Review Article

Information Processing & Information Diffusion in Computing Citation Networks

Ashok Koujalagi* and Mallamma Reddy

Department of Computer Science, Basaveshwar Science College Bagalkot, India

Received: 🗓 July 14, 2018; **Published:** 🗓 August 20, 2018

Corresponding author: Ashok Koujalagi, Department of Computer Science, Assistant Professor & Postdoctoral Researcher, Basaveshwar Science College Bagalkot, India

Abstract

The paper citation network is a traditional social medium for the exchange of ideas and knowledge. In this paper we view citation networks from the perspective of information diffusion. We study the structural features of the information paths through the citation networks of publications in computer science and analyze the impact of various citation choices on the subsequent impact of the article. We find that citing recent papers and papers within the same scholarly community garners a slightly larger number of citations on average. However, this correlation is weaker among well cited papers implying that for high impact work citing within one's field is of lesser importance. We also study differences in information flow for specific subsets of citation networks: books versus conference and journal articles, different areas of computer science, and different time periods.

Keywords: Library Science; Information Processing; Information Manipulation; Data Storage; Network Information Storages System

Introduction

Information diffusion is the communication of knowledge over time among members of a social system. In order to analyze information diffusion, one needs to study the overall information flow and individual information cascades in the networks. Although much recent attention has been focused on new forms of collective content generation and filtering, such as blogs, wikis, and collaborative tagging systems, there is a well-established social medium for aggregating and generating knowledge published scholarly work. As researchers innovate, they not only publish new results, but also cite previous results and related work that their own innovations are based on. This creates a social ecology of knowledge where information is shared and flows along co authorship and citation ties. In this paper, we examine information flow within and between different areas of computer science and its impact. Our basic assumption is many citations are evidence of information flow from one article, and its authors, to another. In order to cite a paper, an author usually, though not always Simkin et al. [1], reads the paper and acknowledges it as being relevant to the subject of their own paper, either by providing information that their work is built upon or by providing information about related approaches to the same problem.

Although not every citation represents the same level of engagement, citation networks provide some of the clearest evidence of information flow. Our work has two primary goals:

first, we are interested in observing the features of information flow in citation networks; and second, we want to know which of these features, such as time spans and community structure representing different fields of research, affect the information flow. Studying citation networks has been the purview of the field of scientometrics, which aims to measure the impact of scholarly publications Dieks Chang [2]. Scientometric data has been available for several decades and so it was already in the 1960s that de Sola Price first observed power laws in scientific citation networks and developed models of citation dynamics De Solla Price [3]. However, the recent emergence of online knowledge sharing has made it particularly easy to study information diffusion on a large scale. Studies of information cascades in blogs Kumar et al. [4,5], social bookmarking sites, and photo sharing have all revealed a highly skewed distribution in the attention a particular post, URL, new story or photo Lerman et al. [6,7] will receive. The attention may be measured through links or tags given to the items.

In separate studies, it has been shown that such networks exhibit strong community structure Tseng et al. [8]; Adamic and Glance [9]; Chin et al. [10], where links or interactions occur more frequently within communities than between them. The role of community structure in information diffusion has also been studied in scientific citation networks. It has been found that there is a longer delay for citations across disciplines than ones within a discipline, implying that information is not only less likely to

diffuse across community boundaries, but when it does, it will do so with a longer time delay Rinia et al. [11]. Information flow between communities is such a relatively small proportion of total information flow, that modeling citation networks without them provides realistic citation distributions and clustering coefficients Borner [12]; Rosvall et al. [13]. The development of efficient network algorithms has lead not just to discoveries of the overall properties of citation networks, but also the detection of changes in citation patterns where a new trend or paradigm emerges Leicht et al. [14]. There has also been interest in visualizing and quantifying the amount of information flow between different areas in science Boyack et al. [15,16] in effect mapping the generation of human knowledge through information flows.

These maps leave open the question, however, of what happens once information has diffused across a community boundary; will it have the same impact as information diffusing within a community? This is an interesting question, because recent empirical work Guimera et al. [17] has shown that new collaborations between experienced authors are more likely to result in a publication in a high impact journal than in collaborations between unseasoned authors or repeat collaborations between the same two authors. The argument is that merging ideas and expertise in a novel way will produce higher impact work. But this work did not address whether the authors were from the same scientific communities or not, or whether the publications cited in the work stemmed from the same field. On the theoretical side, agent-based models of innovation have shown that independent innovation within communities is important, so that the network as a whole does not converge on suboptimal solutions too quickly Lazer et al. [18]. In this paper, to answer the question of the impact of cross-community information flows in computer science, we make empirical observations of citations of computer science articles, focusing specifically on information flow across community boundaries and temporal gaps. In the following sections, we first describe the computer science publication data sets we used and the construction of the citation networks. We then examine the properties of the citation networks and relate the properties of a citing link to subsequent impact of the citing article.

Preliminaries

Definition of Citation Networks: Citation networks are networks of references between documents. In this paper, we focus on paper citation networks, which correspond to information diffusion in the corresponding research areas. From the graph theoretic perspective, citation networks can be thought of as directed graphs with time stamps and community labels on each node:

a)   Nodes: publications;

b)   Edges: one paper citing another;

c)   Edge directions: in order to represent the direction of information flow, we denote the direction of edges from cited papers to citing papers;

d)   Time stamps: years in which the papers were published;

e)   Time spans: the time elapsed between the publication of the cited and citing paper;

f)   Community labels: we classify the papers into different research areas according to their venue information.

Information flows in citation networks can be interpreted as the scientific ideas and knowledge transmitted from publication to publication, which are explicitly indicated by citation relationships. Not all, or perhaps very little, information is preserved from cited to citing paper. Further, the information may be amended in the citing paper. Nevertheless, we assume that the cited paper informed the citing paper. There are two common and significant features of any typical citation network: first, it is directed and almost acyclic; and second, when it evolves over time, only new nodes and edges are added, and none are removed Leicht et al. [19]. The acyclic nature of the graph stems from the simple fact that, with very few exceptions, a paper will not cite a paper published in the future. Although publication delays may lead to such occurrences, most citations are limited to previously published work.

Description of Data Sets: The datasets we study are two large digital libraries encompassing comprehensive scholarly articles primarily in computer science - the ACM1 data set and the CiteSeer2 data set Giles [20]. In the ACM data set, there are several different types of publications, such as books, journal articles, conference proceeding papers, reports, and theses. Books alone account for 113, 089 of the publications in the ACM dataset. Both of the data sets have information about the publication dates and venues; however, some of the information is incomplete or inaccurate. Since our study considers the time evolution and community structure of the networks, we deleted the nodes with an unresolved time or venue information. While ACM data set includes citations to publications outside of the data set, the Citeseer data does not, and so we limit our analysis to citations between articles within each dataset. In addition, some citations between two articles that both reside in the same data set are missing, due to the difficulty in disambiguating and parsing citations from article text Simkin et al. [21].

Even with these limitations, we are left with 346,000 citations for the ACM dataset and 84,000 citations in the Cite Seer dataset, which we use to measure information flows between different computer science communities and the impact of a publication. We will discuss possible biases introduced by missing data below. Even though we are analyzing two separate datasets, they overlap in subject area and time span. It is therefore reassuring that they have a significant, but relatively small overlap in the articles that they contain. There are 613,444 proceedings or journal papers in the ACM dataset that we are studying, and 593,386 of them have distinct titles in the database; while there are 716,774 papers in CiteSeer dataset, and 611,127 have distinct titles. By matching the titles and authors of the 593,386 papers in ACM and 611,127 papers in CiteSeer using a simple cosine similarity measure, we identify 122,978 (20%) papers that are present in both datasets. Finally, Table 1 gives summary statistics of the two data sets and the citation networks we will study.

Structural Features of Citation Networks: Since the structural features of citation networks provide explicit evidence of information flow paths, our study of information diffusion starts with them.

1http://portal.acm.org

2http://citeseer.ist.psu.edu

Degree Distributions: As stated before, we set the direction of an edge to reflect the direction of information flow. The indegree is the number of papers cited in a given paper. In effect, it is the number of papers that may have influenced the paper at hand. The out-degree is the number of papers citing the given paper, reflecting the paper's potential impact and influence. In the previous section, we mentioned that there are different types of publications in ACM, including those with very high in degrees, such as books. Since these comprehensive publications normally have many more references than regular papers, we show the degree distributions separately for books and papers in the ACM data set in Figure 1. The distribution of the length of the reference list for publications (their indegree) is highly skewed; some publications have references from 10 to several hundred papers in the dataset, many have none, or few. In actuality, many of these papers have longer lists of references, but these were not identified, or they fall outside of the dataset. The distributions of out degrees are similarly skewed, an indication of a linear preferential attachment mechanism: already well cited papers are more easily discovered, and subsequently cited: it is the success breeds success phenomenon Burrell [22]. As one might expect, both the in degree distributions and out degree distributions of books in the ACM dataset are significantly heavier tailed with books both citing more and being cited more. It is therefore unsurprising that the in degree and out degree distributions of documents in CiteSeer are more similar to those of ACM papers, as opposed to books.
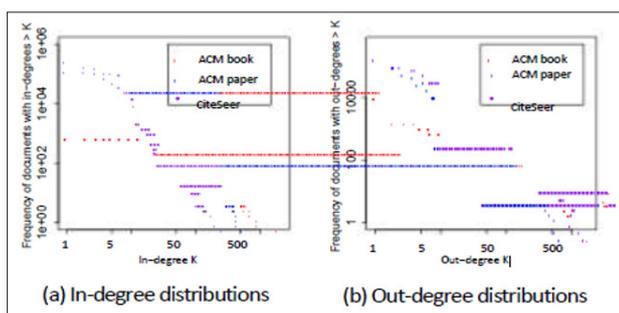

Figure 1: The degree distributions of the ACM and CiteSeer citation graphs.

Connectivity: In order to analyze how information flows through the citation networks, we first study their connectivity. Two vertices A and B are said to be in the same strongly connected component if there exists a path both from A to B and from B to A. For both the ACM and CiteSeer citation graphs with correct time stamps, there are no significant strongly connected components. The absence of large strongly connected components is consistent with the fact that citation graphs are nearly perfect directed acyclic

graphs (DAGs). However, 96.08% of the nodes (publications) are in the largest weakly connected component, in which there is a path between every pair of nodes in the version of undirected graph, in the ACM citation network. Similarly, 96.20% of the nodes are in the largest weakly connected component of the CiteSeer citation network. In the ACM dataset, 5.0% of the papers are published in the years 2004 and 2005. By tracing back their citations, we find that 48.8% of the papers published in previous years are either directly or indirectly cited by the papers in 2004 and 2005.

In the CiteSeer data set, with 3.5% papers published in the years 2003 - 2005, 28.9% of earlier papers are reachable by tracing back the citations. Although the two data sets differ in their cohesion (either due to completeness of data, or other issues of coverage), we observe that each subsequent generation of papers is tied directly or indirectly to a significant portion of the prior work. If we take the 500 most cited articles in the ACM network, we observe that there is a giant component linking a significant fraction of the most influential papers (Figure 2a). In contrast, if we were to select 500 random papers, there would be no giant component as papers are rather unlikely to cite one another (Figure 2b). This observation is consistent with many other networks where the most highly connected nodes tend to be connected to one another Shi et al. [21]. It is also known as the rich-club phenomenon Zhou and et al. [23]; Colizza et al. [24]. Yet it is still striking that such a small number of most influential papers out of tens of thousands in computer science should be connected to one another through one another.
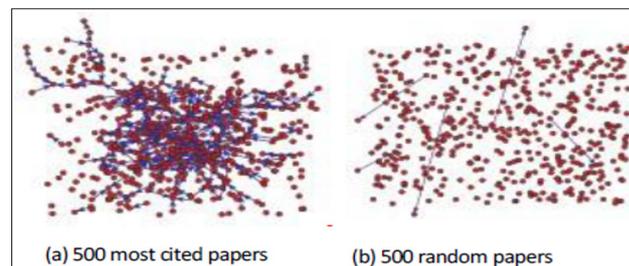

(a) 500 most cited papers   (b) 500 random papers

Figure 2: Subgraphs of the 500 most cited papers and 500 random papers in the ACM citation network.

Average Shortest Directed Path: The shortest paths in graphs (also termed geodesics) relate directly to the accessibility of information. Like many other complex networks, the citation networks we study exhibit the small world phenomenon. The average shortest directed path of the ACM graph is 7.60, and its largest geodesic is 32. Similar to the ACM citation network, CiteSeer has an average shortest directed path of 6.29 and its longest geodesic is 28. However, the reachable pairs of nodes via directed paths in the ACM citation network comprise 0.65% of all possible paths (or node pairs) and 0.41% of all possible paths in the CiteSeer citation network. From the connectivity and shortest directed paths of the citation networks, we can see that, in spite of the largest weakly connected component occupying nearly the entirety of the network, the percentage of reachable pairs of nodes is smaller. But where paths do exist, we observe that the lengths of information flows are generally short. Note that this does not preclude that

there are more circuitous routes involving several papers. In our measurements, we only account for the shortest path. This form of "deep linking", citing original articles, considerably shortens the path between articles.

Sizes of Information Cascades: One may be interested not only in direct citations of a given paper, but subsequent citations of the citing papers, etc. (Figure 3) gives a simple example of an information cascade. An information cascade represents both the direct and indirect influence of a given publication, although the influence is diluted with each subsequent step. As each paper cites several others, it is difficult to attribute influence to any given chain of citations. It may be that the reason that A cites B is unrelated to the reason why B cited C. By taking every node in the citation graph as a root, we run a breadth first search along the outgoing edges and obtain an information cascade tree starting from every paper in the network. The distributions of sizes, depths and numbers of leaves of the information cascade trees are shown in Figure 4. From the figure, we see that all of the distributions of sizes, depths and numbers of leaves have a very a sharp drop in the tail. These drop-offs naturally correspond to the limits of the data sets: there are no more than several hundreds of thousands of papers in each data set, and a cascade can only encompass papers appearing after the given paper. This is different from the power-law distributions of the cascade sizes in the blogosphere observed by Leskovec et al. [25], with slightly different cascade definitions. The blog measurements were unaffected by size limitations in the data, as the largest cascade sizes encompassed no more than a few thousand posts out of millions that were observed (Table 1). The

Spearman correlations of the cascade sizes, depths and numbers of leaves in the information cascades, as well as the out-degrees are shown in Table 2. We see that, in both the ACM and CiteSeer citation graphs, the sizes and depths of the information cascade trees have large correlations, meaning that cascades that encompass several generations of scholarly work are also the largest.
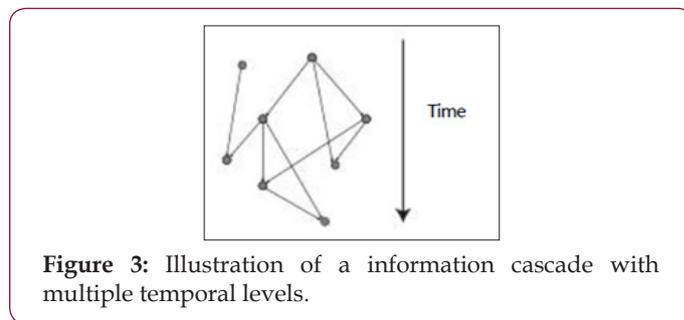

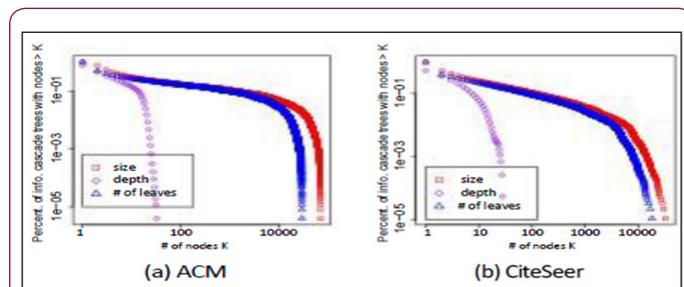**Figure 3:** Illustration of a information cascade with multiple temporal levels.

Figure 4: The distributions of sizes, depths and numbers of leaves of the information cascade trees in ACM and Cite Seer citation graphs.

Table 1: Summary statistics of the citation networks before and after cleaning.

	Orig.		With Publication Date			With Publication Venue		
	Nodes	Edges	Nodes	Edges	Time Range	Nodes	Edges	Communities
ACM	842,422	2,492,503	250,556	861,088	1920 - 2005	119,268	346,289	26
CiteSeer	716,774	1,438,505	93,298	342,657	1958 - 2005	52,411	84,134	23

Table 2: Spearman correlations of the sizes, depths, numbers of leaves of information cascades and ou degrees of all papers in the two citation networks.***, **, and * denote significance at the < 0.05, < 0.01 and ≥ 0.01 levels respectively.

	Size &Out-Deg	Size & Depth	Size &leaves	Depth & Leaves
ACM	0.724***	0.928***	0.885***	0.802***
Cite Seer	0.809***	0.952***	0.884***	0.831***

Information Diffusion and the Effects of Citations: After examining the structural features and the information flow paths between papers in the citation networks, we turn to how information flows between communities, and how different types of citations (from and to various communities and citing old or new papers) would affect the subsequent information diffusion in citation networks.

Information Flows Between Communities: We assign papers to communities according to their venues, using the classification system adopted by Microsoft's, Libra academic search service 3. For example, a paper published in the KDD (Knowledge Discovery and Data Mining) Conference would be classified under "Data Mining", while a paper published in the Journal of Information Professing

and Management would be classified under "Information Retrieval". Because of the incomplete and noisy information in the venues, we are able to classify about 1/3 of the papers with about 80%-90% precision. With this community classification, there are about 205,000 within community citations and 141,000 across community citations in ACM, while 42,000 both within and across community citations in Cite Seer. In order to quantify the densities of information flow from community to community, we first count the number of citations between every pair of communities for each data set separately (e.g. the number of citations of Theory to Theory, Theory to Data Mining, etc.), and get a matrix A with these numbers as its entries. We then compare the number of citations between any pair of communities relative to the rate of citation we

would expect if the volume of inbound and outbound citations were the same, but the citations were allocated at random.

We let Nij be the actual number of ∑citations from i to j, Ni ▪ = Nij bΣhe total number of citations from community i, N ▪ j = i Nij bΣhe total number of citations to community j, and N = ij Nij be the total number of citations in matrix A. Then the expected number of citations, assuming indifference to one's own field and others, from community i to community j is E [Nij] = Ni ▪ ×N ▪ j /N. We define the community weight as a z-score that tells us how many standard deviations above or below expected Nij is. Here we have the observation that N ≫ Ni ▪ and N ≫ N ▪ j, so we approximate the standard deviation by E [Nij]. In this way, for every entry, we get a normalized value, which we call community weight:

$$w_{ij} = (N_{ij} - \sqrt{\frac{N_{i-} \times N_{-j}}{N}}$$

By visualizing the normalized matrix, i.e. matrix of community weights as in Figure 5, we can observe different densities of information flow amongst communities. For example, for each community, as expected, the majority of citations are within the community itself. However, there are some closely related communities. For example, there appears to be considerable information flow from Information Science to Information Retrieval, from Databases to Data Mining, from Information Retrieval to Data Mining and from Computer Vision to Computer Graphics. These flows reflect frequent citations by papers from the second community to those in the first. We also observe that the more theoretical areas such as Algorithms & Theory and Physics are less connected with others, while more applied areas, such as Data Mining, Information Retrieval, and Operating Systems have more information flows two and from other areas. Correlations of information diffusion and citation features If we define information diffusion to occur when a paper is cited, then many factors affect such information diffusion
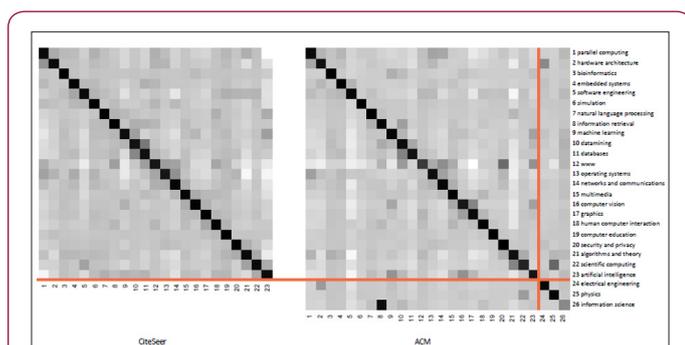

Figure 5: Visualization of the matrices of community weights between different areas of computer science. Darker cells represent more frequent citation than expected if citation were at random, lighter ones depict less frequent citation.

They include the popularity of the research field pertaining to the article in a certain period, the reputation of the authors, the specific innovation reported in the publication, etc. However, there is much we can surmise simply from the citation patterns, time

lapses and community information. Specifically, we examine what kinds of citations would make the citing papers have greater impact, whether it is citing another paper in a related community with strong information flow, or the time elapsed since the publication of the cited paper. As we have stated before, to measure the influence of a particular paper, both directly and indirectly influenced papers may need to be taken into consideration, possibly weighing them differently. However, for both the clarity of the model and lack of consensus in the literature for a particular weighting scheme Aksnes [3], we use the number of citations a paper receives normalized by the average number of citations received by all papers in the same area and year Valderas et al. [25].

This measure allows us to make a fair comparison between articles that may not have finished accumulating citations due to their recency, and to account for differences in the publication cycle for different areas Stringer et al. [26]. Citation networks for all of computer science Since our study focuses mainly on the relationship between information flow and innovation, as opposed to summaries and reviews, we exclude publications that are book chapters and books, and focus on journal articles and papers published in conference proceedings. In the ACM dataset, the articles are already classified according to publication venue type, and so are easily filtered. In the CiteSeer dataset, we find that a majority of publications having 40 or more references tend to be review manuscripts. We exclude such publications from both data sets. Finally, we exclude papers published after 2000, because their recency means that they have not accumulated most of their citations Stringer et al. [23]; Burrell [6]. (Table 3) shows the correlations between community weights and time lapse of the citing and cited paper, and the subsequent impact of the citing paper.

From the table we see that for both citation networks, the weights of in- formation flows between communities (i.e. the community weights) have positive correlations with the influence metric (normalized out-degrees). This means that, on average, a computer science paper will be rewarded for referencing other papers within its own community or proximate communities. More recent papers have had an opportunity to cite more distant papers in time. Since pairs of citations are only recorded between papers in the dataset, older papers will have shorter recorded time lags to the papers they reference, since earlier referenced papers may not be included. The above is reflected in the correlation between the publication year of the citing paper and the time elapsed between the two papers ($\rho = 0.2$, $p < 10-16$). More interestingly, there is a negative between the time elapsed between the papers and the subsequent impact of the citing paper. Note that we are already normalizing by the average citation number of papers in a given year, so that older papers' chance to accumulate more citations is not a factor. The negative correlation between citation time lag and impact could be interpreted as citing more recent work being rewarded by citations. However, it is not uncommon to see some extremely innovative and influential work whose citations reach across communities or draw upon older publications. The overall correlations only reflect the average trend. As we observed

in Figure 1a, a large proportion of the papers receives very few citations, while a few papers garner large numbers of them. We found interesting trends, when, in addition to measuring the overall correlation for all papers, we computed separate correlations for the bottom 90% of the papers according to impact (denoted as ≤ 90% in Table 3 and the top 10% (> 90%). What we can observe is that for less well cited papers, the correlations between impact and community information flow weight are positive, in agreement with the overall trend. This is where the majority of papers lie - they receive few citations and do not lead to large subsequent impact. How- ever, for papers with high impact (dozens to hundreds of citations), the neutral correlations show that citing within one's

own community is less important. Similar patterns are observed for time lags as well. The lower impact articles benefit from citing recent work; but for more influential papers, these correlations are reduced or absent. It may be that a truly innovative article draws upon work that had not been garnering much attention recently, and that is not tied to many other relevant publications. This would imply that the more innovative and more highly cited papers may cross boundaries where information normally does not flow. Subnetworks of papers in different areas We have seen how the weights of information flows between communities affect the subsequent impact of the citing papers in the overall citation networks of computer science.

Table 3: Spearman correlations show the effects of community weights and time differences between the cited and citing papers on the subsequent impacts of citing papers.

	ACM			Cite Seer		
	Overall	**≤ 90%**	**>90%**	**Overall**	**≤ 90%**	**>90%**
time-diff	−0.0659∗∗∗	−0.0581∗∗∗	0.0045∗	−0.0870∗∗∗	−0.0899∗∗∗	0.0124∗
c-weight	0.0889∗∗∗	0.0832∗∗∗	0.0089∗	0.0622∗∗∗	0.0621∗∗∗	0.0314∗

In this subsection, we investigate the correlations in a finer scope. We choose the areas whose papers constitute more than 5% of the total number of papers with community information in both data sets, such as Theory, Distributed and Parallel Computing, Software Engineering, etc. We consider these papers and their sets of references. Again, we compute the correlations of the community weights of those citation edges and the normalized out-degrees of these citing papers. The correlations are given in Table 4. The results show that the correlations are mostly positive or neutral in three areas, except for papers in theory and algorithms in both ACM and CiteSeer. This implies that information diffusion has different impact on publications in theoretical computer science

and applied computer science. In future work we would like to probe these differences in information diffusion for various fields further. Subnetworks of papers in different time periods Instead of grouping papers according to their areas, we can also group them by publication date. In order to reduce the noise introduced by the incompleteness and sparsity of the data sets, we only choose papers in the following four time periods for both the ACM and Cite Seer data sets: 1980 1984, 1985-1989, 1990-1994, and 1995-1999. After grouping the papers according to publication date, same as before, we select the edges with destination papers in the chosen set of papers (e.g. published between 1990 and 1994).

Table 4: Correlations between community weights and normalized out degrees of citing papers grouped by different communities.

Venue	Dataset	Percent	Correlation
Theory & Algorithms	ACM Cite Seer	32.92%	−0.0709∗∗∗
		17.24%	−0.0169∗∗∗
Distributed Computing	ACM	6.39%	0.0223∗
	Cite Seer	10.45%	−0.0018∗
Artificial Intelligence	ACM	5.24%	0.0115∗
	Cite Seer	8.69%	0.0838∗∗∗
Software Engineering	ACM	19.54%	0.0386∗∗∗
	Cite Seer	19.37%	0.1010∗∗∗

We use Pearson correlations of community weights on the citation edges and the logarithm of normalized out degrees of the destination papers, which are shown in Figure 6. Although ACM and CiteSeer have different ranges of confidence intervals for the correlations, the trends of the two sets of correlations are consistent -they are slightly increasing as the time periods grow more recent. Perhaps with the research areas getting finer and deeper, it may be more difficult for researchers to keep up with, understand and cite papers in areas far from their own. At the same time, their own communities have grown and diversified to incorporate information

flows from other areas, so that citing within one's area may provide adequate diversity. However, these are only speculations as to the underlying reasons why citing within one's area would be of greater benefit to more recent papers. Subnetworks of papers versus books We consider one final subset of the citation graph, that of books. As we mentioned before, in the ACM dataset, there are documents labeled as books or book chapters. We select these documents, and study how their citations patterns may be different from those of journal research articles or conference proceedings. Since the datasets did not map books to different fields of computer science,

we just consider the raw out-degree of books as the measure of impact and focus on the time elapsed between the publication of the book and the work it cites. We consider citations from books or book chapters to any type of publication, including papers in journals and conference proceedings. Because books have longer reference lists (Figure 1b), any single citation is less likely to have a strong effect on the impact of the citing article. Indeed, we find that the correlation of time spans and the out-degrees of books is −0.049∗. This is lower than the corresponding correlation between out-degrees and time spans for papers in the same dataset, which is −0.069∗∗∗. This trend is also consistent with the fact that books are expected to cover a substantial amount of material, which may necessitate citing earlier publications. On the other hand, papers may just need to cite the most recent work they are building upon.
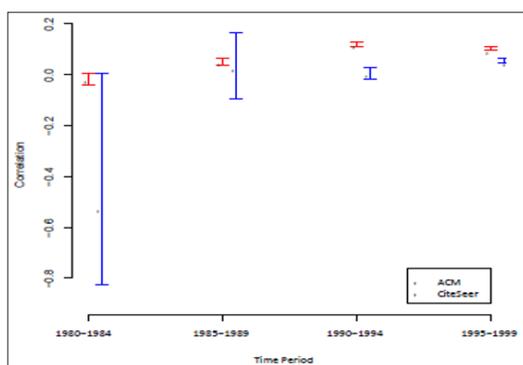

Figure 6: Correlations between community weights and normalized out degrees of citing papers, grouped by different time periods.

Conclusion and Future Work

We analyzed a very old, regimented, and established social medium for knowledge sharing in order to discover patterns of information flow with respect to community structure. Consistent with prior results, we find a wide range in the impact individual publications have. Information cascades, encompassing all chains of citations resulting from a single paper, vary dramatically in size, and only a small proportion of paper pairs are linked via cascades. In contrast the most influential papers are surprisingly interlinked. Many publications go mostly unnoticed, while some garner considerable attention. There are interesting factors, relating to the citation graph, that correlate with the popularity a given publication will enjoy. Our particular interest is on the impact of a particular citation on the success of the citing article. Through intensive study of two data sets of computer science publications, ACM and CiteSeer, we find that citations that occur within community's lead to a slightly higher number of direct citations; and also, citing more recent papers corresponded to receiving more citations in turn. However, our most interesting finding is that for the most influential group of papers, this relationship was reduced or absent, allowing for the possibility that ideas across communities can lead to higher impact work.

Finally, we find that the effect of recency and community on citation structure differs among different areas of computer science and among different time periods. In future work, we would like to expand our study to several additional contexts, including patent citation networks and paper citation networks of various scientific areas, in which the effect of boundary spanning information flows would be investigated. We would also like to extend our analysis to blogs, whose strong community structure has been observed, along with observations of information cascades, but little is known about the effect of this community structure on diffusion properties. However, the sparseness of citation data for blogs, and the loose relationship between them will present additional challenges. We would also like to extend our study using textual analysis, to map specific ideas that are spreading through the citation network. In doing so, we could identify the points at which a particular idea has crossed a community boundary, and measure whether this occasionally leads to large information cascades.

References

1. Simkin M, Roy Chowdhury V (2005) Stochastic modeling of citation slips. Scientometrics 62(3): 367-384.

2. Dieks, D, Chang H (1976) Differences in Impact of Scientific Publications: Some In-dices Derived from a Citation Analysis. Social Studies of Science 6(2): 247-267.

3. De Solla Price D (1965) Networks of Scientific Papers. Science 149(3683): 510-515.

4. Adar E, Zhang L., Adamic LA, Lukose RM (2004) Implicit structure and the dynamics of blogspace. In WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. ACM Press.

5. Rinia E, Van Leeuwen T, Bruins E, Van Vuren H, Van Raan A, et al. (2001) Citation delay in interdisciplinary knowledge exchange. Scientometrics 51(1): 293-309.

6. Lerman K, Jones LA (2007) Social browsing on flickr. In Proceedings of ICWSM.

7. Lerman K (2007) Social Information Processing in News Aggregation. IEEE Internet Computing 11(6): 16-28.

8. Tseng B, Tatemura J, Wu Y (2005) Tomographic clustering to visualize blog communities as mountain views. WWW 2005 Workshop on the Weblogging Ecosystem.

9. Shi X, Bonner M, Adamic LA, Gilbert AC (2008) The very small world of the well connected. In Hypertext '08, NY, USA, ACM p. 61-70.

10. Chin A, Chignell M (2006) A social hypertext model for finding community in blogs. In Hypertext '06, New York, USA p. 11-22.

11. Borner K (2004) The simultaneous evolution of author and paper networks. PNAS 101(supp l1): 5266-5273.

12. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. PNAS 105(4): 1118-1123.

13. Leicht EA, Clarkson G, Shedden K, Newman MEJ (2007) Large-scale structure of time evolving citation networks. The European Physical Journal B 59: 75.

14. Aksnes D (2006) Citation rates and perceptions of scientific contribution. JASIST 57(2): 169-185.

15. Boyack K, Klavans R, Börner K (2005) Mapping the backbone of science. Scientometrics 64(3): 351-374.

16. Guimera R, Uzzi B, Spiro J, Amaral L (2005) Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. Science 308(5722): 697-702.

17. Lazer D, Friedman A (2005) The hare and the tortoise: the network structure of exploration and exploitation. Proceedings of the 2005 national conference on Digital government research pp. 253-254.

18. Kumar R, Novak J, Raghavan P, Tomkins A (2003) On the bursty evolution of blogspace. In WWW New York, ACM Press 8(2): 568-576

19. Giles CL (2004) Citeseer: Past, present, and future. In AWIC p. 2.

20. Adamic L, Glance N (2005) The political blogosphere and the 2004 US election: divided they blog. LinkKDD p. 36-43.

21. Burrell QL (2003) Predicting future cita- tion behavior. JASIST 54(5): 372-378.

22. Zhou S, Mondragon RJ (2004) The Rich-Club Phenomenon in the internet topology. IEEE Commun Lett 8(3): 180-182.

23. Colizza V, Flammini A, Serrano MA, Vespignani A (2006) Detecting rich club ordering in complex networks. Nature Physics 2: 110-115.

24. Leskovec J, Mc Glohon M, Faloutsos C, Glance N, Hurst M (2007) Cascading behavior in large blog graphs. In SIAM International Conference on Data Mining 22:37.

25. Valderas JM, Bentley RA, Buck ley R, Wray KB, Wuchty S, et al. (2007) Why Do Team-Authored Papers Get Cited More? Science 317(5844): 1496-1498.

26. Stringer MJ, Sales Pardo M, Amaral L (2008) Effectiveness of journal ranking schemes as a tool for locating information. PLoS ONE 3(2): 1683.

Submission Link: https://biomedres.us/submit-manuscript.php

Assets of Publishing with us

- Global archiving of articles
- Immediate, unrestricted online access
- Rigorous Peer Review Process
- Authors Retain Copyrights
- Unique DOI for all articles

https://biomedres.us/