Collaboration Over Time: Characterizing and Modeling Network Evolution

Jian Huang[1], Ziming Zhuang[1], Jia Li[2], C. Lee Giles[1]

[1]College of Information Sciences and Technology
[2]Department of Statistics
Pennsylvania State University

Contact: jhuang@ist.psu.edu

Outline

- Introduction and related work
- Characterizing network level collaboration
- Collaboration at the community level
- Collaboration between individuals
  - Models and experiments
- Conclusion and future work

Introduction

- Sociological and computational social network analysis (SNA)
  - Stanley Milgram's *small world experiment* and *six degrees of separation* [Milgram, 1967]
  - Granularities of analysis
  - Scale of network and methodology
- Scientific collaboration network
  - Citation network
  - Coauthorship network
    - Stringent definition: formal documentation of acquaintance and collaboration
    - Implicit information about significance in the collaboration
    - Of interest to SNA researchers and domain practitioners

S. Milgram. The small-world problem. *Psychology Today*, 1:61–67, 1967.

Related Work

- Some recent SNA studies on specific types of SN's
  - Epidemiological network [Liljeros, 2003]
  - Online newsgroup [Borgs, 2004]
  - Blogs and photo sharing websites [Kumar, 2006]
- Studies of coauthorship networks
  - Macroscopic network static properties [Newman, 2001, 2004]
  - Static and dynamic properties [Barabasi et al, 2001]
- Studies of network evolution
  - Network evolution [Doreian et al, 1997; Leskovec et al, 2005]
  - Community evolution [Backstrom et al, 2006; Kumar et al, 2006]
- Studies of network modeling
  - Preferential attachment [Barabási et al, 1999]
  - Link prediction problem [Liben-Novell et al, 2007]

Problems

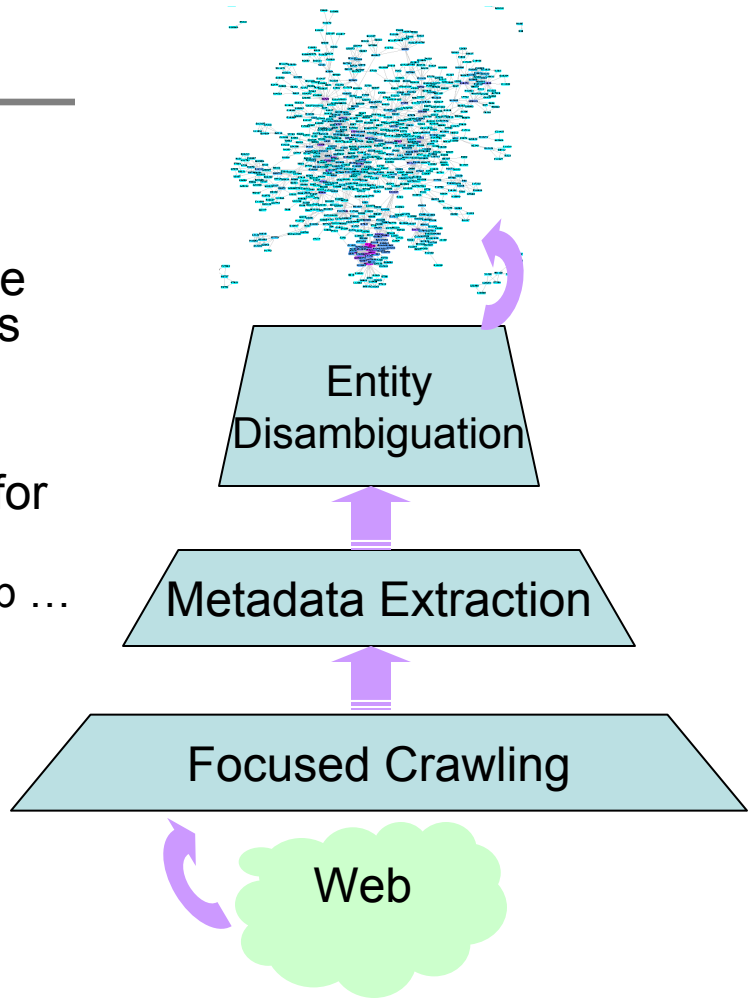
- Study a social network comprehensively
  - Levels of analysis
  - Static vs dynamic
- Generation of high quality large-scale social network data
- Model underlying network evolution mechanism

Contributions

- A comprehensive and focused study of the large-scale scientific network in Computer Science
  - Three levels of analysis: network, community and individual
  - Six topical datasets
- A stochastic model is proposed for characterizing and predicting individual collaboration
  - Exploits neighborhood graph information
- Novel workflow for large-scale SNA
  - Focused crawling → metadata extraction → name disambiguation

Data Collection

- CiteSeer digital library (http://citeseer.ist.psu.edu) [Giles et al, 1998]
  - A popular Computer and Information Science digital library containing over 700,000 papers crawled from the web.
- Automatic metadata extraction [Han et al, 2003]
  - SVM-based Information Extraction (IE) tool for identifying and tagging paper header data
    - Titles, authors, affiliations, emails, year of pub …
- Author name disambiguation [Huang et al, 2006]
  - Distinguish namesakes using extracted metadata
  - Accurately attribute authorship
  - 283,174 authors and 451,305 papers published between 1980 and 2005.



C.L. Giles, K. Bollacker, S. Lawrence, "CiteSeer: An Automatic Citation Indexing System," DL'98 Digital Libraries, 3rd ACM Conference on Digital Libraries, pp. 89-98, 1998.

H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2003.

J. Huang, S. Ertekin, and C. L. Giles. Efficient name disambiguation for large scale databases. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, 2006.

Topical Datasets

- Topical datasets preparation
  - Assign six topics to papers according to the top venues in the computer science conference ranking
    - *http://www-static.cc.gatech.edu/~guofei/CS ConfRank.htm*
  - ~1,700 papers were selected for each topic
    - 11,820 authors, 10,195 papers

Topic	Representative venues	#Authors	#Papers
ai	AAAI, IJCAI, NIPS, KDD ...	2,105	1,666
app	WWW, SIGGRAPH, SIGIR ...	2,087	1,548
arch	DAC, MICRO, HPCA ...	2,589	1,740
db	SIGMOD ,VLDB, ICDE ...	1,559	1,755
system	SIGCOMM, PODC, SOSP ...	1,733	1,785
theory	STOC, FOCS, COLT ...	1,747	1,701

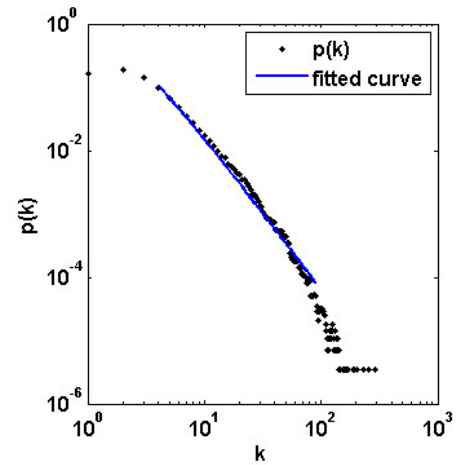
Degree Distribution

- Fitted degree distribution

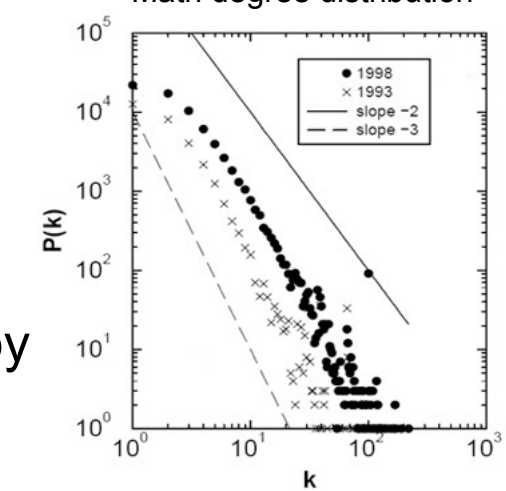
$$p(k) \sim (0.8 + k)^{-2.45}$$

  - Scale-free degree distribution
- Similarities with other coauthorship networks
  - Preferential attachment
  - Much lower counts in small k
    - Loners are uncommon
  - Exponential cutoffs in large k
    - Extremely popular people are rare
  - γ>2 indicates the network is dominated by the `little people'
  - Gradual aging effect [Albert et al, 2002]

CiteSeer degree distribution



Math degree distribution



Albert, R., A. Barabási, 2002. Statistical mechanics of complex networks. Review of Modern Physics 74, 47-97.

A. Barabási, H. Jeong, E. Ravasz, Z. Neda, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Cond-mat/0104162*, 2001.

Comparison of Statistical Network Properties in Coauthorship Networks

- Computer Science
  - NCSTRL (preprints)
  - CiteSeer
    - 2005 snapshot
- Biology
  - MEDLINE and NeuroSci.
- Mathematics
- High energy physics - SPIRES

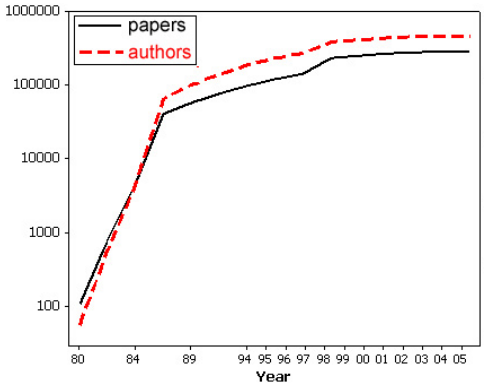
Network	CiteSeer	NCSTRL	Maths	SPIRES	MEDLINE	NeuroSci.
Reference	This paper	[23] [21] [22]	[4]	[23] [21] [22]	[23] [21] [22]	[4]
#Papers	451,305	13,169	70,901	66,652	2,163,923	210,750
#Authors (N)	283,174	11,994	70,975	56,627	1,520,251	209,293
Mean papers/author	4.06	2.55	-	11.6	6.4	-
Mean authors/paper	2.55	2.22	-	8.96	3.75	-
Avg. degree ($< k >$)	5.56	3.59	3.9	173	18.1	11.54
Exponent (γ)	2.45	1.3	2.5	1.2	2.5	2.1
κ	291	10.7	120	1200	5800	400
Diameter (d)	26	31	-	19	24	-
Avg. path length (l_{real})	7.1	9.7	9.5	4.0	4.6	6
Avg. path length, random (l_{rand})	12.14	7.34	8.2	2.12	4.91	5.01
Cluster coefficient (C)	0.634	0.496	0.59	0.726	0.066	0.76
Cluster coefficient, random (C_{rand})	7.8×10^{-6}	3×10^{-4}	5.4×10^{-5}	3×10^{-3}	1.1×10^{-5}	5.5×10^{-5}
Giant Comp. Percentage	65.9%	57.2%	70%	88.7%	92.6%	91%

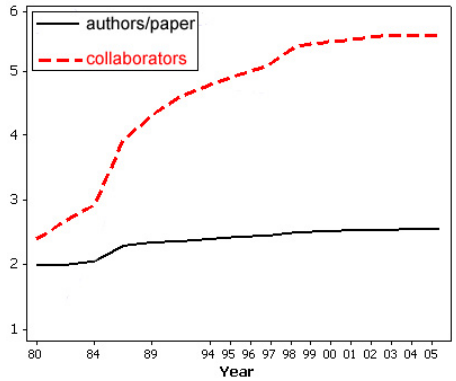
A. Barabási, H. Jeong, E. Ravasz, Z. Neda, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Cond-mat/0104162*, 2001.

M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98:404–409, 2001.

A Growing Small World - CiteSeer

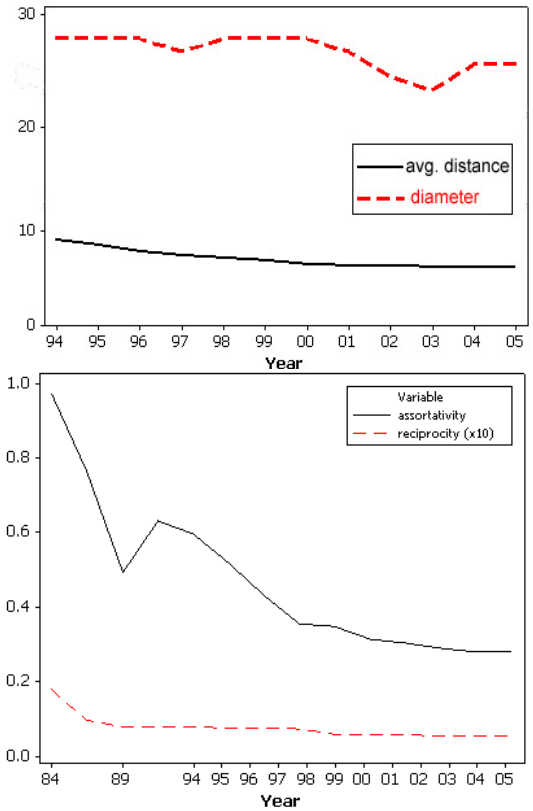
- Growing network of authors and papers manifesting the *small world* phenomenon
  - Scale-free degree distribution
  - High local clustering
    - Existence of small cliques
- Increasing number of collaborators
- Increasing number of authors per paper
  - 2 in 1980 vs 2.55 in 2005





A Growing Small World (cont'd)

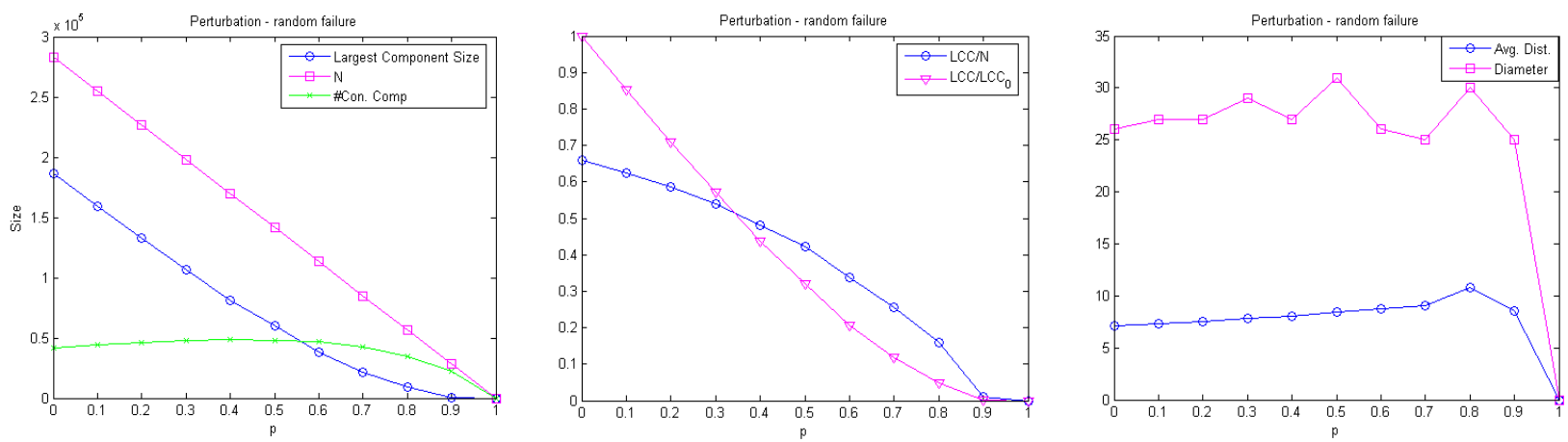
- Average distance shrinks
  - 9.3 in 1994 vs 7.1 in 2005
- Diameter fluctuates and gradually decreases to 26
  - Dips signify merging of giant components
- Shrinking assortativity [Newman, 2003]
  - Assortative degree mixing
    - Nodes tend to associate with others of similar degrees
- Low reciprocity value
  - 0.0055 vs 0.84 in Yahoo! 360 [Kumar et al, 2006]



Metric	Biology	**Comp. Sci.**	Maths	Physics
Assortativity	0.13	**0.28**	0.12	0.36

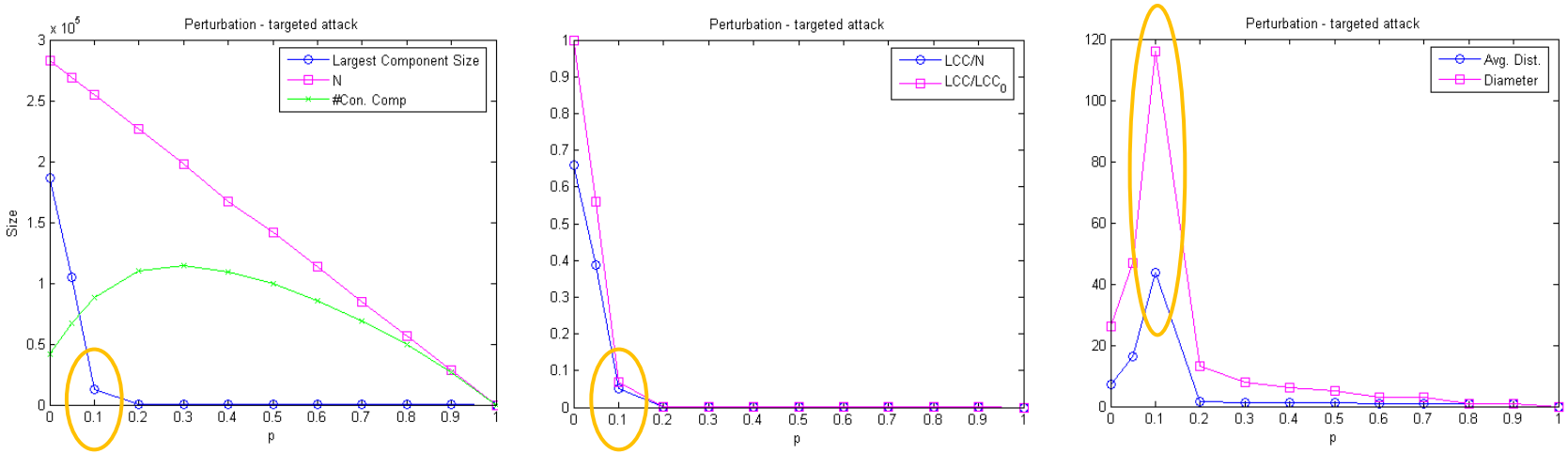
Perturbation Behavior

- Random node removal
  - Infinite systems with γ<3 do not break down under random failure [Cohen et al, 2000]
  - Coauthorship networks are resilient to random node removal
    - Tolerance to inaccurate metadata extraction, name disambiguation, aging



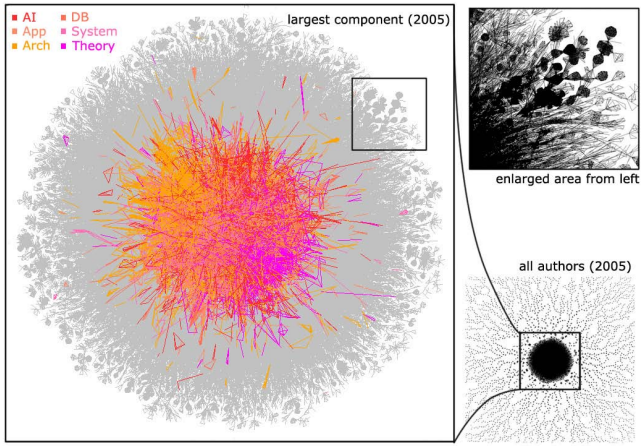
Sensitivity to High Degree Nodes

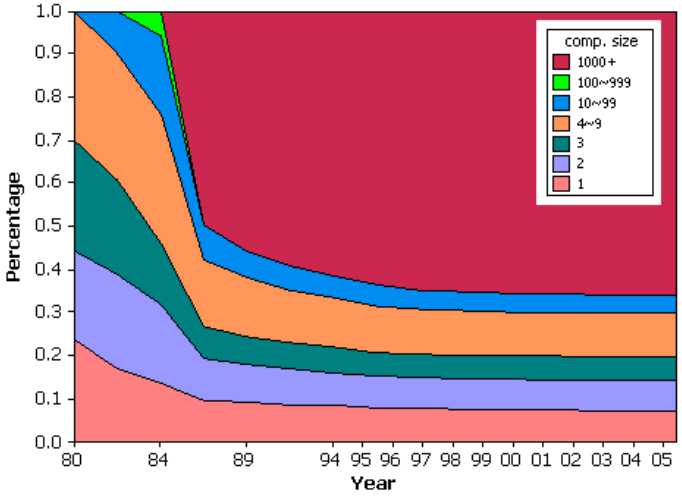
- Targeted node removal
  - Coauthorship networks *are* sensitive to targeted removal of high degree nodes
    - Removing 10% top high degree nodes can break down the network



Collaboration in the Community Level in CiteSeer

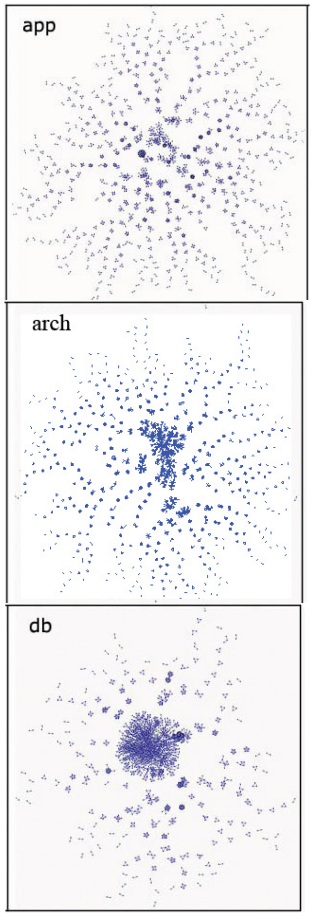
- Largest connected component
  - Visually forms the *core* of the collaboration network
  - Accounts for 66% of nodes in 2005
- Tiers of connected components [Kumar et al, 2006]
  - Giant components
    - Increased by 10% in between 1989 and 1996
    - Plateau at 68%
  - Middle regions
    - Components with 100~999 nodes were absorbed by the giant components
  - Singletons
    - Loners in the collaboration network
    - 7.2% authors, stable over time



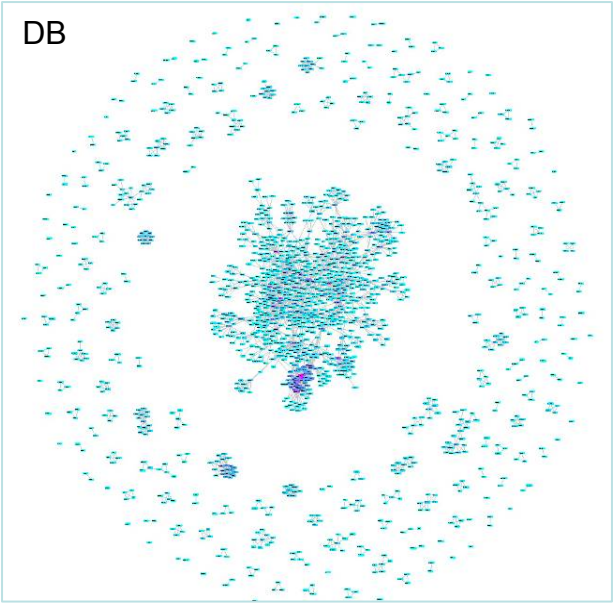


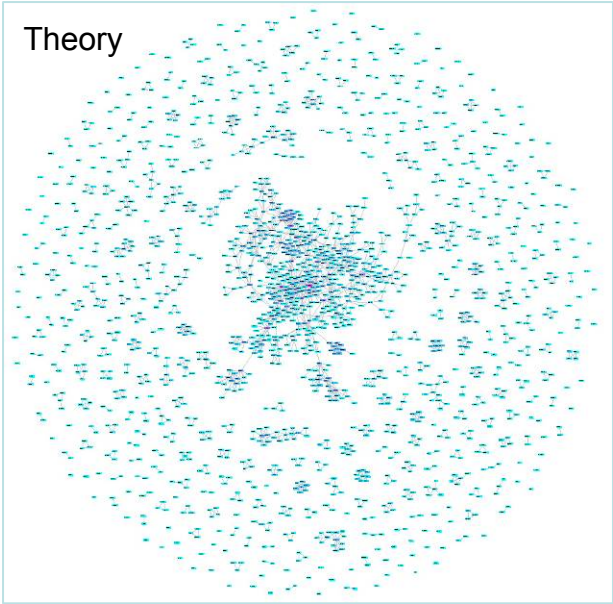
Distinctive Patterns in Topical Communities

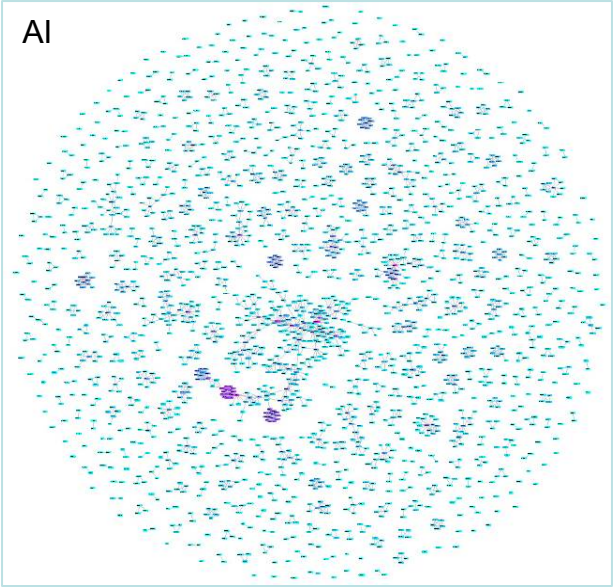
- Database community
  - Most connected, huge largest component (56% nodes), most collaborators
  - More `equal': high reciprocity
- AI community
  - Low reciprocity and high assortativity
- Theory community
  - Most singletons, least collaborators
- Application community
  - Fragmented by topics, no giant components

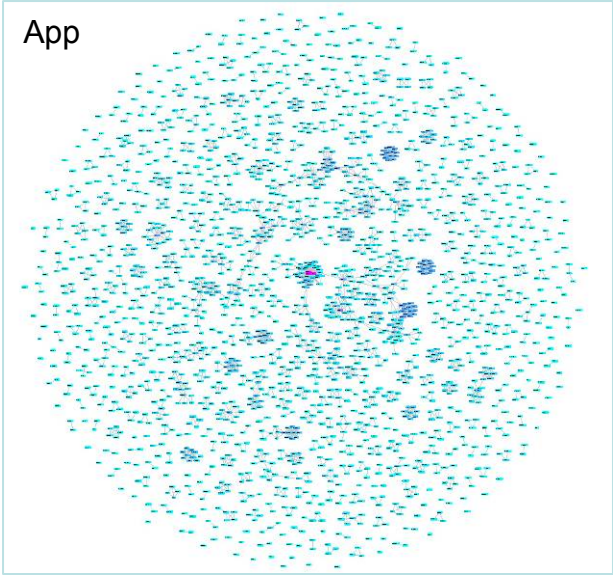


Dataset	Connectivity		Component structure			Collaboration patterns			
	avg. dist.	betweenness	#comp.	largest comp.	singleton	papers/author	collaborator	reciprocity	assortativity
ai	5.5	0.5×10^{-2}	574	11.1%	6.6%	1.9	2.7	$\mathbf{1.7 \times 10^{-3}}$	**0.61**
app	3.5	0.1×10^{-2}	593	**4.9%**	6.3%	2.0	3.0	2.8×10^{-3}	0.29
arch	**8.3**	1.9×10^{-2}	603	21.1%	4.1%	1.9	3.4	3.6×10^{-3}	0.44
db	5.3	$\mathbf{5.9 \times 10^{-2}}$	**205**	**55.9%**	**2.9%**	**3.6**	**4.7**	$\mathbf{7.3 \times 10^{-3}}$	0.35
system	6.0	1.7×10^{-2}	415	24.2%	4.7%	2.6	3.0	1.9×10^{-3}	0.25
theory	6.5	1.9×10^{-2}	461	31.8%	**8.4%**	2.3	2.8	5.2×10^{-3}	0.37


DB


Theory


AI


App

Modeling Individual Collaboration

- Research Question:
  - Given a pair of authors v_i and v_j with existing collaboration $e_{i,j}$, what is the probability that they will collaborate k times within the next Δt time interval?
- A related research question:
  - How does a new collaboration emerge in the network?
    - Preferential linking
    - Disassortative degree mixing
    - …
- Modeling microscopic collaborations
  - Collaboration networks are large but sparse, $\|E\| << \|V\|^2$
  - Predict using the local neighborhood information without resorting to the rest of the network

Stochastic Poisson + Optimization Tree (SPOT)

- SPOT predicts and models the cumulative collaboration rate
  - Uses local neighborhood information
  - Formulates cumulative collaboration as a stochastic Poisson process
  - Optimizes the model by maximizing log-likelihood in a non-parametric manner

Notation

- A series of discrete time snap shots
  $\{\mathcal{G}(t)\}(t = 0, 1, 2, ...)$
- Edge $e_{i,j}(t)$ denote an existing collaboration between nodes v_i and v_j
- Weight $N_{i,j}(t)$ is associated with edge $e_{i,j}(t)$
  - $N(t)$ is discrete and represents the cumulative number of collaborations
    - $N(t)$ is a counting process

Review of Poisson Process

- Poisson process [Ross, 2006]
  - A prominent model for counting processes
  - A pure-birth stochastic process
- Time invariant (homogeneous) Poisson process
  - Rate parameter (intensity): λ
    - Expected number of events that occur per unit time
  - Number of events k in time interval $(t, t+\Delta t]$
  - $Pr_\lambda\{N(t+\Delta t) - N(t) = k\} = \dfrac{e^{-\lambda\Delta t}(\lambda\Delta t)^k}{k!}$
- Non-homogeneous Poisson process
  - The stationary Poisson rate condition does not in general hold in collaboration
    - Various influencing factors such as collaborators, institutions, network structural characteristics, etc
  - Rate parameter can change over time

S. M. Ross. *Introduction to Probability Models*. Academic Press, 2006.

Determine Collaboration Rate

- Leveraging local graph structure to determine the collaboration rate
  - $\mathcal{G}(e_{i,j}(t))$: neighborhood of an existing edge $e_{i,j}(t)$
    - Author pairs v_i , v_j , their coauthors and associated edges
  - Extracted feature vector at time *t*: $\mathbf{a} = (a_1, ..., a_p)$ for characterizing the neighborhood
  - Variable collaboration rate: $\lambda(e_{i,j}(t)) = f(\mathbf{a}_{i,j}(t))$
- Notational shorthand for Poisson probability:

$$Pr_{\lambda(e_{i,j}(t))}\{k_{i,j}\} = Pr_{\lambda(e_{i,j}(t))}\{N(t+\Delta t) - N(t) = k_{i,j} | e_{i,j}(t_0)\}$$

Find Optimal Function *f*

- Log-likelihood maximization

$$
\begin{aligned}
f^* &= \arg\max_f \sum_{e_{i,j}(t)} \log Pr_{\lambda(e_{i,j}(t))}\{k_{i,j}\} \\
&= \arg\max_f \sum_{e_{i,j}(t)} [k_{i,j} \log(\lambda(e_{i,j}(t))) - \lambda(e_{i,j}(t)) - \log k_{i,j}!] \\
&= \arg\max_f \sum_{e_{i,j}(t)} [k_{i,j} \log(f(\mathbf{a}_{i,j}(t))) - f(\mathbf{a}_{i,j}(t))]
\end{aligned}
$$

  - Note: factorial terms not related to the maximization are dropped

Nonparametric Optimization

- *Optimization Tree*
  - Similar to decision tree methods, e.g. CART [Breiman et al, 1984]
  - Grow a tree in a top-down and best-first fashion for nonparametric optimization
    - A splitting rule $a_j \leq h$ in the internal node defines a hyperplane orthogonal to the *j*-th axis in the feature space
- Deriving the optimal decision rule
  - Maximize the sum of log-likelihoods in the two sub-regions

$$(R_1, R_2)^* = \arg\max_{R_1 \cup R_2 = R} \sum_{e_{i,j}(t) \in R_1} [k_{i,j} \log f(\mathbf{a}_{i,j}(t)) - f(\mathbf{a}_{i,j}(t))] + \sum_{e_{i,j}(t) \in R_2} [k_{i,j} \log f(\mathbf{a}_{i,j}(t)) - f(\mathbf{a}_{i,j}(t))]$$

Splitting Rule

- Optimal estimate of the rate in a region
  - Maximizes the log-likelihood

$$\lambda_R = \sum_{e_{i,j}(t) \in R} k_{i,j} \Big/ \parallel R \parallel$$

- Optimal criterion for splitting:

$$\begin{aligned}
(R_1, R_2)^* &= \arg\max_{R_1 \cup R_2 = R} [\sum_{e_{i,j}(t) \in R_1} k_{i,j} \log \lambda_{R_1} + \sum_{e_{i,j}(t) \in R_2} k_{i,j} \log \lambda_{R_2} - \lambda_{R_1} \parallel R_1 \parallel - \lambda_{R_2} \parallel R_2 \parallel] \\
&= \arg\max_{R_1 \cup R_2 = R} [\sum_{R_1} k_{i,j} \log \lambda_{R_1} + \sum_{R_2} k_{i,j} \log \lambda_{R_2} - \sum_{e_{i,j}(t) \in R} k_{i,j}] \\
&= \arg\max_{R_1 \cup R_2 = R} [\sum_{R_1} k_{i,j} \log \lambda_{R_1} + \sum_{R_2} k_{i,j} \log \lambda_{R_2}]
\end{aligned}$$

  - The internal node, which when split causes the maximum increment in log-likelihood, is selected for splitting

Remarks

- Differences with decision trees
  - Optimization tree aims to estimate the parameters in the distribution instead of solving a classification/regression problem
  - Functional form of the splitting rule is derived using likelihood maximization, different from the impurity functions (e.g. *Gini* index) used in decision trees
- Advantages similar to decision trees
  - Smoothes evolving data over time, dampening the effects of outliers in a region
  - Simple (piecewise) linear region boundaries
  - Efficient implementation
    - Sort instances by feature values beforehand and perform linear search per feature to find the maximum
- Entropy maximization (with proper normalization)

Experiment Setup

- Dataset
  - 105,122 collaboration records in 1997~2000 and 2001~2003 in the CiteSeer dataset
  - Split into 10 folds for cross validation
- Feature extraction
  - Three levels of features: individual, pairwise and neighborhood
  - Informative features including network statistics and collaboration patterns

Level	Feature
Individual	- Total number of publications of author v_i - Total number of publications of author v_j
Pairwise	- Whether author v_i and v_j belong to the same affiliation - Total number of collaboration by time t - Total number of shared collaborators by author v_i and v_j - Collaboration rate between v_i and v_j in the previous snapshot
Neighbor-hood	- The fraction of shared collaborators to all the collaborators of author v_i times the fraction of that of author v_j - The fraction of publication with shared collaborators to all publication of v_i times the fraction of that of v_j

Evaluation Metrics

- Sample correlation coefficient

$$r_{y\tilde{y}} = \frac{\sum_i (y_i - \overline{y})(\tilde{y}_i - \overline{\tilde{y}})}{(n-1)s_y s_{\tilde{y}}}$$

- Root mean squared error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_i err_i^2}{n}} = \sqrt{\frac{\sum_i (\tilde{y}_i - y_i)^2}{n}}$$

Empirical Results

- Stochastic Poisson + Optimization Tree (SPOT)
  - A 75-node tree was grown
  - 18.7% increase in correlation coefficient than SVR
  - 8.4% reduction in RMSE compared to SVR
  - Past collaboration rate showed poor performance

Method	Correlation Coefficient (r)	Root Mean Squared Error (RMSE)
Past Collaboration	0.227	4.74
SVR	0.673	1.53
SPOT	**0.782**	**1.42**

- Comparators
  - Support Vector Regression [Vapnik, 1999]
    - C=10 and γ=0.001 with RBF kernel
  - Past Collaboration [Newman, 2001]
- An example of the learned model for rate prediction

```
IF (total collaboration is between 9 and 18)
    AND (collaboration rate in the previous
         interval is lower than 7)
    AND (fraction of shared authors among
         author i's coauthors times that among
         author j's is lower than 0.025) ...
THEN IF (fraction of shared papers in author i's
         papers times that in author j's is
         lower than 0.0124)
     THEN predict rate=2.38
     ELSE predict rate=3.51
```

Conclusion and Future Directions

- Longitudinal studies of a Computer Science coauthorship network
  - Static and dynamic network properties
  - Community structural characteristics and distinct topical communities patterns
  - Stochastic model for cumulative collaboration
- Future work
  - Investigate interdisciplinary collaborations
  - Model new collaboration
    - E.g. focal closure and cyclic closure [Kossinets et al, 2006]

Acknowledgments

- This work was partially supported by the National Science Foundation (NSF).
- The authors would like to thank Reka Albert and Isaac Councill for insightful discussions.
- Visualization was performed with Large Graph Layout (LGL) and Prefuse.

References

R. Albert and A. Barabási. Statistical mechanics of complex networks. *Cond-mat/0106096v1*, 2001.

L. Amaral, A. Barthelemy, and H. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97:11149–11152, 2000.

L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan. Group Formation in Large Social Networks: Membership, Growth, and Evolution. In KDD, 2006.

A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

A. Barabási, H. Jeong, E. Ravasz, Z. Neda, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Cond-mat/0104162*, 2001.

C. Borgs, J. Chayes, M. Mahdian, and A. Saberi. Exploring the community structure of newsgroups. In *Proc. of 10th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, 2004.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

C. Chang and C. Lin. *LIBSVM: a library for support vector machines*, 2001.

P. Doreian and F. N. Stokman, editors. *Evolution of Social Networks*. Gordon and Breach, New York, 1997.

E. Elmacioglu and D. Lee. On six degrees of separation in DBLP-DB and more. *ACM SIGMOD Record*, 34:33–40, 2005.

H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2003.

J. Huang, S. Ertekin, and C. L. Giles. Efficient name disambiguation for large scale databases. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, 2006.

G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 331:88–90, 2006.

R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proc. of the 12th ACM International Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 611–617, 2006.

J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of 11th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, pages 177–187, 2005.

D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

References

L. Licamele and L. Getoor. Social capital in friendship-event networks. In *Proceedings of Sixth IEEE International Conference on Data Mining (ICDM)*, pages 959–964, 2006.

F. Liljeros, C. Edling, and L. Amaral. Sexual networks: implications for the transmission of sexually transmitted infections. *Microbes and Infection*, 5:189–196, 2003.

S. Milgram. The small-world problem. *Psychology Today*, 1:61–67, 1967.

M. E. J. Newman. Who is the best connected scientist? a study of scientific coauthorship networks. *Physical Review E*, 64:016132, 2001.

M. E. J. Newman. Clustering and preferential attachment in growing networks. *Physical Review Letters E, 64(025102)*, 2001.

M. E. J. Newman. Scientific collaboration networks: I. network construction and fundamental results. *Physical Revivew E*, 64, 2001.

M. E. J. Newman. Scientific collaboration networks: II. shortest paths, weighted networks, and centrality. *Physical Revivew E*, 64, 2001.

M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98:404–409, 2001.

M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67:026126, 2003.

M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101:5200–5205, 2004.

M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E*, 68:036122, 2003.

S. Redner. Citation statistics from more than a century of physical review. *APS Meeting Abstracts*, 2004.

S. M. Ross. *Introduction to Probability Models*. Academic Press, 2006.

J. Ruan and W. Zhang. Identification and evaluation of weak community structures in networks. In *Proceedings of National Conference on Artificial Intelligence (AAAI)*, 2006.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1999.

Questions and Comments?