# Collaboration Over Time: Characterizing and Modeling Network Evolution

Jian Huang
College of Information Sciences and Technology
Pennsylvania State University
University Park, PA 16802, US
jhuang@ist.psu.edu

Ziming Zhuang
College of Information Sciences and Technology
Pennsylvania State University
University Park, PA 16802 US
zzhuang@ist.psu.edu

Jia Li
Department of Statistics
Pennsylvania State University
University Park, PA 16802, US
jiali@psu.edu

C. Lee Giles
College of Information Sciences and Technology
Pennsylvania State University
University Park, PA 16802, US
giles@ist.psu.edu

## ABSTRACT

A formal type of scientific and academic collaboration is coauthorship which can be represented by a coauthorship network. Coauthorship networks are among some of the largest social networks and offer us the opportunity to study the mechanisms underlying large-scale real world networks. We construct such a network for the Computer Science field covering research collaborations from 1980 to 2005, based on a large dataset of 451,305 papers authored by 283,174 distinct researchers. By mining this network, we first present a comprehensive study of the network statistical properties for a longitudinal network at the overall network level as well as for the intermediate community level. Major observations are that the database community is the best connected while the AI community is the most assortative, and that the Computer Science field as a whole shows a collaboration pattern more similar to Mathematics than to Biology. Moreover, the small world phenomenon and the scale-free degree distribution accompany the growth of the network. To study the individual collaborations, we propose a novel stochastic model, *Stochastic Poisson model with Optimization Tree (*SPOT*)*, to efficiently predict any increment of collaboration based on the local neighborhood structure. SPOT models the non-stationary Poisson process by maximizing the log-likelihood with a tree structure. Empirical results show that SPOT outperforms Support Vector Regression by better fitting collaboration records and predicting the rate of collaboration.


*WSDM'08,* February 11–12, 2008, Palo Alto, California, USA.


## Categories and Subject Descriptors

I.5.1 [**Pattern Recognition**]: Models—*statistical*; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries

## General Terms

Algorithms, Experimentation, Measurement

## Keywords

Social Network Analysis (SNA), network evolution, stochastic network modeling

## 1. INTRODUCTION

We study the evolution of real-world networks by characterizing and modeling a specific type of social network, the scientific collaboration network. Scientific collaboration leverages the intellectual and material resources in academia and research and greatly benefits the community. Because a predominant form of scientific collaboration is coauthorship, a collaboration network can be portrayed by vertices representing researchers and edges corresponding to coauthored papers. Over time, the evolution of such a network records and reflects the growth of an academic field, shedding light on its future development.

Our work investigates three levels of analysis of networks at different scales. The highest **network level** analysis characterizes the pattern of network evolution using various network statistics. Such measurements enable us to show that the underlying growth pattern manifests the well-known *small world phenomenon*. At the intermediate **community** level, we study the evolution of the structure of autonomously-formed connected components. Visualization of topical communities reveals the distinctive collaboration patterns in each community. At the lowest **individual** level, we propose a novel stochastic model, *Stochastic Poisson model with Optimization Tree (*SPOT*)*, to efficiently predict future collaboration between individuals based on their local neighborhood structure.

**Contributions**: Our contributions are as follows:

Table 1: Comparison of the statistical properties of several coauthorship networks.

| 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 ($\gamma$) | 2.45 | 1.3 | 2.5 | 1.2 | 2.5 | 2.1 |
| $\kappa$ | 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 ($C_{rand}$) | $7.8 \times 10^{-6}$ | $3 \times 10^{-4}$ | $5.4 \times 10^{-5}$ | $3 \times 10^{-3}$ | $1.1 \times 10^{-5}$ | $5.5 \times 10^{-5}$ |
| Giant Comp. Percentage | 65.9% | 57.2% | 70% | 88.7% | 92.6% | 91% |

1. To the best of our knowledge, this paper is the first comprehensive and focused study of the large-scale scientific collaboration network in Computer Science. It investigates networks at three different granularities and provides insights into the distinct collaboration patterns in six topical communities.

2. A novel stochastic model (SPOT) is proposed for individual collaboration prediction, which shows superior performance over Support Vector Regression [30]. SPOT is efficient by exploiting local neighborhood information and is generalizable to other types of networks.

**Organization**: We first briefly review related work, and present our findings on the static and dynamic properties of the Computer Science collaboration network on both the **network level** and the **community level**. We then focus on the **individual level**, proposing the SPOT model to investigate how the collaboration between a particular pair of authors evolves over time. Finally we conclude with plans for future work.

## 2. RELATED WORK

Static properties of various social networks have been studied in the context of epidemiological networks [17], online newsgroups [5], blogs and photo sharing websites [13], citation networks [14], the database research community [9], etc. Many of these networks follow a power-law degree distribution and exhibit the *"small world phenomenon"* [18, 2]. Recent studies [8, 27, 14] on the network dynamics and evolution reveal some interesting growth patterns, e.g. a gradually shrinking diameter and degree densification in the citation networks which can be explained by the Forest Fire model [14]. Social capital can be quantified for predicting event participation in a friendship-event network [16].

Recently, Newman applied modern network analysis techniques to study the static network properties of several coauthorship network, such as MEDLINE (biomedicine), SPIRES (high energy physics), NCSTRL (Computer Science preprints) [19, 23, 25]. Barabási et al [4] studied the dynamics of coauthorship networks using the Maths and Neuroscience networks as prototypes [4]. These studies however only focused on the macroscopic network properties. Other studies on these networks have also revealed some distinctive properties, e.g. the mixing pattern (i.e. positive or negative *assortativity*, which indicates the tendency of nodes linking to others that have a similar or dissimilar degree distribution) [24]. The component structure of such networks represents smaller research communities of different sizes, and communities can be discovered through clustering [29].

Preferential attachment [3] (i.e. the tendency of a new node connecting to an existing node in the network is proportional to its degree) is probably the best known mechanism for depicting the scale-free networks. For collaboration (coauthorship) networks, Barabási et al [4] extended preferential attachment to model the web of science. Although the proposed model can simulate networks in terms of global network statistics (e.g. exponents) by only taking into account node degrees, one can not generally infer the probability that a pair of nodes in the network will in the future collaborate. Recently, Newman found that the probability of scientists to collaborate increases with the number of common collaborators and there is a positive correlation between the number of past collaborations and future (repeat) collaborations [20]. Liben-Nowell and Kleinberg [15] formalized this as the *link prediction* problem, and studied the effect of various network proximity measures for predicting the addition of edges in the coauthorship network. These informative structural features in the node's neighborhood, combined with the available exogenous extracted information in our datasets, will be used in the statistical model for predicting the number of collaborations in this paper.

## 3. DATA COLLECTION

We studied the evolution of scientific collaboration using data drawn from the *CiteSeer* Digital Library[1]. *CiteSeer* is a popular digital library with a focus on computer and information sciences and consists of more than 700,000 academic papers. From *CiteSeer*, we extracted the metadata of these papers [10], including titles, authors, affiliations, physical/email addresses, publication years, etc. To minimize the impact of the author name ambiguity problem, we used the author disambiguation techniques proposed in [11] so that each vertex in this network represents a

[1]The CiteSeer Digital Library: *http://citeseer.ist.psu.edu*.

distinct author. We obtain a large dataset (denoted by *CS*) consisting of 283,174 unique authors and 451,305 papers published between 1980 and 2005 (Figure 1:left). In Table 1 we compare our network with several other coauthorship networks in the literature [1].

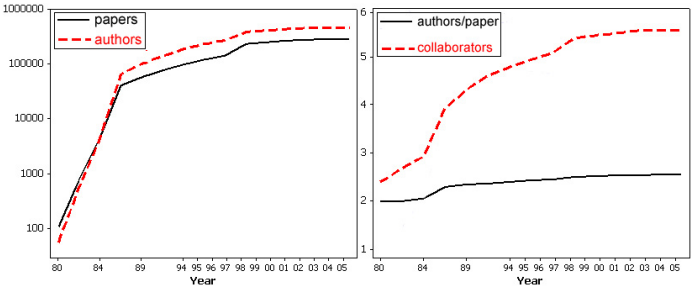


**Figure 1: Left: number of authors and papers from 1980 to 2005. Right: the increasing trend of collaboration.**

To study collaboration in a more fine-grained fashion, we decomposed and mapped the data into six topical datasets (Table 2): Artificial Intelligence (ai), Applications (*app*), Architecture (*arch*), Database (*db*), System (*system*), and Theory (*theory*). We obtained a list of representative conferences in each topic from Computer Science Conference Rankings[2]. We matched the names of these conferences with the venue information extracted from the paper metadata and constructed six datasets. Each dataset contains papers published in the representative conferences corresponding to a particular topic. For consistency, we intentionally selected approximately 1,700 papers for each topic, yielding 11,820 authors and 10,195 papers in total.

To capture the dynamics of the growing network, we used a *"snowball sampling"* approach. Using the publication timestamps of the papers, we generated a series of 25 year-by-year snapshots for the *CS* dataset. Each snapshot at time $t$ is represented by $\mathcal{G}(t) = (V, E)$, where every vertex $v \in V$ represents an author, and a pair of authors $u$ and $v$ are connected by an edge $\langle u, v \rangle \in E$ if and only if they have coauthored one or more papers. For any given timestamp $t$, $\mathcal{G}(t)$ contains all the vertices and edges that have appeared up to time $t$. For example, the snapshot $\mathcal{G}(2001)$ contains all the authors who have published between 1980 and 2001, and $\mathcal{G}(2005)$ represents all the author and paper information in the *CS* dataset.

[2] *http://www-static.cc.gatech.edu/~guofei/CS_ConfRank.htm*

**Table 2: Overview of the six topical datasets.**

| 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 |

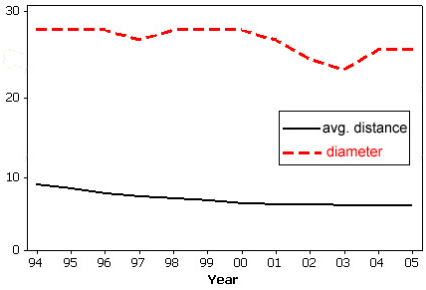


**Figure 2: The average distance of the network gradually decreases over time, indicating that the collaboration *"world"* in fact *gets smaller over time*.**

# 4. CHARACTERIZING NETWORK EVOLUTION

## 4.1 More Collaboration in a Smaller World

Collaboration among authors becomes increasingly popular as evidenced in our 25 snapshots (Figure 1:right). Over the past 25 years, the average number of collaborators per author has been steadily increasing, and so has the average number of authors per paper but at a much slower pace (from 2 authors per paper in 1980 to 2.55 in 2005).

The growth pattern of the network can be also investigated by examining the diameter and the average distance between a pair of authors in the collaboration network over time. Distance is defined here as the number of *hops* from one vertex to another and diameter as the longest distance between two vertices. As shown in Figure 2, the average distance between pairs of authors continuously decreases from 9.3 in 1994 to 7.1 in 2005, which is another evidence of scientific collaboration gaining popularity. On the other hand, the network diameter fluctuates around 26, with two dips in $\mathcal{G}(1997)$ and $\mathcal{G}(2003)$ which we believe are due to the merge of the giant components which we further explore in the next subsection.

Further investigation reveals the intrinsic growth pattern of the network. First of all, the coauthorship network manifests the *small world phenomenon* [18, 2]. As shown in Table 1, all the coauthorship networks show a very high clustering coefficient compared to Erdös-Rényi random networks of the same size, suggesting that there is strong local clustering in these networks. This partially explains why the coauthorship network grows through the addition of edges between a pair of nodes having a common node that forms a closed triangle. In other words, the introduction between authors by the same coauthor accounts for the increasing number of collaborations. Also, despite the stringent definition of coauthorship, the CS network has average distance 7, implying that the famous six degrees of separation can also be valid in the scientific world. Less obviously, the average distance scales logarithmically with respect to the number of researchers in the network:

$$\frac{logN}{log < k >} = 7.32 \approx l_{real} \quad (1)$$

further confirming that the coauthorship network is a growing small world.

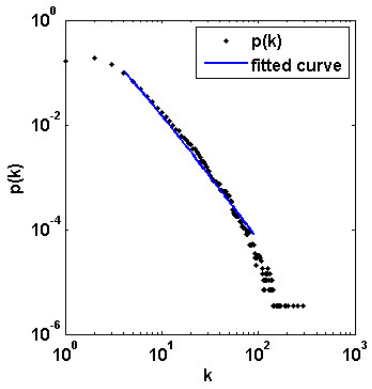


**Figure 3: The degree distribution of the CiteSeer co-authorship network in 2005.**

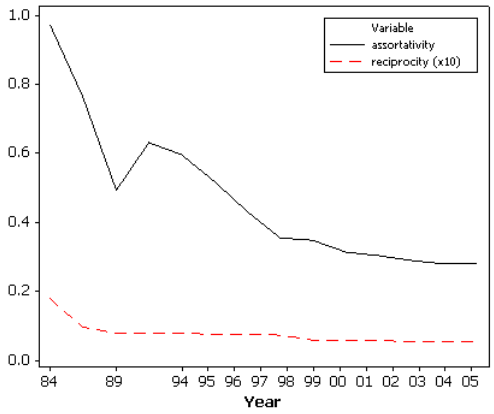


**Figure 4: Assortativity and reciprocity are both shrinking over time. Note that assortativity dropped significantly.**

The scale-free degree distribution also characterizes the network growth pattern. Figure 3 shows the degree distribution in $\mathcal{G}(2005)$[3]. Qualitatively, the degree distribution follows the power law especially in the middle degree region. Similar to other coauthorship networks [4], the counts in small degrees are substantially lower and the curve tapers off more quickly in the large degrees resembling an exponential distribution. The best fit for the (scale-free) degree distribution in the middle region is obtained by minimizing the sum of squares criterion and is (as shown with the straight line)[4]:

$$p(k) \sim (0.8 + k)^{-2.45} \tag{2}$$

In Table 1 we compare our findings on $\mathcal{G}(2005)$ with existing work on the coauthorship networks in other domains, e.g. biology, physics, and maths. Based on these metrics, the way in which computer scientists collaborate is much more similar to that of mathematicians and physicists than to biologists. For instance, the average distance of the computer science collaboration network lies between that of mathematics and physics.

## 4.2 Shrinking Assortativity and Reciprocity

Two interesting findings are made regarding the *assortativity* and *reciprocity* of the network. The first one is a quantitative measure of the tendency for vertices to be connected to other nodes that are similar (or dissimilar). In our particular case of the collaboration network, we consider *assortativity mixing by degree* - the tendency of nodes to link to others that have a similar degree distribution (e.g. researchers with many collaborators tend to coauthor with other researchers who also have many collaborators). This is defined in [24] as:

$$a = \frac{\sum_i j_i k_i - N^{-1} \sum_i j_i \sum_{i'} k_{i'}}{\sqrt{[\sum_i j_i^2 - N^{-1}(\sum_i j_i)^2][\sum_i k_i^2 - N^{-1}(\sum_i k_i^2)]}} \tag{3}$$

where $j_i$ and $k_i$ are the excess in-degree and out-degree of the vertices that the $i^{th}$ edge leads into and out of respectively, and $N$ is the total number of edges. Note that we consider the collaboration network as symmetric and ignore the order of authorship.

We measure *reciprocity* of the network as the tendency for a pair of coauthors to exchange their positions in the authors list, formally defined as:

$$r = \frac{\sum e^t_{vu}}{\sum e^{t'}_{uv}}, \tag{4}$$

where $e^t \in G_t$, $e^{t'} \in G_{t'}$, $u, v \in G_t$, $t < t'$. Intuitively, it is the fraction of directional edges $< v, u >$ such that $< u, v >$ also exists in the graphs with earlier time-stamps.

Shown in Figure 4 are *assortativity* and *reciprocity* from 1984 to 2005. The positive *assortativity* could be corroborated by the observation [2] [26] that real-world social networks, unlike most artificial networks, are usually strongly **assortative**. *Assortativity* eventually dropped to 0.28 in $G_{2005}$, smaller than that of the Physics community (see Table 3). *Reciprocity* was also significantly lower than those observed in online social network: the *reciprocity* in $G_{2005}$ was 0.0055 in sharp contrast to 0.702 in the Flickr social network and 0.84 in the Yahoo 360 social network [13], as actors in these online social networks generally exchange links and linking back is much less costly than that in coauthorship networks.

# 5. COLLABORATION AT THE COMMUNITY LEVEL

In this section, we present our findings as we scale down the level of analysis to the **community level**, with focus on two different types of communities.

## 5.1 Component Structure Evolution

The first type of communities are those autonomously

[3]To reduce the effect of outliers in the middle and high degrees, we use the median of a window of five points (the median is resilient to outliers) to plot the degree distribution.
[4]The curve may be better fitted with a piece-wise scale-free distribution with lower exponent for low degrees and larger exponent for high degrees.

**Table 3: Assortativity coefficient in Computer Science and other research domains.**

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

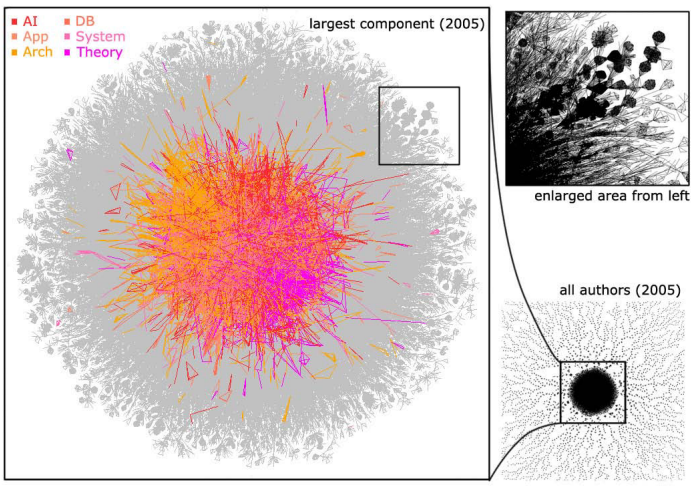


**Figure 5: The largest component of $\mathcal{G}(2005)$, enlarged on the left, visually forms the *core* of the network shown in the bottom-right inset. Edges showing collaborations in different topics are rendered in different colors. Visualization is performed by the Large Graph Layout (LGL) package.**

formed through collaboration in the entire Computer Science field. Such communities are represented by *components* of different sizes in the network. Formally defined, a *component* is a connected subgraph, i.e. two vertices are in the same component if and only if there is a traversal path between them.

Let $|C|$ denote the size of a component $C$. We are interested in three types of components:

- **Giant components** represent the *large* ($|C| \geq \tau$, where $\tau$ is a threshold) groups of researchers who are connected to each other either directly as coauthors or indirectly through a chain of collaborators. Collectively, the giant components form the *"core of productivity"* in the network, usually containing the most prolific and active collaborators. The *largest component* (Figure 5) typically covers a significant portion of the network [25]. In our case the *largest component* in $\mathcal{G}(2005)$ covers about 65.95% of the network.
- **Singletons** are the individuals who never publish with other researchers. They are the loners in the community, represented by vertices with zero degree. About 7.2% of the vertices in $\mathcal{G}(2005)$ are singletons.
- **Middle region** is the remaining section in the collaboration spectrum, typically consisting of relatively isolated groups of researchers who seldom collaborate with *outsiders*. In $\mathcal{G}(2005)$ the middle region comprises about 26.85% of the authors.

The dynamics of how components of different sizes evolve over time are even more interesting (Figure 6). The band at the top represents the giant components ($\tau = 10^3$) and the bottom band corresponds to the singletons. The remaining bands constitute the middle region. Most notably, the number of vertices that belong to the middle region

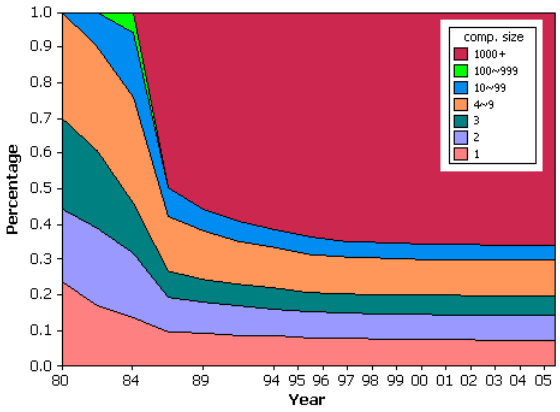


**Figure 6: The evolution of component structure, shown as the fraction of vertices in components of different sizes. The middle region gradually lost ground as the giant components came into dominance. The percentages became constant after reaching a steady state.**

drops significantly, and then quickly reaches a steady state where the width of each band remains almost constant since $\mathcal{G}(1989)$. The components of sizes between 100 and 999 (the green band) disappear after 1987 and we find that they merged with the giant components[5]. Similar patterns of connected component evolution were also observed in other types of social networks [13].

## 5.2 Collaboration in Topical Communities

The second type of communities correspond to different research topics. We study the collaboration in topical communities using the six topical datasets described earlier and present the findings in Table 4. One interesting observation is that the database community has the best connected collaboration network, with the lowest average distance between pairs of authors and the highest average betweenness. Its component structure (see Figure 7) has the fewest components and the largest component covers a significant portion ($\sim$ 60%) of the network. Compared with other topics, database researchers also have the largest number of collaborators on average, and are more reciprocal to coauthors [13]. On the other hand, AI researchers have the weakest tendency to reciprocate in publication. They however have the largest assortativity among all six topics (Table 4), showing the strongest tendency to collaborate with someone who are *in the same league*, i.e. researchers with many collaborators tend to coauthor with other researchers who also have many collaborators. Figure 7 shows a comparison between the database and the applications community. The much lower betweenness of the applications topic might be due to the fact that it contains authors from several rather disjoint communities, such as multimedia research and information retrieval, and is thus more heterogeneous than the other topics.

[5]We note the merging effect may be confounded by the relatively low availability of online publication in the 1980s; the constitution since mid 1990s corresponding to the majority of papers in CiteSeer, however, remains relatively stable.

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

**Table 4: Summary statistics of the six topical communities. Cells with particularly high/low values are highlighted in bold font.**

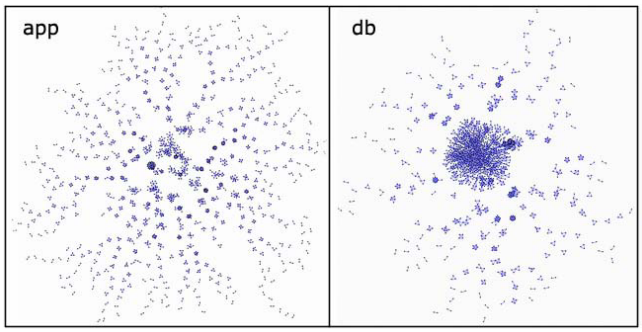


**Figure 7: The database community (right) is more cohesive than the applications community (left), and its largest component forms a significantly larger *core* in the network.**

# 6. SCIENTIFIC COLLABORATION BETWEEN INDIVIDUALS

We now investigate how a collaboration between a pair of authors evolves over time, which serves as the finest level collaboration in the network. Specifically, a fundamental research question is: *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 $\Delta t$ time interval.*

As shown above, real world social networks (coauthorship networks in particular) are typically very large in terms of the number of vertices $|V|$, but are very sparse in terms of the number of edges $|E|$, i.e. $|E| \ll |V|^2$. Algorithms that examine all pairs of vertices in the network result in the computational complexity $\Omega(|V|^2)$, and as such are generally intractable for large $|V|$. Nonetheless, we can model the microscopic collaboration patterns between all pairs of vertices with existing edges, due to the sparsity of the edges in such networks. In the coauthorship network, existing edges represent previous collaboration between a pair of authors. Intuitively, past collaboration patterns between two authors and the structural information in the neighborhood of the vertices involved would suffice for the prediction of future collaboration. Hence, we propose a novel and efficient ($\Theta(|E|\log(|E|))$) method for learning and predicting the collaboration patterns between author pairs with existing collaboration and without resorting to the rest of the network.

We note that there is a related yet separate question: how does a new collaboration appear in the network? The answer to this depends highly on the macroscopic network characteristics. For instance, artificial networks are usually dissortative [24], whereas preferential linking is common in social networks for which the well known Barabási-Albert (BA) model [3] can be used to predict $P\{e_{i,j}\}$. Also, in real world scenarios, there is usually insufficient information in the data to predict intercollegiate or interdisciplinary collaboration.

We first propose the *Stochastic Poisson model with Optimization Tree* (SPOT) method, consisting of a stochastic model for collaboration over time and an Optimization Tree for optimizing the model. Then we evaluate the performance of the model on longitudinal coauthorship network data. Although we derive the model in the coauthorship network, our learning algorithm can be readily generalized to other types of evolutionary networks.

## 6.1 A Stochastic Poisson Model for Collaboration

We denote an evolutionary collaboration network by a series of discrete time snapshots $\{\mathcal{G}(t)\}(t = 0, 1, 2, ...)$. In the graph $\mathcal{G}(t)$, an existing collaboration between authors $v_i$ and $v_j$ is denoted by an edge $e_{i,j}(t)$, associated with its weight equal to the number of collaborations $N_{i,j}(t)$ up to time $t$. Since $N(t)$ is discrete and represents the cumulative number of collaborations, $N(t)$ is a counting process. The Poisson process [28] is a prominent stochastic model for counting processes, appearing in many real world phenomena. In the time invariant Poisson process with rate $\lambda$, the probability of the number of events $k$, occurring in the $\Delta t$ time interval, is given by:

$$Pr_\lambda\{N(t + \Delta t) - N(t) = k\} = \frac{e^{-\lambda \Delta t}(\lambda \Delta t)^k}{k!} \qquad (5)$$

In our case, however, the stationary Poisson rate condition typically does not hold for scientific collaboration due to various influencing factors. For example, collaborating with prolific authors may lead to a faster rate. Thus this requires the use of the more general nonstationary (nonhomogenous) Poisson process [28], where the rate is a variant function rather than a constant. As noted above, the local collaboration structure provides valuable information for determining the underlying rate function. For instance, while international or intercollegiate collaboration is not uncommon, two authors may collaborate more when they work in the same laboratory; authors collaborated assortatively before are likely to collaborate in the same way in the future. Specifically, for an existing collaboration $e_{i,j}(t)$ at time t, the subgraph $\mathcal{G}(e_{i,j}(t))$ is defined as the **neighborhood** of $e_{i,j}(t)$, consisting of the author pairs $(v_i, v_j)$, their immediate neighbors (coauthors) and the associated edges. A feature vector $\mathbf{a} = (a_1, ..., a_p)$ is computed with respect to the neighborhood $\mathcal{G}(e_{i,j}(t))$. For instance, features

concerning the local neighborhood can be the fraction of the number of shared coauthors to the total number of collaborators for each author, the collaboration rate in the previous snapshot, the cumulative number of publications, etc. Therefore, for a specific edge $e_{i,j}(t)$, the collaboration rate $\lambda$ is defined as a function of the feature vector of the neighborhood, i.e. $\lambda(e_{i,j}(t)) = f(\mathbf{a}_{i,j}(t))$. Our goal is to learn the optimal function $f$ from the past collaboration patterns for the prediction of future collaboration.

We obtain a training instance for each edge between two adjacent snapshots $e_{i,j}(t)$ and $e_{i,j}(t+\Delta t)$. We then compute the increment of publication by $k_{i,j} = N(t+\Delta t) - N(t)$ and the feature vector $\mathbf{a}_{i,j}(t)$ from the neighborhood $G(e_{i,j}(t))$. For simplicity of notation, we let $Pr_{\lambda(e_{i,j}(t))}\{k_{i,j}\}$ denote

$$Pr_{\lambda(e_{i,j}(t))}\{N(t+\Delta t) - N(t) = k_{i,j} | e_{i,j}(t_0)\} \qquad (6)$$

Note that the probability conditions on the existence of an edge (denoting collaboration) prior to the counting process. Once the counting process starts, it is modeled as a non-homogeneous Poisson process and thus in any time interval $\Delta t$ the increment $k_{i,j}$ can be any non-negative number. We also assume $\Delta t$ to be one unit of time in the sequel. Given the observations, the optimal function $f$ should maximize the log-likelihood with respect to the training instances:

$$\begin{aligned} f^* &= \arg\max_f \log \prod_{e_{i,j}(t)} Pr_{\lambda(e_{i,j}(t))}\{k_{i,j}\} \\ &= \arg\max_f \sum_{e_{i,j}(t)} \log Pr_{\lambda(e_{i,j}(t))}\{k_{i,j}\} \\ &= \arg\max_f \sum_{e_{i,j}(t)} [k_{i,j} \log(\lambda(e_{i,j}(t))) - \lambda(e_{i,j}(t)) - \log k_{i,j}!] \\ &= \arg\max_f \sum_{e_{i,j}(t)} [k_{i,j} \log(f(\mathbf{a}_{i,j}(t))) - f(\mathbf{a}_{i,j}(t))] \qquad (7) \end{aligned}$$

where the function $f$ is plugged into Eq. (7) and the factorial terms $k_{i,j}!$ are dropped as they are not related to the maximization. In general, there is no closed-form solution for Eq. (7). We now propose a nonparametric solution to this optimization problem.

## 6.2 The Optimal Tree for Estimating the Rate Function

In an evolutionary network, we model the collaboration, or more generally the addition of weights to the edges, as a non-stationary Poisson process. The collaboration rate function depends on the neighborhood structure. We refrain from making *a priori* parametric assumptions about the rate function, as such our model can be applied to a wide variety of real-world networks. Furthermore, we opt to solve the optimization problem using a nonparametric approach, the **Optimization Tree**, inspired by decision tree methods such as CART [6].

Similar to decision trees, the Optimization Tree is grown in a top-down and best-first fashion. Our goal is to derive the decision rule criterion in each split in order to find the function $f^*$ in Eq. (7). Each internal node in the tree represents a decision: $a_j \leq h$, where $a_j$ is the j$th$ attribute and $h$ is the splitting value. Geometrically, a splitting hyperplane is orthogonal to an axis in the feature space and splits the feature space into hypercubes (regions). As such it dramatically reduces the search space and consequently the training time as well. We note that the terms in the summation in Eq. (7) are not related to one another. Thus for any internal node, the optimal decision rule should maximize the log-likelihood in the two sub-regions:

$$\begin{aligned} &(R_1, R_2)^* = \\ &\arg\max_{R_1 \cup R_2 = R} \sum_{e_{i,j}(t) \in R_1} [k_{i,j} \log f(\mathbf{a}_{i,j}(t)) - f(\mathbf{a}_{i,j}(t))] \\ &\quad + \sum_{e_{i,j}(t) \in R_2} [k_{i,j} \log f(\mathbf{a}_{i,j}(t)) - f(\mathbf{a}_{i,j}(t))] \qquad (8) \end{aligned}$$

Since all the instances in a node are assumed to be drawn from an identical Poisson distribution, the optimal estimate for the rate which maximizes the log-likelihood for the training samples in region $R$ is the mean:

$$\lambda_R = \sum_{e_{i,j}(t) \in R} k_{i,j} \Big/ \parallel R \parallel \qquad (9)$$

Substituting Eq. (9) into Eq. (8), we derive the optimal splitting criterion for the region R:

$$\begin{aligned} (R_1, R_2)^* &= \arg\max_{R_1 \cup R_2 = R} [\sum_{e_{i,j}(t) \in R_1} k_{i,j} \log \lambda_{R_1} \\ &\quad + \sum_{e_{i,j}(t) \in R_2} k_{i,j} \log \lambda_{R_2} - \lambda_{R_1} \parallel R_1 \parallel - \lambda_{R_2} \parallel R_2 \parallel] \\ &= \arg\max_{R_1 \cup R_2 = R} [\sum_{R_1} k_{i,j} \log \lambda_{R_1} + \sum_{R_2} k_{i,j} \log \lambda_{R_2}] \qquad (10) \end{aligned}$$

where the linear terms in Eq.(10) are dropped, since they sum up to $\sum_{e_{i,j}(t) \in R} k_{i,j}$ and thus are not related to the optimization.

In each step of growing the Optimization Tree, the internal node, which when split causes the maximum increment in the log-likelihood, is selected as the splitting node. Techniques used in training decision trees can also be adopted here for efficient implementation. For instance, training instances can be sorted by feature value beforehand and a linear search can be performed per feature to obtain the maximum in Eq.(10). After the tree is grown, each leaf in the Optimization Tree corresponds to a fixed rate Poisson distribution and can therefore be used for prediction.

## 6.3 Remarks on the Optimization Tree Method

Before turning to the experimental studies, it is worthwhile to remark on the proposed Optimization Tree method. We first offer an alternative viewpoint of the Optimization Tree method as **entropy maximization**. Substituting $\lambda$ in Eq. (9) into Eq. (10), we note that with proper normalization each term in Eq. (10) can be regarded as a proxy of the entropy of the corresponding sub-region. Therefore, although we aim to maximize the log-likelihood, the Optimization Tree method can in fact be regarded as maximizing the proxy of the entropy of the samples in a greedy fashion.

Although the proposed Optimization Tree method shares some structural similarities with decision trees, we emphasize the key differences between them. First, the Optimization Tree aims to estimate a probability distribution or more precisely its parameters with respect to the input feature vector, rather than solving a classification or regression problem as in classification or regression trees. Second, the Optimization Tree solves the optimization problem by maximizing the log-likelihood of the training instances,

**Table 5: Features used for collaboration prediction.**

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

instead of using the impurity function such as Gini index [6] to determine the decision rules. We formally derive the functional form of the decision rule in each split which is fundamentally different from the impurity functions as used in classification/regression trees.

Despite these differences, the Optimization Tree shares some **advantages** with other decision tree methods. First, the Optimization Tree can be regarded as a smoothing technique for samples within a region, as such it dampens the effect of outliers in the training data. Moreover, since SPOT learns from a non-homogeneous Poisson process rather than a fixed distribution, it has the effect of smoothing evolving data over time. Second, with a linear region boundary the Optimization Tree provides a simple and stable model for estimating the rate function. Also, the training computation complexity is $O(p\,|E|\log|E|)$, where $p$ is the dimension of the feature space. It is thus more efficient than other methods such as SVMs. These features are highly desirable for our problem.

## 6.4 Empirical Studies on the Collaboration Model

To evaluate the accuracy of the prediction, we collected a total of 105,122 collaboration records in 1997 to 2000 and 2001 to 2003 from the *CS* dataset. The goal is then to predict the increment in collaboration between the snapshots[6]. Table 5 shows the different features used in each collaboration record. These features are organized in three levels: the first level relates to the individual author *per se*, the second level concerns with pairwise collaboration information and the third level pertains to the collaboration structure in the neighborhood. Since we only use local information, an efficient solution can be easily achieved. Although Table 5 is not a complete list of all possible features, we intentionally select the most informative features in each level to showcase the prediction power of our method.

We use two models for comparison. The Past Collaboration Rate model is used as a baseline which predicts the number of collaborations to be the same as that in the previous time interval, regardless of other information. In [20], Newman has shown by using a relative probability that the number of past collaborations is a good indicator of the probability of future collaboration. We also choose Support Vector Machines [30] as a sophisticated model for comparison, due to its efficacy in various classification and regression problems. Specifically, given the feature vectors and the corresponding number of collaborations as training samples, Support Vector Regression (SVR)[7] [30] is used as a strong baseline comparator to learn the regression model. The SVR model then predicts the number of collaborations in the test data.

Note that the above two baselines are however inherently different from SPOT, which predicts the number of collaboration with a Poisson probability distribution. In practice, it is desirable to obtain a discrete value as the number of collaborations. The mean of the Poisson distribution is a natural choice, since it maximizes the log-likelihood in the training data. In the sequel, we use the Poisson mean corresponding to the distribution predicted by the SPOT method to compare with the values predicted with the baseline methods. To give an actual example, rules of the following form could be condensed from the tree for predicting the distribution:

```
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
```

Thus in the example above, the leaf nodes predict the collaboration rate. If the predicted rate is 2.38 for the respective Poisson distribution, the probabilities of incremental collaborations of 0, 2 and 3 are 0.093, 0.262 and 0.208 respectively.

We evaluate the performance of these models by measuring two goodness of fit statistics for the data. The sample correlation coefficient (Pearson correlation coefficient) measures the linear relationship between the predicted and the true number of collaboration. Let $y_i$ and $\tilde{y}_i$ $(i = 1, ..., n)$ denote the true and predicted increment of collaboration, the sample correlation coefficient is defined as,

$$r_{y\tilde{y}} = \frac{\sum_i (y_i - \overline{y})(\tilde{y}_i - \overline{\tilde{y}})}{(n-1)s_y s_{\tilde{y}}} \tag{11}$$

where $\overline{y}$ and $\overline{\tilde{y}}$ are sample means, $s_y$ and $s_{\tilde{y}}$ are sample standard deviations. The correlation coefficient is between -1 and 1, and the higher the absolute value the stronger the linear relationship between the predicted and true value. We also measure the absolute magnitude of the residuals (the difference between the predicted and the true value) with

[6]Publications have been aggregated into a three year interval. The increment in publications between the intervals ranges from 0 to 42. We did not choose a one year interval since the number of publications by two authors is relatively small and unstable.

[7]The classical implementation $\epsilon$-SVR in LIBSVM [7] is used in our experiments.

**Table 6: Comparison of prediction performance of Past Collaboration (baseline), Spot and SVR using 10-fold cross validation.**

| 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** |

the Root Mean Squared Error (RMSE),

$$\text{RMSE} = \sqrt{\frac{\sum_i err_i^2}{n}} = \sqrt{\frac{\sum_i (\tilde{y}_i - y_i)^2}{n}} \qquad (12)$$

Lower RMSE indicates a better fit to the data as the predicted values deviate less from the true values.

We held out a portion of the training data to determine the parameters in the methods. The only parameter in SPOT is the number of internal nodes which determines the size of the tree. Experiments on the hold-out datasets suggest that a tree with 75 internal nodes fits the data well. This parameter is relatively insensitive as the correlation coefficient ranges from 0.73 to 0.82 in the hold out dataset. Similarly, we determine the parameters C=10 and $\gamma$=0.001 (with RBF kernel) in the SVR model using the same hold-out dataset.

We conducted 10-fold cross validation in the remaining data in all our experiments, namely, parameters are trained on 90% of the samples and tested on the remaining data in a rotation manner. The results are summarized in Table 6. The baseline Past Collaboration model scores only 0.227 coefficient correlation, while both SPOT and SVR score much higher, suggesting that there is significant information in the neighborhood structure that can be used for prediction. Using the Poisson mean as the predicted value, SPOT increases the correlation coefficient by 18.7% compared to SVR. Considering the contribution of other complicating factors in future collaboration, SPOT's high correlation coefficient (0.782) suggests that there is strong correlation between the predicted and true value and thus SPOT can be a good predictor. Similar results are found in the RMSE metric. The RMSE is 4.74 when using only the number of collaboration in the previous time interval for prediction. SVR and SPOT significantly outperforms the baseline by accounting for the information existed in the neighborhood. Compared to SVR, SPOT incurs 8.4% less RMSE. These results demonstrate that SPOT fits a better model for collaboration than SVR.

# 7. CONCLUSIONS AND FUTURE WORK

This work adds to the literature of social network analysis of collaborating authors by characterizing and modeling the growth of a large Computer Science collaboration network over a period of 25 years. By conducting a longitudinal analysis at the **network** and the **community** levels, our study presents the first comprehensive picture of the evolving trends of collaboration in the field. In addition, we also quantify, compare and contrast the distinctive collaboration patterns in six topical communities. On the **individual** level, we proposed a new model, the SPOT method, for learning and predicting collaborations between pairs of authors. In an evolutionary network, we model the collaboration as a non-stationary Poisson process and propose the Optimization Tree method to learn the collaboration rate function. Since only local neighborhood information is used, our model is able to learn from a heterogeneous network, is efficient, and is flexible to the set of features. Our experimental results show that SPOT outperforms the popular SVR method for collaboration prediction.

There are several opportunities for future studies. For instance, one could study the creation and evolution of *bridge nodes* in different topical communities since they serve as brokers of information between different research areas. The identification of such nodes will further our understanding of the interaction between collaborations at the individual level and community level. Another direction, as alluded earlier, is to extend our model with the Bayes formula to predict the joint probability $P\{N(t+\Delta t)-N(t) = k, e_{i,j}\}$ for new edges, where the prior of new edges may be modeled by topological and topical information such as cyclic closure and focal closure [12].

# 8. ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation (NSF).

The authors would like to thank Réka Albert and Isaac Councill for insightful discussions.

# 9. REFERENCES

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

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

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

[4] 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.

[5] 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.

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

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

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

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

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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