Weaving the Topics of CHI: Using Citation Network Analysis to Explore Emerging Trends

Clement Lee
School of Mathematics, Statistics and Physics
Newcastle University
Newcastle upon Tyne, UK
clement.lee@newcastle.ac.uk

Andrew Garbett
Junyan Wang
Bingzhang Hu
Dan Jackson
Open Lab
Newcastle University
Newcastle upon Tyne, UK
andy.garbett@newcastle.ac.uk
j.wang81@newcastle.ac.uk
bingzhang.hu@newcastle.ac.uk
dan.jackson@newcastle.ac.uk

ABSTRACT

This paper provides a comprehensive and novel analysis of the annual conference proceedings of CHI to explore the structure and evolution of the community. Self-awareness is healthy for a diverse and dynamic community, allowing it to anticipate and respond to emerging themes. Instead of using a traditional topic modelling approach to analyze the text of the papers, we adopt a social network analysis approach by analyzing the citation network of papers. After constructing such a citation network, community detection is applied in order to cluster papers into different groups. The keywords of these groups are found to represent different research themes in human-computer interaction,

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3312776

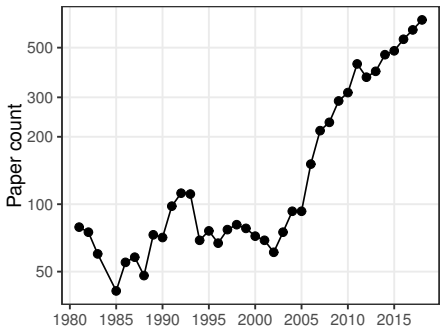

Figure 1: Number of CHI papers, plotted on logarithm scale, by year.

while the growth or decline of these groups is visualized by their paper shares over time. Lastly, we contribute a visualization tool for exploring emerging trends within our community, which can be used to predict likely topics at future CHI conferences.

KEYWORDS

Bibliometrics; citation network analysis; community detection; trend prediction; topic modelling; visualization

INTRODUCTION AND BACKGROUND

CHI is a diverse, multi-disciplinary, and cross-cultural community – one which has changed significantly over its lifetime. The annual conference proceedings are the essential means of disseminating contributions, and can inform us about the structure and evolution of the CHI community. An analysis of these publications can provide important insights, and self-awareness is healthy for a community, allowing it to anticipate and respond to emerging themes.

Bibliometrics are useful in understanding the big picture of an academic field. The landscape of human-computer interaction (HCI) has already been studied in this way, in this very conference [1, 6, 8]. Although these analyses have been useful, existing approaches have focused on analyzing only the publication text. While Padilla *et al.* [6] have incorporated a network approach, these networks are constructed from the word frequencies, which lead to the results being heavily influenced by the most commonly used terms in the field. Other existing research has analyzed intra-conference publishing trends through authorship analysis [5] to understand collaboration within CHI. Henry *et al.* [4] attempt to understand the wider HCI community through inter-conference citation networks for the purposes of visualization. Matejka *et al.* [7] use citation network data to explore affordances of visual design when presenting genealogical citation network data. However, existing attempts have yet to explore citation networks for the purposes of topic analysis for emerging trends.

We consider citation network analysis as an alternative to clustering of papers, which has the same goal of topic analysis, as it removes two problems: the difficulty that words may have multiple, changing meanings; and the need to create an artificial network. In addition, it directly links relevant papers through references.

METHODOLOGY

Our methodology consists of: obtaining information about SIGCHI papers for each year of publication, including the papers they reference; creating a network of the citations; and, performing community detection within the citation network.

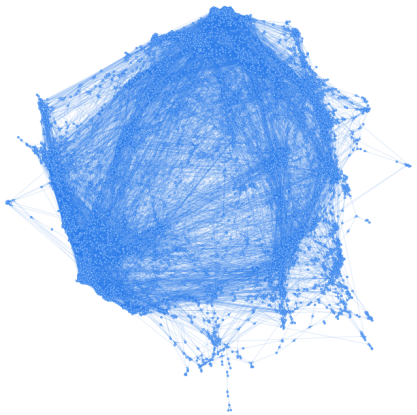

Figure 2: The giant component of the citation network of the CHI papers. Each of the 6239 nodes represents a paper, while each of the 26662 edges represents a reference/citation.

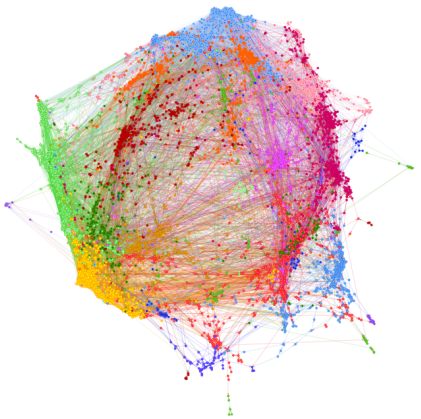

Figure 3: Groups detected by the spin glass algorithm and visualized in different colors.

Data Collection

The *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* for the years of its publication (1981 to 2018, excluding 1984) are available in *The ACM Digital Library* (ACM DL, dl.acm.org). We automated fetching of the *Table of Contents* from each of the proceedings to obtain a list of publications for that year and, for each publication: a unique ACM DL identifier, author names, title, abstract (if available), and DOI (if available). This process resulted in the details for 6935 *CHI papers*, which break down over the years as plotted in Figure 1. Only papers in the main tracks are included.

For each CHI paper, we automated the download of the references page for the entry in the ACM DL to obtain the text of references for the publication and, where the referenced work is identified in the library, its unique identifier. Note that the matching for references within the ACM DL is imperfect and that, in addition, the earlier years may contain errors from the OCR of scanned pages. Where the list of references in paper *A* contains paper *B*, paper *B* is a *reference* of *A* and, conversely, *A* represents a *citation* of *B*. Also, each is mutually considered as a *neighbor* of the other.

Citation Network

Our approach to detecting communities within the realm of HCI requires us to first construct a citation network: a graph of papers connected by edges formed by references/citations. When considering the bounds of our network, we have the references *between* the CHI papers themselves, and we could choose to extend this by finding citing papers outside the field by scraping citations (rather than references) for these papers. However, references external to the ACM DL remain *unresolved*: they are not reliably uniquely identified, with only partial and uneven availability of DOIs, and we have no readily available means to collect the equivalent metadata for those papers. We could consider all papers within the ACM DL as these references can be *resolved*, yet this approach would form an arbitrary domain across all the fields within the ACM. Instead, we choose to draw a justifiable line and stay entirely within the realm of CHI papers and the internal references/citations within that community.

From the nature of citations within an annual conference, we expected to create a *directed acyclic graph* (DAG). However, 60 pairs of papers (120 references) were found to mutually cite one another. These pairs were broken systematically by removing one reference (preferring references to an earlier year, otherwise arbitrarily by unique numeric identifier). No circular references (with a longer path length) remained. In the complete citation network, only 6239 are fully connected, meaning that a path through neighbors can be found between any two papers. The network of these papers, which amount to 90% of the 6935 papers, has 26662 references/citations in total, and is called the *giant component* hereafter. This giant component is plotted in Figure 2, and will be the focus of our analysis.

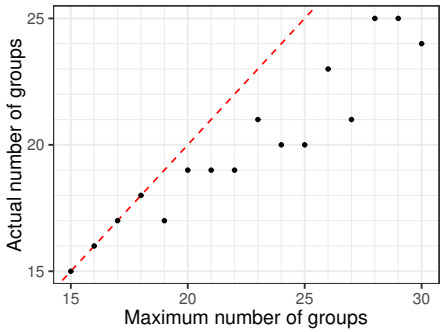

Figure 4: Actual number of groups detected by the spin glass algorithm, against the maximum number of groups permitted. The chosen actual- and maximum number of groups are 19 and 20, respectively.

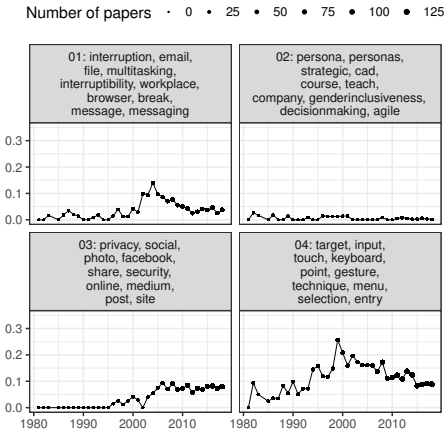

Figure 5: Paper share of the groups detected each year, for groups 1 to 4. The dot size represents the number of papers.

The results should not be meaningfully affected by discarding the remaining 10% of the total number of papers as, among the connected graphs they form, the largest consists of only four papers.

Community Detection

Based on the giant component, we can cluster the papers so that connectivity is high within each group, and low between different groups. This is achieved by *community detection* methods, designed specifically for clustering network data. Over the last two decades, numerous community detection algorithms have been developed using different heuristic rules, and are detailed in the comprehensive review by Fortunato [2]. We use the algorithm by Reichardt and Bornholdt [9], which is based on the *spin glass* model in statistical physics, for three reasons. First, it gives more representative results in the sense that the groups are less unequal in size, compared to other algorithms. Second, it is flexible as it allows the maximum number of groups to be specified beforehand. Finally, empirically it usually results in high modularity, which is the main criterion for judging how good the clustering is [3].

RESULTS

Upon applying the spin glass algorithm to this data set, we found that the actual number of groups detected does not increase linearly with the maximum number of groups permitted; see Figure 4. Further increasing the maximum number only results in smaller groups splitting from larger groups, instead of significantly re-clustering a large proportion of the papers. The results of 19 actual groups (for a maximum of 20) are reported here. With the same layout as Figure 2, we visualize the groups in Figure 3.

Topic Words

To illustrate the usefulness of the clustering, we examine the words used in the papers' abstracts by calculating their *term frequency-inverse document frequency* (TFIDF). The top ten words for each group, which can be seen as the topic words, are reported in Figures 5, 6 and 9. Note that the words come into the analysis post-clustering, compared to being the main vehicle of clustering in the traditional topic modelling.

The topic words found are highly representative of different research themes in the HCI community, especially in the larger groups. On the other hand, group 11 is quite isolated with 14 papers only. This is one drawback of community detection algorithms in general, which are usually probabilistic in nature, and may not be able to merge smaller groups with larger ones.

Topic Trends

We can explore the dimension of publication year post-clustering by computing the percentage of papers of each group in each year. This measure is called the paper share hereafter, and is plotted

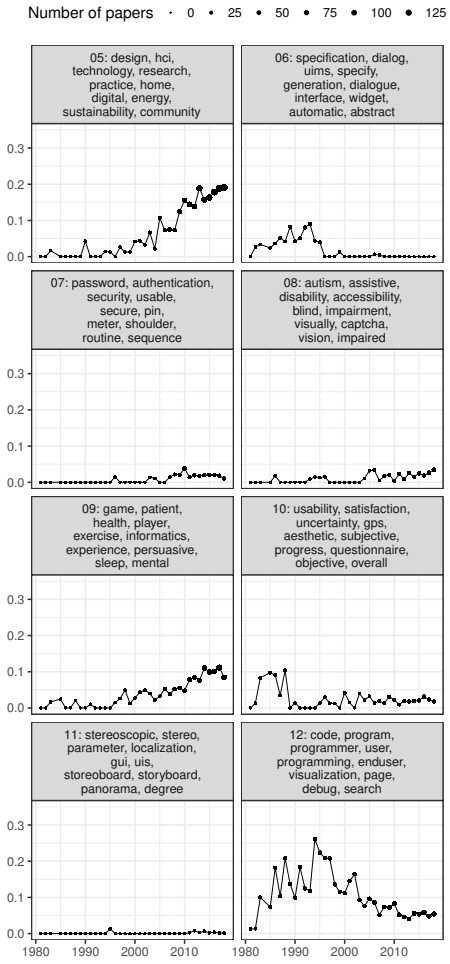

Figure 6: Paper share of the groups detected each year, for groups 5 to 12. The dot size represents the number of papers.

over publication year in Figures 5, 6 and 9. This is essentially splitting the total number of papers each year in Figure 1 by group. The rise and fall of various topics can be seen clearly, with the two corresponding to groups 5 and 18 being the major themes in recent years.

DISCUSSION AND FUTURE WORK

Within this paper we have explored CHI proceedings through its paper citation network to understand longer term trends and derive meaningful descriptors to these groupings. We contribute a visualization tool[1] for exploring these emerging trends within our community (Figures 7 and 8).

Given the evolving nature of technology, these trends are reflected in the emergence of new topics within CHI. Despite this, subcommittees at CHI change infrequently which can lead to some becoming larger and amorphous, with new and emerging research failing to find a suitable, more specific committee. Given this, we argue that bibliometrics might well be used to establish new data-driven subcommittees to support emerging themes at CHI. However, we must undertake additional processing to collate topic keywords into single, defined themes. In future work we will also model our citation network data set in order to understand group distributions, quantify how their interconnectedness evolves over time, and identify emerging groups for future CHI proceedings.

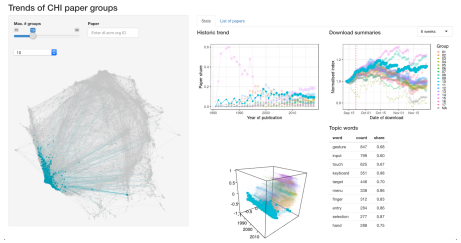

Figure 7: Visualization tool to explore citation network and associated keywords from TFIDF.

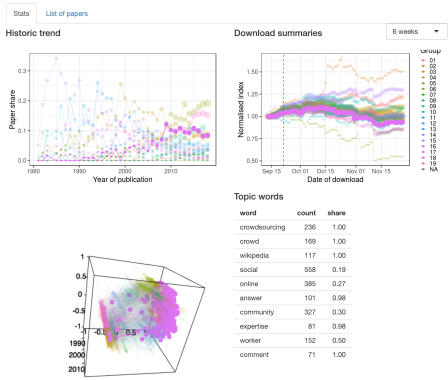

Figure 8: Explore group historic trends (paper shares), download summaries, and topic keywords.

[1] https://clement-lee.shinyapps.io/chi_topics

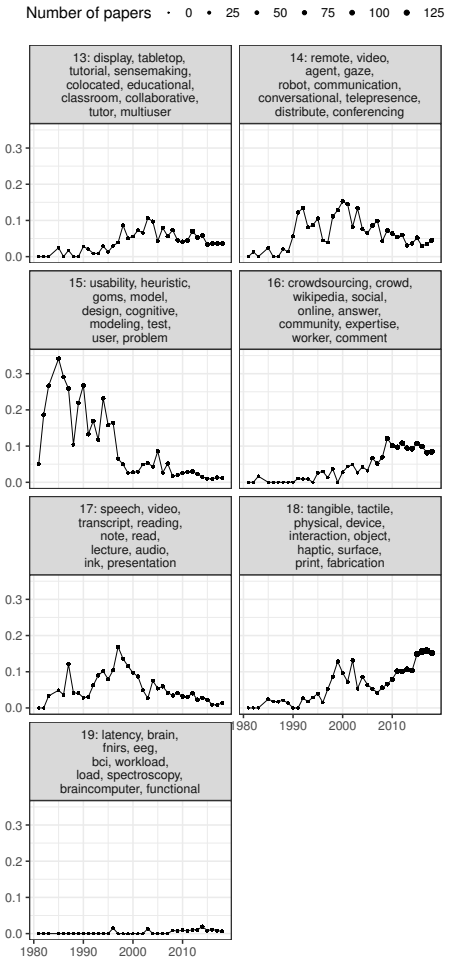

Figure 9: Paper share of the groups detected each year, for groups 13 to 19. The dot size represents the number of papers.

Our approach allows us to understand the impact of publications beyond simply identifying highly cited papers and towards determining those that cut across domains, or are highly influential within smaller topic areas. Similarly, we can also use this analysis to highlight *topic setting* papers that introduce new concepts to the CHI community.

Future work will explore predictive aspects of the data and use download statistics to strengthen our predictive models in order to understand the distribution of topic areas in upcoming CHI proceedings.

Our novel method can also be applied to other ACM conference proceedings and we call upon other researchers to explore their domains accordingly.

We hope that this paper, through leveraging our novel approach to topic modelling and community detection within the proceedings of CHI, inspires new research questions and begins to explore the potential of these forms of community self-reflection.

ACKNOWLEDGMENTS

This research was funded by the EPSRC Digital Economy Research Center (EP/M023001/1). Data supporting this paper is available under an 'Open Data Commons Open Database License'. Contact Newcastle Research Data Service at rdm@ncl.ac.uk.

REFERENCES

[1] Christoph Bartneck and Jun Hu. 2009. Scientometric Analysis of the CHI Proceedings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 699–708. https://doi.org/10.1145/1518701.1518810

[2] Santo Fortunato. 2010. Community detection in graphs. *Physics Reports* 486 (Feb 2010), 75–174. Issue 3-5. https://doi.org/10.1016/j.physrep.2009.11.002

[3] Benjamin H. Good, Yves-Alexandre de Montjoye, and Aaron Clauset. 2010. Performance of modularity maximization in practical contexts. *Physical Review E* 81, 4 (2010), 046106.

[4] Nathalie Henry, Howard Goodell, Niklas Elmqvist, and Jean-Daniel Fekete. 2007. 20 years of four HCI conferences: A visual exploration. *International Journal of Human-Computer Interaction* 23, 3 (2007), 239–285.

[5] Joseph 'Jofish' Kaye. 2009. Some Statistical Analyses of CHI. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems (CHI EA '09)*. ACM, New York, NY, USA, 2585–2594. https://doi.org/10.1145/1520340.1520364

[6] Yong Liu, Jorge Goncalves, Denzil Ferreira, Bei Xiao, Simo Hosio, and Vassilis Kostakos. 2014. CHI 1994-2013: Mapping Two Decades of Intellectual Progress Through Co-word Analysis. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3553–3562. https://doi.org/10.1145/2556288.2556969

[7] Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2012. Citeology: Visualizing Paper Genealogy. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems (CHI EA '12)*. ACM, New York, NY, USA, 181–190. https://doi.org/10.1145/2212776.2212796

[8] Stefano Padilla, Thomas S. Methven, David W. Corne, and Mike J. Chantler. 2014. Hot Topics in CHI: Trend Maps for Visualising Research. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems (CHI EA '14)*. ACM, New York, NY, USA, 815–824. https://doi.org/10.1145/2559206.2578867

[9] J. Reichardt and S. Bornholdt. 2006. Statistical mechanics of community detection. *Physical Review E* 74, 1 (2006), 016110.