Some Statistical Analyses of CHI

Joseph 'Jofish' Kaye
Nokia Research Center
955 Page Mill Road
Palo Alto CA 94301
USA
jofish.kaye-at-nokia.com

CHI 2009, April 4 – 9, 2009, Boston, Massachusetts, USA
ACM 978-1-60558-247-4/09/04.

Abstract
In this paper I show a variety of ways to represent and think about statistical aspects of CHI and its sister conferences. In particular, I look at author counts, gender analysis, and representations of repeat authors. I use these visualizations to motivate questions about what the preferred state of CHI should be. For example, should we strive for gender equality at CHI, and if so, why, and if not, why not? Should we encourage the current trend of increasing number of authors per paper, or might we be losing something in that process? I do not hope to answer these questions, but rather to encourage their discussion.

Keywords
Visualizations, statistical analysis, bibliometric analysis, gender, authorship.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction
A certain amount of introspection is healthy for a field. Looking at ourselves can provide insights into the ways we do work and can help understand the field and the ways it is changing over time.

In HCI, this has taken a number of forms. For example, Grudin has been instrumental in encouraging consideration of the role that a narrative historical approach can bring to a better understanding of the field (i.e. [4]). These are characterized by discussions of the history of the field in terms of the actors, institutions, conferences and trends that were important at various times.

A second strand of analysis comes from content analysis of the publications themselves For example, Barkhuus & Rode studied selections of papers at four-year intervals over the history of the CHI conference to show the increasing importance of evaluation as a necessary component of a published CHI paper [1].

A third strand of self-analysis comes from bibliometric analyses of works, and analysis of the meta-data around publications. For example, Oulasverta used bibliometric analysis to determine statistic like the most cited first authors, most influential sites of research, the most cited papers and the most prolific authors [6]. Similarly, Wania et al. used citation analysis and multivariate analyses to cluster the field of HCI into seven clusters and to demonstrate the centrality of various authors to the field [7] Diakopulous analyzed and charted the geographical distribution of authors at CHI 2006 [3] and in this year's proceedings, Bartneck & Hu use a bibliometric approach to identify trends including, among others, geographical origin of authors and prolific institutions [2].

The work I present here was first presented on my personal blog, http://jofish22.livejournal.com, and, in one case, the visualization of repeat authors at CHI and other conferences, through my personal Facebook page. Self-publishing this work in this semi-public forum prior to its publication here has had both advantages and disadvantages. The disadvantage is that to some people this work will not be novel, as they may have seen some or all of it in passing, perhaps due to it showing up in searches, or because they are personal friends who read my blog. However, this disadvantage is significantly outweighed by the advantages of the public discussion, critique, and subsequent revisions of the work. Particularly in the analysis of gender I will discuss later, the public nature of the analysis was important in improving the quality of the results.

Methods

All of the data studied in this paper were downloaded from the ACM Digital Library (www.acm.org/dl), either through custom-written Python scripts or through existing tools such as Flashgot. The ACM DL does not provide data to the public in any structured format, requiring custom scripts to analyze scraped data. Data was then saved as simple text files that were analyzed using more Python scripts and Microsoft Excel.

All data in these examples used only the *Proceedings of the SIGCHI Conference* of the year in question, not the extended abstracts; furthermore I have no access to lists of either attendees or authors of submitted papers, which would be interesting points of comparison.

Author Counts

A comparatively simple metric for looking at changes in papers submitted to CHI over the course of its 25-year history is to look at the number of authors for each paper. The simplest way to represent this change is just to look at the raw mean (average) and mode (most

frequently-occurring result). An analysis of the 1983 proceedings shows that the average was 2.2 authors and the mode was 2 authors. By 2007, this had climbed to an average of slightly over 3.5 authors with a mode of 3 authors per paper. However, this does not tell the whole story. For example, the effect could be produced by just one paper with the kind of authorship pattern common in the high-velocity particle physics community, where hundreds of authors are common. This comparatively simple story – along with interim datapoints – is shown in Figure 1.

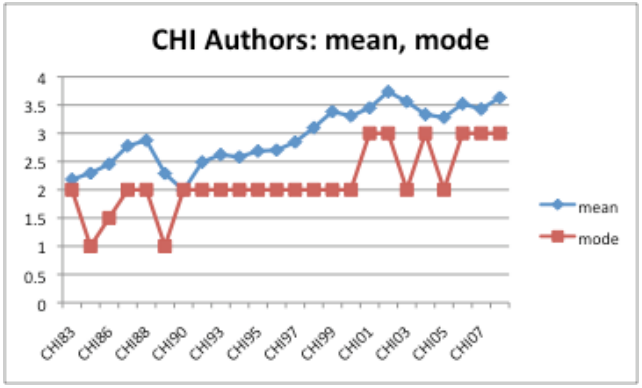

Figure 1. Number of authors. In 1986, the same number of papers had 1 and 2 authors, thus the 'mode' of 1.5.90-

In the bubble plot in Figure 2, we can see a different representation of the same information. This clearly shows the decrease in single author papers and the corresponding increase in 2-to-4 author papers.

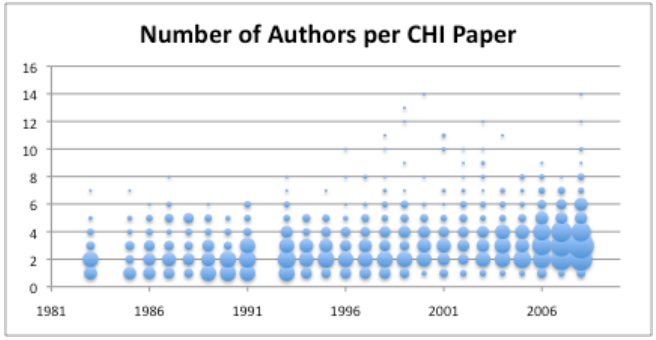

Figure 2. Frequency of number of authors per year. Thanks to Seth 'Beemer' McGinnis for assistance with this visualization.

A third representation of this same data is shown in Figure 3. This representation demonstrates the increasing frequency of papers with more than six authors, making up some 10% of the papers in 2006 and yet is nearly completely absent for the first 15 years of CHI's history.

I include these three different representations of the same information (four if one counts the numerical representation in the first paragraph) to emphasize the difficulties in choosing ways to represent data in meaningful ways. I could, for example, have emphasized the number instead of the percentage of authors to show the growth of the conference. Instead, I have chosen to emphasize factors like the increasing absence of single-authors papers in the field.

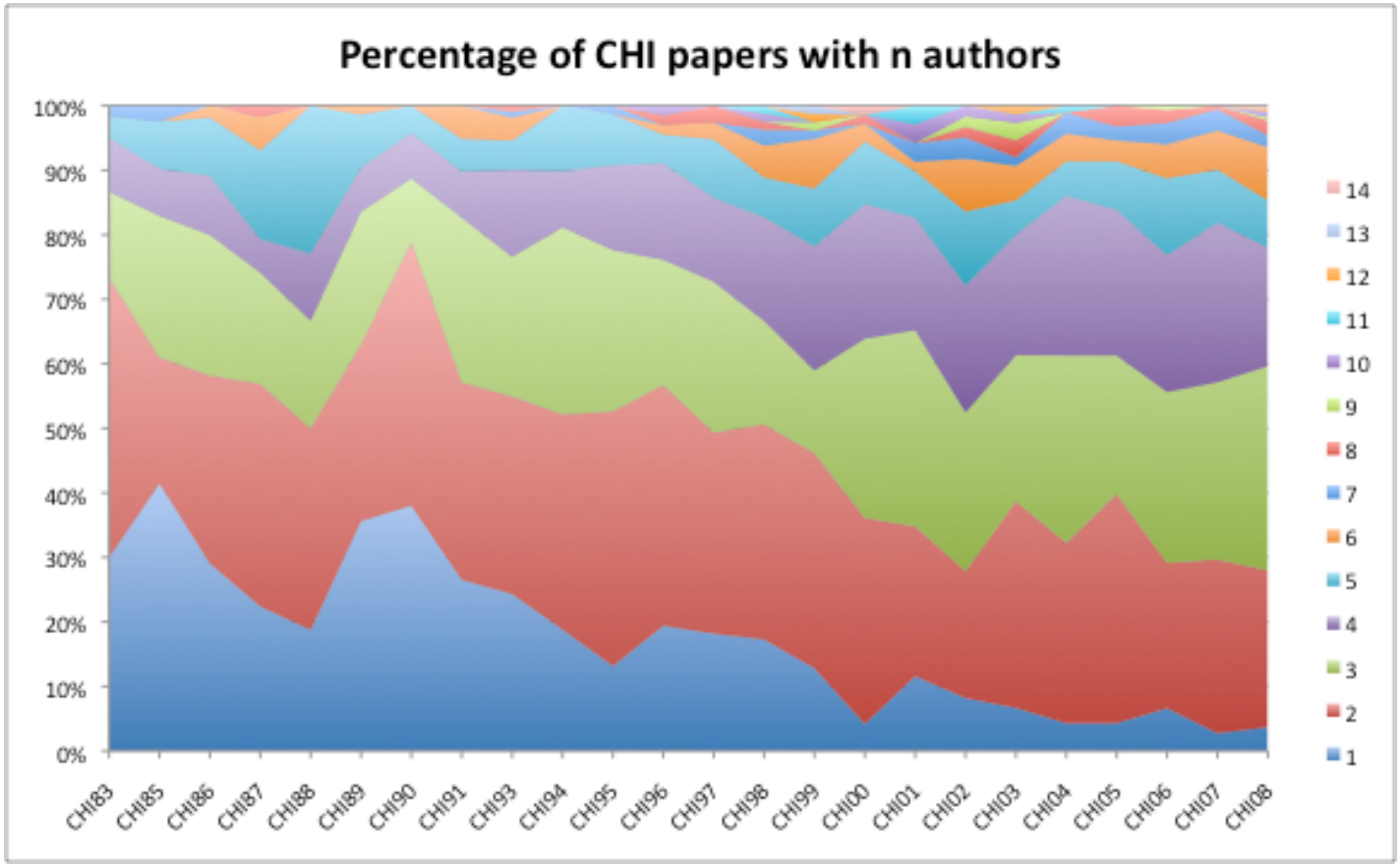

Figure 3: Number of authors as a percentage of the total papers for that year.

Gender analysis

My second set of graphs is about the percentage of women at CHI. This was prompted by a visit in 2006 to NordiCHI, a conference which felt different to me from many others I have been to. I hypothesized that this may have been due to the increased number of women authoring papers – an effect, I had thought, that might be due to Scandinavia's history and culture of female equality. I decide to test this hypothesis by extracting names from the proceedings of NordiCHI and other conferences, and running them through a simple gender analyzer. This last clause turned out be significantly trickier than I anticipated.

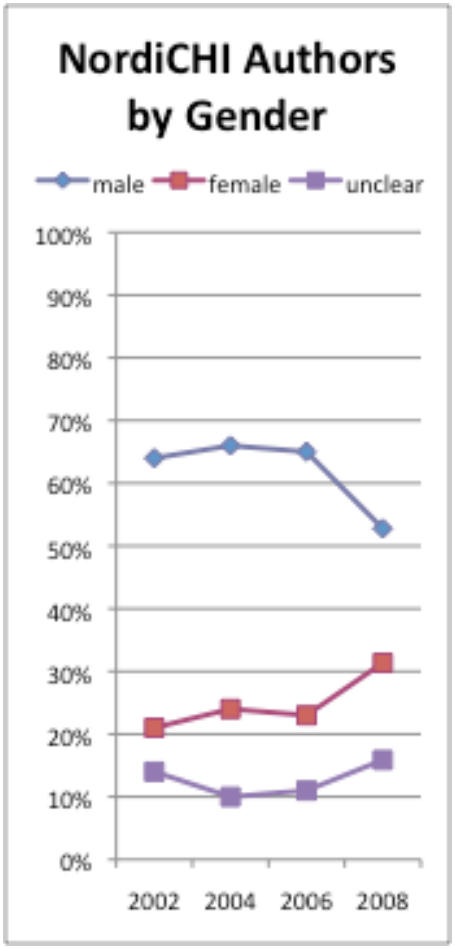

Figure 5: Nordichi Authors by Gender. This graph is position to facilitate comparison to Figure 4.

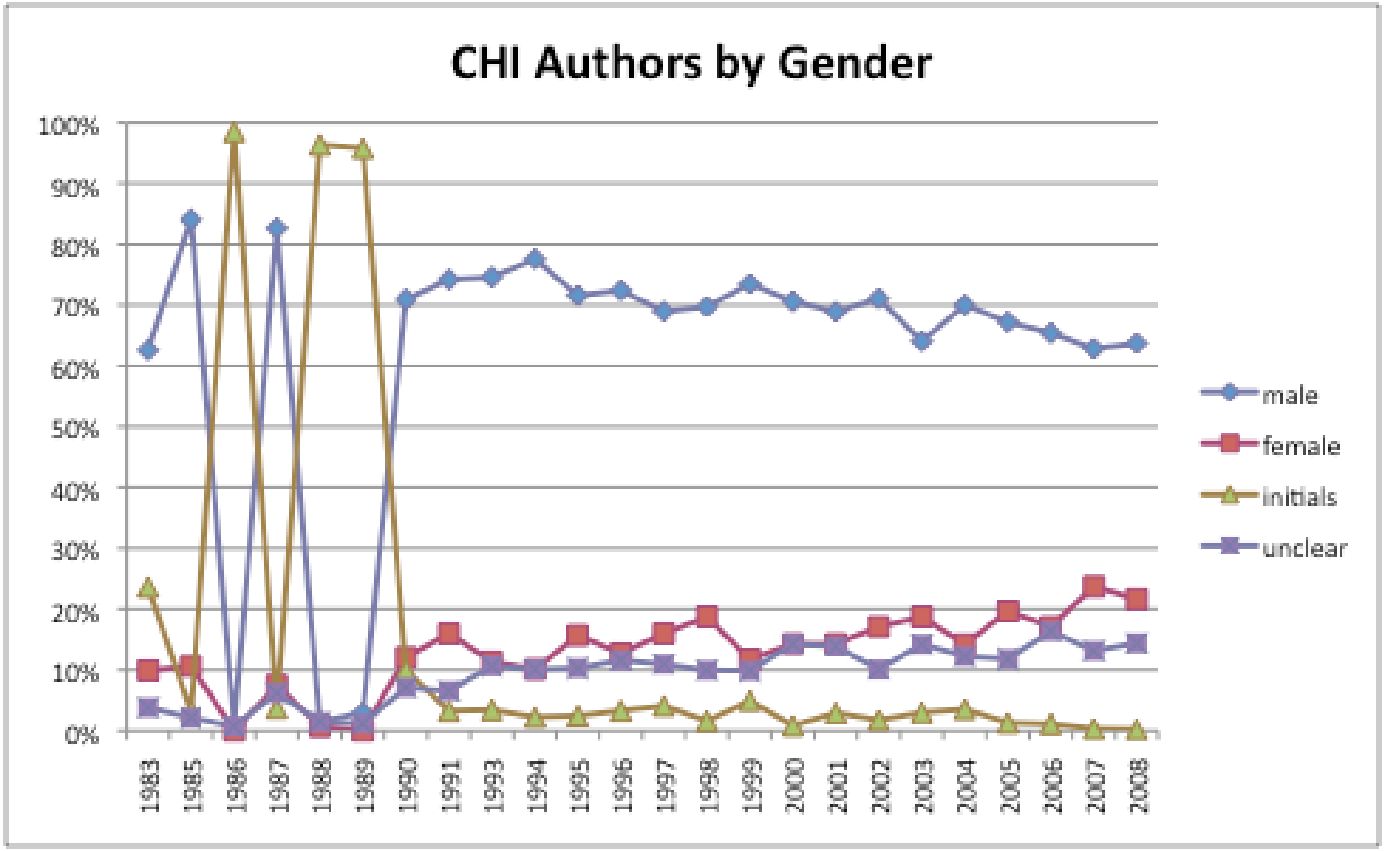

Figure 4. CHI Authors by Gender

Determining gender from a name is not a trivial task, even for humans. As data sources, I used a variety of official lists, including census data and official lists from which babies' names must be chosen, which are common in, for example, several Scandinavian countries. However, even with these lists I still found I was unable to automatically determine the gender of nearly a third of my datapoints. Luckily, I was able to augment these data sources with a variety of lists scraped from several baby-naming websites. The results of such an analysis for CHI are shown as Figure 4. By comparison, the system showed that NordiCHI'06 had a nearly identical gender make-up to CHI as shown in Figure 5 (although NordiCHI 2008 does show a

abhishek andrés aniket aniket benoît bernheim bongwon bongwon boram brien brynjar chang chao-ju chen chia crysta danaë darshan duen dzimitry elad erum esko eytan françois géry guozhong hao-hua hirokazu hongan hongbin hsin-yen hyunjung ido ing-marie jacky jarke ji jianzhuang jin-ling kandha kari-jouko kayur lishuang lujo lujo malte masatomo michie mwajuma nalini nan-yi naoto neeraja nirmal pei-yu philippas piyou poika poika poika pourang pourang pourang pourang rami rock rosta sageev saleema sasi seoktae shamsi sheizaf shouichi shumeet siân sin-hwa sriram sriram stinne sumit svetlin taemie taowei tapan tek-jin thorsten tuck udai vaishnavi vassilis woohun xiangshi xianhang xiaolong xiaolong xiaoou xinyong ya-lin yedendra yee-yin yiwen yu-chen yuan-chi yuanyuan yuna

Figure 6. Names the genderzyer program did not initially recognize from the CHI 2008 proceedings.

considerable difference.) I also compared first author statistics to nth authors in an analysis I do not present here; there is no significance difference in either CHI or NordiCHI.

There are a few points I would like to make about Figure 4 before continuing. The first is the about the label "initials". Proceedings for CHI'86, CHI'88 and CHI'89 in the ACM DL include only authors' initials, making gender identification difficult or impossible. The other instances of initials are authors who habitually publish with their initials, such as m.c. schraefel. (To increase the number of identified by a name followed by a surname, such as T. Scott Saponas, I took the second word as being the name to determine gender.)

The graph shows a gradual decrease in the percentage of authors' names identifiable as men over the history of the conference. At CHI'85 (the second conference; the first has a large number of initials in the proceedings) approximately 85% of authors were identifiable as male; that figure is approximately 65% in 2006. That decrease is in part attributable to the increase in female authors: as the graph shows, approximately 10% of authors at CHI'85 were female, rising to a high of approximately 20% at CHI'05. What is perhaps more interesting is the rise in the percentage of names identified as "unclear". One aspect of this is names that are, in fact, of ambiguous gender, such as "Pat", which can be short for "Patrick" or "Patricia", or "Leslie". This is further confused by cultural differences: the name "Jean" is male in French but female in British English.

There are other implications for the difficulty of determining gender caused by the increasing geographical and cultural diversity of CHI. For example, Chinese names are often not as strongly gendered as English names. Another aspect of this is that I am somewhat skeptical of the consistency of various sources for representing characters outside of the standard 26 letters of the English alphabet: for example, after adding the baby name lists, the system was still unsure of the (male) name François, perhaps due to the cedilla (ç). As an example, the other names that the system was unable to identify from the CHI 2006 Proceedings are shown in the sidebar, Figure 6.

This suggests, first, that the name lists I was able to find are biased towards names that are more common in the English-speaking world. This also suggested that a distributed solution to this problem would be optimal, relying on the knowledge of many people from many cultures to analyze names. I implemented this by including check boxes next to unknown names to enable users of the system (publicly available at http://jofish.com/genderyzer) to fill in the answers to names with which they were familiar.

Repeat Authorship & Conference Growth

The final analysis I wish to discuss is that of repeat authorship in conferences. This is an interesting longitudinal question because it changes the character of a conference. One could imagine, for example, a conference in which there was a complete turnover of authors every year – perhaps a conference only open to college seniors, for example. The other extreme would be a conference in which the same individuals showed up year after year, perhaps in the manner of a class reunion. Most conferences, of course, fall between

these two extremes. One way to represent this data is as a percentage. For example, we can state that 45% of authors at CHI 2006 had previously published at CHI; on average, 37% of authors at any given CHI had previously published at CHI. A more detailed way to look at this data is to look at the changes over the course of the conference, as shown in Figure 7. It is hard to see the detail in this figure at the size necessary for publication; full versions are available at http://posters.jofish.com. However, some information is apparent even at this scale.

The poster is arranged left to right, with one column per conference from 1983 at the far left to 2007 at the far right. Each column is made up of the names of authors who published at that conference. Authors who published at the first conference, in 1983, are shown in red. If they published at the second conference, then they are still shown in red at the bottom of the second column followed by the authors who first published at the second conference in orange, and so on. The colors repeat every six years (as yellow is left out of the spectrum as it is hard to read); there is no significance to the fact that, for example, authors who first published in 1983 and in 1990, 1996, 2002 and 2008 are all in red.

Within each cohort of new authors, names are in alphabetical order by last name, which makes finding names much easier. In addition, any name only appears once per year, regardless of the number of papers authored.

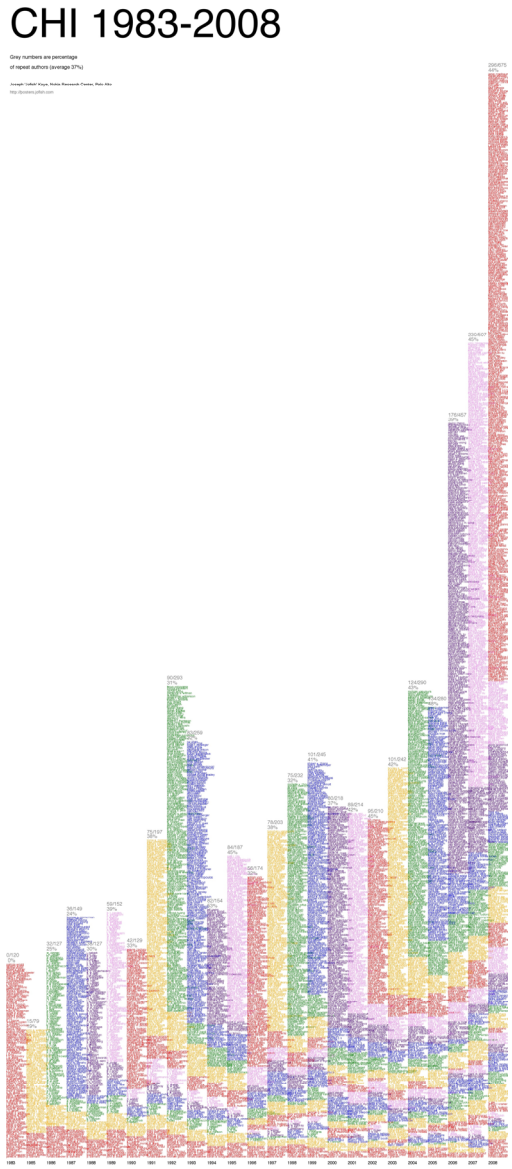

Figure 7. Repeat authorship at CHI.

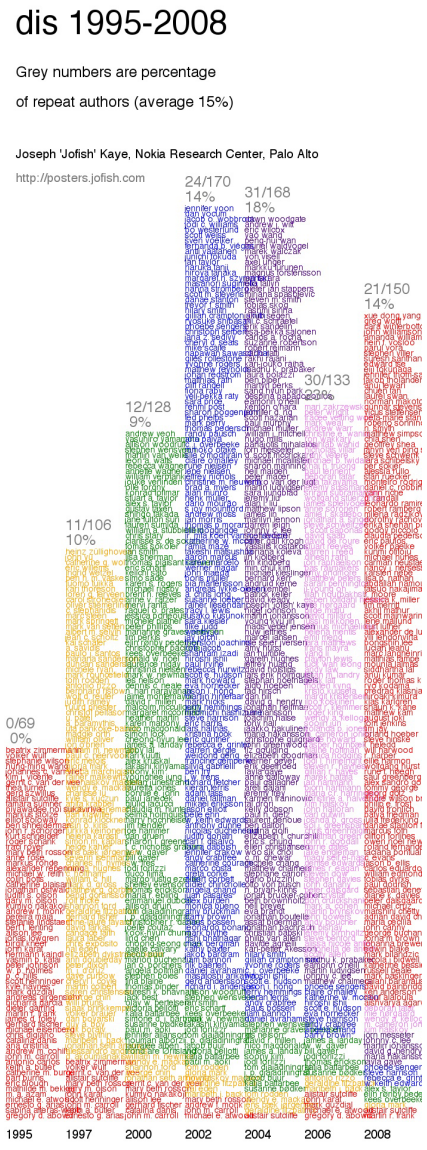

Figure 8: Repeat authorship at DIS

A few other details are also apparent. Perhaps the most significant change is the increase in the size of the conference that took place in 2006. It is also interesting to note the number of authors who have been publishing at CHI for over twenty years: looking at the higher resolution version available on the website, it is apparent that Ronald Baecker, Susan T. Dumais, Kate Ehrlich and 8 more of their colleagues were all authors at both CHI'83 and CHI'07. Perhaps more interesting, however, is to compare the CHI repeat authorship graph to other conferences. For example, the graph of DIS, a much younger conference, shows much lower repeat participation rates (Figure 8), although also a far less drastic increase in the size of the conference as reflected in the number of papers in the proceedings. A similar trend can be noted in the CSCW graph. Other graphs for comparative but not as immediately similar conferences are available on the website, including SIGIR (the ACM Special Interest Group on Information Retrieval) and JCDL (the Joint Conference on Digital Libraries).

These graphs also demonstrate the unique name problem that is common in bibliometric analyses. For example, "William W. Gaver", "William Gaver" and "Bill Gaver" are all the same person. To try and mitigate this problem, name recognition in these graphs was performed on only the last name and the first initial. This was also necessary to provide any correlations to the years during which only initials were recorded in the ACM DL. Of course, it is possible that some names may have been accidentally conflated in this process – if, for example, there was a "Tim Rodden" as well as a

Figure 9: Repeat authorship at CSCW

"Tom Rodden" – but I am yet to locate any instance of this actually happening.[1]

Discussion

I am submitting this paper to alt.chi entirely because I am aware that this is not a standard CHI paper. There is no particular trend connecting the three kinds of information visualized in this paper, and there is no single overarching message except a desire to engage with and discuss the data represented here. However, it is my hope that it can contribute to the ongoing discussion within the field about the nature of the HCI community.

These visualizations raise questions about the way the world is and, thus encourage us to think along these new axes for ways to understand the qualities we wish to encourage in a conference. For example, is the dearth of single-author papers a function of the increasingly collaborative nature of research, or does it represent a form of intellectual discomfort with individual commitment? Should CHI actively strive to have a conference that is more open to female authors as a way to counteract the "incredible shrinking pipeline" of female computer scientists? Is CHI growing so fast as to lose a sense of shared identity? Should the field strive to encourage more repeat authorship to encourage a richer sense of identity? It is my hope that this paper will both contribute to our self-analysis as a discipline and also encourage us to think in more detail about the way we would like our field to become.

[1] To encourage further ongoing critical analysis of this work, I will buy an ice cream at CHI for the first person to identify any given such error in these graphs.

References

I would like to thank my colleagues and the readers of my blog for their thoughtful discussion and analysis of these visualizations.

References

[1] Barkhuus, L., & Rode, J. (2006). From Mice to Men – 24 years of Evaluation in CHI. In alt.chi. Retrieved February 28, 2008, from http://www.viktoria.se/altchi/index.php?action=showsubmission&id=78.

[2] Bartneck, C. and Hu, J. Scientometric Analysis Of The CHI Proceedings. *Proc. CHI 2009.*

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

[4] Grudin, J. (2005). Three faces of human-computer interaction. IEEE Annals of the History of Computing, 27(4), 46-62.

[5] Horn, D, Finholt, T.A., Birnholt, J.P., Motwani, D., and Jayaraman, S. Six degrees of jonathan grudin: a social network analysis of the evolution and impact of CSCW research. *Proc. CSCW 2004* 582-591.

[6] Oulasverta, A. A bibliometric exercise for SIGCHI Conference on Human Factors in Computing Systems http://www.hiit.fi/node/290 2006.

[7] Wania, C.E., Atwood, M. E., and McCain, K.W. How do Design and Evaluation Interrelate in HCI Research? *Proc. DIS 2006* 90-98.