What makes a good researcher?

Abhishek Arora
Stanford University
arorabhi@stanford.edu

Anshul Mittal
Stanford University
anmittal@stanford.edu

Raghav Pasari
Stanford University
rpasari@stanford.edu

ABSTRACT

In this paper we investigate the characteristics of a good researcher and contrast them with that of other researchers which would give further insights into how to become a good researcher. For this we analyze the collaboration and citation pattern of an author. Furthermore, we propose a new metric which takes into account just the local characteristics of an author and show how well it performs with respect to a global metric like pageRank in determining the goodness of an author. For our experiments, we use the DBLP citation data which has more than four hundred and fifty thousand papers and close to a million authors.

1. INTRODUCTION

Scientific collaboration and citation practices have been an important focus for social scientists, seeking to provide insight into scienctific research as an inherently team-based endeavour. Several studies [4] [8] have also shown the significant effect that such practices are known to have on scientific productivity.

Scientists have traditionally used author collaboration and paper citation networks to study such practices. These networks are very common in the research community and share common properties with other well known networks. More specifically, they are both scale free networks following a power law degree distribution, having one big central community etc. among other things. While collaboration networks are relevant for understanding the network structure of scientific fields (Jones et al., 2008), citation networks are central in providing insight into the hierarchies within a field and among fields.

However, besides finding general trends in the practices of the research community as a whole, it will also be interesting to observe how such practices vary from one researcher to another and how they are related to one's scientific merit. Every advisor has some tips for his/her students with regard to the research practices he/she ought to follow, for example, *always do thorough literature search and cite many papers*, *focus on quality rather than quantity* etc., however, little work has been done to evaluate the correlation between such practices and the scientific merit or influence. In this paper, we attempt to analyse and throw light on this side of the story. We consider several such widely accepted standards regarding the collaboration and citation practices, and evaluate them by developing some novel metrics. In order to correlate the metric values with an author's scientific influence, we experiment with several different influence metrics and pick the best one for our analysis. For the purpose of our study, we also develop a new kind of network which we call the ”author citation network”. We explore the general properties and community structure of this novel network and compare it with the more standard author collaboration and paper citation networks.

Our study reveals some interesting trends. For example, we observe that ”good” authors tend to collaborate more with other ”good” authors emphasizing the importance of peer group in research. Similarly, our studies reveal that although the successful authors tend to focus more on the quality of research, they never seem to compromise on quantity either! Here, we must note that we don't have information about the 1st and 2nd authorship, etc. The following sections reveal these and many more captivating patterns in the citation and collaboration practices.

The rest of the paper is organized as follows. The next section presents a critical comparison of the previous work in this domain. In Section 3, we briefly discuss the dataset that we have used and the different graphs that we have constructed for analysis. Section 4 discusses the different metrics used for ranking the authors. We validate these metrics against some well known facts about these researchers and then use the best metric for analysis in the rest of the paper. Section 5 explores the citation practices, patterns of the researchers and their papers, diversity of the communities they publish papers in, how many and what kind of papers they cite. Section 6 performs similar analysis for an author's collaboration practices. In Section 7, we attempt to learn some important lessons from the career graphs of such researchers, that is, how the quality and quantity of their papers tends to change over time and whether there is any difference in the temporal patterns between them and others.

2. RELATED WORK

Citation networks have been studied quantitatively almost from the moment citation databases first became available, perhaps most famously by Derek de Solla Price [3]. Since then, several researchers have studied the topological properties and graph statistics of paper citation networks and proposed generalised models for their formation [1] [12]. Most recently, Leicht et al. [5] has explored important temporal characteristics of such networks. However, all these papers deal with citation networks where the nodes are scientific papers/documents and not the authors themselves. Therefore, even though they reveal interesting trends about scientific research, little information can be obtained directly about the authors themselves.

Similarly, collaboration networks have been the subject of extensive study to find patterns in the collaboration practices of researchers in various fields. The most notable work in this domain is by M.E.J. Newman in 2001 [9] [10] where he answers some pertinent questions concerning such networks such as the number of collaborators per scientist, size and properties of giant component, typical distances and centrality measures. In much the same way as citation networks, other researchers like Bettencourt et al. [7] have also studied the temporal evolution and properties of such networks.

Most recently, Ding [4] and Wallace et al. [8] have studied the influence of collaboration networks on citation practices. More specifically, they try to reveal trends in citations using degrees of separation in collaboration network (including self-citations). However, all of the above mentioned papers present an average analysis for all authors and do not present an analysis of such properties with respect to the influence of the authors in the scientific community. Moreover, none of the papers explore the temporal variation in such characteristics.

3. INPUT GRAPHS AND THEIR PROPERTIES

The data set used in the paper is the DBLP papers and citation relation data set available at [6]. It consists of all database related papers in the DBLP network up to October 2010 (even though the dataset has been previously advertised as consisting of all computer science researchers, our results have shown that the dataset is heavily biased towards the field of databases). The data includes the following attributes for each paper -

- Unique index id of the paper
- Title of the paper
- List of authors (comma separated)
- Year of publication
- Venue of publication
- List of paper ids cited by this paper

In this paper, we restrict our analysis to the paper id, list of authors, year of publication and other papers cited by the paper but do not use the title and the venue of publication. Using this data set, we contruct three different kinds of graphs-

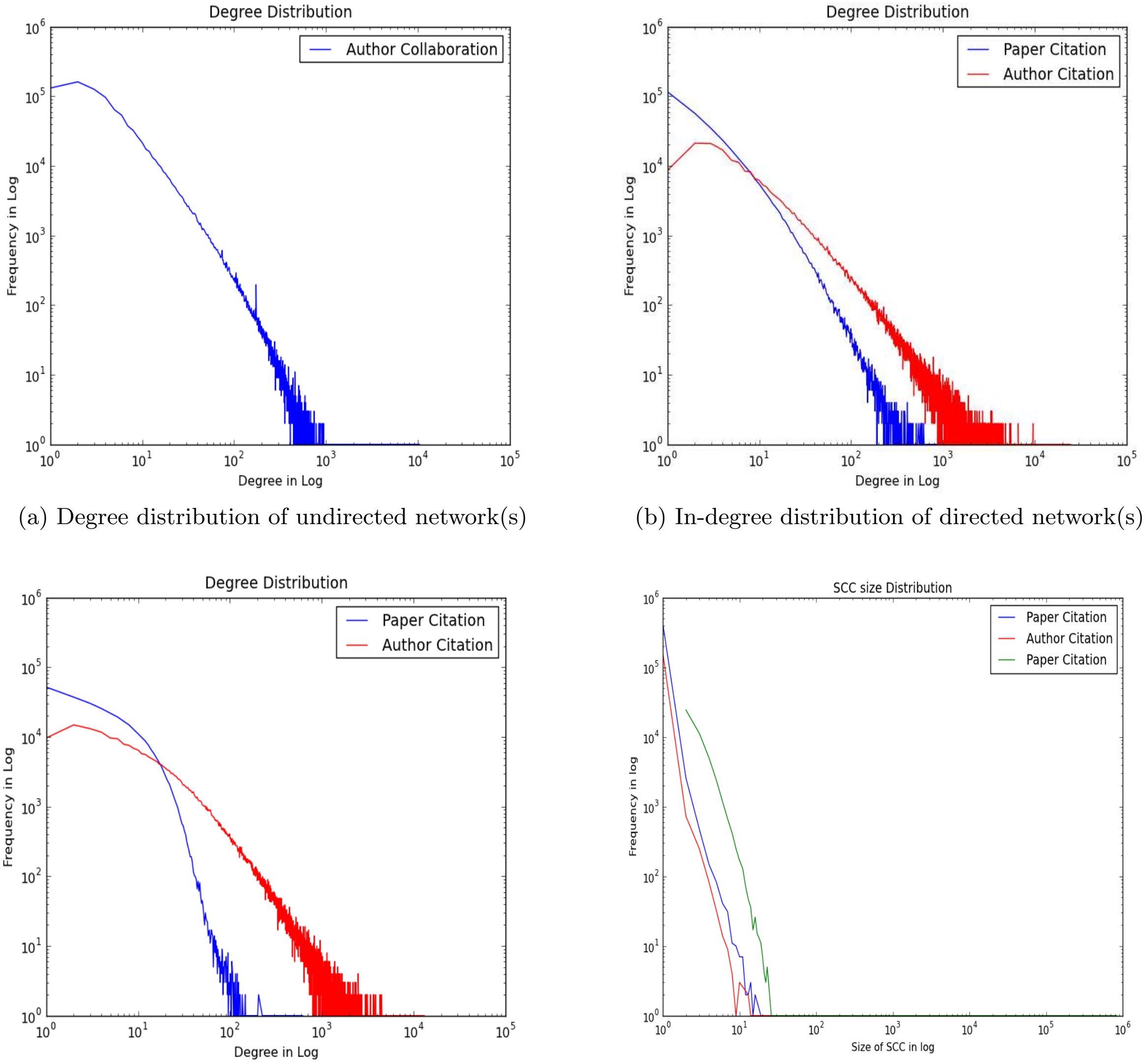
1. Author Collaboration
2. Paper Citation
3. Author Citation

The first two are the more traditional author collaboration and paper citation graphs. In the author collaboration graph, every author is a node and an edge between two authors indicates that the two collaborated on some paper at some point in time (please note that such a formulation will result in multiple small cliques in our graph because all authors of a particular paper will have all possible edges amongst themselves). In the paper citation graph, every paper represents a node and an edge from node a to node b paper a cites paper b. While such a network gives valuable information about the citations of the papers, we cannot extract information about the citation practices of authors directly from this network (there can be multiple heuristics for deriving this information indirectly, for e.g. summing over all papers of one author etc.). For this reason, we construct a third network in which every author is a node and an edge(directed) denotes that one author was cited by the other in his/her paper. Such a network, hereafter referred to as the author citation network, can be directly used to investigate important characteristics of citation practices of authors. It is important to note here that the author and paper citation networks are directed whereas the author collaboration network is undirected. Moreover, the author citation and collaboration networks are multi-graphs (multiple edges can exist between two nodes) whereas the paper citation graph is a simple DAG (directed acyclic graph).

Table 1 lists some properties of the three graphs (the graphs are shown in Figure 1). In these graphs, the clustering coefficient and SCCs are generated after removing the multi-edges. It can be seen that the author collaboration graph has a high clustering coefficient, small number of SCCs and large number of nodes in the biggest SCC, all due to the presence of multiple small cliques in the network. From the table, it is also evident that the author citation graph is an aggregator of paper citation graph for every author. It has higher clustering coefficient and higher number of nodes in largest SCC and also has lesser number of nodes as compared to paper citation graph (because one author can have multiple papers). However, it has much more edges than the paper citation graph due to the presence of multi edges. The author collaboration graph has much more nodes than the other two networks because the data set has many papers about which no citation information is given.

It can also be seen that in each of the plots, the curve first increases for small values and then declines. This is because almost every publication tends to have a few collaborators and some other papers that it cites. Moreover, most papers also tend to be cited by at least a few papers, leading to the decline in indegree of citation networks.

Let us now discuss the community structure of these graphs.



(a) Degree distribution of undirected network(s)

(b) In-degree distribution of directed network(s)

(c) Out-degree distribution of directed network(s)

(d) SCC distribution of all networks

Figure 1: Properties of graphs

Table 1: Properties of graphs

Property	**Paper Citation**	**Author Citation**	**Author Collaboration**
Number of Nodes	475886	334764	975001
Number of Edges	2327450	10746625	5644866
Average Clustering Coefficient	0.11439	0.35763	0.630
Number of SCCs	419193	159840	46021
Percentage of Nodes in Largest SCC	.1081	.5171	.8576

3.1 Community structure of the input graphs

In order to determine the community structure of the networks involved, we use the fast agglomerative clustering method for large networks described in [11]. The method starts with considering every node to be in a different community. It then performs two phases repeatedly. In the first phase, we consider the neighbours j of i and we evaluate the gain of modularity that would take place by removing i from its community and by placing it in the community of j. The node i is then placed in the community for which this gain is maximum, but only if this gain is positive. If no positive gain is possible, i stays in its original community. In the next phase, the graph is collapsed so that each community acts as a node in the new graph. While performing the experiments, we ignore all communities with membership below a certain threshold.

Using this method, we obtain the community structure of

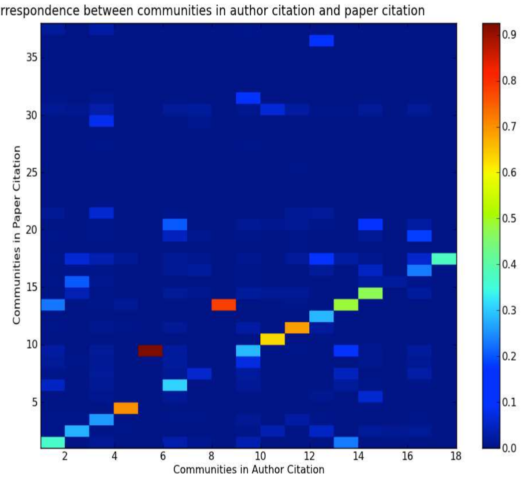


Figure 2: Correspondence between community structure of paper and author citation networks

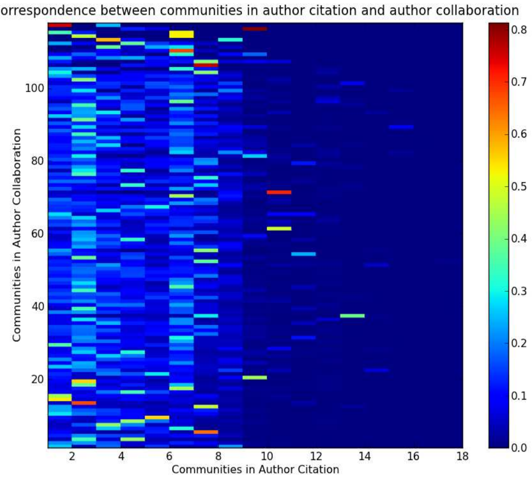


Figure 3: Correspondence between community structure of author collaboration and author citation networks

the three graphs involved. It is important to note that the communities in the citation networks will correspond to the researchers publishing on a particular topic or sub branch of computer science whereas the communities in the collaboration network may correspond to all researchers working in a particular institution or a research lab. Figure 2 shows the correspondence between the communitites of the author and paper citation networks which are plotted on the X and Y axis, respectively. The metric,m_1 used here is-

$$m_1(x,y) = \frac{\sum_{\forall authors \in x} \text{fraction of papers of the author in y}}{\text{\#of authors in x}}$$

While evaluating this metric, we consider only the top 80% authors (based on the fraction of their papers lying in the paper community y) in the author community to remove any outliers. As is evident from the graph, there is an almost one to one correspondence between the community structure of the two networks confirming our hypothesis that both networks reveal content based communities. The one to one correspondence is visible as the bright diagonal which says that the communities have a high score. The majority blue area shows that most communities do not correspond to each other.

Figure 3 shows the correspondence between the communitites of the author citation and author collaboration networks which are plotted on the X and Y axis, respectively (while constructing this plot we consider only those authors which occur in both networks). The metric plotted here is the number of authors in the intersection of the two communities divided by the number of authors in the collaboration network (which is always the smaller of the two communitites). A careful look at the plot reveals that one community in the citation network matches with several communities in the collaboration graph whereas one community in the collaboration graph has a significant overlap with only a few citation communities. This corresponds to a row in the color map having very few (usually 1) bright value but a column can have multiple bright values. This is expected because one research area in databases will have researchers from many institutions publishing in that area whereas the database group in one institution is expected to have focus in a few research areas only.

Having discussed the different charateristics of our networks, we next move to the measures for finding the influential authors using these networks.

4. INFLUENCE METRICS

Ranking of the scientific influence or productivity of authors has been a long standing area of research. Several researchers have proposed metrics to evaluate a researcher's measure of influence in the scientific community. Some of these are number of papers, number of citations, average number of citations, H-index, G-index etc. Out of these, H-index is widely considered to be the best and most standard metric. These are examples of local metrics which can be evaluated with only the local information about every node. However, in our case where we are additionally given the graph of researchers, a better ranking may be obtained by obtaining an influence ranking of the nodes in the network.

The next questions, therefore, are that which metric should be used for such an evaluation and which networks to evaluate it on. First, let us consider the question of which metric to use. There are several measures which can be used for such an evaluation, namely, betweenness centrality, closeness centrality, degree centrality, HITS (hubs and authorities), pageRank etc. In this paper, we use the pageRank [2] to find influential nodes in the network. The pageRank algorithm is an influence metric where a node is considered important if it is pointed to by other important nodes. If r is the *rank vector* of the network and M is the *stochastic adjacency matrix* (as defined in [2]), the rank r_j for node j is given by the equations-

Table 2: Pearson correlation of different influence metrics

	H-index	pRankAuCollab	pRankAuCit	numPapers	numCitations	avgCitations
H-index	1.0	0.49	0.65	0.64	0.72	0.28
pRankAuCollab	0.49	1.0	0.35	0.70	0.39	0.05
pRankAuCit	0.65	0.35	1.0	0.39	0.90	0.28
numPapers	0.64	0.70	0.39	1.0	0.50	0.06
numCitations	0.72	0.39	0.90	0.50	1.0	0.29
avgCitations	0.28	0.05	0.28	0.06	0.29	1.0

Table 3: Top 5 authors using different metrics

Rank	pageRank in author collaboration	pageRank in author citation	H-index
1	Alberto L. Sangiovanni-Vincentelli	Jeffrey D. Ullman	Hector Garcia Molina
2	Hans-Peter Seidel	Jim Gray	Jeffrey D. Ullman
3	Thomas s. Huang	E. F. Codd	David J. Dewitt
4	Donald F. Towsley	C. A. R. Hoare	Rakesh Agrawal
5	Ron Kikinis	Donald D. Chamberlin	Scott Shenker

$r_j = \sum_{i->j} r_i/d_{out}(i)$
$r = Mr$

Coincidentally, the pageRank algorithm derives its inspiration from the citation practices in scientific communities making it the most relevant evaluation metric for our analysis. It is also among the fastest in the many metrics discussed above (much faster than betweenness) and therefore, more suitable for the present scenario where we have more than a million nodes in the graph. We evaluate this metric on two networks - author collaboration and author citation network (we don't use the paper citation network because, as noted in the previous section, the author citation network is an aggregator of paper citation network and is much more suitable for evaluating properties of authors).

Table 2 shows the pearson correlation between the rankings obtained by using different metrics- pRankAuCollab measures the pageRank in author collaboration network, pRankAuCit measures the pageRank in author citation network, numPapers ranks on the basis of number of papers, numCitations uses the total number of citations obtained by the author and avgCitations uses the number of citations obtained per paper. It is clearly evident that the pageRank in author citation network is biased towards the number of citations and average number of citations whereas the pageRank in collaboration network performs poorly vis-a-vis these metrics and tends to give more weight to the number of papers (traditionally not considered to be a good metric of scientific productivity although one that has sadly become quite omnipresent these days!). An interesting observation is that H-index has higher correlation with the number of papers than the citation pageRank because the H-index is necessarily bound by the number of papers of an author. H-index also has a lower correlation with the number of citations than the citation pageRank (both have same correlation with respect to average citations). Finally, H-index is better correlated with citation pageRank than the collaboration pageRank.

Table 3 lists the top 5 authors obtained using three metrics - pageRank in author collaboration network, pageRank in author citation network and the widely used H-index. By looking at the profiles of these researchers manually and taking into consideration the honours and recognition bestowed upon these researchers, it can be seen that using pageRank in the author citation network gives the best indication of an author's prominence in the scientific community (3 turing awardees in the top 5). We therefore, use this metric as a measure of influence in the rest of the paper.

In the next section, we investigate some patterns in the citation practices of the authors with respect to their measured influence.

5. CITATION PRACTICES

We want to study the citation practices of author's in the author citation network. Here we hypothesize that highly ranked authors tend to cite one another more, whereas low ranked authors do not cite each other much - they also primarily cite highly ranked authors. This is fairly obvious as highly ranked authors are highly ranked because they receive many citations. For this we construct a sub-graph of the author citation graph by considering only the nodes with ranks within a certain fixed range of a given node's rank so that only those nodes which have pageRank very close to the given node's remain in the network. We then find the clustering coefficient of the resulting sub-graph. In Figure 4, we plot a graph of this clustering coefficient of sub-graph for different rank nodes. For all the plots where we have page rank on the x-axis (in all further sections too), we sort the nodes by their pageRank in descending order, take points spaced by some fixed value(50), centre a bin around the points with bin size(60) and averaged the y-values in that bin.

From this graph we can see that highly ranked authors tend to cite one another a lot (high clustering coefficient) whereas the low ranked authors don't cite each other much. This is in line with our hypothesis.

5.1 Trends on Papers the author writes

Here, we would like to answer how uniformly are the papers of an author are cited. We have used Jain's fairness index as a measure of uniformity. Given n buckets and xi as the number of balls in bucket i, the uniformity in the ball distribution across the buckets can be measured using Jain's

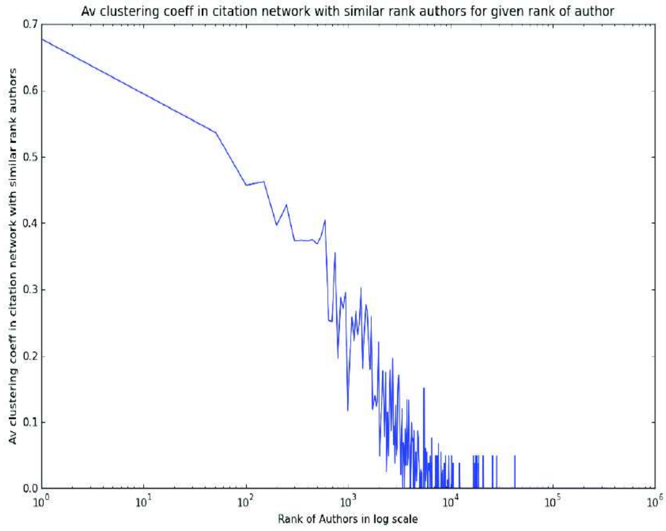


Figure 4: Average clustering coefficient in citation network with similar rank authors

fairness index as:

$J(x_1, x_2, \ldots, x_n) = \frac{(\sum_{i=1}^{n} x_i)^2}{n.\sum_{i=1}^{n} x_i^2}$

In our context, buckets are the papers of an author and balls are the number of citations of the paper. Figure 5 shows the plot of the uniformity in paper citation versus page rank of the author in the author citation network. We observe that the paper citation uniformity shows an increasing trend with rank.

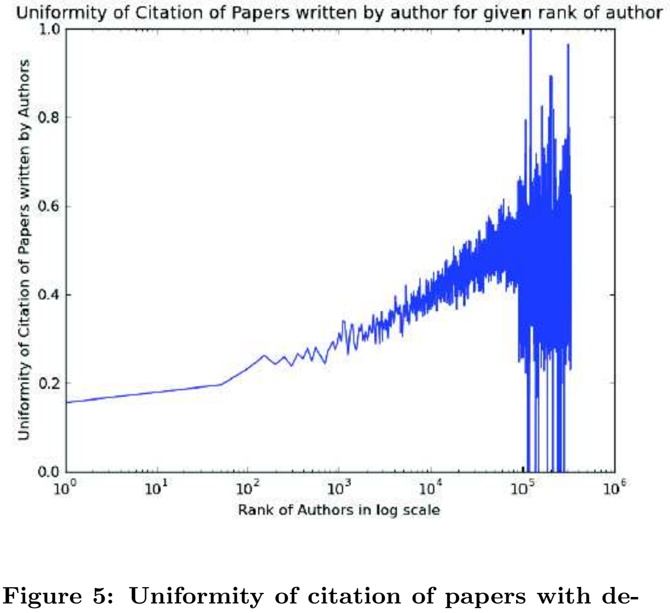


Figure 5: Uniformity of citation of papers with decreasing author ranks

The trend in the previous plot will become more clear if we also plot the number of papers the author writes versus the rank of the author. An author having large number of papers is more likely to have low uniformity in their citation as only a few papers would be quality papers. Also, we don't have information about the 1st author, 2nd author, etc about a paper in our dataset. A reputed author is likely to have many papers as last author(mostly with his PhD students and other collaborators in the community). Figure 6 verifies the fact highly ranked authors indeed have relatively large number of papers.

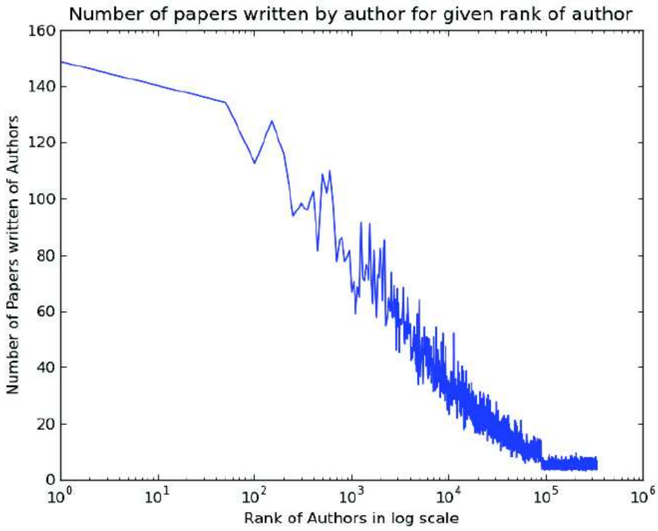


Figure 6: Number of papers with decreasing author ranks

Above, we could have also used entropy for measuring uniformity instead of Jain's fairness index. It is just that we wanted to try out different metrics. Now, we explore how diversely are the communities in which author publishes papers.

The entropy can be calculated as:

$H = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$

where x_i = fraction of papers the author publishes in community i and n = number of communities the author publishes papers in. Community here refers to the community is paper citation network. The graph below shows the plot for the entropy of paper publications in different communities versus rank. Thus, we see that highly ranked authors publish papers in very diverse communities and may be that is why they became highly ranked. Or it may be that, since the highly ranked authors have large number of papers, they are likely to have more collaborators and that too in different communities in paper citation network. Note that communities in paper citation network represent a research area as indicated in the community structure section.

6. COLLABORATION PRACTICES

In this section, we explore the collaboration practices to find any significant pattern for the top researchers. We begin by investigating the trend in the number of collaborators and then go on to explore any possible biases in the kind of collaborators.

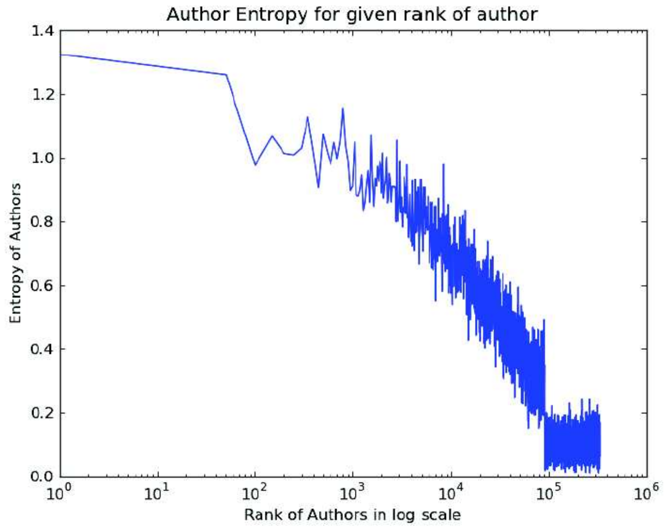


Figure 7: Author entropy with decreasing author ranks

Figure 8 shows the log-plot for the number of unique collaborators for the researchers in decreasing order of their ranks. As is seen from the plot, the number of collaborators decreases as the rank of the author decreases. However, the decline is not very prominent, which is understandable because the number of collaborators will be proportional to the number of papers and, as noted in Section 3, the pageRank in author citation network does not have a very good correlation with the number of papers.

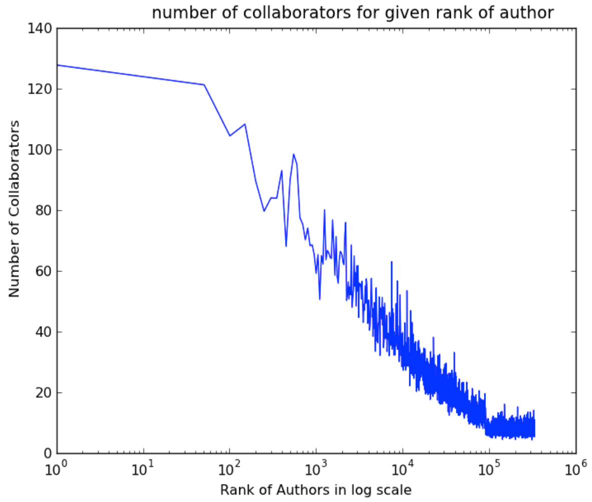


Figure 8: Number of collaborators with decreasing rank

Having observed the number of collaborators, we now try to investigate the kind of collaborators preferred by the more influential researchers, i.e. whether highly influential researchers tend to prefer similar researchers for collaboration. While the observation seems obvious on count, it must also be noted that several behavioral theories argue for the presence of complications (due to personality conflicts or personal idiosyncracies) when two highly noted individuals collaborate. Therefore, it is not obvious what the answer should be. In order to answer this question, we perform two different experiements. In the first experiment, we plot the average pageRank of the collaborators of a particular author in decreasing order of his/her rank. The plot is shown in Figure 9. The curve decreases almost linearly with decreasing influence denoting that the high influence nodes tend to prefer other high influence nodes for collaboration. It is important to note here that the pageRank was calculated in the citation network and not the collaboration network, so it was not obvious that the collaborators of a high pageRank individual will also have a high pageRank, although this does turn out to be the case.

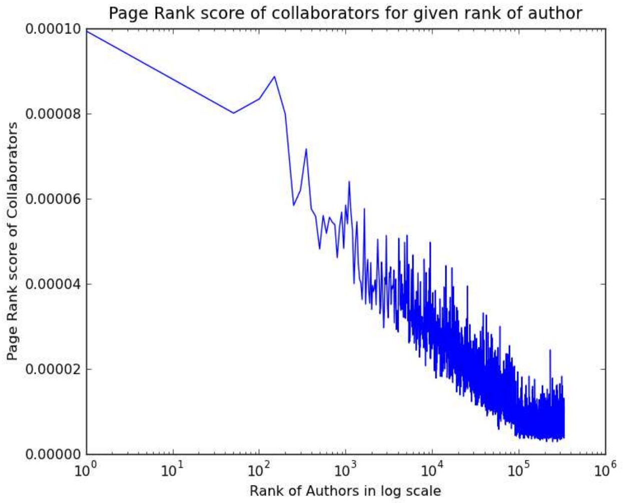


Figure 9: Average pageRank score of collaborators with decreasing rank

In order to confirm our observation, we perform another experiment. We construct a sub-graph of the collaboration graph by considering only the nodes with ranks within a certain fixed range of a given node's rank so that only those nodes which have pageRank very close to the given node's, remain in the network. We then find the clustering coefficient of the resulting sub-graph. The clustering coefficient in the sub-graph will give us a measure of how likely the nodes are to collaborate amongst themselves. We then plot the clustering coefficient for all nodes in decreasing order of their rank. The plot is shown in Figure 10. It can be seen from the plot that the clustering coefficient decreases steadily with decreasing influence. The more influential authors have a high clustering coefficient, once again signifying that they tend to prefer each other when it comes to collaborating for research. Note that even though the original graph had a lot of cliques and hence, a high clustering coefficient, the subgraph will break all those cliques and the clustering coefficient in the subgraph is therefore not biased because of it. In case it still retains those cliques, that just reinforces the strong collaboration relationship among those authors. Besides, we are interested in the declining trend of clustering coefficient rather than the absolute value.

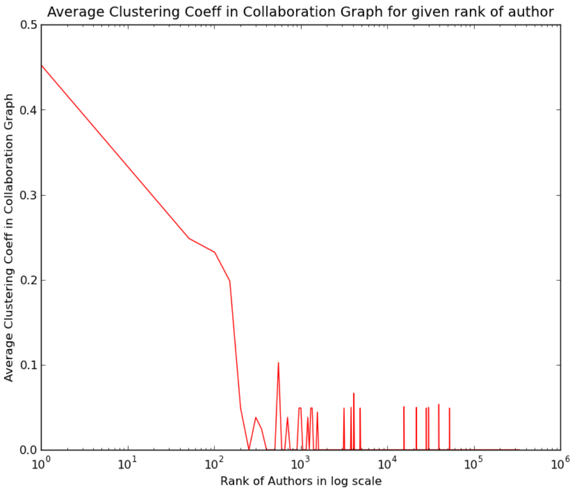


Figure 10: Clustering coefficient in collaboration network with decreasing rank

Note that there are a couple of caveats in the reasoning presented above. Firstly, while the above experiments do exhibit high collaboration between the influential researchers, the given data is insufficient to establish causality, i.e. we cannot definitively say, with just the given data, whether they collaborate because they are influential or whether they became influential because they collaborated. Secondly, the above said correlation can also be due to the fact that people tend to limit their collaboration to their own institute (i.e. some geographical factors are at play) and the institute has some very stringent entry barriers, resulting in the good researchers collaborating more amongst themselves.

6.1 Uniformity in paper citations of the collaborators

Here we analyze the paper citation uniformity (as in previous section) of the collaborators of an author. The motivation behind studying this is to see whether the collaborators a highly ranked author (who generally has large number of papers) also have similar uniformity as that of the author. For this, we plot the mean and standard deviation (Figure 11) of the uniformity in paper citation of the collaborators of an author versus the page rank of the author.

From the plots, we see that the mean uniformity of the collaborators of an author shows an increasing trend with rank, while the standard deviation of uniformity of the collaborators shows a decreasing trend with rank. Compare these plots with the Figure 5. If we look at the values where the curves start, the standard deviation(0.25) and the uniformity(0.25) in Figure 5 and 0.43 $0.25 + 0.25$ in mean uniformity curve, it seems that the set of collaborators of highly ranked author constitutes of some other highly ranked authors but most of the collaborators having very high uniformity(these authors have low ranks Figure 5). These low rank collaborators are most likely the PhD students.

7. TEMPORAL CAREER PATTERNS

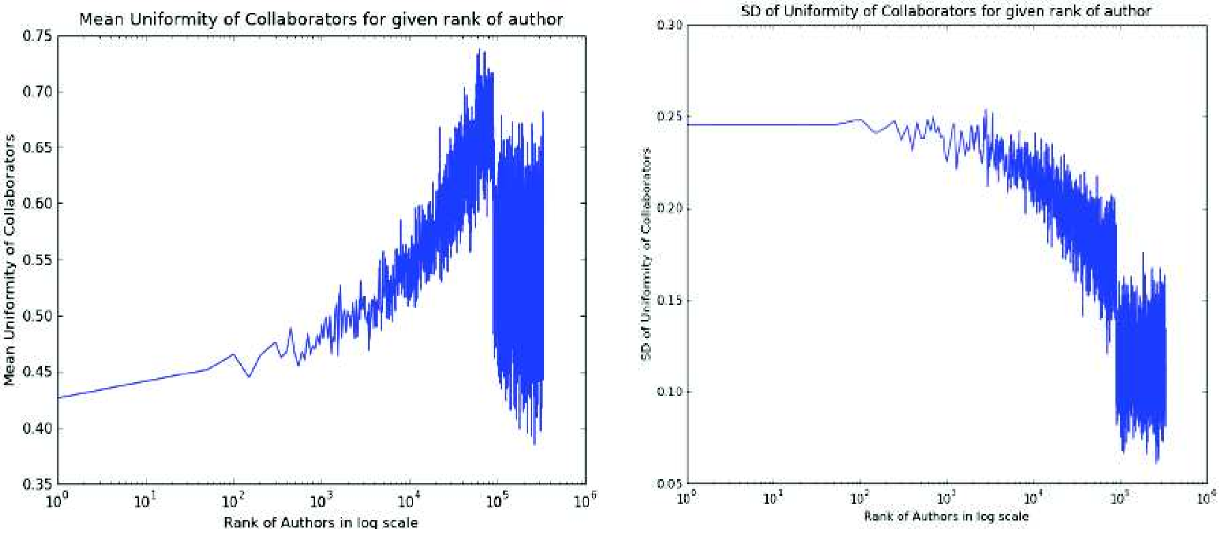
In this section, we investigate the career graph of a researcher as it evolves over time. The motivation behind this is to see if there is a specific pattern in the quality and/or quantity of research of a star researcher with respect to time. In order to perform this analysis, we first plotted the number of papers and total number of citations for all papers published in a particular year for some top researchers. The sample plots for Jeffrey Ullman and Jim Gray are shown in Figure 12. Similar plots were obtained for other top researchers as well (which we omit here for the sake of brevity).

Some very interesting and useful observations can be made from these graphs. As is clear from the plots, there seems to be no definite pattern in the temporal graph of the number of papers published in an year. However, there is a very interesting pattern in the number of citations in an year (which is basically the sum of citations received by all papers published in that year). The plots tend to have a prominently "peaky" nature, i.e., each of the curves is characterised by the presence of very prominent peaks. This corresponds to a real world scenario where the researcher tends to focus on a harder, more time consuming problem which reaps more dividends as against easier ones which may lead to a flatter curve. Another very interesting observation is that there is no corresponding pattern in the number of papers curve denoting that even though the researcher may be focussing on a harder problem, he/she does not stop publishing other papers (a very promising strategy tip for any newbie!).

The given plots can also be interpreted to reconfirm the notion that the citation practices of researchers are influenced by the quality of papers and not by the influence of the researcher. This is because, if the citation network had been influenced by the influence of the researcher, we would have expected a continuously decreasing power law graph here (owing to the preferential attachment of every upcoming researcher to the influential node's papers). This contradicts with the "peaky" nature observed which can only be explained by the argument that people tend to vote for content over influence.

Although the observed pattern is indeed very useful in drawing some important conclusions, it is difficult to comment just based on the plots for a few researchers. We need to develop a metric and observe the general trend to be have statistically significance in our analysis. We observe that the higher the peak strength(in the temporal citation plot) and the more early is the peak in the author's career, the higher is the rank of the author. This trend we found out by plotting some sample authors. This is intutuive in the sense that good researches usually take off quite early in their career. So, we need to design a metric that captures the strength of the peak and when that peak(in the temporal citation pattern) occured in the author's career. Thus, we developed a metric for an author in the following manner and plotted it pattern for all the researchers:

1. We first need to smoothen out the citation pattern of an author to remove small random peaks that are not significant form the view point of an author's influence(noise). For smoothing, we take a running average (we take the running average for 4 years in our



(a) Mean uniformity of collaborators

(b) Standard deviation of uniformity of collaborators

Figure 11: Mean and Standard deviation in uniformity values

measurements). Let the original curve be denoted by the function $f(t)$. Then, the smoothened curve, $g(t)$ is given by, $g(t) = \sum_{i=0}^{3} f(t-i)$

2. Now, we capture the peaks by taking the difference function of the smoothened curve (note that we are dealing with a discrete curve here). We took difference function to capture the peaks(filter out low frequencies). So, the difference function, $h(t)$ correponding to the function obtained in Step 1 will be, $h(t) = |g(t) - g(t-1)|$

3. Multiply every point for the autor with the number of years since the first publication to obtain the value of the metric, i.e., the value of the metric m is given by, $m = \sum_{t=startT}^{t_{Max}} h(t)(t - startT)$
   where startT is the year of first publication and t_{Max}denotes the last year plotted(2010)

Not only, does this metric capture the "peaky" nature of the curves, it also accounts for the fact that different researchers may belong to different eras by taking their first publication year into account. Also, it captures how early an author takes off in his career by publishing good papers. In Figure 13, the authors are mapped to the X-axis in decreasing order of their influence and the Y-axis shows the value of the metric for the authors (we follow the binning technique used in the previous section here as well). As can be seen from the plot, the value of the metric steadily declines as the influence of the author declines, thus, confirm our observation on the pattern in the behavior.

In table 4, we list the correlation of our metric with other influence indices. As is seen from the table and is clear from the graph, our metric has a high correlation with the citation pageRank and can be used for measuring a node's influence in the network as an alternative to H-index. Also note that, just like H-index, our metric also depends 'only' on the local attributes of the author and is therefore, fast and easy to compute locally.

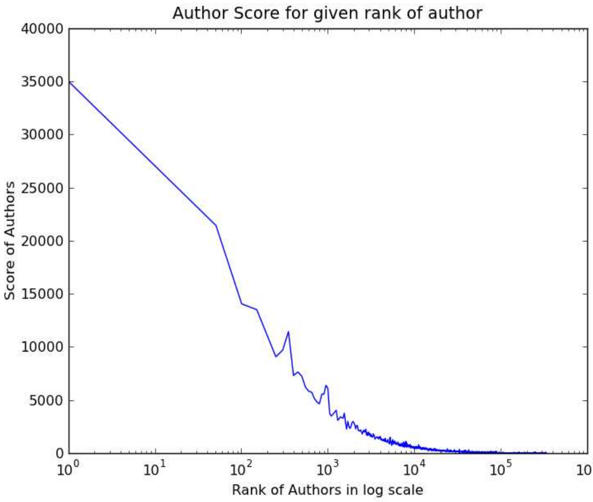
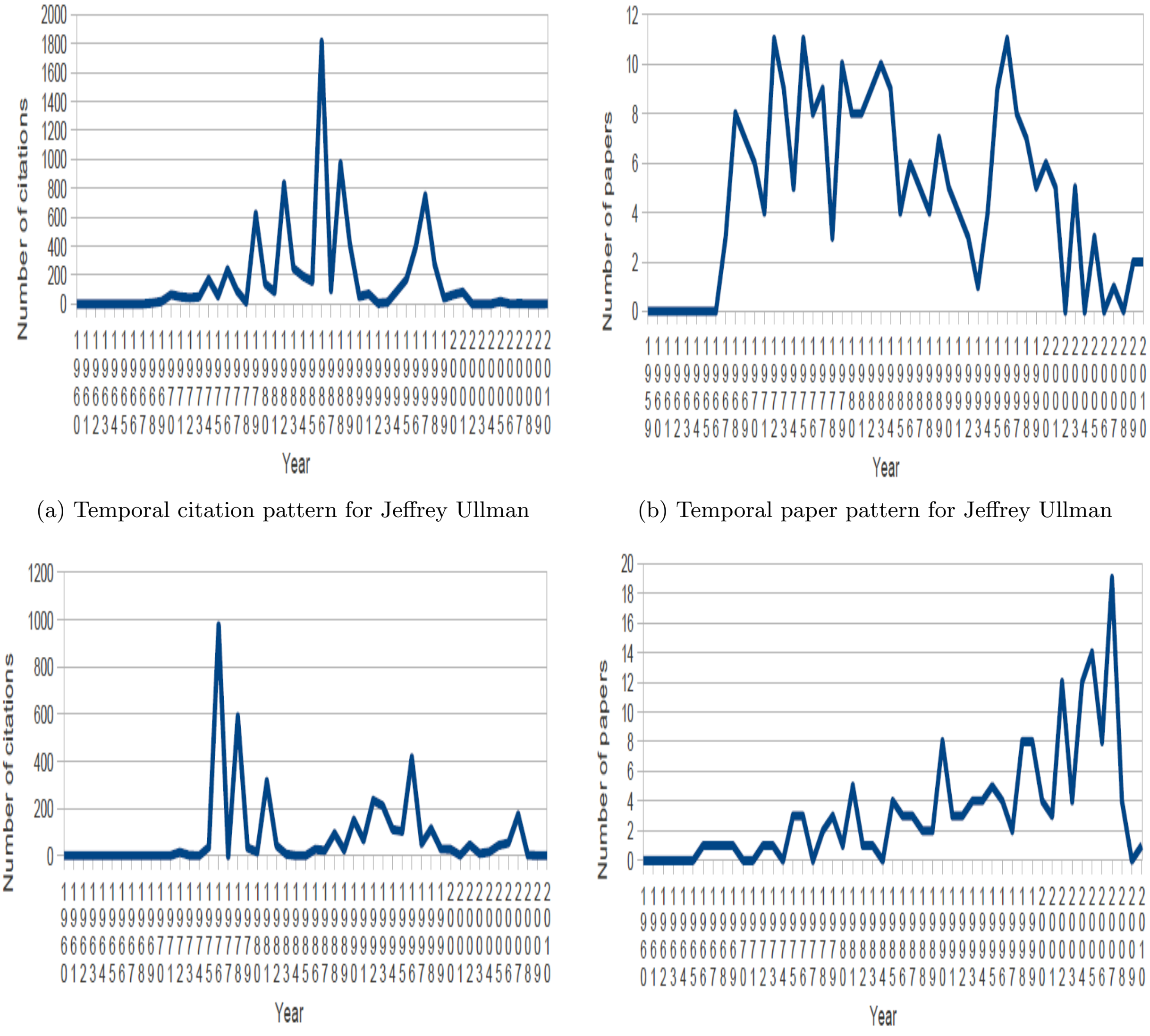


Figure 13: Metric score with decreasing author ranks

8. CONCLUSION AND FUTURE WORK

Through the analysis presented in this paper, we can conclude that a successful researcher tends to be a prolific collaborator who collaborates and publishes papers in several different research topics. Peer group also plays an important role in a researcher's productivity and good researchers tend to collaborate and cite each other more. Moreover, a successful researcher focusses on challenging problems which require time and garner a lot of citations, however, at the same time, he/she does not compromise on the quantity of research as well. Our metric captures this trend very neatly



(a) Temporal citation pattern for Jeffrey Ullman

(b) Temporal paper pattern for Jeffrey Ullman

(c) Temporal citation pattern for Jim Gray

(d) Temporal paper pattern for Jim Gray

Figure 12: Temporal patterns for two top researchers

Table 4: Pearson correlation of different influence metrics with our proposed metric

Index	Person Correlation
H-index	0.51
pRankAuCollab	0.27
pRankAuCit	0.79
numPapers	0.32
numCitations	0.75
avgCitations	0.22

and can be a good alternative to the H-index.

For future work, we feel that a better metric for an author's ranking (than H-indeX) can be learned on the basis of the various metrics considered in the paper, namely, the number of papers, the number of citations and the average number of citations per paper . It will be interesting to see how such a metric performs and whether the learned values generalize beyond one network and if yes, then to which kind of networks. For this purpose we can use some machine learning technique like ranked SVM to learn some ranking function which takes into account only local factors and is able to predict the global ranking of the author. It would be interesting to study some more temporal citation and collaboration patterns of an author to determine some sort of causality - does more citation lead to more collaboration or the other way around.

9. ACKNOWLEDGMENTS

We are extremely grateful to the TAs of the course, especially Cherian Mathew and Chenguang Zhu, for their guidance and support.

10. REFERENCES

[1] S. Bilke and C. Peterson. Topological properties of citation and metabolic networks. *Phys. Rev.*, 64.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Seventh International World-Wide Web Conference*, pages 990–998. WWW, April 1998.

[3] D. J. de Solla Price. Networks of scientific papers. *Science*, 149.

[4] Y. Ding. Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *J. Informetrics*, 5.

[5] K. S. E. A. Leicht, G. Clarkson and M. E. J. Newman. Large-scale structure of time evolving citation networks. *Eur. Phys. J.*, 59.

[6] L. Y. J. L. L. Z. Jie Tang, Jing Zhang and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 990–998. ACM SIGKDD, August 2008.

[7] D. I. K. Luis M. A. Bettencourt and J. Kaur. Scientific discovery and topological transitions in collaboration networks. *J. Informetrics*, 3(3):210–221.

[8] V. L. Matthew L. Wallace and Y. Gingras. A small world of citations? the influence of collaboration networks on citation practices. *CoRR abs/1107.5469*.

[9] M. E. J. Newman. Scientific collaboration networks: I. network construction and fundamental results. *Phys. Rev.*, 64.

[10] M. E. J. Newman. Scientific collaboration networks: Ii. network construction and fundamental results. *Phys. Rev.*, 64.

[11] R. L. Vincent D. Blondel, Jean-Loup Guillaume and E. Lefebvre. Fast unfolding of communites in large networks. In *Journal of Statistical Mechanics: Theory and Experiment*, page 10008. IOP and SISSA, October 2008.

[12] B. L. T. Xiaolin Shi and L. A. Adamic. Information diffusion in computer science citation networks. In *ICWSM*. AAAI ICWSM, May 2009.