Information Diffusion in Computer Science Citation Networks

Xiaolin Shi
Dept. of EECS
University of Michigan
Ann Arbor, MI
shixl@umich.com

Belle Tseng
Yahoo Inc.
3420 Central Expwy
Santa Clara, CA
belle@yahoo-inc.com

Lada Adamic
School of Information
University of Michigan
Ann Arbor, MI
ladamic@umich.edu

Abstract

The paper citation network is a traditional social medium for the exchange of ideas and knowledge. In this paper we examine information diffusion in citation networks by analyzing the correlations between various citation choices and the subsequent impact of the articles. 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.

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 & Roychowdhury 2005), 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. The primary goal of our work is to investigate which features of citation networks, such as time spans and community structure representing different fields of research, affect the information flow.

The role of community structure in information diffusion has 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.* 2001). 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 2004; Rosvall & Bergstrom 2008). The development of efficient network algorithms has lead not just to discoveries of the 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.* 2007). There has also been interest in visualizing and quantifying the amount of information flow between different areas in science (Boyack, Klavans, & Börner 2005), 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.* 2005) 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. 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.

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 correlate the properties of a citing link to subsequent impact of the citing article.

Preliminaries

Definition of citation networks

From the graph theoretic perspective, citation networks can be thought of as directed graphs with time stamps and community labels on each node:

- *Nodes*: publications;
- *Edges*: one paper citing another;
- *Edge directions*: in order to represent the direction of information flow, we denote the direction of edges from cited papers to citing papers;
- *Time stamps*: years in which the papers were published;
- *Time spans*: the time elapsed between the publication of the cited and citing paper;
- *Community labels*: we classify the papers into different research areas according to their venue information i.e. the conferences or journals where they were published.

Information flows in citation networks can be interpreted as the scientific ideas and knowledge transmitted among publications, which are indicated by citation relationships. Not all 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.* 2007).

Description of data sets

The datasets we study are two large digital libraries encompassing comprehensive scholarly articles primarily in computer science — the ACM[1] data set and the CiteSeer [2] data set (Giles 2004). In the ACM data set, there are several different types of publications, such as books, journal articles, conference papers, reports, etc. 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 the ACM dataset includes citations to publications outside of it, 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 & Roychowdhury 2005). Even with these limitations, we are left with 346,000 citations for the ACM dataset and 84,000 citations in the CiteSeer dataset, which we use to measure information flows between different computer science communities and the impact of a publication.

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.

[1] http://portal.acm.org

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

Information diffusion and the effects of citations

We now examine 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 CiteSeer.

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 N_{ij} be the actual number of citations from i to j, $N_{i\cdot} = \sum_j N_{ij}$ be the total number of citations from community i, $N_{\cdot j} = \sum_i N_{ij}$ be the total number of citations to community j, and $N = \sum_{ij} N_{ij}$ 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 $E[N_{ij}] = N_{i\cdot} \times N_{\cdot j}/N$. We define the community weight as a z-score that tells us how many standard deviations above or below expected N_{ij} is. Here we have the observation that $N \gg N_{i\cdot}$ and $N \gg N_{\cdot j}$, so we approximate the standard deviation by $\sqrt{E[N_{ij}]}$. In this way, for every entry, we get a normalized value, which we call *community weight*:

$$W_{ij} = (N_{ij} - \frac{N_{i\cdot} \times N_{\cdot j}}{N})/\sqrt{\frac{N_{i\cdot} \times N_{\cdot j}}{N}}$$

[3] http://libra.msra.cn

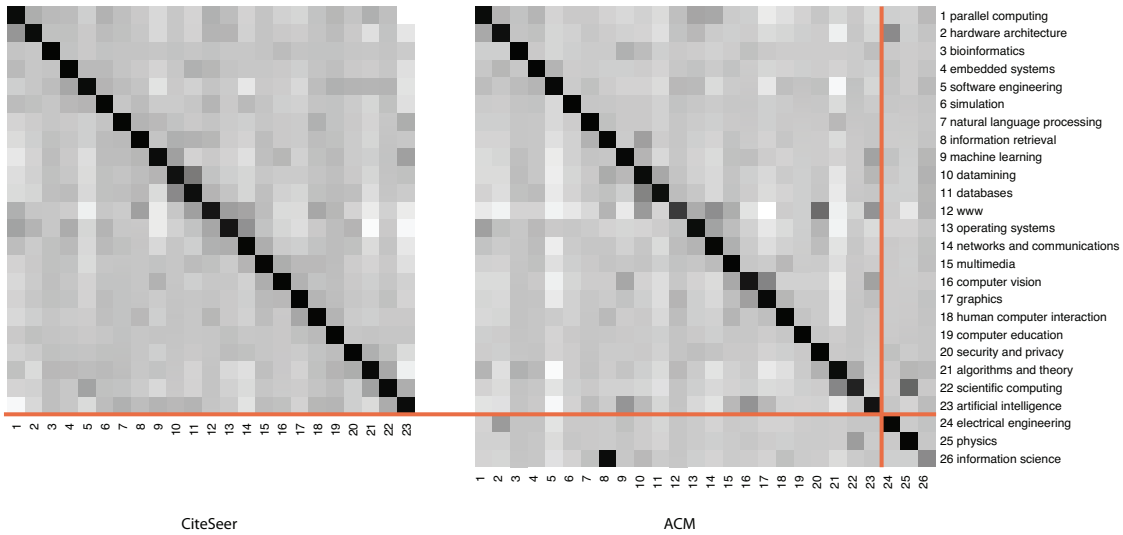

Figure 1: Visualization of the matrices of community weights between different areas of computer science. Darker cells represent more frequent citation than expected if citation were at random, lighter ones depict less frequent citation.

By visualizing the normalized matrix, i.e. matrix of community weights, as in Figure 1, 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. 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 2006), 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.* 2007). 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, Sales-Pardo, & Amaral 2008).

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, Sales-Pardo, & Amaral 2008; Burrell 2003).

Table 1 shows the correlations between community weights and time lapse of the citing and cited paper, and the subsequent impact of the citing paper. From it we see that for both citation networks, the weights of information 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 timelags 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

	ACM			CiteSeer		
	Overall	$\leq 90\%$	$>90\%$	Overall	$\leq 90\%$	$>90\%$
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^{*}

Table 1: Spearman correlations show the effects of community weights and time differences between the cited and citing papers on the subsequent impacts of citing papers.

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. In fact, we have observed that 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 $\leq 90\%$ in Table 1) 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. However, 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.

Conclusions 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. 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 communities 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.

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.

Acknowledgements

We would like to thank Eytan Bakshy for helpful comments and suggestions. This research was supported in part by a grant from NEC.

References

Aksnes, D. 2006. Citation rates and perceptions of scientific contribution. *JASIST* 57(2):169–185.

Borner, K. 2004. The simultaneous evolution of author and paper networks. *PNAS* 101(suppl_ 1):5266–5273.

Boyack, K.; Klavans, R.; and Börner, K. 2005. Mapping the backbone of science. *Scientometrics* 64(3):351–374.

Burrell, Q. L. 2003. Predicting future citation behavior. *JASIST* 54(5):372–378.

Giles, C. L. 2004. Citeseer: Past, present, and future. In *AWIC*, 2.

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

Leicht, E. A.; Clarkson, G.; Shedden, K.; and Newman, M. E. J. 2007. Large-scale structure of time evolving citation networks. *The European Physical Journal B* 59:75.

Rinia, E.; Van Leeuwen, T.; Bruins, E.; Van Vuren, H.; and Van Raan, A. 2001. Citation delay in interdisciplinary knowledge exchange. *Scientometrics* 51(1):293–309.

Rosvall, M., and Bergstrom, C. T. 2008. Maps of random walks on complex networks reveal community structure. *PNAS* 105:1118.

Simkin, M., and Roychowdhury, V. 2005. Stochastic modeling of citation slips. *Scientometrics* 62(3):367–384.

Stringer, M. J.; Sales-Pardo, M.; and Amaral, L. 2008. Effectiveness of journal ranking schemes as a tool for locating information. *PLoS ONE* 3(2):e1683.

Valderas, J. M.; Bentley, R. A.; Buckley, R.; Wray, K. B.; Wuchty, S.; Jones, B. F.; and Uzzi, B. 2007. Why Do Team-Authored Papers Get Cited More? *Science* 317(5844):1496b–1498.