Following Bibliometric Footprints: The ACM Digital Library and the Evolution of Computer Science

Shion Guha[1], Stephanie Steinhardt[2],
Syed Ishtiaque Ahmed[1]
[1]Department of Information Science
[2]Department of Communication
Cornell University, Ithaca, NY 14850
Email: {sg648, sbg94, sa738}@cornell.edu

Carl Lagoze
School of Information
University of Michigan
Ann Arbor, MI 48109
Email: clagoze@umich.edu

ABSTRACT

Using bibliometric methods, this exploratory work shows evidence of transitions in the field of computer science since the emergence of HCI as a distinct sub-discipline. We mined the ACM Digital Library in order to expose relationships between sub-disciplines in computer science, focusing in particular on the transformational nature of the SIG Computer-Human Interaction (CHI) in relation to other SIGs. Our results suggest shifts in the field due to broader social, economic and political changes in computing research and are intended as a prolegomena to further investigations.

Categories and Subject Descriptors

K2. History of: Theory.

Keywords

scientometrics; CHI; digital libraries

1. INTRODUCTION

In a recent article, Jon Kleinberg articulates that the past decade has seen a major transition within the field of computer science as a result of drivers like the internet and social networking [12]. This is a trend that is shaped by technical factors (e.g., increasing ubiquity of mobile devices) and by the growing socially embedded nature of computing that has developed with the pervasive internet and popular social media applications. As computing is increasingly entwined with our lives, the discipline of computer science has incorporated human factors into it.

In our research we quantitatively analyze and examine the Association of Computing Machinery (ACM) Digital Library (DL) to uncover evidence of this transition. The ACM is the "world's largest educational and scientific computing society" [1] and the ACM DL contains "every article ever published by ACM and bibliographic citations from major publishers in computing." [2] Almost all the articles in the ACM DL are affiliated with Special Interest Groups (SIG), each of which roughly corresponds to a subfield of the computer science discipline. The coverage and partitioning of this digital library therefore provides an excellent data source for an analysis of the shape of the discipline.

We analyzed all ACM SIG publications over time, their relative proportion to total publication output in the digital library, and their citation linkages to publications in other SIGs to mine evidence of the activity level of each subfield and the collaborations among them. The ACM's SIG in Computer-Human Interaction (SIGCHI) served as our lens. SIGCHI describes itself as "the world's largest association of professionals who work in the research and practice of computer-human interaction"[3] defined as an interdisciplinary group "composed of computer scientists, software engineers, psychologists, interaction designers, graphic designers, sociologists and anthropologists." [3] In addition to this qualitative reasoning, SIGCHI also has the largest number of publications of any SIG (27,964) with a very high growth rate in recent years (4.7%).

Our data shows an emergence of CHI as a top SIG coincident with the materialization of the Internet as a public resource, further emphasized in trends leveraging its massive amounts of online data and phenomena such as social media. We suggest that this emergence corresponds to a transition of computer science from traditional topics such as algorithms and complexity theory through an increasingly interdisciplinary flavor integrating designers, social scientists and artists.

The exploratory study detailed within this note is an important precursor toward understanding emerging forms of scientific work. The quantitative field-wide analysis leveraging the digital library provides a gateway to uniquely address issues of social and technical concern, such as developing theories to support interdisciplinary scientific collaboration, or defining relationships between funding and publication output.

JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.
Copyright © ACM 978-1-4503-2077-1/13/07...$15.00.

2. BACKGROUND

Much research has explored paradigm shifts in science and computing, emphasizing the connection between government and trends in computer science. For example, the U.S. National Research Council Committee on Innovations in Computing and Communications published a report entitled “Funding a Revolution” [13] detailing the role of the federal government in the rise of major trends in computer science such as relational databases, the development of the internet, theoretical computer science, artificial intelligence, and virtual reality.

Further research explores the embedding of the computer and networking as part of daily social interaction. Scholars use field introspection for evaluating conference practices and field directions, organizational policies and structure, and for using this knowledge to think critically about the field’s successes and failures in current structures. In particular, the CHI community has recently seen an increased attention to bibliometric methods as a self-reflection tool for understanding trends within the sub-discipline. Bibliometric is used as an approach that defines “What is CHI?” [5] and “How can HCI be evaluated?” [4]. Previous analyses of the ACM SIGCHI have been completed through a series of approaches: clustering CHI publications into thematic categories [11, 14, 17]; geographical analysis of the distribution of authors in one year at the CHI conference [6]; or authorship propensities within its sub-communities [10]. Barkhuus [4] sought to define the changes in evaluation methods published in the CHI conference to date and ground these approaches in historical trends of the subject of CHI papers: from a focus on the technical elite user and scientific technologies, to global widely-accessible internet tools and the layman user.

Shi et. al. [16] performed a comprehensive analysis of the ACM Digital Library and JSTOR to find that recent citation trends in computer science are biased towards interdisciplinary research and bridging different groups. We extend these inferences and investigate temporal citation patterns of sub-groups within computer science.

3. THE PRESENT STUDY

We performed a bibliometric analysis of all publications since SIGCHI’s inception in order to reveal evidence of the evolution of the computer science field.

3.1. Data Processing and Analysis

We extracted metadata records from the ACM Digital Library that includes information on all of the conference proceedings through 2012. We used Digital Object Identifier (DOI) as the primary key to differentiate between two publications. We used Python’s BeautifulSoup library to parse the XML metadata files; and then we made a number of manual adjustments to cope with any inconsistencies in the data (any auto-entry errors, non-standard characters, duplications, missing information etc.). We extracted citations from the metadata of each publication. For this iteration of the research, citations external to ACM were included in our analysis.

We then classified the publications in different SIGs by mining the Series ID of each of the conference proceedings. Each of the ACM conference proceedings falls under a series, and each series belongs to one or more SIGs. Because some conferences contained multiple series IDs in which multiple SIGs laid claim to them, our resulting sets of publications were not mutually exclusive by SIG. After preparing the dataset, we performed a statistical analysis on the intra- and inter-SIG citations. Intra-SIG is the average percentage of citations in a single publication of that SIG that refer to a publication internal to the SIG (intra-SIG). The inter-SIG citations refer to publications outside of its sub-discipline: SIG-to-CHI is the average percentage of citations in a single publication, which refers to SIGCHI; CHI-to-SIG is the average number of citations within a single publication of SIGCHI to that SIG.

3.2. Results

We focus our statistical analysis on average growth rates (Fig. 1) of citations per publication by year to distinguish evolving patterns and to compare these growth rates.

3.2.1 CHI-to-SIGs

When considering citations in CHI publication to other SIGs, we note some interesting differences.

A particularly salient relationship is found between SIGGRAPH (Computer Graphics and Interactive Techniques) and CHI. The inter-SIG citation rate between SIGGRAPH and CHI has been growing linearly ($R^2=0.998, p<0.001$, GR= 0.036) indicating a constant, strong, temporal relationship.

Conversely, SIGSOFT (Software Engineering) and CHI had a weak relationship in the early 1980s, but have since exhibited quadratic growth ($R^2=0.966, p<0.001$, GR= 0.021) indicating a strengthened relationship in the last 15 years. Towards the other extreme, SIGWEB (Hypertext and the Web) had no citation relationship with SIGCHI until the year 1997, despite that SIGWEB was born prior to CHI in 1987. At this point, there is an exponential growth pattern in citations between SIGWEB and CHI ($R^2=0.974, p<0.001$, GR= 0.12). SIGKDD (Knowledge Discovery and Data) holds an increasingly strengthened relationship with CHI since its inception in 1999, exhibiting a quadratic growth rate ($R^2=0.997, p<0.001$, GR= 0.092).

3.2.2 SIGs-to-CHI

On the other hand, CHI has also shown influence within the citation patterns of other SIGs. SIGIR (Information Retrieval) and SIGMOD (Management of Data) hold distinctly strong relationships in the SIG-to-CHI direction.

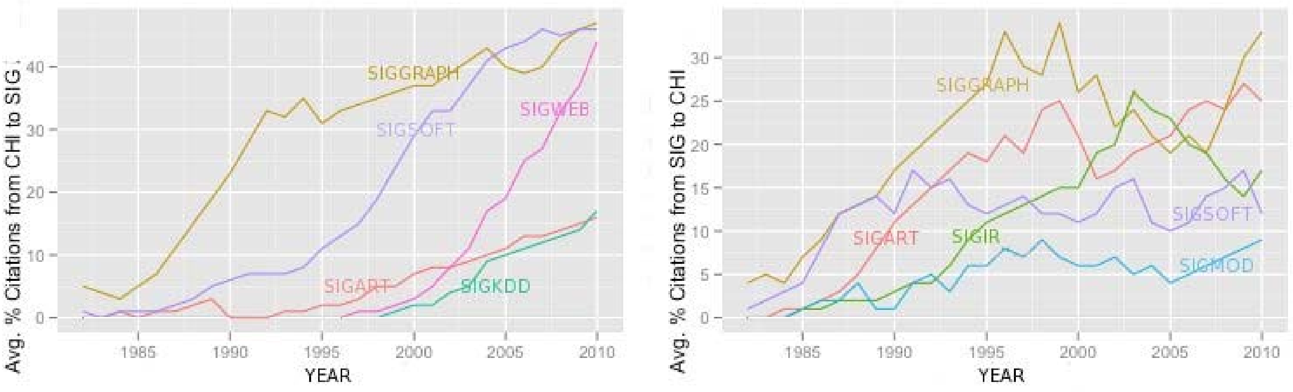

Figure 1: Top 5 SIG-to-CHI and CHI-to-SIG citation percentages per article from 1983 to 2010

These are not particularly strong in the direction of CHI and citing outward. Both groups exhibit linear relationships with CHI (SIGIR: R2=0.918, p<0.001, SIGMOD: GR= 0.021; R2=0.953, p<0.001, GR= 0.018) indicating a mild but constant amount of influence from CHI over the years. Perhaps somewhat unsurprisingly, SIGGRAPH citations show a reciprocal linear relationship with CHI (R2=0.993, p<0.001, GR= 0.042). There is also a constant linear growth exhibited by SIGART (Artificial Intelligence) citing CHI (R^2=0.977, p<0.001, GR= 0.034), suggesting a moderately strong influence from CHI over the years.

3.2.3 An Aggregative & Egocentric Vision of CHI

Aggregating citation patterns throughout CHI's lifespan and normalizing by raw publication count, developed an egocentric representation of the citation relationships between CHI and other SIGS. We refer to this calculation as ACP (Average Number of Citations per Publication). CHI and SIGGRAPH hold a very strong reciprocal relationship, as seen in their closeness within the CHI-centric model in Figure 2. Not only did SIGMOD and SIGIR's intertwining with SIGCHI grow over time as evidenced by Figure 1, but these two groups also represent the next most-cited SIGs by CHI after SIGGRAPH (1.73 ACP, 10%; SIGIR 1.68, 10%). While SIGOPS (0.14 ACP, 1%) and SIGWEB (3.2 ACP, 19%) represent the furthest nodes from CHI in Figure 2, we limited this visualization to only 10 total nodes, selected by largest average percentage of citations from CHI per publication.

SIGGRAPH (12.87 ACP, 54%), SIGART (12.64 ACP, 79%) and CHI (16.76 ACP, 51%) exhibit similarly dense self-citation patterns. In contrast, publications within SIGWEB on average exhibit a very weak self-citation percentage (1.28 ACP, 79%)

4. DISCUSSION

Our results show the patterns of evolution of the field of computer science during the 29 years of SIGCHI's existence, largely through positive citation relationships between SIGCHI and other SIGs. Not only has there been a movement within CHI, but also a shift in the focus of all SIGs as seen by citations into CHI by core computer science sub-disciplines which is presented in Figure 1.

We believe this is evidence of at least three significant paradigm changes in the field of computer science since its inception in the post-World War II period. First, in the 1980s there was a shift from a focus on algorithms toward use cases, and the introduction of human factors into the field of computer science. From this, the field of HCI was built as the interface between computer science and social science and bloomed as an interdisciplinary sub-discipline of computer science, which recognizes the importance of the human in technology development. Barkhuus [4] outlines the trends in early HCI work as being concerned with the technical elite: how technical users use command

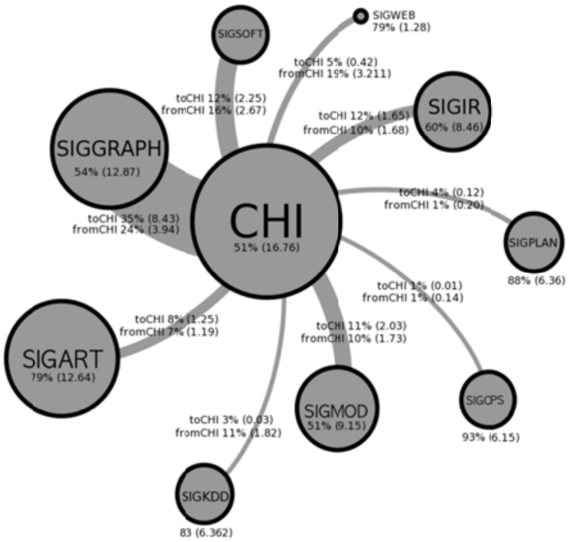

Figure 2: An aggregative snapshot of ACM SIGCHI's lifespan of and nine (9) SIGs with largest percentage of CHI citations per publication normalized by the total number of publications: node size is the average percentage of in-SIG citations, edge is the average percentage of SIG-to-CHI citations and edge length is 1 /average percentage of CHI-to-SIG citations.

lines [8], use their tacit programming knowledge [15] and handle an interface [9].

During the first decade of its existence, CHI reached its highest level of self-citation, indicating the emergence of HCI as a new thread of computer science with a distinct identity from the algorithm-focused research. At the time of CHI's origins, SIGGRAPH produced the largest output of any SIG in the ACM.

The second paradigm shift in computer science occurred in the late 1990s when the internet and personal computer redefined the end-user from the technical elite and brought with it studies of how the layman interacts with new technologies [4, 7]. As a result of the increased necessity to understand how to build new technologies for public consumption, the graphical user interface (GUI) also received increased attention during this time. Our data suggests that it was here when HCI research shifted from the command line toward a closer integration of SIGGRAPH concerns (interaction, design and information) and solidified a reciprocally strong intertwining between the two sub-disciplines. This period also coincides with the CHI's large publication output, surpassing in raw numbers such "core" computer science sub-disciplines as programming languages, operating systems, software development.

In the early 2000s, we see what appears to be another shift in the field of computer science, as social media integrated even more closely the relationship between technology and the everyday user. During this period, the interchange between CHI and SIGKDD intensified, as large-scale data mining of online social systems became one of the foundations of HCI research.

5. CONCLUSIONS AND FUTURE WORK

Major implications from this work are two-fold. First, we have quantitatively shown the complex, temporal citation relationships between different sub-areas of computer science with a slant towards human-computer interaction. Second, we have sought to ensconce these relationships in a broader, historical context to show the evolution of computer science over the last few decades from being algorithm-driven to being driven by human data. The analysis presented here describes high-level trends that could be identified through our data set. To the best of our knowledge, our exploratory work is the first to show this evidence using bibliometric reasoning. Future work intends to address more acute shifts with less obvious relation to the dialog and rhetoric surrounding computer science.

We hope to spark interest in the politics that shape computer science, such as relationships of funding agencies to these paradigm shifts in field focus; the roles of universities, industry and government in moving forward computing; and potential future directions for the field and sub-disciplines. In the future, we intend to probe deeper into the citation relationships between other disciplines (sociology, communication, psychology) and computer science and the trichotomous relationship between funding, productivity and impact in computer science.

6. ACKNOWLEDGMENTS

We wish to thank the ACM Digital Library for providing their corpus to us for our research.

7. REFERENCES

[1] ACM. 2013. http://www.acm.org/

[2] ACM Digital Library. 2013. https://dl.acm.org/

[3] ACM Special Interest Groups. 2013. https://dl.acm.org/sigs.cfm?CFID=273142187&CFTOKEN=24062479

[4] Barkhuus, L and Rode, J. 2007. From Mice to Men - 24 Years of Evaluation in CHI. In Proc. CHI'07. DOI=10.1145/1240624.2180963

[5] Bartneck, C and Hu, J. 2009. Scientometric analysis of the CHI proceedings. In Proc. CHI'09, pg. 699-708.

[6] Diakopulous, N. Geographical distribution of CHI authors. http://www.cc.gatech.edu/~nad/Projects/CHIViz

[7] Grier, D.A. 2007. When Computers Were Human. Princeton University Press.

[8] Haggett, A, McFadden, J and Newsted, P. 1981. Naive user behavior in a restricted interactive command environment. SIGSOC Bull. 13, 2-3, 139-.

[9] Hammond, N, Jørgensen, A, MacLean, A, Barnard, P and Long, J. 1983. Design practice and interface usability: Evidence from interviews with designers. In Proc. CHI'83, 40-44.

[10] Horn, D, Finholt, T, Birnholtz, J, Motwani, D and Jayaraman, S. 2004. Six degrees of jonathan grudin: a social network analysis of the evolution and impact of CSCW research. In Proc. CSCW'04, 582-591.

[11] Kaye, J. 2009. Some statistical analyses of CHI. In Proc. CHI EA'09, 2585-2594.

[12] Kleinberg, J. 2011. The Human Texture of Information. Is the Internet Changing the Way You Think? Eds. Brockman, J. Harper Perennial.

[13] Nat. Res. Council Com. on Inn. in Comp. and Comm.1999. Funding a Revolution: Government Support for Computing Research. Nat. Academy Press.

[14] Newman, W. 1994. A preliminary analysis of the products of HCI research, using pro forma abstracts. In Proc. CHI'94, 278-284.

[15] Soloway, E, Ehrlich, K & Bonar, J. 1982. Tapping into tacit programming knowledge. In Proc. CHI'82, 52-57.

[16] Shi, X, Leskovec, J and McFarland, D. 2010. Citing for high impact. In Proc. JCDL'10, 49-58.

[17] Wulff, W and Mahling, D. 1990. An assessment of HCI: issues and implications. SIGCHI Bull. 22, 1 (June 1990), 80-87.