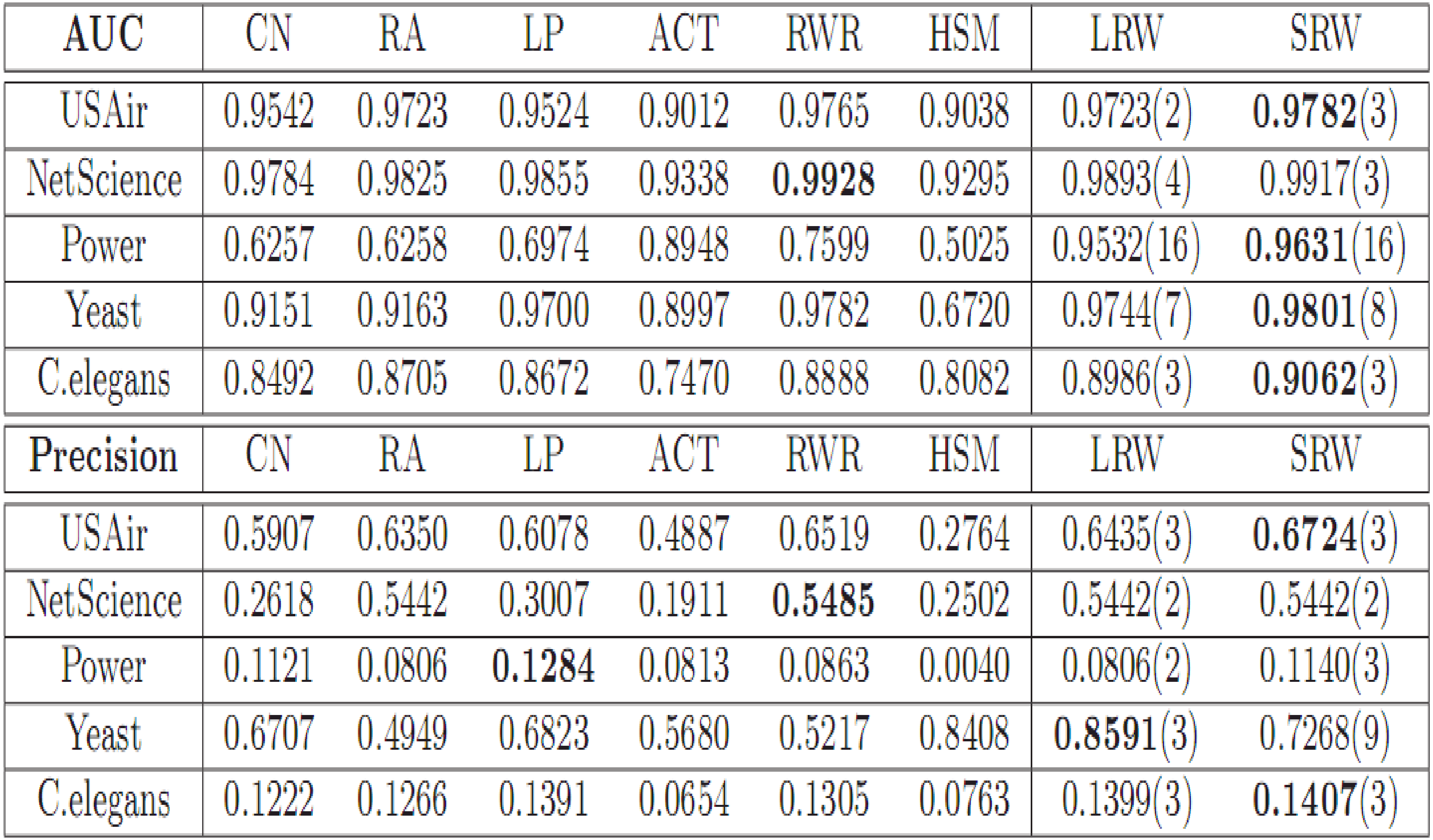


Measures	PPI	NS	Grid	PB	INT	USAir
CN	0.889	**0.933**	**0.590**	0.925	**0.559**	0.937
Salton	0.869	0.911	0.585	0.874	0.552	0.898
Jaccard	0.888	**0.933**	**0.590**	0.882	**0.559**	0.901
Sørensen	0.888	**0.933**	**0.590**	0.881	**0.559**	0.902
HPI	0.868	0.911	0.585	0.852	0.552	0.857
HDI	0.888	**0.933**	**0.590**	0.877	**0.559**	0.895
LHN1	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	0.939	0.938	0.639	0.936	0.632	0.945

RA>AA>CN

T. Zhou, L. Lü, Y.-C. Zhang, EPJB 71 (2009) 623

Quasi-Local and Global Similarity Indices

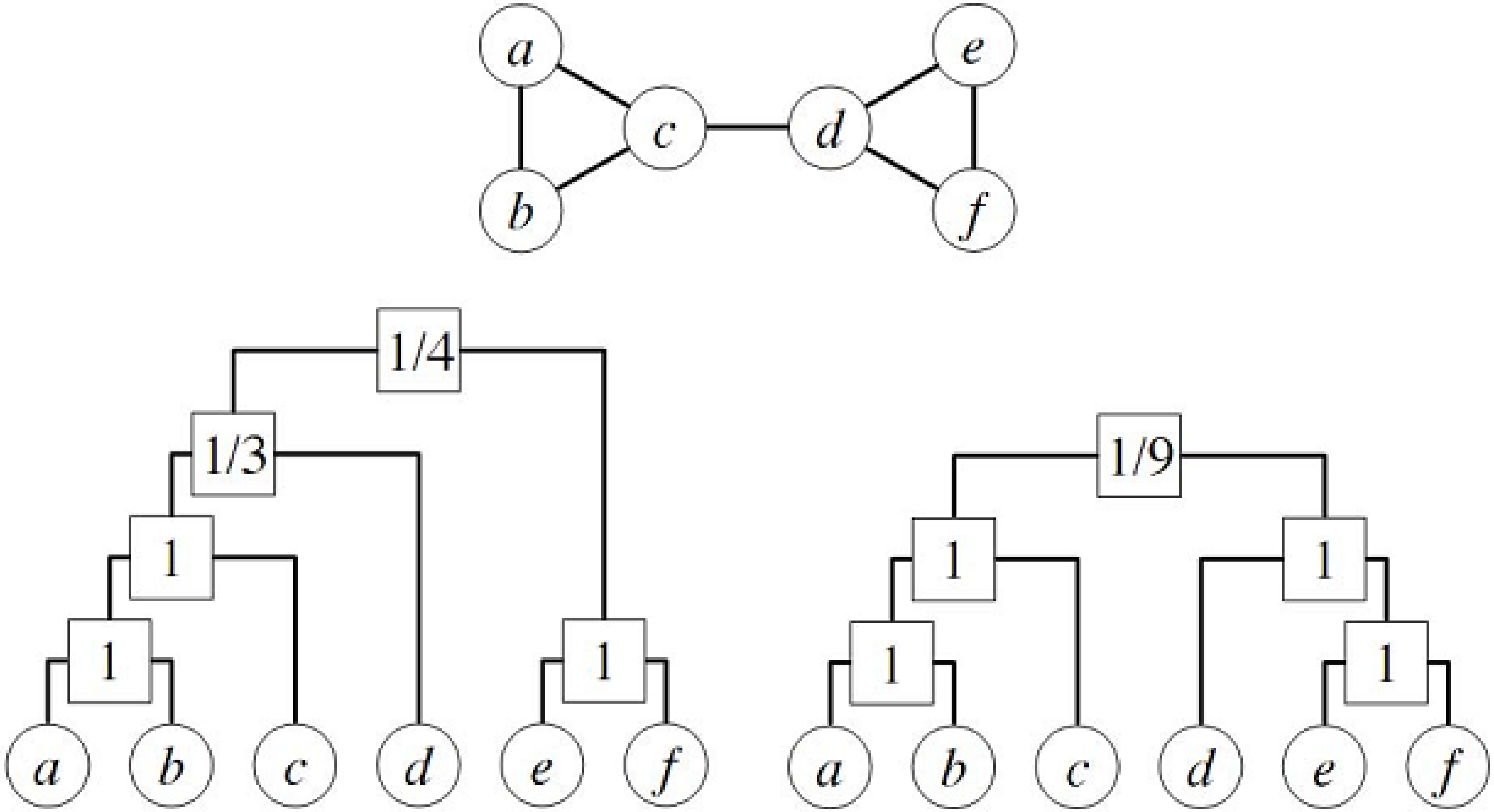


AUC	CN	RA	LP	ACT	RWR	HSM	LRW	SRW
USAir	0.9542	0.9723	0.9524	0.9012	0.9765	0.9038	0.9723(2)	**0.9782**(3)
NetScience	0.9784	0.9825	0.9855	0.9338	**0.9928**	0.9295	0.9893(4)	0.9917(3)
Power	0.6257	0.6258	0.6974	0.8948	0.7599	0.5025	0.9532(16)	**0.9631**(16)
Yeast	0.9151	0.9163	0.9700	0.8997	0.9782	0.6720	0.9744(7)	**0.9801**(8)
C.elegans	0.8492	0.8705	0.8672	0.7470	0.8888	0.8082	0.8986(3)	**0.9062**(3)

Precision	CN	RA	LP	ACT	RWR	HSM	LRW	SRW
USAir	0.5907	0.6350	0.6078	0.4887	0.6519	0.2764	0.6435(3)	**0.6724**(3)
NetScience	0.2618	0.5442	0.3007	0.1911	**0.5485**	0.2502	0.5442(2)	0.5442(2)
Power	0.1121	0.0806	**0.1284**	0.0813	0.0863	0.0040	0.0806(2)	0.1140(3)
Yeast	0.6707	0.4949	0.6823	0.5680	0.5217	0.8408	**0.8591**(3)	0.7268(9)
C.elegans	0.1222	0.1266	0.1391	0.0654	0.1305	0.0763	0.1399(3)	**0.1407**(3)

L. Lü, C.-H. Jin, T. Zhou, PRE 80 (2009) 046122;
W.-P. Liu, L. Lü, EPL (submitted)

Hierarchical Model

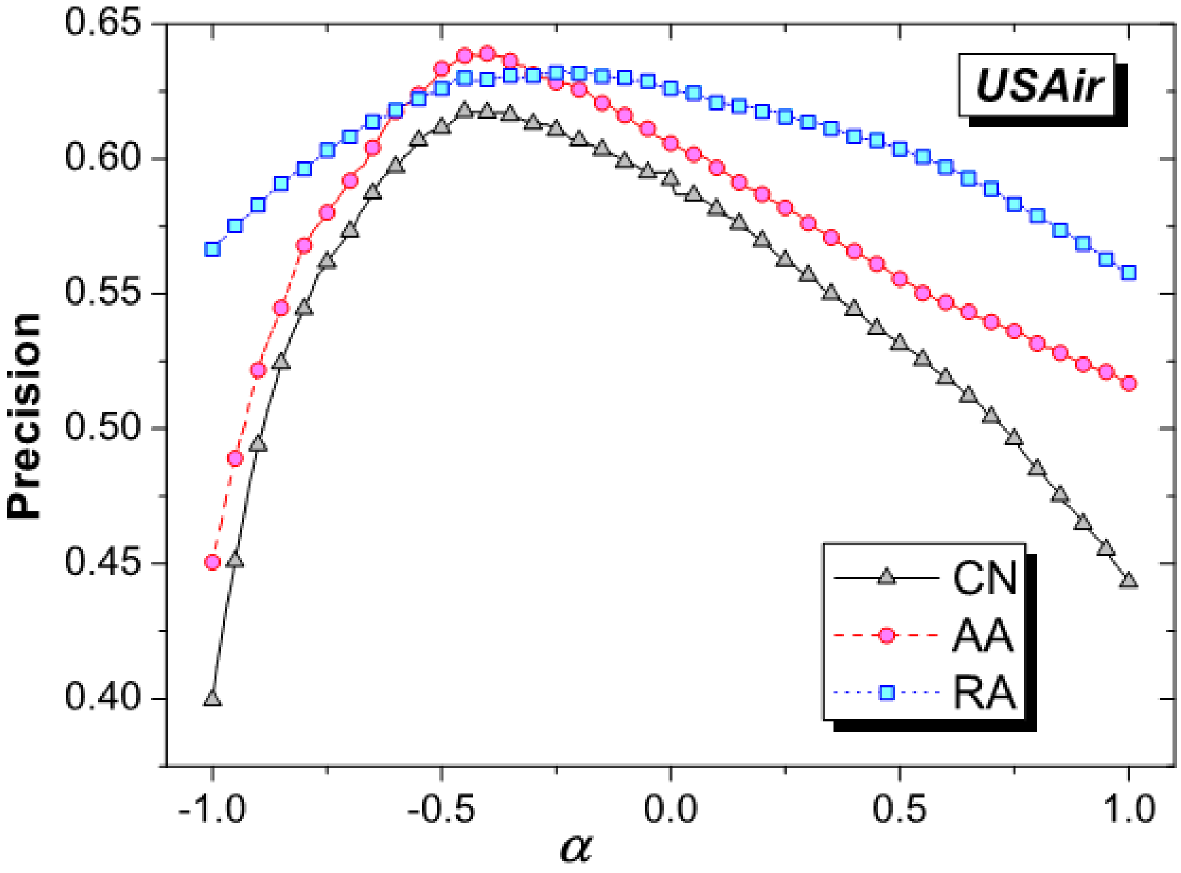


This method is slow, less accurate, yet can give some structural insights of networks organization.

A. Clauset, C. Moore, M. E. J. Newman, Nature 453 (2008) 98

Link Prediction in Weighted Networks

——Role of Weak Ties



Similar Observation for Weighted Katz Index: **D. Liben-Nowell, J. Kleinberg, JASIST 58 (2007) 1019**
Local Weighted Indices: **T. Murata, S. Moriyasu, IEEE/WIC/ACM Intl. Conf. Web Intell. 2007**
Weak Ties Analysis: **L. Lü, T. Zhou, Europhys. Lett. 89 (2010) 18001**

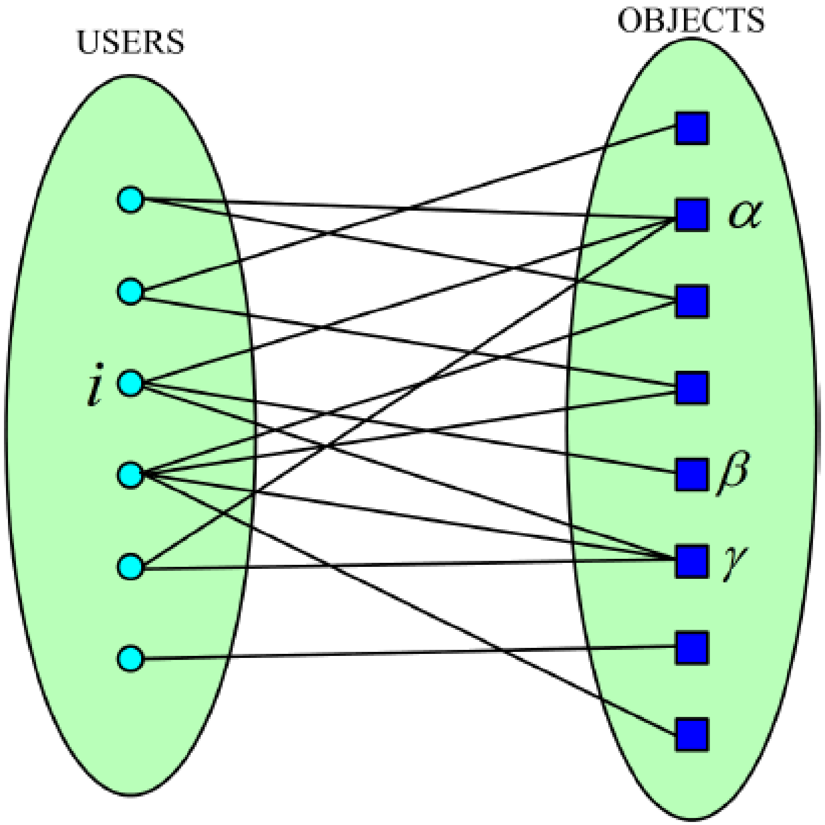
Personalized Recommendation

- Personalized recommender systems use the personal information of a user (the historical record of his activities and possibly his profile) to uncover his habits and to consider them in the recommendation.
- Personalized recommender systems provide a promising way to solve the information overload problem.
- Personalized recommender systems have already been successfully applied in many e-commerce web sites, such as *Amazon.com*.

Problem Description

Known information: the record of interactions between users and objects, the users’ profiles, the objects’ attributes, the content, the time stamps, the user-user relationships, etc.

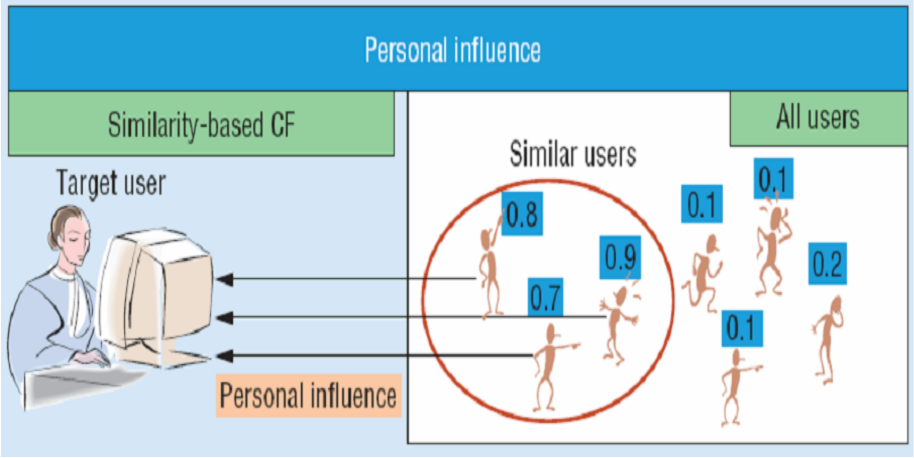
Required information: whether a targer user will like an unselected object, and if so, to what extent he/she likes it. Basically, a personalized recommender system should provide an ordered list of unselected objects to every target user.

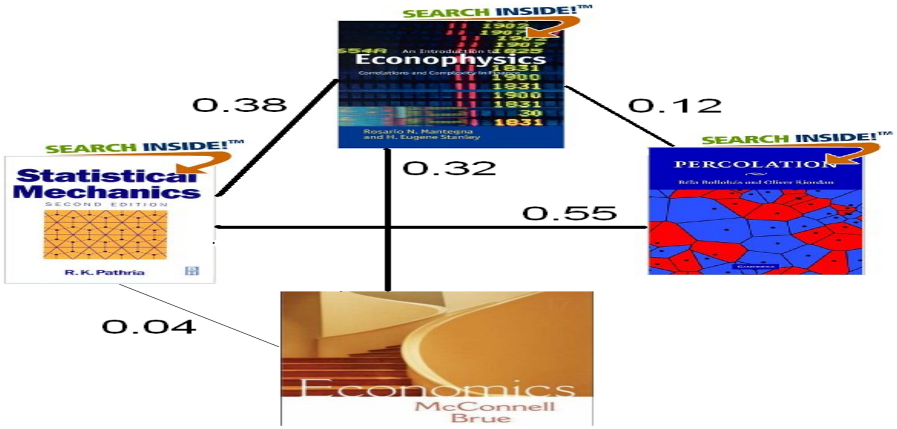


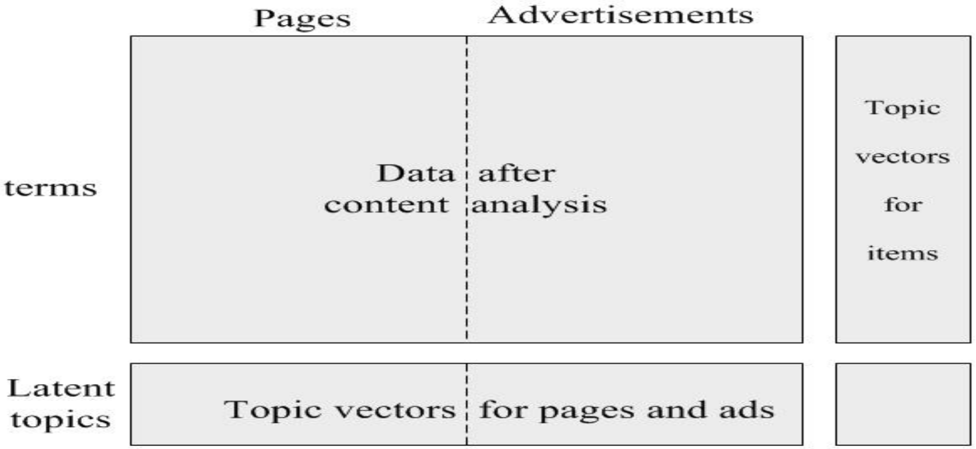
Target user: *i*

Main Methods

- Collaborative filtering
- Iterative refinement
- Diffusion/Local Diffusion
- Principle component analysis
- Latent semantic model
- Content-based analysis
- Latent Dirichlet allocation
- Hybrid algorithm and ensemble approach
- Matrix factorization
- ……







Our Work

- How to evaluate algorithmic performance
- Local diffusion methods: energy/mass and heat
- Accuracy of local diffusion methods
- Improved Algorithms: two examples
- Hybrid algorithm: solving the accuracy-diversity dilemma
- Adaptive algorithm: real-time response to changing data

Evaluating Algorithmic Performance

- Accuracy

  ** Overall Ranking: AUC, Ranking Score

  ** Top Recommended Objects: Precision, Recall, F-Measure

- Diversity

  ** Intra-Similarity: Recommendations to a user are diverse

  ** Inter-Similarity Recommendations to different users are diverse

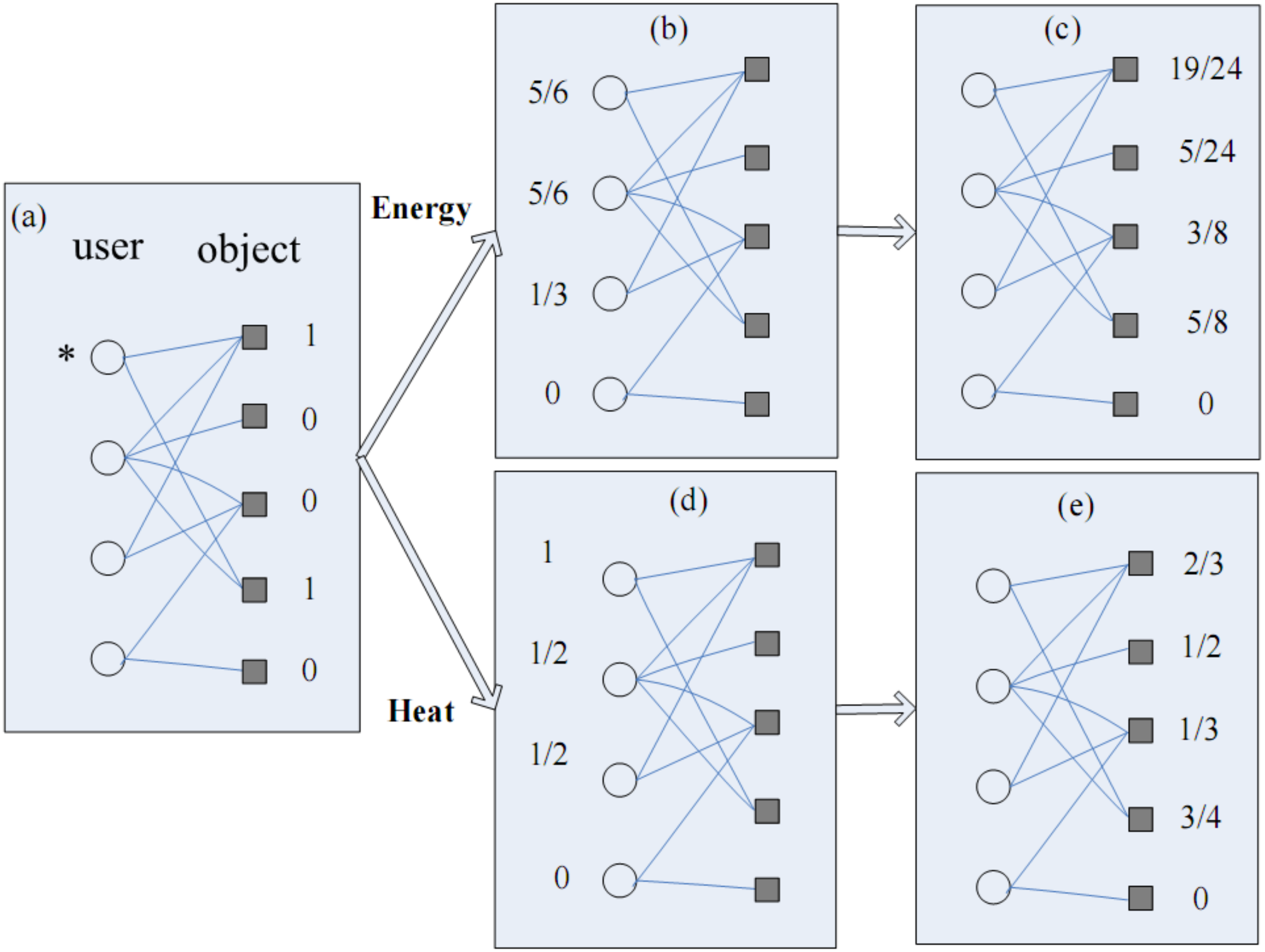
- Novelty

  ** Popularity

  ** Self-information

J. L. Herlocker *et al.,* ACM Trans. Inf. Syst. 22 (2004) 5
T. Zhou *et al.,* EPL 81 (2008) 58004
T. Zhou *et al.,* NJP 11 (2009) 123008
T. Zhou *et al.,* PNAS (Accepted, March 2010)

Basic idea on local diffusion

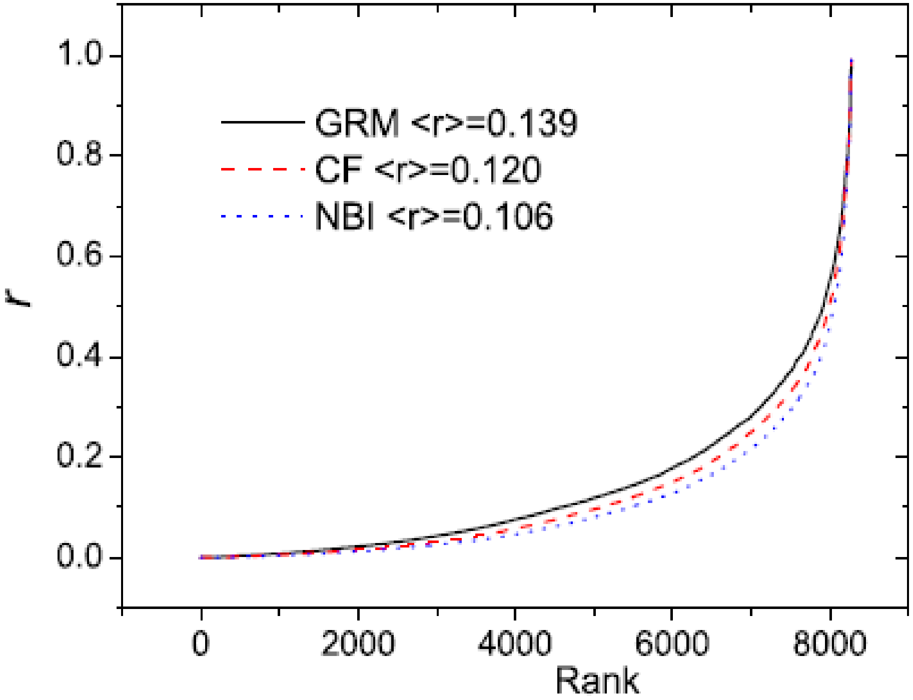


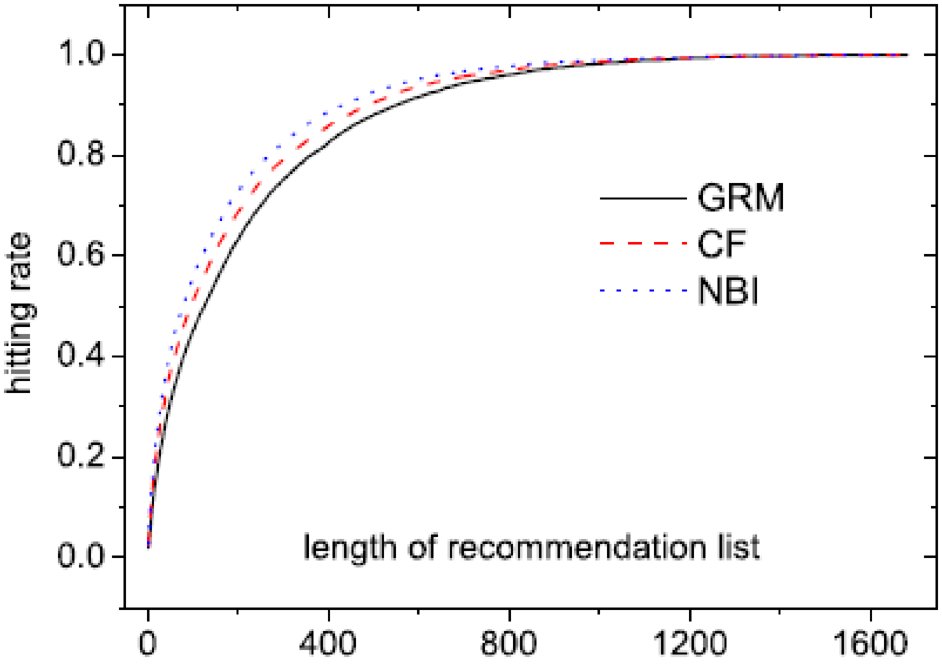
T. Zhou *et al*., PRE 76 (2007) 046115; PNAS (Accepted)

For Global Diffusion, see:

Y.-C. Zhang, *et al*., PRL 99 (2007) 154301; EPL 80 (2007) 68003

Accuracy of Mass Diffusion and Heat Conduction

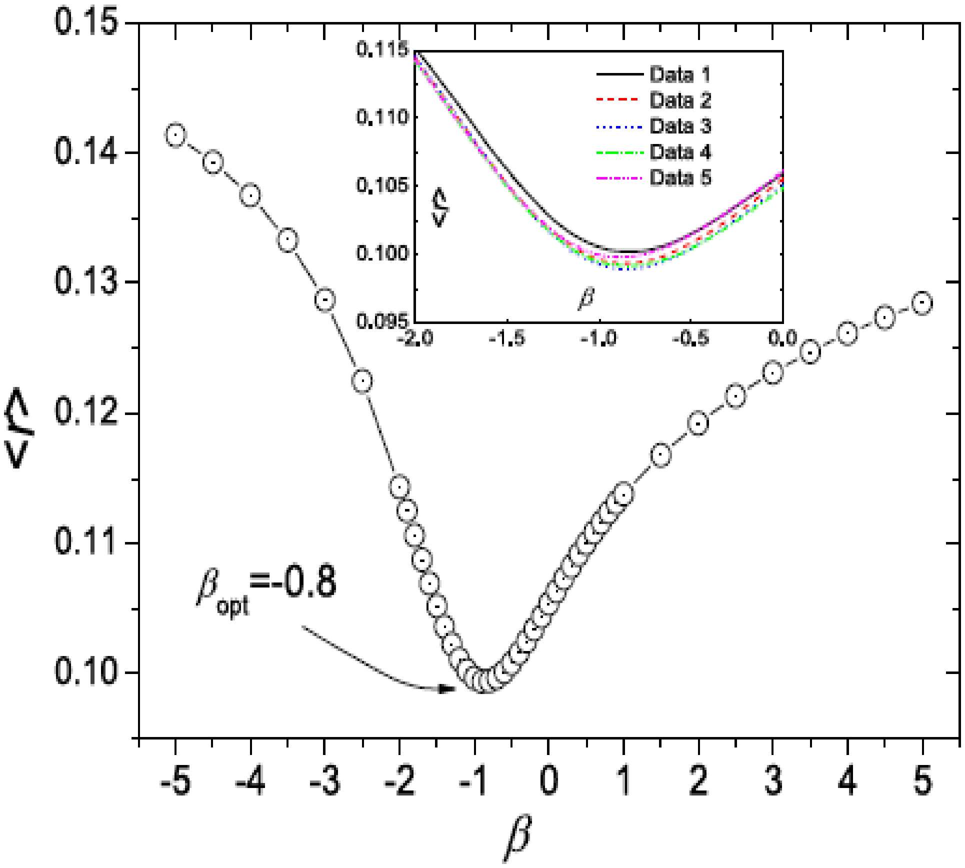


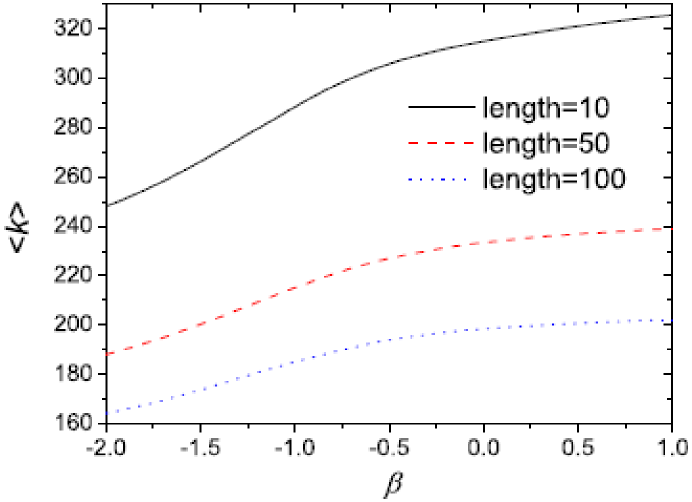


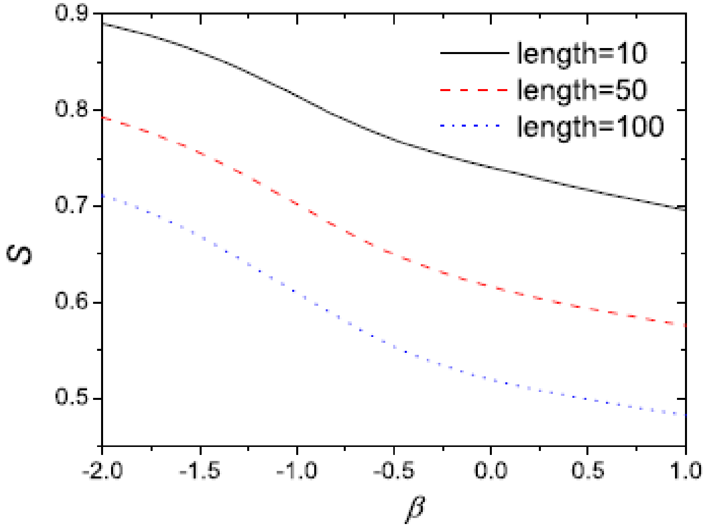
Heat Conduction is even less accurate than the collaborative filtering (CF) and global ranking method (GRM).
Order of accuracy: MD > CF > GRM > HC.

Improved Algorithm I: Initial Resource Depends on the Degree

$$\vec{f}_j = a_{ij} d_j^{\beta}$$

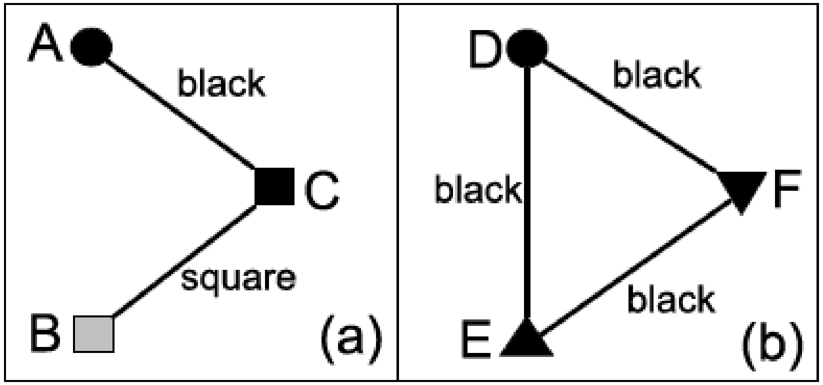




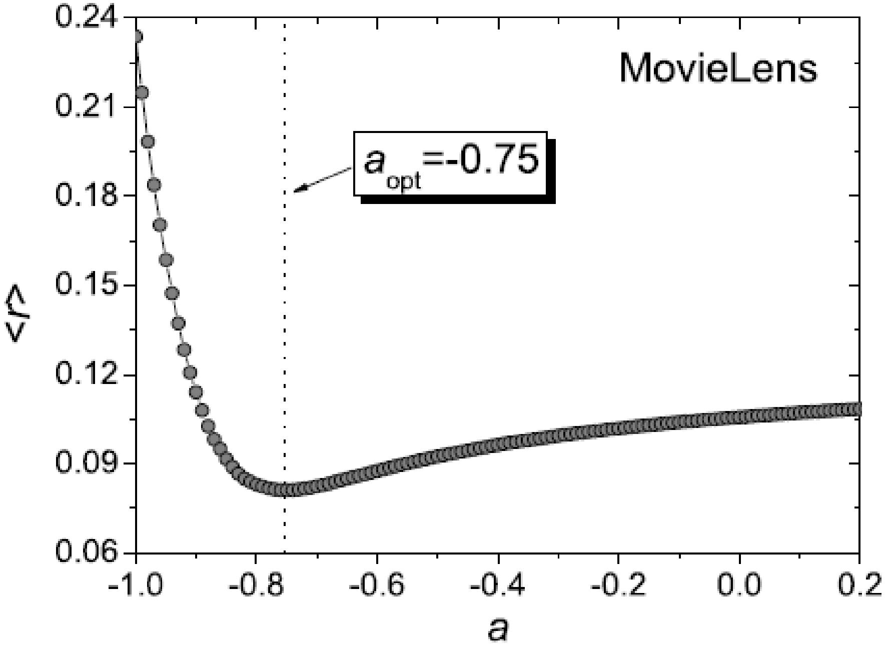


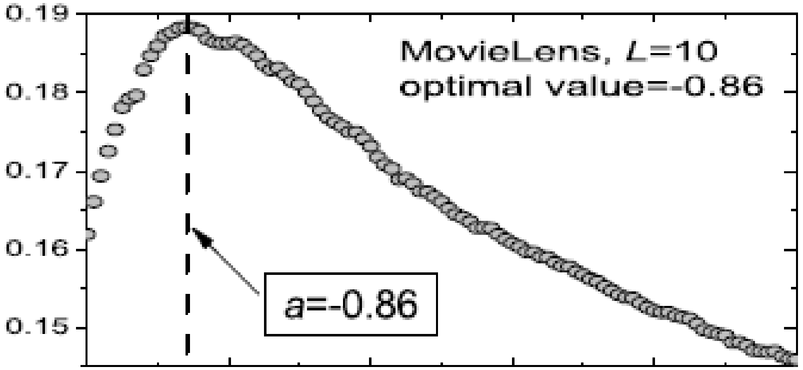
T. Zhou *et al.*, EPL 81 (2008) 58004

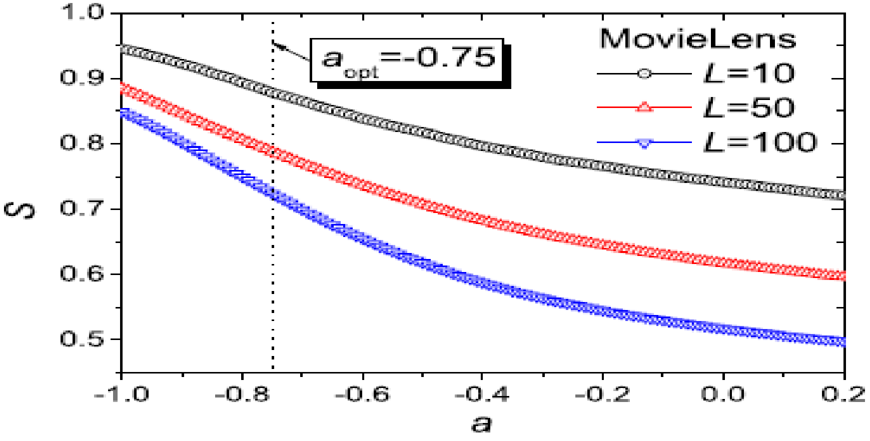
Improved Algorithm II: Depressing Higher-Order Correlation to Eliminate redundancy



$$\vec{f}' = (W + aW^2)\vec{f}$$



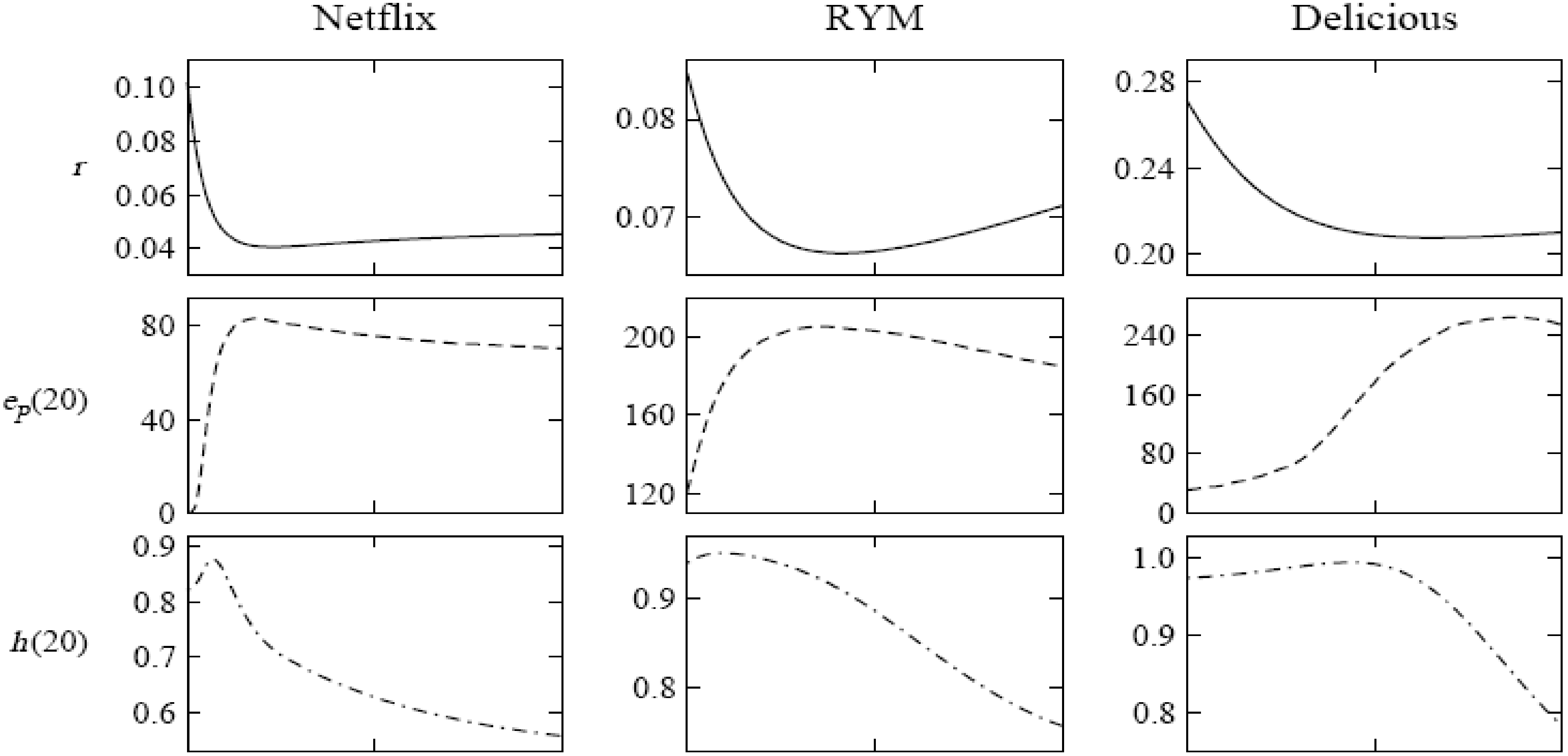




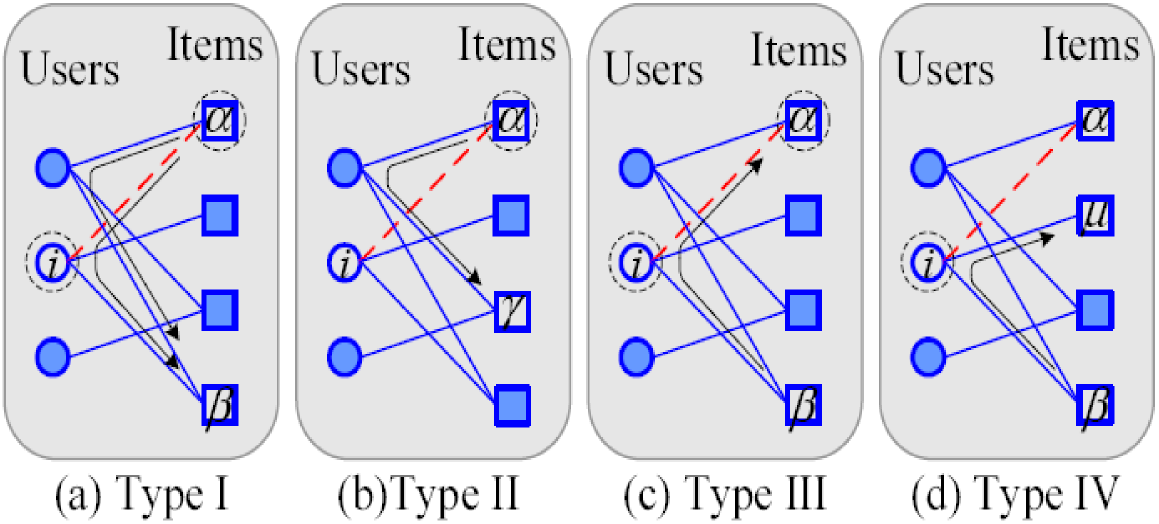
T. Zhou *et al.*, NJP 11 (2009) 123008

Hybrid Method: $w_{ij} = \frac{1}{d_i^{1-\lambda} d_j^{\lambda}} \sum_{l=1}^{N} \frac{a_{li} a_{lj}}{k_l}$

Data	N	M	Sparsity	GRM	CF	MD	HC	Hybrid	λ^*
Netflix	10000	6000	0.0117	0.057	0.051	0.045	0.102	0.040	0.23
RYM	33786	5381	0.00337	0.119	0.087	0.071	0.085	0.066	0.41
Delicious	10000	232657	0.00053	0.314	0.223	0.210	0.271	0.207	0.66



Adaptive Algorithm: Real-Time Response to Changing Data



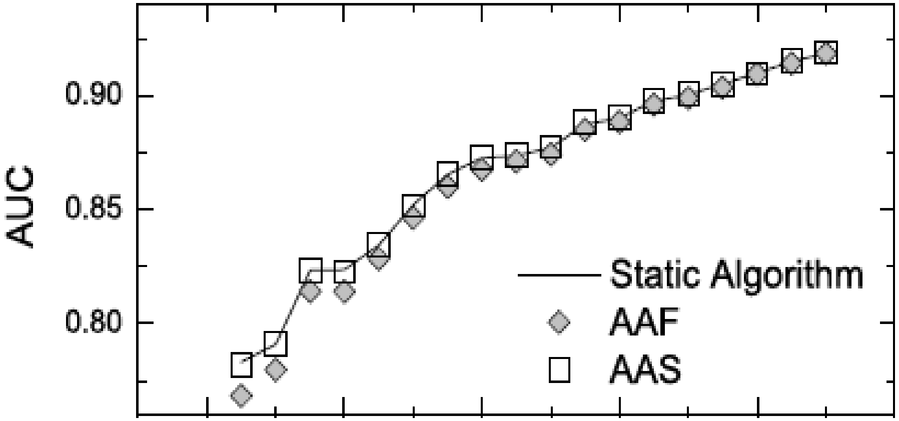
(a) Type I (b)Type II (c) Type III (d) Type IV

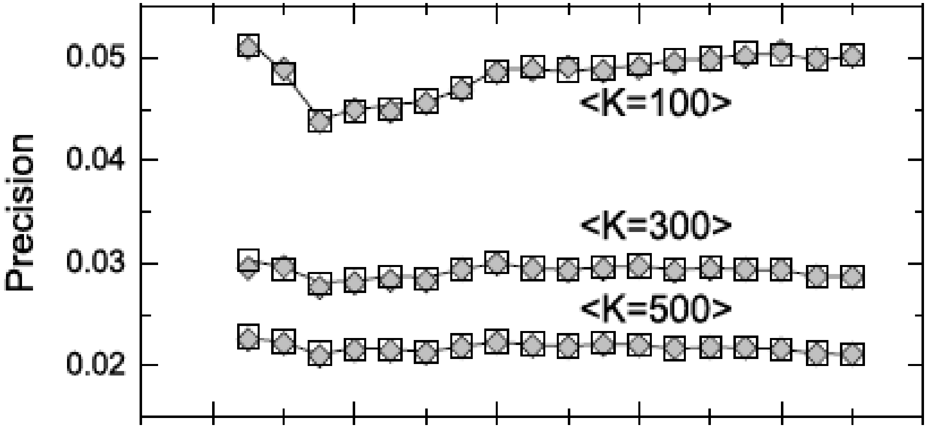
$$\delta_{\beta\alpha}^{(l+1)} = \frac{-1}{k_\alpha^{(l+1)}} m_{\beta\alpha}^{(l)} + \frac{1}{k_\alpha^{(l+1)} k_i^{(l+1)}}.$$

$$\delta_{\gamma\alpha}^{(l+1)} = \frac{-1}{k_\alpha^{(l+1)}} m_{\gamma\alpha}^{(l)}.$$

$$\delta_{\alpha\beta}^{(l+1)} = \frac{1}{k_\beta^{(l+1)} k_i^{(l+1)}}.$$

$$\delta_{\mu\beta}^{(l+1)} = \frac{-1}{k_\beta^{(l+1)} k_i^{(l)} k_i^{(l+1)}}.$$





C.-H. Jin, J.-G. Liu, Y.-C. Zhang, T. Zhou, arXiv: 0911.4910

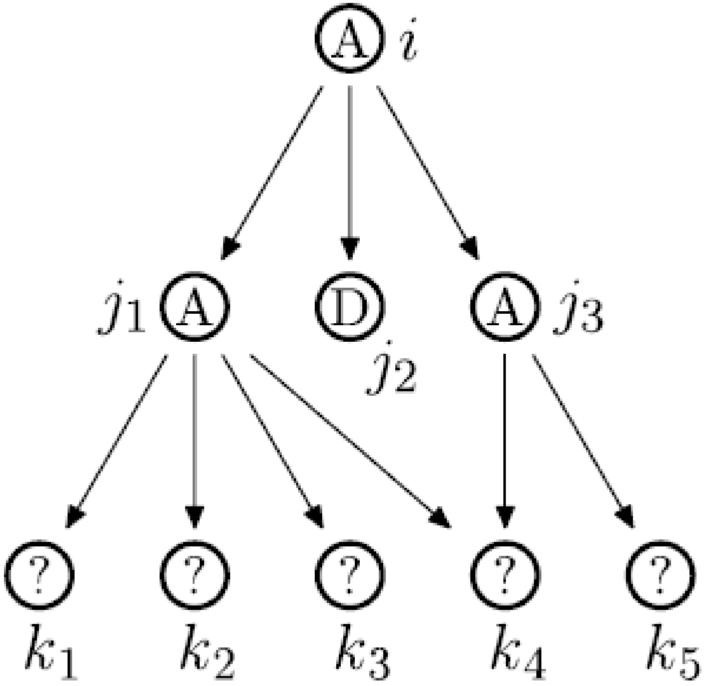
Social Filtering

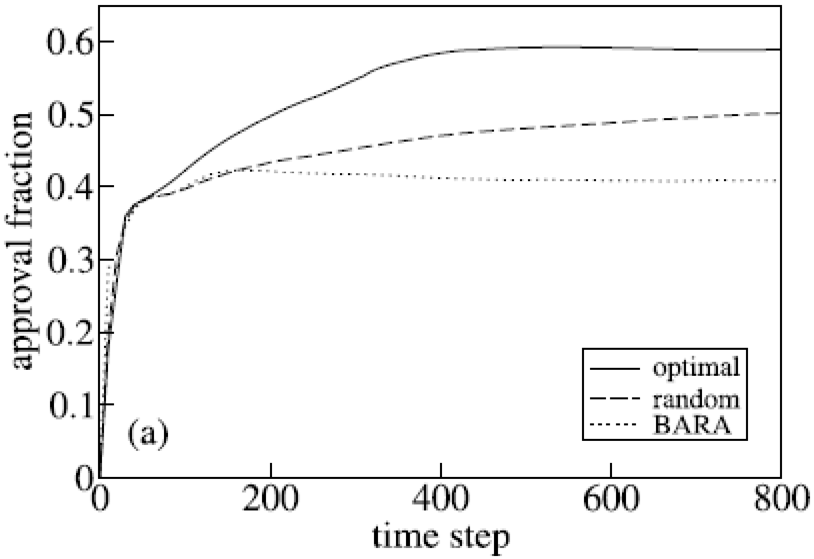
- Most of the current recommender systems still adopt the centralized way where the systems analyze the data and decide which should be recommended to whom.
- Such a paradigm is challenged by the fact that the social influence plays a more important role than similarity of past activities, and the recommendations made by a system is less preferred than by a friend.
- The fast development of online social communities make the social filtering techniques a promising tool in the next-generation recommender systems.
- Users could receive additional value by social recommendations.

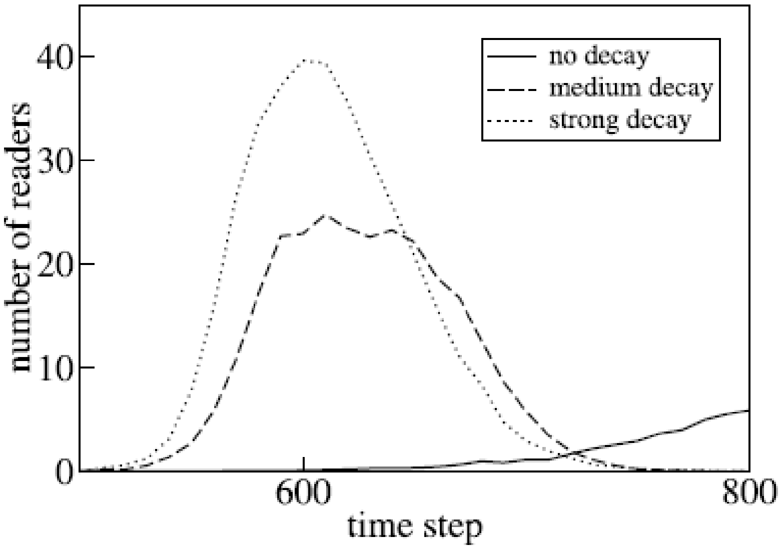
Lai *et al*., Proc. 5th Intl. Conf. E-Commerce (ACM Press, 2003)
Pon *et al*., ACM SIGKDD 2007; Ahn *et al*., WWW 2007

Adaptive Model for News Recommendation

A leader-follower network with identical in-degree is built, where news can only flow from leaders to followers.



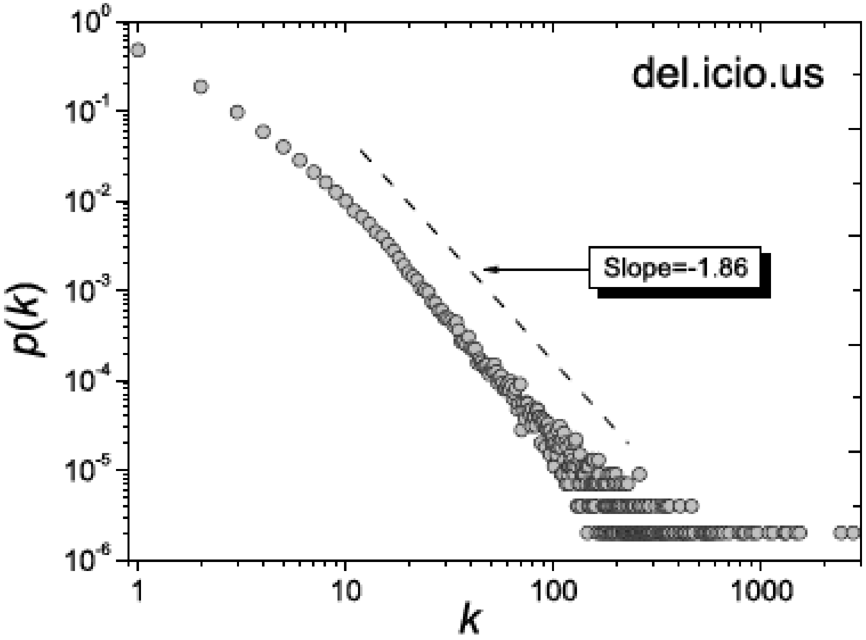




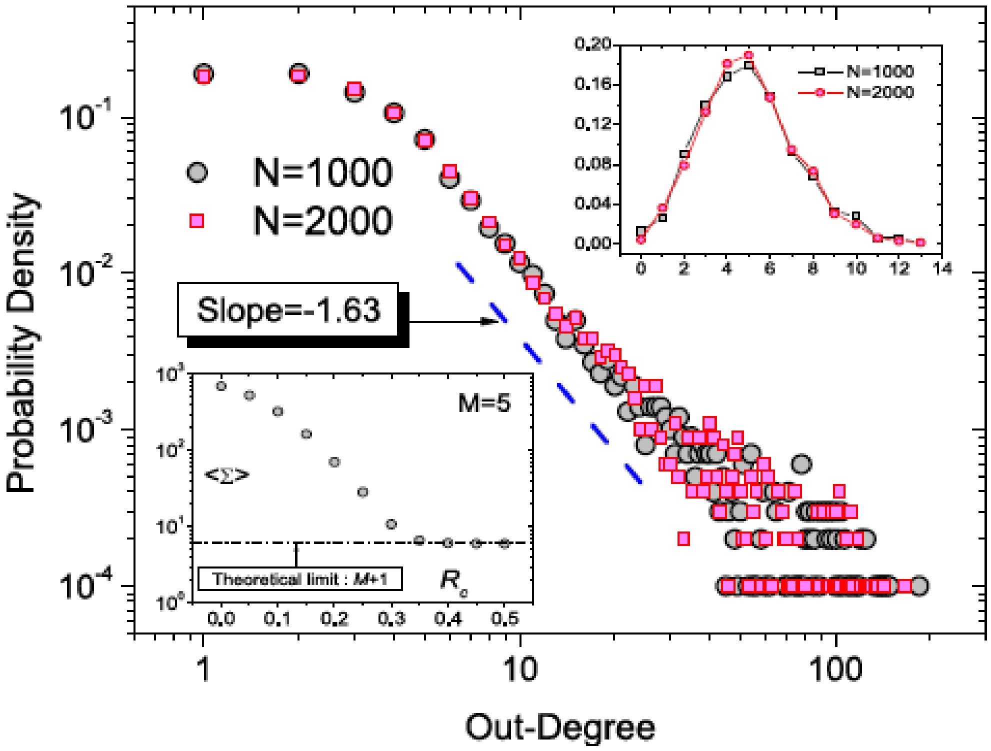
M. Medo, Y.-C. Zhang, T. Zhou, EPL 88 (2009) 38005

Scale-Free Leadership Structure

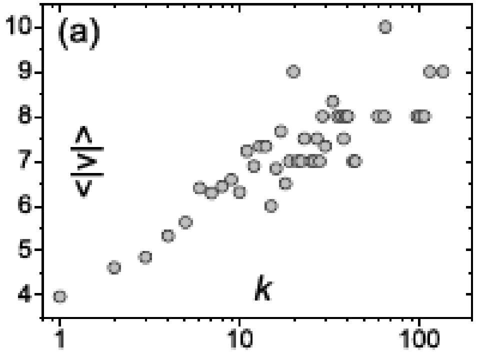
Real Distribution

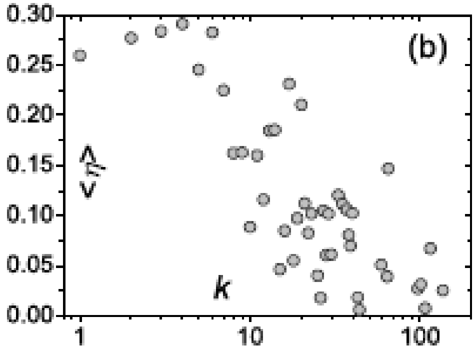


Result of the Model



Factor-I: Wider Scope of Interests





Factor-II: Better judgment

T. Zhou, et al., Working Paper

Look into the future

What about the next-generation information filtering techniques ?

- Static -> Adaptive
- Centralized -> Decentralized
- Design Algorithm -> Guide Users
- Personalized Recommendation -> Personalized Algorithms
- Accuracy Only -> Comprehensive Evaluation
- Spam-Sensitive -> Spam-Robust
- Predict -> Drive and Control

10 Selected Publications During the PhD Study

T. Zhou, J. Ren, M. Medo, Y.-C. Zhang, “Bipartite network projection and personal recommendation”, *Phys. Rev. E* **76**, 046115 (2007)

Y.-C. Zhang, M. Medo, J. Ren, **T. Zhou**, T. Li, F. Yang, “Recommendation model based on opinion diffusion”, *EPL* **80**, 68003 (2007)

T. Zhou, L.-L. Jiang, R.-Q. Su, Y.-C. Zhang, “Effect of initial configuration on network-based recommendation”, *EPL* **81**, 58004 (2008)

T. Zhou, H.-A. T. Kiet, B. J. Kim, B.-H. Wang, P. Holme, “Role of Activity in Human Dynamics”, *EPL* 82, 28002 (2008)

J. Ren, **T. Zhou**, Y.-C. Zhang, “Information Filtering via Self-Consistent Refinement”, *EPL* **82**, 58007 (2008)

T. Zhou, L. Lü, Y.-C. Zhang, “Predicting missing links via local information”, *Eur. Phys. J. B* **71**, 623 (2009)

M. Medo, Y.-C. Zhang, **T. Zhou**, “Adaptive model for recommendation of news”, *EPL* **88**, 38005 (2009)

T. Zhou, R.-Q. Su, R.-R. Liu, L.-L. Jiang, B.-H. Wang, Y.-C. Zhang, “Accurate and diverse recommendations via eliminating redundant correlations”, New J. Phys. **11**, 123008 (2009)

L. Lü, **T. Zhou**, “Link Prediction in Weighted Networks: The Role of Weak Ties”, *EPL* **89**, 18001 (2010)

T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, Y.-C. Zhang, “Solving the apparent diversity-accuracy dilemma of recommender systems”, *PNAS* (March 2010)

10 Highly Cited Papers

G. Yan, **T. Zhou**, B. Hu, Z.-Q. Fu, B.-H. Wang, “Efficient Routing on Complex Networks”, *Phys. Rev. E* **73**: 046108 (2006) **Time Cited 85**

T. Zhou, G. Yan, B.-H. Wang, “Maximal planar networks with large clustering coefficient and power-law degree distribution”, *Phys. Rev. E* **71**, 046141 (2005) **Time Cited 80**

W.-X. Wang, B.-H. Wang, C.-Y. Yin, Y.-B. Xie, **T. Zhou**, “Traffic dynamcis based on local routing protocol on scale-free networks“, *Phys. Rev. E* **73**: 026111 (2006) **Time Cited 66**

G. Yan, **T. Zhou,** J. Wang, Z.-Q. Fu, B.-H. Wang “Epidemic spread in weighted sacle-free networks”, *Chin. Phys. Lett.* **22**, 510-513 (2005) **Time Cited 53**

M. Zhao, **T. Zhou**, B.-H. Wang, W.-X. Wang, “Enhance synchronizability by structural perturbations”, *Phys. Rev. E* **72**, 057102 (2005) **Time Cited 43**

C.-Y. Yin, B.-H. Wang, W.-X. Wang, **T. Zhou**, H.-J. Yang, “Efficient routing on scale-free networks based on local information”, *Phys. Lett. A* **351**, 220 (2006) **Time Cited 38**

P.-P. Zhang, K. Chen, Y. He, **T. Zhou**, B.-B. Su, Y.-D. Jin, H. Chang, Y.-P. Zhou, L.-C. Sun, B.-H. Wang, D.-R. He, “Model and empirical study on some collaboration networks”, *Physica A* **360**, 599 (2006) **Time Cited 36**

W.-X. Wang, B. Hu, **T. Zhou**, B.-H. Wang, Y.-B. Xie, “A mutual selection model for weighted networks”, *Phys. Rev. E* **72**, 046140 (2005) **Time Cited 33**

T. Zhou, J.-G. Liu, W.-J. Bai, G.-R. Chen, B.-H. Wang, “Behaviors of susceptible-infected epidemics on scale-free networks with identical infectivity”, *Phys. Rev. E* **74**, 056109 (2006) **Time Cited 32**

T. Zhou, B.-H. Wang, “Catastrophes in scale-free networks”, *Chin. Phys. Lett.* **22**, 1072 (2005) **Time Cited 29**

Statistics on All Publications: Total Citations=1082,H-Index=18

Full Publications

1. Physical Review E (28)

   71 (2005) 046135; 71 (2005) 046141; 72 (2005) 016702; 72 (2005) 046139; 72 (2005) 046140; 72 (2005) 057102; 72 (2005) 066702; 73 (2006) 026111; 73 (2006) 037101; 73 (2006) 046108; 73 (2006) 058102; 74 (2006) 046103; 74 (2006) 056109; 75 (2007) 021102; 75 (2007) 036106; 76 (2007) 037102; 76 (2007) 046115; 76 (2007) 057103; 76 (2007) 061903; 77 (2008) 021920; 78 (2008) 066109; 79 (2009) 016113; 79 (2009) 026113; 79 (2009) 052102; 80 (2009) 017101; 80 (2009) 031144; 80 (2009) 046108; 80 (2009) 046122.

2. Physica A (25)

   354 (2005) 505; 360 (2006) 599; 367 (2006) 337; 367 (2006) 613; 368 (2006) 607; 371 (2006) 773; 371 (2006) 814; 371 (2006) 861; 374 (2006) 864; 375 (2007) 355; 375 (2007) 687; 375 (2007) 709; 376 (2007) 215; 384 (2007) 656; 387 (2008) 1683; 387 (2008) 3025; 387 (2008) 6391; 388 (2009) 462; 388 (2009) 1237; 388 (2009) 1713; 388 (2009) 2949; 388 (2009) 4867; 389 (2010) 179; 389 (2010) 881; 389 (2010) 1259.

3. Europhysics Letters (11)

   80 (2007) 68003; 81 (2008) 58004; 82 (2008) 28002; 82 (2008) 58007; 83 (2008) 40003; 86 (2009) 40011; 87 (2009) 68001; 87 (2009) 68002; 88 (2009) 38005; 88 (2009) 68008; 89 (2010) 18001.

4. European Physical Journal B (7)

   47 (2005) 587; 53 (2006) 375; 60 (2007) 89; 62 (2008) 101; 65 (2008) 251; 66 (2008) 557; 71 (2009) 623.

5. Physics Letters A (6)

   351 (2006) 220; 362 (2007) 115; 364 (2007) 189; 366 (2007) 14; 368 (2007) 431; 372 (2008) 1725.

6. New Journal of Physics (5)

   10 (2008) 023006; 10 (2008) 073010; 10 (2008) 123027; 11 (2009) 103001; 11 (2009) 123008.

7. Others (5)

   Prog. Nat. Sci. 16 (2006) 252; Chaos, Solitons & Fractals 31 (2007) 772; Ars Combinatoria 83 (2007) 289; IEEE Circuits and Systems Magazine (Feature Article) 2008(3) 67; PNAS (to be published in March 2010).

Thanks for your attention!