Collaboration Signatures Reveal Scientific Impact

Yuxiao Dong, Reid A. Johnson, Yang Yang, Nitesh V. Chawla
Interdisciplinary Center for Network Science and Applications (iCeNSA)
Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556
{*ydong1,rjohns15,yyang1,nchawla*}*nd.edu*

***Abstract*—Collaboration is an integral element of the scientific process that often leads to findings with significant impact. While extensive efforts have been devoted to quantifying and predicting research impact, the question of how collaborative behavior influences scientific impact remains unaddressed. In this work, we study the interplay between scientists' collaboration signatures and their scientific impact. As the basis of our study, we employ an ArnetMiner dataset with more than 1.7 million authors and 2 million papers spanning over 60 years. We formally define a scientist's collaboration signature as the distribution of collaboration strengths with each collaborator in his or her academic ego network, which is quantified by four measures: sociability, dependence, diversity, and self-collaboration. We then demonstrate that the collaboration signature allows us to effectively distinguish between researchers with dissimilar levels of scientific impact. We also discover that, even from the early stages of one's researcher career, a scientist's collaboration signature can help to reveal his or her future scientific impact. Finally, we find that as a representative group of outstanding computer scientists, Turing Award winners collectively produce distinctive collaboration signatures throughout the entirety of their careers. Our conclusions on the relationship between collaboration signatures and scientific impact give rise to important implications for researchers who wish to expand their scientific impact and more effectively stand on the shoulders of "collaborators."**

***Keywords*-Collaboration Signature; Scientific Impact; Academic Social Network; Science of Science; Scientific Success.**

I. Introduction

Since its inception, science has benefited from the synergy of collaboration. As it has matured, science has become increasingly interdisciplinary, with researchers constantly forging new collaborations to solve ever larger, and ever more complex problems [1]. The effects of collaboration on the scientific endeavor can be studied via the "science of science," an emerging discipline wherein scientists use science to study the scientific process, employing quantitative assessment to arrive at a better understanding of its dynamics and, ultimately, to improve the outcomes it effects [2], [3], [4], [5], [6]. But it can be difficult to quantitatively discern the degree to which collaboration influences the impact of a researcher's work. Further complications arise due to the difficulty of formulating an objective measure of a researcher's long-term scientific impact, which requires the evaluation of his or her entire body of work.

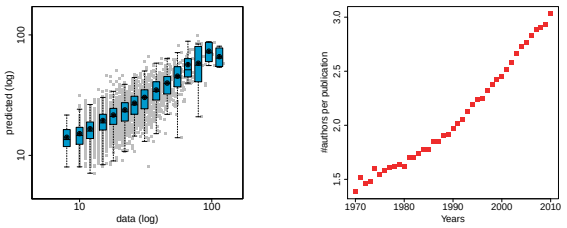
One measure of scientific productivity and impact that has received significant attention is the *h*-index. Proposed by J. E. Hirsch in 2005, a researcher's *h*-index is defined as h if h of his or her papers receive at least h citations and the number of citations of the remaining papers is at most h each [7]. Aside from the *h*-index, there are other factors that may, if properly quantified, be used to measure a researchers' impact. For example, a scientist may be recognized for having authored a particularly influential publication, or what might be colloquially termed a "*big-hit*" paper. Or, as scientists are inclined to publish scientific results in prestigious venues (e.g., *Nature*, *Science*, *KDD*, and *ICDM*) to effectively disseminate their findings, a researcher might be recognized as a frequent contributor to such "top" venues.

A significant amount of work has been devoted to using quantifiable measures of scientific impact, such as citation counts and *h*-indices, to predict the future impact of publications and researchers [2], [5], [8], [9], [10]. These prediction strategies, however, are fundamentally limited by the heavy-tailed distributions of these measures, whereby the majority of publications collect few citations and the low-*h*-index researchers dominate the total number of scientific practitioners [11]. To circumvent these limitations, which arise when citation counts and *h*-indices are estimated directly, Dong et al. have investigated scientific impact by addressing the question of whether a publication will contribute to its authors' *h*-indices within a given timeframe [12].

Despite these and other extensive investigations, knowledge concerning the interplay between collaboration behaviors and scientific impact is sorely lacking. Yet, it has been observed that social interactions play an essential role in human society at the levels of both collective phenomena and individual behaviors, with recent work having found that the concept of an "ego" social network can be used to distinguish individuals in human communication networks [13]. A particular type of ego network, known as a collaboration ego network, can be used to capture collaboration patterns. In this work, *we study how researchers' collaboration ego networks influence the nature and progression of their scientific impact throughout their research careers.*

Contributions. Our study is performed on an academic dataset comprised of over 1.7 million authors and 2 million papers spanning more than 60 years from the premier online

ASONAM '15, August 25-28, 2015, Paris, France
 ISBN 978-1-4503-3854-7/15/08 $15.00
DOI: http://dx.doi.org/10.1145/2808797.2808846



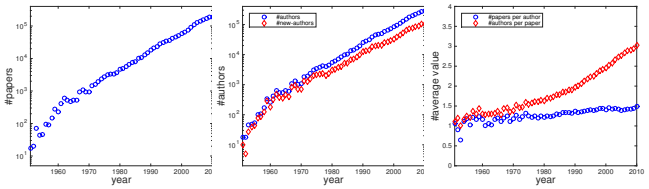
(a) Collaborations reveal impact (b) Ascending collaboration trend

Figure 1. **Collaborations in computer science.** (a) x-axis: h-indices in data; y-axis: predicted h-indices from collaboration signatures. (b) x-axis: year; y-axis: the number of authors per publication. *The average number of collaborators per published work doubled between 1970 and 2010.*

academic service ArnetMiner [14]. We formally define a researcher's *collaboration signature* as the distribution of the fraction of collaboration strengths with each of his or her collaborators in the collaboration ego network. We then associate the researcher's collaboration signature with four quantified measures—sociability, dependence, diversity, and self-collaboration—that demonstrate how the collaboration signature reveals scientific impact. We demonstrate that researchers with different levels of scientific impact produce significantly different collaboration signatures, irrespective of the way that scientific impact is quantified (e.g., h-index, big-hit papers, or top-venue publications). Additionally, we find that recipients of the Turing Award, an accolade for outstanding contributions to the field of computer science, collectively produce unique collaboration signatures that are characterized by a relatively low and stable level of sociability and a relatively high level of self-collaboration, with both trends persisting the entirety of their careers.

To further explore the extent to which scientific impact can be revealed by collaboration signatures, we employ prediction-based case studies. Given only four features present in the collaboration signature, we find surprisingly strong performance for predicting future scientific impact (see Figure 1(a)). Our case studies further find that the collaboration signature of a scientist's first fifteen years of research is highly correlated with influence after thirty years (Pearson correlation coefficient = 0.75).

To the best of our knowledge, this work is the first to study collaboration behavior across a researcher's academic career. It is also the first to find the collaboration signature, which describes how a scientist distributes his or her collaboration efforts and investment, can serve as a powerful indicator of the evolution of a researcher's scientific impact. For this reason, we conclude that the collaboration signature has important implications for our understanding of the mechanisms that underlie the progression of scientific impact. This finding is particularly relevant to the field of modern computer science, wherein collaborations have in recent decades become increasing prevalent (see Figure 1(b)).



(a) (b) (c)

Figure 2. **Characteristics of academic data.** (a) The number of papers at each year (y-axis is log scale). (b) The number of authors at each year (y-axis is log scale). (c) The average #papers by each author and #authors for each paper.

II. Data

The data used in this work is sourced from Arnet-Miner [14], [15], which is the premier free online service for academic social network analysis and mining. This academic dataset consists of 1,712,433 computer scientists and 2,092,356 papers from computer science venues held between 1950 and 2012[1]. In total, we extracted 4,258,615 collaboration relationships from this dataset. The size and quality of this dataset enable us to systematically investigate the interplay of scientists' collaboration behaviors and scientific impact. Below, we briefly explore and characterize this academic collaboration social network data.

First, we examine the evolution of the computer science community. Figures 2(a) and 2(b) show the yearly number of computer science publications and authors between 1950 and 2010. We observe an exponential development in the number of computer science publications and researchers during this period. Figure 2(b) also provides the number of authors whose first published paper was in the corresponding year. Figure 2(c) shows the average number of papers that each author publishes and the average number of authors for each scientific publication. We observe that, between 1950 and 2010, average publication output remained roughly constant (blue circle), while collaboration gradually but substantially expanded (red square).

III. Collaboration Signatures

We suspect an interplay between a researcher's collaboration network and his or her scientific impact. Accordingly, from each researcher's academic publication records, we extract an ego collaboration network, which we then use to define a unique, personalized collaboration signature.

Conceptually, the ego network of an individual in a social network is defined as the set of ties that this individual (i.e., ego) has to his or her friends [16]. The concept of an ego network has attracted significant attention, largely due to the important role they play in benefiting individuals. Recently, it has even been argued that a sufficient understanding of ego networks can reveal previously undiscovered mechanisms that underlie network dynamics [13]. Yet despite an

[1]The dataset is publicly available at https://aminer.org/AMinerNetwork.

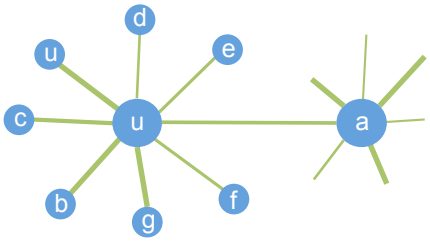


Figure 3. **A collaboration signature.** u's collaboration ego network consists of the ego u and u's collaboration relationships, including the self-collaboration with u. Collaborations are indicated by lines, with collaboration strength denoted by line thickness. u's collaboration signature is the distribution of the fraction of collaboration weight with each collaborator.

expanding literature on ego networks [17], [13], there is a dearth of published work that investigates the self-tie (the tie between the individual and her- or himself). This is particularly unfortunate, as self-ties, arguably unlike online social networks, are non-neglectable components in the academic collaboration network when studying single-author publication. Here we consider a single-author publication as a self-collaboration and define a collaboration ego network as below:

Definition **Collaboration Ego Network.** The collaboration ego network of a researcher u consists of u as the focal ego and represents all of u's collaboration relationships, including self-collaboration.

Figure 3 shows an illustrative example of a collaboration ego network for researcher u, where $a, \cdots, g$ and u are located around u in the ego network. We further define tie weight in collaboration social networks. In the context of social network analysis, tie weight is usually referred to as the emotional closeness and social investment between two people [13]. More precisely, the tie weight between two researchers u and v in a collaboration network is defined as $w_{uv} = \sum_{p \in P} \frac{1}{n_p - 1}$, where P is the set of publications that u and v coauthored and n_p is the number of authors of each publication $p \in P$ [18], [19], [20]. However this definition is not applicable to the single-author situation, where $n_p = 1$. To situate tie weight in our definition of the collaboration ego network, we redefine tie weight as $w_{uv} = \sum_{p \in P} \frac{1}{n_p}$. By this definition, $w_{uv} = w_{vu}$.

We focus on the way that a researcher divides collaboration effort among collaborators and how these patterns observed in the collaboration ego network have an effect on his or her scientific impact. In particular, we define a "collaboration signature" to capture the collaboration patterns encoded in a collaboration ego network.

Definition **Collaboration Signature.** Given a researcher u, the researcher's collaboration ego network, and the associated tie weight of each collaboration relationship, we define researcher u's collaboration signature as the distribution of the fraction of collaboration weight with each of u's collaborators.

We refer to the fraction of collaboration weights as tie strength, namely the normalized measurement of tie weights in an ego network. Tie strength is formally defined by $s_{uv} = \frac{w_{uv}}{\sum_{k \in \Gamma(u)} w_{uk}}$, where $\Gamma(u)$ is u's collaborators. Note that it is not necessary that s_{uv} is equal to s_{vu}. To study the collaboration signature in u's collaboration ego network, we further define the following four measures that are based on our definition of the collaboration signature:

- *Sociability*: the number of collaborators, $|\Gamma(u)|$;
- *Dependence*: the fraction of collaborators fulfilling $s_{uv} > s_{vu}$, $\frac{\sum_{v \in \Gamma(u)} I(s_{uv} > s_{vu})}{|\Gamma(u)|}$;
- *Diversity*: the Shannon entropy of collaboration strength distribution, $-\sum_{v \in \Gamma(u)} s_{uv} \times \log(s_{uv})$;
- *Self-collaboration*: the fraction of self-collaboration, s_{uu}.

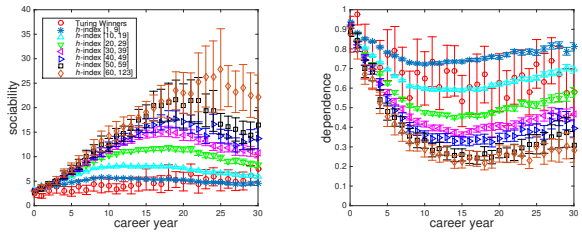
Sociability is derived from Dunbar's number, which is the suggested number of social connections that an individual can comfortably maintain due to cognitive limitations [21]; it thus provides a means of examining the number of collaboration relationships that researchers can maintain throughout their academic careers. *Dependence* indicates the level of one's research dependence. For example, when $s_{uv} > s_{vu}$, papers coauthored by u and v have a relatively larger contribution to u's publications in the mutual collaboration relationship between u and v than they do to v's, which means that u relies more on his/her collaboration relationship with v than v relies on his/her relationship with u. *Diversity* is defined as the Shannon entropy of a researcher's collaboration behaviors; it thus provides a means of investigating how researchers distribute scientific collaborations among different collaborators. A higher *diversity* score implies that a researcher splits his/her collaboration effort more evenly among collaborators. *Self-collaboration* is formulated as a researcher's independence in collaboration. A higher *self-collaboration* score indicates that an individual researcher devotes more effort to independent research than to collaborative endeavors.

IV. Collaboration Signatures and Scientific Impact

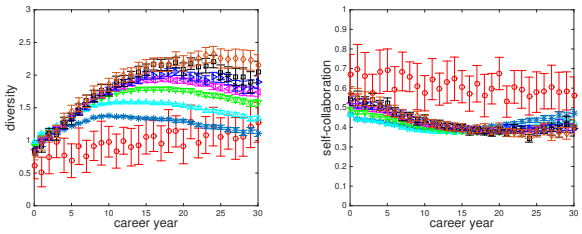
In this section, we investigate the correspondence between the collaboration signatures of researchers in the academic social network and their scientific impact in academia. We quantify researchers' scientific impact based on three intuitive measures: *h*-index, top-venue papers, and big-hit publications.

A. Collaboration Signatures vs. h-index

As a function of both the number of publications and the number of citations per publication [7], the *h*-index is designed to improve upon single-dimension measures such as citation or publication counts. Despite debate over whether the *h*-index is an effective measure of scientific impact, it



(a) Sociability (b) Dependence



(c) Diversity (d) Self-Collaboration

Figure 4. **Collaboration Signature vs. *h*-index.** x-axis: the x^{th} year of one's research career. y-axis: (a) sociability; (b) independence; (c) diversity; (d) self-collaboration. The minimum career length is set to 10 and all characteristics are observed at a 95% confidence interval.

has become a *de facto* standard for measuring academic performance and has been applied to several widely used academic evaluation systems, including ArnetMiner[2] and Google Scholar[3]. Hirsch has also suggested that the *h*-index has the potential to predict academic honors and awards. In this work, we use the *h*-index as a measure of scientific impact. By examining the connection between researchers' collaboration signatures and their respective *h*-indices, we gain insight into the relationship between these collaboration signatures and scientific impact.

Figure 4 shows how collaboration signatures reveal scientific impact as measured by *h*-index. In this figure, the *h*-index is discretized into seven intervals, each representing a group. A researcher is a member of the group represented by the interval in which his or her *h*-index is contained. The resulting intervals from [1, 9] to [60, 123] contain 21393, 23434, 5901, 1849, 647, 247, and 172 researchers, respectively. In addition to these groups, we also plot the collaboration signatures of Turing Award winners from 1966 to 2010. In Figure 5 we illustrate the *h*-indices (as of 2012) of Turing Award winners with respect to the length of their research careers when they received the Turing Award. The length of the researchers' careers prior to winning the award ranges from 11 to 43 years, while their *h*-indices range from 25 to 83.

We characterize a researcher's collaboration signature at each year of his or her research career, where the beginning of an individual's research career is defined as date of the earliest publication attributed to the researcher in our dataset. To construct each researcher's collaboration ego network and

[2]https://aminer.org/

[3]http://scholar.google.com/

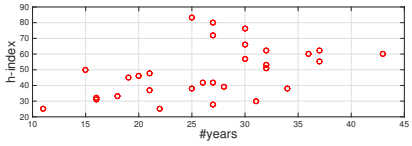


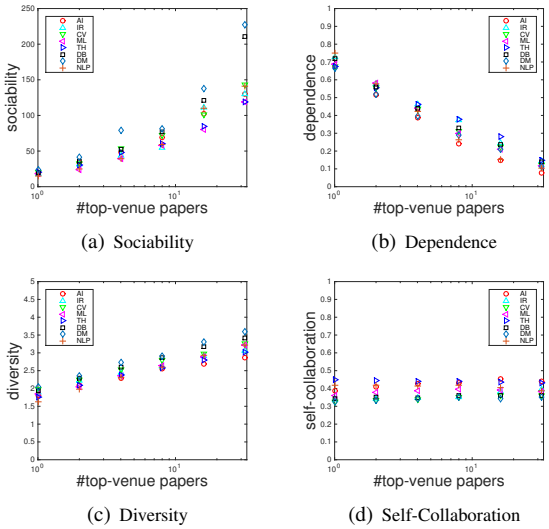
Figure 5. **Turing Award Winners.** x-axis: the number of years since the researcher first published a paper when presented with the Turing Award; y-axis: the researcher's *h*-index as of 2012.

to calculate his or her collaboration signature, we extract all of the researcher's publications spanning the entirety his or her entire research career. We see that researchers with different levels of *h*-indices exhibit significant differences in their collaboration signatures. Additionally, Turing Award winners have the most distinctive collaboration signatures from among the researchers in all four measures.

Figure 4(a) shows the evolution of researchers' sociability across their research careers. Generally, we observe that researchers with higher *h*-indices have greater sociability than those with lower *h*-indices, regardless of career stage. We also observe that sociability tends to increase monotonically for all groups of researchers until it reaches a peak value. The particular peak value of sociability differs for each group, indicating that researchers with different *h*-indices reach their sociability peaks at different points in their research careers. Through the first five years of a researcher's career—the time that may be characterized by his or her graduate studies—each group of researchers shows the same increasing trend in sociability. From the fifth to tenth years, the *h*-index groups ([1, 9], [10,19], and [20,29]) start to follow different trends from other groups. The 30s, 40s, 50s, and 60s *h*-index groups reach their peak sociability values at their 17^{th}, 19^{th}, 22^{th}, and 25^{th} year, respectively. Surprisingly, we find that as a noteworthy group of researchers in computer science, Turing Award winners demonstrate a far lower and more stable level of sociability across their careers than the other groups of researchers.

Figure 4(b) shows the evolution of researchers' dependence across their academic careers. By the fifth year—the typifying end of a budding researcher's graduate studies—researchers with different *h*-indices exhibit different degrees of collaboration dependence. Also by their fifth year, the 50s and 60s groups exhibit dependence scores around 0.5. In general, researchers' dependence scores decrease at the beginning of their careers and only begin to increase after they have been a member of the research community for more than fifteen years. We also observe that high-*h*-index researchers maintain consistently lower dependence scores.

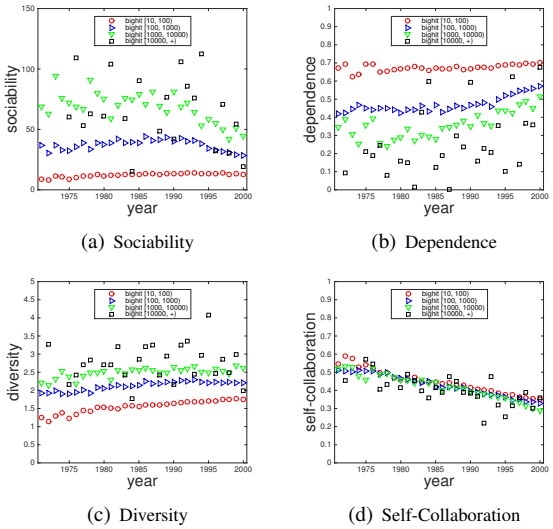
Figure 4(c) shows the diversity of researchers' collaboration behaviors. Researchers start to display different diversity values at around the 8^{th} year of their careers. The diversity values of the 10s, 20s, 30s, and >40s groups stop increasing and stabilize around the 8^{th}, 12^{th}, 15^{th}, 18^{th} year of their research careers, respectively.



(a) Sociability (b) Dependence

(c) Diversity (d) Self-Collaboration

Figure 6. **Collaboration Signatures vs. Top Venues.**



(a) Sociability (b) Dependence

(c) Diversity (d) Self-Collaboration

Figure 7. **Collaboration Signatures vs. Big Hits.**

Figure 4(d) shows the evolution of researchers' self-collaborations. We see that the long-term difference in self-collaboration between groups of scholars with different *h*-indices can be identified even at the very early stage of their careers. We also observe that the fraction of researchers' self-collaborations decreases gradually over time, though Turing Award winners exhibit much higher self-collaboration values than other researchers. Overall, the fraction of self-collaboration reaches a stable state after one has been a member of the research community for more than fifteen years.

Summary. Based on the above analysis, we arrive at the following conclusions. First, collaboration signatures can serve to distinguish researchers of different levels of scientific impact (as evaluated by *h*-index). Second, researchers tend to produce stabilized collaboration signatures only after about fifteen years from the start of their research careers. Third, researchers form different collaboration signatures during the first fifteen years of their research careers, which can significantly affect their future scientific impact. Finally, as a representative group of outstanding computer scientists, Turing Award winners produce the most distinguishable collaboration signatures, even from the very early stages of their careers.

B. Collaboration Signatures vs. Top Venues

Researchers aim to frequently publish influential scientific work in prestigious venues, whereby their work can be effectively disseminated and their influence accumulated. In turn, these influential publications serve to maintain or even elevate the prestige of their respective venues. Thus when assessing scientific impact, it is important to account not only for the influence of a paper itself, but also for the prestige of the venue in which it is published. To this end, we count the number of "top-venue" papers as a measure of a researcher's scientific impact. Specifically, we extract eight focus areas in computer science that are considered to have top publication venues, namely Artificial Intelligence (AI), Information Retrieval (IR), Computer Vision (CV), Machine Learning (ML), Theory (TH), Databases (DB), Data Mining (DM), and Natural Language Processing (NLP). For each area, we choose three top venues[4]. Finally, we use the number of top-venue papers in each area to quantify a researcher's scientific impact.

Figure 6 shows the collaboration signatures of researchers with different scientific impact (as evaluated by the number of top-venue papers). The x-axis of the figure represents the number of top-venue publications in each area. Note that, if a researcher publishes in top venues for more than one area, the area with the majority of the researcher's publications is designated as his or her research area.

From Figure 6(a), we find that, regardless of research area, the degree of sociability exhibited by researchers tends to increase as they produce more top-venue publications. We observe a similar trend for diversity scores (see Figure 4(c)), where research diversity tends to increase with the number of publications that reach top-venues. By contrast, Figure 4(b) shows that research dependence decreases with the number of publications that reach top-venues. Thus as researchers achieve greater scientific impact (as measured by a greater number of top-venue papers), their research dependence decreases while their collaboration diversity increases. In Figure 6(d), we observe that there is no obvious trend between the degree of self-collaboration and the number top-venue publications. All of these observations are hold true for all of the computer science research areas considered.

[4]**AI:** IJCAI, AAAI, ECAI. **IR:** SIGIR, ECIR, TREC. **CV:** CVPR, ICCV, ECCV. **ML:** ICML, NIPS, ECML. **TH:** FOCS, STOC, SODA. **DB:** SIGMOD, VLDB, ICDE. **DM:** KDD, ICDM, SDM. **NLP:** ACL, EMNLP, COLING.

C. Collaboration Signatures vs. Big Hits

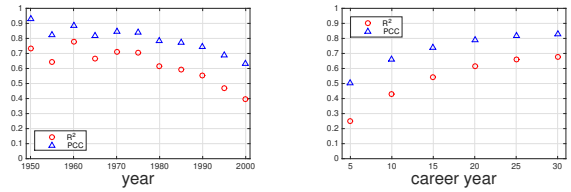
Famous researchers, such as Turing Award winners, are often recognized for their most-influential work. For example, the Nobel Prize in Physics is usually awarded to researchers in recognition of their outstanding contributions to science disseminated through a landmark publication [22]. As another example, the ACM Infosys Foundation Award in 2013 was presented to Dr. David Blei "*for pioneering the area of topic modeling*," with the accolade explicitly referencing his landmark paper on Latent Dirichlet Allocation[5]. Accordingly, in this work we consider the "big-hit" paper (i.e., the most cited one among a researcher's publications) as a measure of scientific impact.

Figure 7 shows the collaboration signatures of researchers with different levels of big-hit papers. First, each researcher is associated with the number of citations of his or her big-hit publication. For each year, we then classify each new researcher into one of four groups based on the number of citations accumulated by his or her big-hit paper, namely [10, 100), [100, 1000), [1000, 10000), and [10000, $+\infty$). The x-axis shows the year from 1970 to 2000 and the y-axis represents the four properties of collaboration signature.

We observe that researchers with different levels of big-hit papers display significantly different collaboration signatures. Similar to the observations above, on the one hand researchers with high scientific impact have high sociability and diversity, while on the other hand they have low dependence in collaborations. We also observe that the self-collaboration ratio is not indicative of the scientific impact that is quantified by big-hit papers. Finally, we observe that, in general, while different groups of researchers maintain different collaboration signatures, each group of scholars at each year has smoothly equal collaboration signatures from 1970 to 2000 in terms of sociability, dependence, and diversity from Figures 7(a), 7(b), and 7(c), respectively.

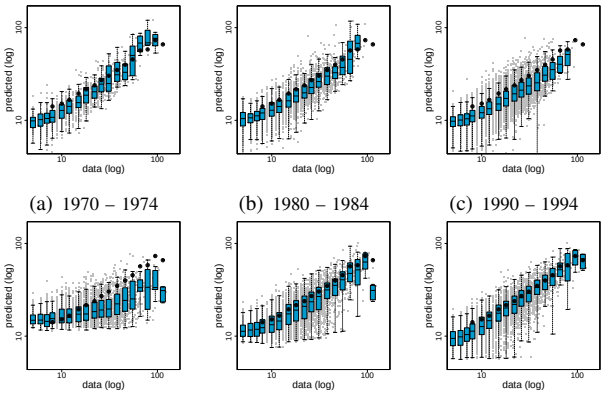
Summary. In the above sections, we empirically explore the correspondence between collaboration signatures and scientific impact as measured by h-index, number of top-venue papers, and number of citations for big-hit papers. Regardless of which measure of scientific impact is employed, researchers of a given level of scientific impact appear to retain similar collaboration signatures. Furthermore, we find that researchers that exhibit different levels of scientific impact can be distinguished by four signature measures of collaboration, and that these signatures may even collectively serve as an indicator of their future scientific impact. Consequently, our observations engender important implications for applications that require the understanding of scientists' collaboration behaviors, as well as for the study of the mechanisms that underlie the development and progression of scientific impact.

[5]http://awards.acm.org/award_winners/blei_3974465.cfm



(a) (b)

Figure 8. **Extent to which scientific impact can be revealed from collaboration signatures, conditioned on (a) the start year of the researchers' academic careers (x-axis) or (b) the first x years of the researchers' academic careers (x-axis).**



(a) 1970 – 1974 (b) 1980 – 1984 (c) 1990 – 1994

(d) First 5 years (e) First 15 years (f) First 25 years

Figure 9. ***h*-indices in data vs. predicted *h*-indices from collaboration signatures.** x-axis: h-indices in data; y-axis: predicted h-indices. A diagonal line from (0, 0) to (123, 123) would denote perfect prediction results.

V. Can Scientific Impact Be Predicted from Collaboration Signatures?

In this section, via two prediction-based case studies, we further explore the extent to which research scholars' scientific impact can be inferred from their collaboration signatures. Please note that the focus of these studies is to demonstrate how scientific impact can be revealed from collaboration signatures.

A. Predicting for Different Generations

Our first case study is to use the collaboration signatures of researchers who enter academia in different periods to predict their respective h-indices. For example, by setting five year intervals, we divide researchers into 11 distinct groups, ranging from 1960 to 2004, based on the first year they joined academia (e.g., [1960, 1964], [1965, 1969], $\cdots$, [2000, 2004]). We then collect the researchers' collaboration signatures, which we use as features to predict their h-indices in 2012 (the last year represented in our dataset).

As our goal is to provide further evidence of the correspondence between collaboration signatures and scientific impact, we use linear regression to report the results, primarily due to the method's predictive power and simplicity. The only features used are the four values previously elaborated

upon that are based on researchers' collaboration signatures, namely: sociability, dependence, diversity, and self-collaboration. To demonstrate the extent to which scientific impact can be revealed by collaboration signatures, we employ two methods to evaluate the results of linear regression, R^2 [23] and Pearson correlation coefficient (PCC).

To illustrate the use of collaboration signatures, we demonstrate how they reveal the scientific impact (as quantified by *h*-index) generated by researchers whose careers began in different years. Figure 8(a) reports the prediction performance for 11 groups of researchers as measured by R^2 and PCC. We observe that between 1960 and 1980, R^2 ranges between 0.6 and 0.8, while PCC ranges between 0.8 and 0.93. This performance indicates that researchers' scientific impact can be reasonably inferred from our four simple collaboration signatures. We also note that the prediction performance declines with time. The reason for this decline is trivial: if some researchers join academia later than others, then less data—and thus less information—can be extracted from their respective signatures.

Further, we compare the actual *h*-indices with our predicted ones. In Figures 9 (a,b,c), the black circles correspond to the average predicted *h*-index per bin. These figures clearly illustrate the decline in inferential power associated with limited time constraints.

B. Predicting from the Early Stages

Our second case study is to examine how our ability to infer a researcher's scientific impact from his or her collaboration signatures is affected by the length of his or her research career (and hence the length of the extracted signatures). To this end, we first choose researchers who have publication records in the dataset that extend at least 30 years. We then extract the researchers' yearly collaboration signatures from their publications. The cumulative collaboration signatures beginning from the first career year are used as the input of the regression model.

Figure 8(b) demonstrates that with longer (in terms of years) collaboration signatures, future scientific impact can be predicted with increasing fidelity, as measured by R^2 and PCC. By using the first year data, the R^2 and PCC are 0.25 and 0.5, respectively. By using 15-year collaboration data, however, the performance reaches 0.55 and 0.75 in terms of R^2 and PCC, respectively. These prediction results serve to validate our observations that collaboration signatures can reveal scientists' scientific impact (see Figure 4).

Similarly, we also provide the comparisons between the actual *h*-indices and the predicted *h*-indices in Figures 9 (d,e,f). The performance improves when additional years of signatures are employed for the inference.

Overall, these two predictive case studies provide evidence that scientists' collaboration signatures can serve as indicators of their scientific impact. The experimental results also demonstrate the effectiveness of collaboration signatures for predicting future scientific impact.

VI. Related work

Science has developed a merit-driven career process whereby an individual is promoted through various career stages based on the evaluation of his or her past achievements and the perceived potential for future achievement. Correctly assessing past scientific impact and the potential for future impact is, therefore, absolutely essential for the effective evaluation of individual scientists.

Several measures have been proposed for assessing a scholar's scientific impact [24]. Perhaps the most widely used impact measure is the number of citations a scholar has accumulated. However, citation counts are contingent on the length of one's scientific career, and thus favor elder scientists. To mitigate this and other related effects, several alternate measures have been proposed that in some way rely on citations. One such measure, the *h*-index, attempts to measure both the productivity and impact of the published work of a research scholar [7]. In this work, we propose to measure the scientific impact of research scientists by using not only the *h*-index, but also the number of top-venue publications and the citations of big-hit publications.

Despite debate over the suitability of scientific impact measures, predicting a scholar's future impact has nonetheless received significant attention [5], [8], [25]. As a focal effort, the KDD Cup 2003 held a data mining competition to estimate citation counts [10], and many efforts to predict the number of future citations for scholarly work have followed [2], [9], [5], [8]. More recently, Dong et al. formulate a scientific impact prediction problem of inferring whether a given publication will increase its authors' *h*-indices in the future [12]. Although a tremendous amount of work has explored the prediction of future scientific impact, the study of the effects of collaboration patterns in academia on scholars' scientific impact has received scant attention.

Yet, it has been repeatedly demonstrated that networks can have a profound impact on our lives. Shi et al. study the interplay between a publication's citation network and its scientific impact [26]. Yang et al. propose to infer the number of future collaborators in academic social networks by using historical data [27]. In the work of [28], [29], the authors measure researchers' scientific impact using their coauthorship network centralities. These discoveries, among others, inspire our exploration into the connection between collaboration networks and scientific impact.

To that end, we demonstrate that research scientists exhibit distinguishable collaboration signatures. We find that these collaboration signatures can serve as powerful indicators of real-world scientific impact, and can be applied to other tasks in academic network mining, such as name disambiguation [30] and collaboration prediction [31], [32].

VII. Conclusion

In this work, we study the interplay of collaboration behaviors in academic social networks and research scholars' scientific impact. We focus on the collaboration ego network and formally define the collaboration signature of a scientist.

We discovered that scientists with similar level of scientific impact appear to retain similar collaboration signatures. We also observed that scientists' collaboration signatures are indicative of the progression of their scientific impact. Additionally, we found that the collaboration signature is a strong indicator of researchers' *h*-indices. Overall, our findings provide empirical evidence that scientific impact can be inferred from collaboration signatures, and also offer important implications for understanding the mechanisms that underlie the progression of scientific impact.

Despite the promising results, there is still much room left for future work. First, while the function of collaboration signatures has been identified in the area of computer science, we have limited knowledge about whether these observations are equally valid in other scientific disciplines. Second, in this work we refrain from examining the causality between scientific impact growth and collaboration signature development, which could be expanded in future work.

Acknowledgments. This work is supported by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA) grant #FA9550-12-1-0405, and the National Science Foundation (NSF) Grant OCI-1029584.

References

[1] J. Tang, S. Wu, J. Sun, and H. Su, "Cross-domain collaboration recommendation," in *KDD '12*, 2012, pp. 1285–1293.

[2] D. E. Acuna, S. Allesina, and K. P. Kording, "Future impact: Predicting scientific success," *Nature*, vol. 489, no. 7415, pp. 201–202, Sep. 2012.

[3] J. A. Evans, "Future science," *Science*, vol. 342, no. 6154, pp. 44–45, 2013.

[4] B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones, "Atypical combinations and scientific impact," *Science*, vol. 342, no. 6157, pp. 468–472, 2013.

[5] D. Wang, C. Song, and A.-L. Barabási, "Quantifying long-term scientific impact," *Science*, vol. 342, no. 6154, pp. 127–132, 2013.

[6] X. Shi, L. A. Adamic, B. L. Tseng, and G. S. Clarkson, "The impact of boundary spanning scholarly publications and patents," *PLoS ONE*, vol. 4, no. 8, p. e6547, 08 2009.

[7] J. E. Hirsch, "An index to quantify an individual's scientific research output," *PNAS*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[8] H. Shen, D. Wang, C. Song, and A.-L. Barabási, "Modeling and predicting popularity dynamics via reinforced poisson processes," in *AAAI'14*, 2014, pp. 291–297.

[9] R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li, "To better stand on the shoulder of giants," in *JCDL '12*, 2012, pp. 51–60.

[10] J. Gehrke, P. Ginsparg, and J. M. Kleinberg, "Overview of the 2003 kdd cup," *SIGKDD Explorations*, vol. 5, no. 2, pp. 149–151, 2003.

[11] F. Radicchi, S. Fortunato, and C. Castellano, "Universality of citation distributions: Toward an objective measure of scientific impact," no. 45, 2008.

[12] Y. Dong, R. A. Johnson, and N. V. Chawla, "Will this paper increase your *h*-index?: Scientific impact prediction," in *WSDM '15*. ACM, 2015, pp. 149–158.

[13] J. Saramki, E. A. Leicht, E. Lpez, S. G. B. Roberts, F. Reed-Tsochas, and R. I. M. Dunbar, "Persistence of social signatures in human communication," *PNAS*, vol. 111, no. 3, pp. 942–947, 2014.

[14] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: Extraction and mining of academic social networks," in *KDD '08*, 2008, pp. 990–998.

[15] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in *KDD'09*, 2009, pp. 807–816.

[16] L. C. Freeman, "Centered graphs and the structure of ego networks." *Mathematical Social Sciences*, vol. 3, no. 3, pp. 291–304, 1982.

[17] S. Gupta, X. Yan, and K. Lerman, "Structural properties of ego networks," in *SBP'15*, April 2015.

[18] M. E. J. Newman, "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality," *Phys. Rev. E*, vol. 64, p. 016132, 2001.

[19] A. Barrat, M. Barthlemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *PNAS*, vol. 101, no. 11, pp. 3747–3752, 2004.

[20] Q. Ke and Y.-Y. Ahn, "Tie strength distribution in scientific collaboration networks," *Phys. Rev. E*, vol. 90, p. 032804, 2014.

[21] R. Dunbar, "Coevolution of neocortical size, group size and language in humans," *Behavioral and Brain Sciences*, vol. 16, pp. 681–735, 1993.

[22] H.-W. Shen and A.-L. Barabási, "Collective credit allocation in science," *PNAS*, vol. 111, no. 34, pp. 12 325–12 330, 2014.

[23] L. Magee, "R^2 measures based on Wald and likelihood ratio joint significance tests," *The American Statistician*, vol. 44, no. 3, pp. 250–253, 1990.

[24] E. Garfield, "Citation indexes for science: A new dimension in documentation through association of ideas," *Science*, vol. 122, no. 3159, pp. 108–111, 1955.

[25] M. Newman, "Prediction of highly cited papers," *Europhysics Letters (EPL)*, vol. 105, no. 2, p. 28002, 2014.

[26] X. Shi, J. Leskovec, and D. A. McFarland, "Citing for high impact," in *JCDL '10*. ACM, 2010, pp. 49–58.

[27] Y. Yang, Y. Dong, and N. V. Chawla, "Predicting node degree centrality with the node prominence profile," *Scientific reports*, vol. 4, 2014.

[28] E. Yan and Y. Ding, "Applying centrality measures to impact analysis: A coauthorship network analysis," *JASIST*, vol. 60, no. 10, pp. 2107–2118, 2009.

[29] S. Servia-Rodriguez, A. Noulas, C. Mascolo, A. Fernandez-Vilas, and R. Diaz-Redondo, "The evolution of your success lies at the centre of your co-authorship network," *PLoS ONE*, pp. –, 2014.

[30] J. Tang, A. C. Fong, B. Wang, and J. Zhang, "A unified probabilistic framework for name disambiguation in digital library," *IEEE TKDE*, vol. 24, no. 6, pp. 975–987, 2012.

[31] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo, "Mining advisor-advisee relationships from research publication networks," in *KDD'10*, 2010, pp. 203–212.

[32] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla, "When will it happen?: Relationship prediction in heterogeneous information networks," in *WSDM '12*, 2012, pp. 663–672.