Mind the Gap
Gender and Computer Science Conferences

Sytze Van Herck[1] [0000-0001-5627-042X]
and Antonio Maria Fiscarelli[2] [0000-0003-0287-4388]

[1] Supported by the Luxembourg National Research Fund (FNR) DTU-DHH,
C2DH, University of Luxembourg, 11, Porte des Sciences,
L-4366 Esch-sur-Alzette, Luxembourg
sytze.vanherck@uni.lu,
[2] Supported by the Luxembourg National Research Fund (FNR) DTU-DHH,
C2DH, CSC-ILIAS, University of Luxembourg, 11, Porte des Sciences,
L-4366 Esch-sur-Alzette, Luxembourg
antonio.fiscarelli@uni.lu

Abstract. Computer science research areas are often arbitrarily defined by researchers themselves based on their own opinions or conference rankings. First, we aim to classify conferences in computer science in an automated and objective way based on topic modelling. We then study the topic relatedness of research areas to identify isolated disciplinary silos and clusters that display more interdisciplinarity and collaboration. Furthermore, we compare career length, publication growth rate and collaboration patterns for men and women in these research areas.

Keywords: Gender, Bibliometrics, Topic Modelling

1 Introduction

Diane Jackaki described the feeling of most Digital Humanists in her keynote "Jack of all trades, master of One" at the Digital Humanities Summer School of 2017 in Oxford [1]. Even in interdisciplinary areas of research, people still cling to their disciplinary backgrounds. In our title "Mind the Gap" we refer to the continuing split between disciplinary areas, even when interdisciplinarity is actively encouraged in many fields. We also discuss the gender gap and the consistently low number of female researchers in computer science. An example of the disciplinary schism can be found in cryptography where researchers attend and publish at conferences such as the International Cryptology Conference (CRYPTO). A title such as "New and Improved Key-Homomorphic Pseudorandom Functions" [2] already suggests that cryptography research can be very specific. Our aim is to identify disciplinary areas in Computer Science (CS) through topic modelling and study the interdisciplinary overlap between these research 'silos'. Furthermore we look into collaboration patterns and in particular gender distribution for each research area [3]. The example from cryptography illustrates our assumption that conference papers published throughout the years can be grouped to represent a CS area.

2 Topic modelling

The main idea is to define a topic (or CS area) as a set of conferences and each conference as the set of all papers that have been published. A paper is defined as the set of keywords contained in its title. In the next step, similar conferences are grouped together in CS areas using text mining. Other examples of automated topic modelling are provided in [4, 5] where a probabilistic model for topic modelling is proposed. On the other hand, Biryukov and Dong [6] arbitrarily define fourteen different CS subareas based on top ranked conferences, while in [7] the CS areas are defined according to experts' opinion and the conferences' impact ratings.

2.1 Dataset

The dataset is publicly available at `https://data.mendeley.com/datasets/3p9w84t5mr/1` and contains 148512 papers published at 81 different CS conferences since 1960 [8]. We limited the data subset to papers published between 2000 and 2015 since the DBLP bibliographic library had a narrow focus at the start which broadened over the years to include most CS conferences [9]. By 2005 DBLP covered 65% of conferences from the aggregated list Reitz and Hoffmann created [9]. As a result of our selection the conferences CCC, Digital Libraries, EuroCOLT, ISTCS, MFDBS and PDIS were excluded either because they changed names or ended before 2000, leaving us with 104680 papers and 75 conferences.

2.2 Data retrieval

The dataset contains the URL associated to each paper instead of the full title. Therefore we used a scraping tool to retrieve the title of a paper based on the URL and the associated xpath selector. About 19% of the papers did not contain any URL and we simply removed them. We discovered that all the other papers were published on 4 main websites: the AAAI conference website (`https://www.aaai.org/ocs/`), the Springer website (`https://link.springer.com`), the Computer Science Digital Library of the IEEE Society `https://www.computer.org/csdl` and the ACM Digital Library `https://dl.acm.org/`. Making use of the R package "rvest" we were able to scrape the Springer website and the Computer Science Digital Library of the IEEE Society, collecting data for about 50% of the papers and 51 conferences. Unfortunately, the AAAI conference website and the ACM library make use of dynamic content that rendered the scraping process impossible. These two websites contain the remaining 50% of papers and 24 conferences. ACM kindly provided us with a collection of all proceedings published in their digital library. The collection consists of XML files related to a specific conference proceeding of a certain year and contain information about the conference itself and all papers published afterwards. The title of each paper can be easily retrieved using its XML tag. Considering only conferences held between 2000 and 2015, we were able to collect data related to 23 conferences.

The ACM library did not contain any proceedings for the IEEE International Conference on Computer Communications (INFOCOM). In the end we were able to retrieve data for 74 conferences.

2.3 Text mining

The corpus contains a document for each conference, and a document lists the keywords of all papers published at the conference. We used text mining methods to process the data and extract CS areas, but only after cleaning up the data. For example, we converted each letter to lower case, removed numbers, punctuation, stopwords (words that do not contain any significance such as articles and prepositions) and performed stemming by reducing the words to roots. The R package "tm" provides all these functionalities and a collection of stopwords for the English language. At the end of this phase, the corpus contained 73 documents with 21158 distinct terms. We then built a term frequency-inverse document frequency (tf-idf) matrix, where the number or rows is equal to the number of documents in the corpus and the number of columns is equal to the number of terms in the corpus. Each entry of the matrix is computed as follows:

$$w_{ij} = tf_i \times \log \frac{N}{df_i} \tag{1}$$

where tf_i is the frequency of term j in document i, N is the number of documents in the corpus and df_i is the number of documents that contain the term j. Each row entry of the tf-idf matrix represents a document in the corpus, where each element represents the relevance of a certain word in the document.

2.4 The clustering algorithm

We used a clustering algorithm to group conferences and define CS areas [10]. This specific algorithm requires a similarity matrix. Therefore we used the cosine measure to compare documents, defined as follows:

$$sim(d_i, d_j) = \frac{\sum_{k=1}^{n} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2} \sqrt{\sum_{k=1}^{n} w_{jk}^2}} \tag{2}$$

Based on this measure we can define a similarity matrix S, a square matrix whose number of rows and columns are equal to the number of documents in the corpus. Each entry of the matrix defines how similar two documents are, where $S_{ij} = 0$ indicates that the documents do not have any term in common and $S_{ij} = 1$ indicates that the two documents consist of the same set of terms. After creating the similarity matrix, we can introduce a clustering algorithm. We selected the Degenerate Agglomerative Hierarchical Clustering Algorithm (DAHCA) for several reasons. First it is hierarchical and agglomerative, meaning that it can recognize hierarchies of clusters from a large number of very specific clusters to a small number of generic clusters. In addition, unlike other hierarchical clustering algorithms, it allows the merging and formation of clusters of different sizes.

2.5 Computer Science Areas

As a result of the DAHCA, we ended up with a four-level hierarchy containing 34, 19, 8 or 3 clusters. The first level of 34 clusters was too specific, while the last level of 3 clusters was too generic. In the end we decided to visualise the third level containing 8 clusters because it would be most readable and meaningful. A part of the hierarchy is shown in table 2, where we assigned a meaningful name to the eight clusters chosen from the second level. Based on the conference titles and keywords, the clusters were named: Databases and Information Systems (DBIS), Knowledge Engineering (KE), Software Engineering (SE), Artificial Intelligence (AI), Interdisciplinary/Networks/Web (I/N/W), Cryptography (CRYPTO), Theoretical Computer Science/Concurrency (TCS/C), and Algorithmic Theory (ALGO). Table 1 contains the the most relevant words for each CS area. More specifically, SE includes conferences on software engineering, software maintenance and code analysis. AI falls into four categories: learning theory, machine learning, optimisation and computer vision. I/N/W includes interdisciplinary conferences on human computer interaction, computational biology and computational linguistics, as well as conferences on networks and the web. CRYPTO focuses on cryptography and software encryption. TCS/S treats logic, theoretical computer science, and concurrent and parallel systems. Finally, ALGO contains algorithms and theoretical computer science. In table 2, the CRYPTO cluster is identified at the first level and remains separate at the next levels, demonstrating how how particular and specific cryptography conferences are. Furthermore, the cluster I/N/W contains very diverging topics ranging from human-computer interaction to world wide web, but these topics might be grouped because of a similar application or theoretical foundation.

Table 1: Top ten most relevant keywords in each CS area

CS Area	Keywords
DBIS	"retriev", "entiti", "topic", "xml", "recommend", "relev", "queri", "learn", "databas", "text"
KE	"conceptu", "prefac", "ontolog", "schema", "busi", "uml", "olap", "warehous", "xml", "semant"
SE	"softwar", "mainten", "workshop", "engin", "refactor", "revers", "reengin", "evolut", "legaci", "comprehens"
AI	"imag", "stereo", "recognit", "scene", "motion", "pose", "segment", "camera", "face", "video"
I/N/W	"poster", "genom", "gene", "protein", "wireless", "transcript", "demo", "regulatori", "acl", "ancestr"
CRYPTO	"cryptanalysi", "cipher", "encrypt", "attack", "secur", "signatur", "rsa", "multiacparti", "zeroacknowledg", "relatedackey"
TCS/C	"announc", "brief", "logic", "automata", "bisimul", "firstacord", "azaiaalqcalculus", "quantum", "schedul", "concurr"
ALGO	"automata", "approxim", "planar", "problem", "quantum", "bound", "algorithm", "game", "graph", "minimum"

Table 2: List of conferences included in each cluster. The outer left column contains the eight cluster names and the central column contains the conferences included in each cluster. The outer right column contains the same conferences but they are listed in different rows that represents the cluster in the previous (more specific) level for a total of 19 clusters.

DBIS	ADBIS CIKM SIGMOD DASFAA ICDE DBPL ICDT PODS EDBT SSDBM	ADBIS CIKM SIGMOD DASFAA ICDE DBPL ICDT PODS EDBT SSDBM
KE	CoopIS ER KDD UIST RIDE DOLAP	CoopIS ER KDD UIST RIDE DOLAP
SE	CSMR ICSM ASE WCRE FASE ICSE ICPC SCAM MSR	CSMR ICSM ASE WCRE FASE ICSE ICPC SCAM MSR
AI	ALT COLT ICML CADE PLDI POPL IPCO SOCG SIGMETRICS SOSP STOC CVPR ICCV SIGCOMM LFCS DISC/WDAG ISSAC	ALT COLT ICML CADE PLDI POPL IPCO SOCG SIGMETRICS SOSP STOC CVPR ICCV SIGCOMM LFCS DISC/WDAG ISSAC
I/N/W	CHI IGPCE RECOMB MOBICOM OSDI IJCAI WWW PODC SIGIR VLDB NSDI AAAIDEA ACL WIDM	CHI IGPCE RECOMB MOBICOM OSDI IJCAI WWW PODC SIGIR VLDB NSDI AAAIDEA ACL WIDM
CRYPTO	CRYPTO EUROCRYPT FSE	CRYPTO EUROCRYPT FSE
TCS/C	CONCUR LICS FSTTCS SP SPAA SWAT WADS	CONCUR LICS FSTTCS SP SPAA SWAT WADS
ALGO	ESA ICALP ISAAC SODA FOCS MFCS LATIN STACS	ESA ICALP ISAAC SODA FOCS MFCS LATIN STACS

3 Visualisations

After outlining the creation of the eight clusters, we will interpret some of our results in this section. Several visualisations serve both as a tool for analysis and as a communication method. First of all, we study topic relatedness to understand the interdisciplinarity and overlap between different research areas based on a visualisation created in R. Secondly, we create several more visualisations to analyse gender in relation to career length, publication growth and collaboration. For these visualisations we generated the data source through SQL queries executed per cluster. We ran all queries using Python 3's pymysql library to connect to the MySQL database and combined the results of each cluster in a single csv file. The final visualisations were created in Tableau and allow interactions such as details-on-demand and filtering.

3.1 Topic relatedness

Authors do not always publish at the same conference and especially in interdisciplinary areas of CS, authors tend to collaborate more often and publish in different conferences. Figure 1 displays how different areas of CS relate to each

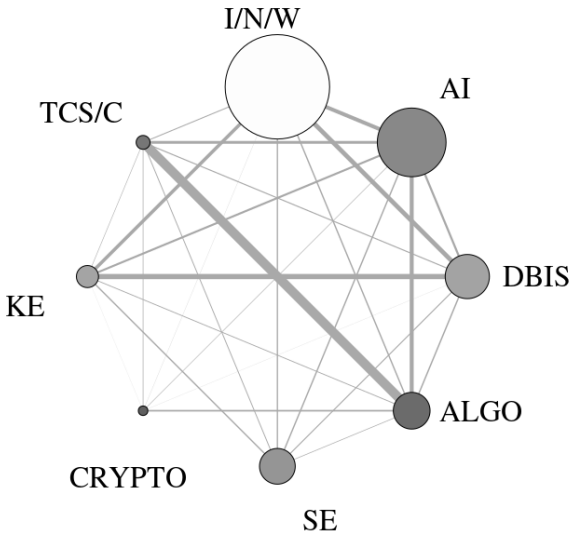

Fig. 1: Topic relatedness graph for eight conference clusters linked based on the percentage of authors who published in both research areas.

other and to what degree authors publish in different CS areas. We first built a topic relatedness graph where each node represents a CS area and edges represent how related they are to each other. The node size indicates the number of papers published and different shades of gray indicate the average number of authors per paper. Light shades of gray represent a high number of authors, while dark shades of gray indicate a low number of authors. The edge thickness is based on the percentage of authors that have published at least one paper in two linked CS areas. Even though all areas are connected, when an edge is not visible this indicates a very loose connection.
As you can deduce from the visualisation, CRYPTO contains the lowest number of papers and the lowest average number of authors per paper. The CRYPTO cluster therefore represents a niche area where researchers are less inclined to collaborate, probably because their research is very specific as we assumed initially. TCS/C and ALGO also show similar results. As a complete opposite, I/N/W contains a much higher number of papers and average number of authors per paper than any other area. In fact, it contains conferences about human-computer interaction, bioinformatics and computational linguistics; all areas which promote collaboration between researchers. If you then look at the relatedness of different areas, CRYPTO is the least related to other CS areas. This isolation of CRYPTO indicates that researchers specialised in cryptography do not publish in any other research area. SE is also rather isolated, yet it closely links to six other research areas, whereas CRYPTO only shows some links to three other areas. ALGO has the highest general relatedness, probably because algorithmic theory is the foundation of many CS areas. The areas that show the highest relatedness and share the highest numbers of authors are TCS/C and ALGO, followed by KE and DBIS.

3.2 Career Length

We should first define gender and explain how this information was generated in the dataset. Gender as a psychological and sociological term originated in the United States and refers to "the state of being male or female as expressed by social or cultural distinctions and differences, rather than biological ones; the collective attributes of traits associated with a particular sex: or determined as a result of one's sex" [11]. Even in this definition a binaristic understanding of gender is implied, where a person can only identify as being male or female. The same binary distinction occurs in the dataset we use. The gender of an author was determined based on their first name using the Genderize API [12] which returns either"Male, Female or NA" for 'not assigned' based on a list of identified first names, returning a confidence score for each result. The Genderize API performs relatively better than other methods inferring gender from names on the web, yet this approach mostly works well for western industrialized countries [13]. Agarwal et al. also decided to assign "NA" to any author name with a confidence score below 60% [14]. In general, the gender was not determined for 14,2% of authors, while 69,1% of authors were identified as male, and only 16,7% were identified as female [14]. Science careers often show the trend of a leaky pipeline,

where people drop out at various segments of their career and women drop out more often than men as their career progresses [15]. Based on this finding in literature we decided to compare the career length for men and women within the different clusters. The career length represents the period between the first and last year that an author published at a conference included in the dataset.

Fig. 2: Average career length of female and male authors in years for each cluster. The clusters are numbered as follows: (1) DBIS, (2) AI, (3) I/N/W, (4) TCS/C, (5) KE, (6) CRYPTO, (7) SE and (8) ALGO.

Figure 2 represents the average career length of men and women in the eight fields of research we defined earlier. In general the average career length for women is lower than for men, which confirms the leaky pipeline issue even for the brief span of fifteen years. The biggest difference in career length between men and women can be found in CRYPTO, the research area with the second lowest average percentage of female authors (at just 12,6%). On the other hand there is almost no difference in the career length of male and female researchers for the KE and SE research areas, which is generally shorter than in other research areas. We assume that this short career length is related to the rapidly changing nature of these research areas. The longest career length occurs in the ALGO cluster, probably because this is a research discipline that remains fairly stable and discipline specific.

3.3 Publication Growth Rate

The publication growth rate (PGR) indicates the growing or diminishing popularity of a research area for each cluster over the course of sixteen years. We look into both the absolute and the relative publication growth for male and female authored papers in each cluster to assess whether publications authored by women grew at a faster rate than those of men. Finally, we compare publication growth of two theoretical and disciplinary research areas (TCS/C and ALGO) to the PGR of rather interdisciplinary clusters (DBIS, I/N/W and KE).

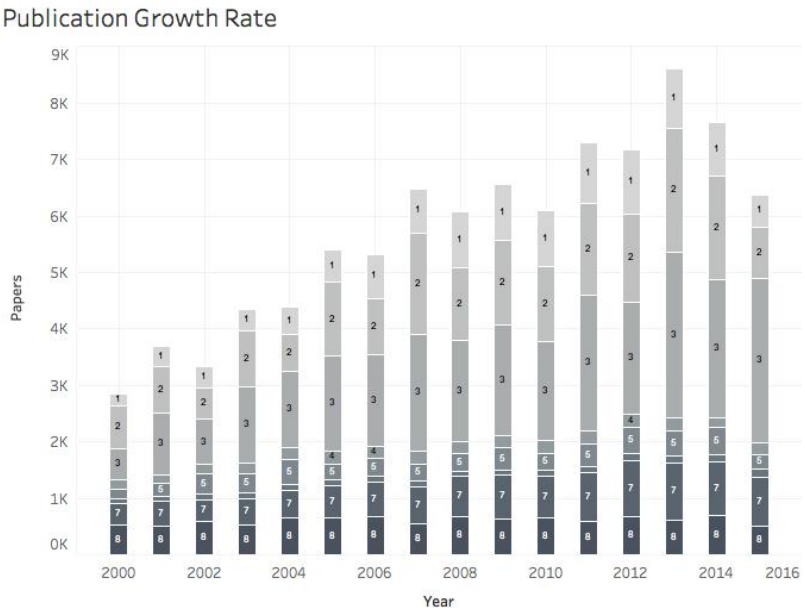

Fig. 3: Publication growth rate per year for each cluster. The clusters are numbered as follows: (1) DBIS, (2) AI, (3) I/N/W, (4) TCS/C, (5) KE, (6) CRYPTO, (7) SE and (8) ALGO.

In figure 3 the largest research area in terms of the number of publications is the I/N/W cluster, whereas CRYPTO contains the lowest number of publications overall. Table 2 shows that CRYPTO only contains three conferences in a single cluster. That same explanation does not hold true for I/N/W because this research area contains fourteen conferences, compared to seventeen conferences in the AI cluster. When we study the I/N/W cluster in detail however, it becomes clear that the Computer-Human Interaction (CHI) conference causes the high number of publications for this research area (see figure 6).

In order to compare relative growth rates, we calculated the growth rate comparing the number of papers in 2014 to the number of papers in 2000. Since the dataset uses a snapshot from the DBLP Dataset on September 15th of 2015, we decided to calculate the growth rate based on the last full year contained in the dataset.

$$PubGrowRate_{i,y2,y1} = \frac{P_{i,y2} - P_{i,y1}}{P_{i,y2} + P_{i,y1}} \tag{3}$$

where $P_{i,y}$ is the number of papers published during the year y

In figure 4 we decided to study the publication growth for each research area split by gender. We simply counted the number of papers written by either female or male authors and left out the unidentified authors entirely. Some papers were counted twice in case they were written by a man and a woman so the total number of papers does not equal the sum of male and female authored papers. The I/N/W cluster has the highest number of both male- and female-authored papers, with 2322 papers authored by women in 2013 compared to 6452 papers authored by men in 2015. The lowest number of male- and female-authored papers on the other hand can be found in CRYPTO with only 20 papers written

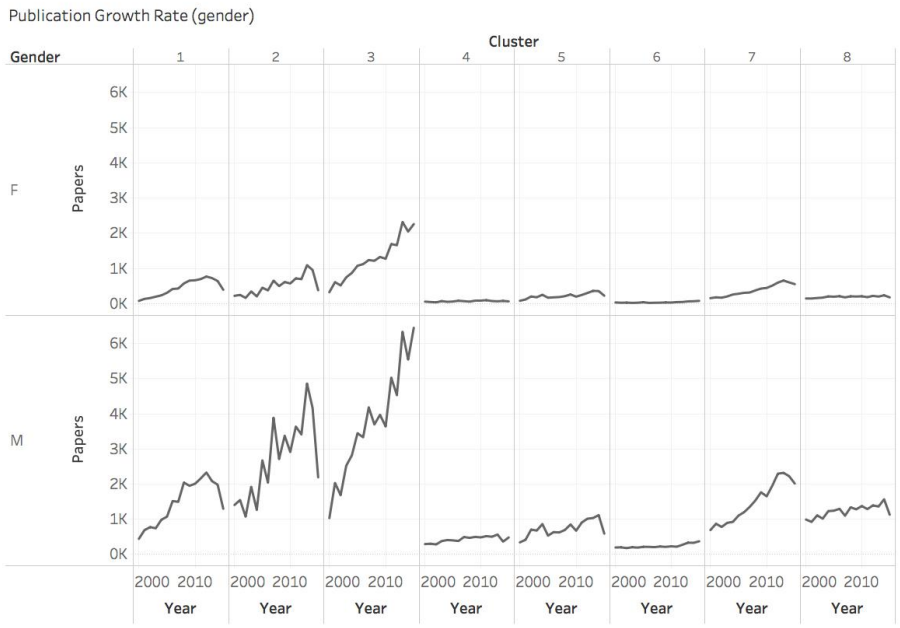

Fig. 4: Publication growth rate for female and male authors per year for each cluster. The clusters are numbered as follows: (1) DBIS, (2) AI, (3) I/N/W, (4) TCS/C, (5) KE, (6) CRYPTO, (7) SE and (8) ALGO.

by women in 2003, and 168 papers written by men in 2002. Even though the highest and lowest number of papers can be found in the same clusters, women consistently authored fewer papers than their male counterparts in the dataset. When we look at the relative PGR however, the rise in the number of female-authored publications is higher for all clusters. The largest difference in PGR for men and women can be found in the DBIS research area where female-authored papers grew by 78,6% and male-authored papers by only 64,1%. In ALGO on the other hand the difference in PGR for men and women was lowest, with 23,6% for women and 22,7% for men. The rather interdisciplinary clusters such as DBIS, I/N/W and KE showed the highest relative PGR with 62,5%, 63,3% and 49,9% respectively. The theoretical and disciplinary conference clusters such as TCS/C (0,3%) and ALGO (14%) on the other hand showed the lowest relative PGR. The smallest growth rate of 0,3% for TCS/C might be due to a strict limitation of the number of papers accepted for conferences in this CS research area.
Figure 5 displays two theoretical clusters (TCS/C and ALGO) on conference

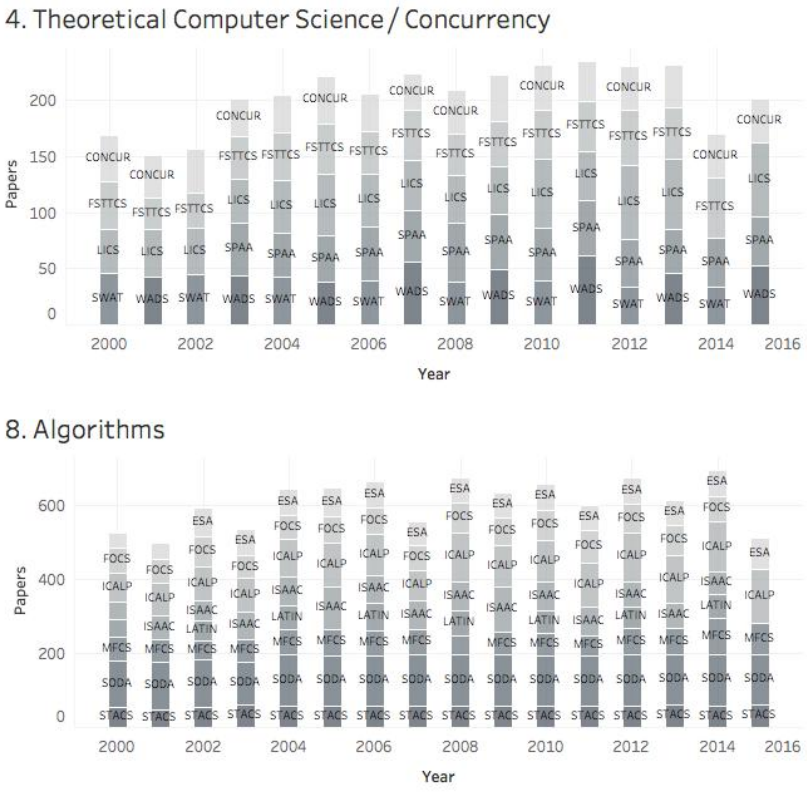

Fig. 5: Comparison of the publication growth rate per year of TCS/C and ALGO.

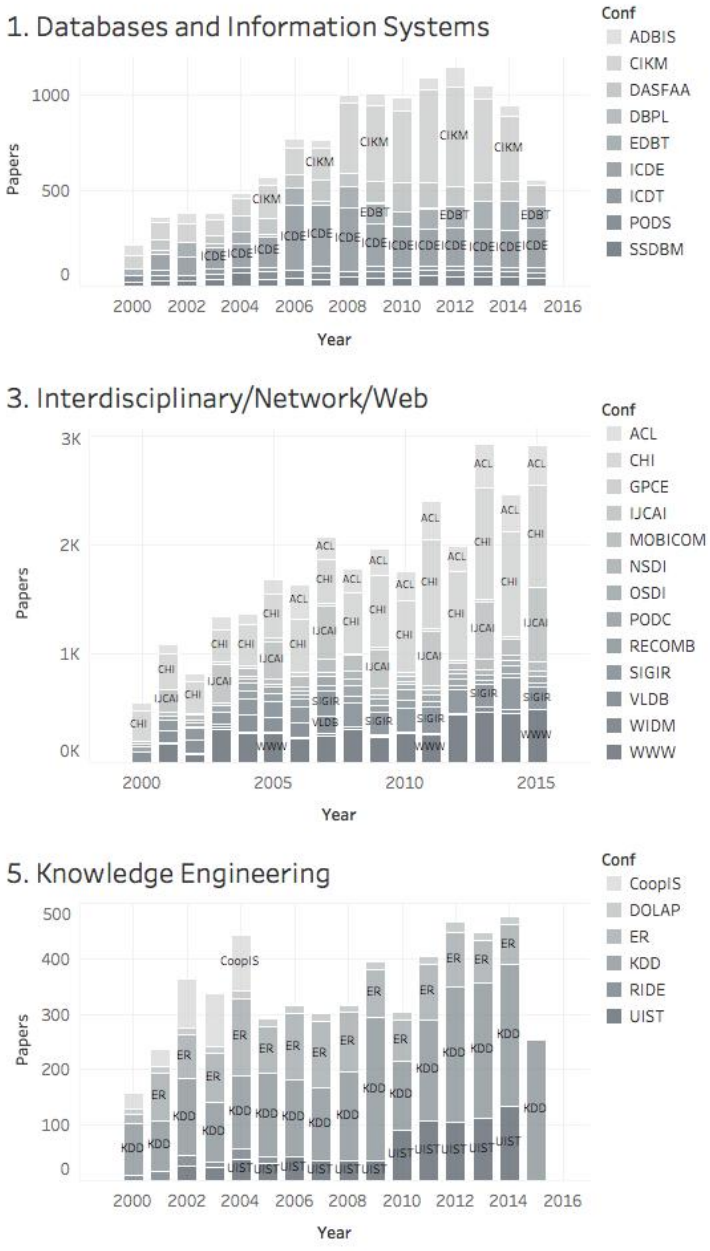

Fig. 6: Comparison of the publication growth rate per year of DBIS, I/N/W and KE.

level. Both TCS/C and ALGO show little variation in the number of papers accepted for each single conference. Even so, some conferences are considerably more popular (in terms of PGR) than others. In the TCS/C cluster for example, the International Conference on Concurrency Theory Symposium on Logic in Computer Science (LICS) has a PGR of 22,8% between 2000 and 2013, whereas the Scandinavian Symposium and Workshops on Algorithm Theory (SWAT) dropped 16,5% in PGR between 2000 and 2014. In the ALGO research area the International Colloquium on Automata, Languages, and Programming (ICALP) gained 27,8% of publications, compared to an increase of only 1,5% in publications for the IEEE Symposium on Foundations of Computer Science (FOCS). In addition, conferences in these research areas publish a relatively low number of papers varying between 28 and 66 papers in the TCS/C cluster, and 40 to 139 publications in the ALGO cluster.
The interdisciplinary research areas of clusters such as DBIS, I/N/W and KE displayed in figure 6 all contain at least one conference with a steep publication growth rate: from 66,7% for the Conference on Information and Knowledge Management (CIKM) in DBIS, over 69,9% for Mobile Computing and Networking (MOBICOM) in I/N/W, to 62,3% for the ACM User Interface Software and Technology Symposium (UIST) in KE. The conferences with the highest number of publications are the Conference on Knowledge Discovery and Data Mining (KDD) with 296 papers published in 2014 for KE, 1028 papers published in 2013 for the CHI conference in the I/N/W cluster, and CIKM with 520 papers in 2012 within DBIS. In spite of the overall high PGR in these interdisciplinary research areas, some annual conferences ended before 2015.
Although the disciplinary clusters TCS/C and ALGO display the smallest difference in PGR based on gender, the opposite is not entirely true for the interdisciplinary clusters DBIS, I/N/W and KE. Both DBIS and KE are part of the top three research areas where the difference between PGR for men and women is highest, yet I/N/W finished third to last with a difference of only 3,8% PGR between men and women. However, if we sort the clusters based on the relative PGR for female-authored papers, the distinction between disciplinary and interdisciplinary becomes clear. Disciplinary conferences have the lowest PGR for female-authored papers with 14,3% (TCS/S) and 23,6% (ALGO) respectively, whereas interdisciplinary conferences show a clear trend towards a higher percentage of female-authored papers every year including increases of 63,4% (KE), 72,6% (I/N/W) and even 78,6% (DBIS).

3.4 Collaboration Patterns

Academic researchers are often evaluated, funded and hired based on their publications. Whether or not they collaborate with other researchers makes a difference in the value that is attributed to their contributions. Furthermore, when co-authoring, the position of a researcher in the list of authors is often considered a good indication of exactly how much of the work they performed. Given the limited scope of this research paper, we cannot look into the author's position but we do have information regarding the percentage of female authors

per paper and the number of authors they collaborated with. We would also like to test our assumption that collaboration occurs more often at interdisciplinary conferences and whether or not this makes a difference in gender balance for single papers. Conferences present the composition of an academic field of research on an international level and form the ideal case to study gender balance. This distribution of men and women in research institutions and specific areas of research has become increasingly important to obtain government funding.
We decided to visualize the gender balance within the larger clusters and at specific conferences in the tree map format introduced by Ben Schneiderman [16]. Tree maps convey information through both area and colour, which allows researchers to locate outliers and identify cause-effect relationships [17]. Because we wanted to display as many variables as possible in a limited space, we decided to use tree map visualisations. In order to interpret tree maps correctly, we will first explain the variables assigned to area size, color and clustering. In figure 7 the larger groupings represent the clusters, whereas the divisions within the cluster visualize a specific year. The size of each rectangle represents the number of papers in a given year. The clusters are ordered left to right and top to bottom starting from the largest to the smallest number of papers and within a cluster the years are ordered in the same way. The color represents the percentage of female authors and varies from 7,9% to 21,5% yet to allow comparison with figure 8 the color scale ranges between 0% and 35% from light to dark gray. The tree map comparing the conferences with the highest and lowest percentage of female

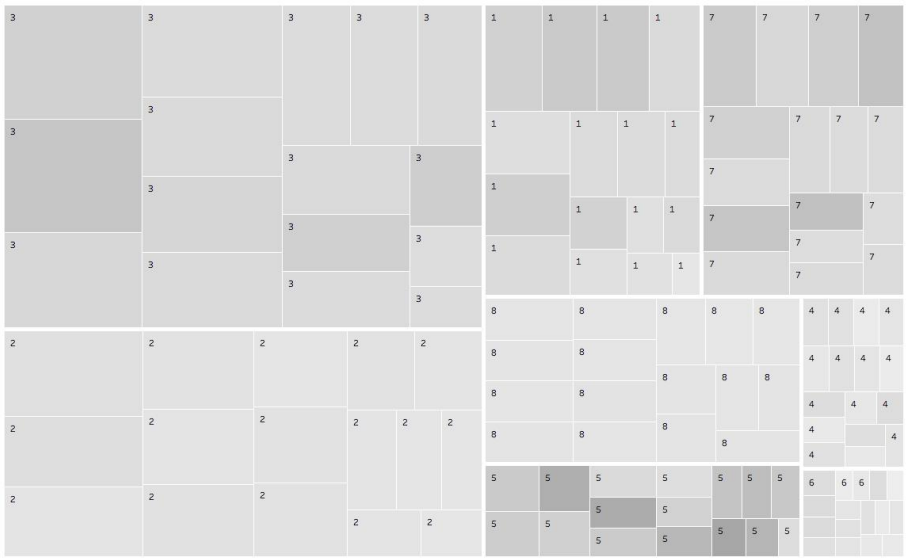

Fig. 7: Average percentage of female authors per paper grouped by cluster and split by years. The clusters are numbered as follows: (1) DBIS, (2) AI, (3) I/N/W, (4) TCS/C, (5) KE, (6) CRYPTO, (7) SE and (8) ALGO.

authors are structured in the same way, but the highest clustering represents a single research area and at the second level each conference is grouped together, split into specific conference years. Tableau also enables details-on-demand, so when you hover over a single data entry, the conference acronym, conference year, percentage of female authors and total number of papers are displayed.
In figure 7 the fifth conference cluster (middle cluster at the bottom) represents KE and contains the highest percentage of female authors overall. CRYPTO (bottom right cluster) on the other hand shows the lowest percentage of female authors. The larger the number of papers in a cluster, the higher the percentage of female authors (with the exception of KE). The clusters we labeled interdisciplinary in our previous sections (DBIS, I/N/W, KE) and SE all contain a higher percentage of female contributors, whereas the disciplinary clusters (TCS/C, ALGO) and AI have fewer female authors overall. In KE the years with the highest participation of women with more than 20% female-authored papers include 2001 (20,3%), 2006 (20,1%), 2009 (21,1%), 2013 (20,8%) and 2015 (21,5%). The CRYPTO research area contains fewer than 10% female authors in conference years 2003 (8,9%), 2006 (7,9%) and 2007 (9,97%) and is the only conference

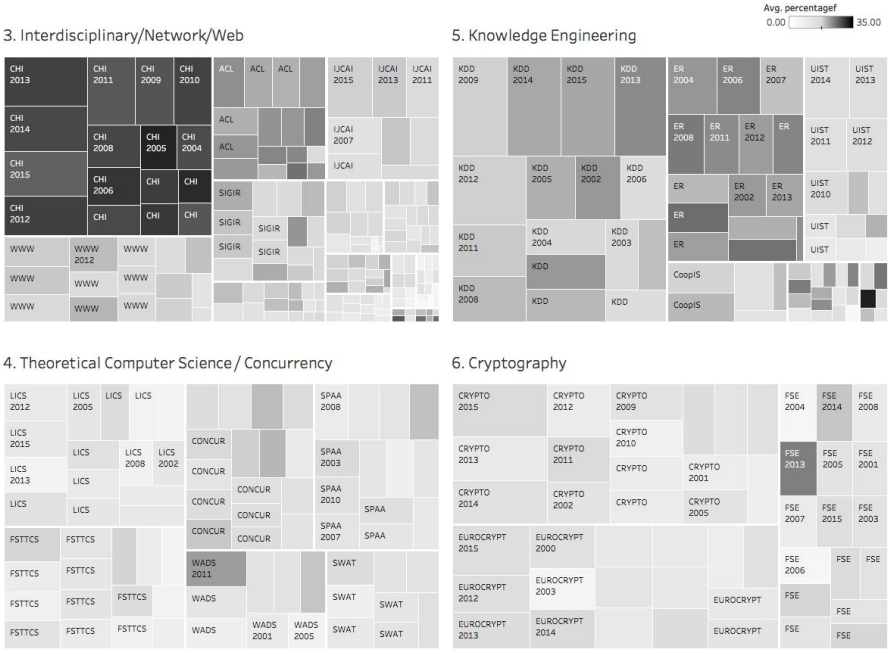

Fig. 8: Comparison of the gender distribution in four research areas split into single conferences. Top: I/N/W and KE with the highest average percentage of female authors per paper. Bottom: TCS/C and CRYPTO with the lowest average percentage of female authors per paper.

cluster including fewer than 10% female authors. Higher percentages of female authors did therefore not necessarily occur at the end of the period studied here. In order to gain more insight into the gender balance at disciplinary and interdisciplinary conference clusters, we will now compare the interdisciplinary I/N/W and KE to the disciplinary TCS/C and CRYPTO in figure 8.
If we look at figure 8 certain conferences have a higher percentage of female authors than others even within the same research area. The highest percentage of female authors overall can be found at the CHI conference with over 30% of female-authored papers for four years, including 2001 (30,0%), 2002 (31,6%), 2005 (32,21%) and 2006 (30,7%). Besides the Association for Computational Linguistics (ACL) and WIDM conference, very few other conferences within I/N/W contain more than 20% of female-authored publications. Within KE both the ACM User Interface Software and Technology Symposium (UIST) and CoopIS conference contain fewer than 20% female authors for most years, with the exception of 20,6% female authors at UIST in 2002, and 20,07% of female-authored papers at CoopIS in 2003. In the disciplinary clusters the lowest percentage of female authors can be found in the CRYPTO cluster at EUROCRYPT in 2003 (4,6%) and at FSE in 2006 (4,4%) although FSE also contains the highest percentage of female authors within CRYPTO in 2013 (24,7%). In TCS/C the WADS conference in 2011 has an exceptionally high percentage of female authors at 22,4%. This comparison demonstrates that the gender balance for an entire cluster does not account for high differences in female authorship at the conference level. Although the CHI conference in the I/N/W research area contains the highest percentage of female authors, not all other I/N/W conferences have such a high level of female participation. On the other hand, the disciplinary research areas also showcase some outliers where they reach the 20% female-authored publications threshold. One explanation for a higher female collaboration rate in interdisciplinary fields could stem from the assertion that women are more attracted to research areas that emphasise social issues, which perhaps appear more often in interdisciplinary research [18].

4 Conclusion

In literature on bibliometrics such as [6, 7] we discovered manual categorisation of conferences in clusters. In order to avoid this rather subjective approach, we decided to first define conferences based on the text mining of paper titles, secondly we grouped conferences using a clustering algorithm and thirdly we manually assigned names to the clusters based on the conference titles. The algorithm provided four levels of clustering and we decided to use a grouping of eight clusters at the basis of our visualisations.
The topic relatedness graph illustrated the isolation of CRYPTO and lower levels of collaboration in CS areas such as CRYPTO, TCS/C and ALGO. In I/N/W on the other hand we discovered more collaboration between authors. In general the career length for women was lower than for men, illustrating the leaky pipeline. In both engineering clusters (KE and SE) authors had a shorter career length

in comparison to the longest career length in ALGO, probably because ALGO remains more stable whereas KE and SE depend on rapidly changing technology. Studying the publication growth rate we found that interdisciplinary research areas such as DBIS, I/N/W and KE had a faster publication growth than disciplinary research areas such as TCS/C and ALGO. Furthermore, the publication growth rate for female authors is positively correlated with the interdisciplinarity of a research area, whereas the difference between male and female publication growth rate is less indicative. When we then look at collaboration patterns, all interdisciplinary clusters mentioned previously as well as SE contain a higher percentage of female authors than the disciplinary clusters. However, the breakdown of clusters into individual conferences shows that in some cases the higher percentage of female authors was caused by a single conference (such as CHI). Overall the definition of a research area as disciplinary or interdisciplinary lies at 20% female authors.

First of all, we would like to point out that the overall percentage of female authors is low in computer science. The percentage of female PhD students in CS lies at a similar percentage with 22,1% in 2004 and a decrease to 20,8% in 2014 [19]. When we split research areas into disciplinary and interdisciplinary fields, we do find that female authors are mostly concentrated in interdisciplinary research areas. This begs two related but contradictory questions: Why are women more attracted to interdisciplinary fields of research? Or, why do interdisciplinary conferences publish more female-authored papers? Although we cannot answer these questions here, we do think it is in every conference organisers best interest to encourage both female authors and collaboration. Janet Abbate points out that women should receive the same respect, mentoring and help with research, grants and publications as their male peers [20]. Furthermore, conference organisers should provide accommodations such as childcare at conferences [20]. Most importantly, women should not be isolated or marginalized, but rather included and represented as keynote speakers, editorial board members and conference organisers [20].

References

1. Jakacki, D.: Jack of all trades, master of one: the promise of intermethodological collaboration. https://player.fm/series/digital-humanities-at-oxford-summer-school/2017-opening-keynote-jack-of-all-trades-master-of-one-the-promise-of-intermethodological-collaboration (2017) [Online; accessed January 11, 2018].
2. Banerjee, A., Peikert, C.: New and improved key-homomorphic pseudorandom functions. In: International Cryptology Conference, Springer (2014) 353–370
3. Van Herck, S.: Visualising gender balance: Ten computer science conferences and the digital humanities conference compared (2017)
4. Paul, M., Girju, R.: Topic modeling of research fields: An interdisciplinary perspective. In: Proceedings of the International Conference RANLP-2009. (2009) 337–342
5. Daud, A., Li, J., Zhou, L., Muhammad, F.: Conference mining via generalized topic modeling. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer (2009) 244–259

6. Biryukov, M., Dong, C.: Analysis of computer science communities based on dblp. Research and advanced technology for digital libraries (2010) 228–235
7. Bird, C., Devanbu, P., Barr, E., Filkov, V., Nash, A., Su, Z.: Structure and dynamics of research collaboration in computer science. In: Proceedings of the 2009 SIAM International Conference on Data Mining, SIAM (2009) 826–837
8. Swati, A., Ashish, S., Nitish, M., Rohan, K., Denzil, C.: Dblp records and entries for key computer science conferences. https://data.mendeley.com/datasets/3p9w84t5mr/ (2017) [Online; accessed January 11, 2018].
9. Reitz, F., Hoffmann, O.: An analysis of the evolving coverage of computer science sub-fields in the dblp digital library. In: International Conference on Theory and Practice of Digital Libraries, Springer (2010) 216–227
10. Fiscarelli, A.M., Beliakov, A., Konchenko, S., Bouvry, P.: A degenerate agglomerative hierarchical clustering algorithm for community detection. In: Asian Conference on Intelligent Information and Database Systems, Springer (2018) 234–242
11. Gender, n.: OED Online. Oxford English Dictionary (2011) [Online; accessed January 24, 2018].
12. GenderizeAPI. https://genderize.io [Online; accessed January 24, 2018].
13. Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., Strohmaier, M.: Inferring gender from names on the web: A comparative evaluation of gender detection methods. CoRR **abs/1603.04322** (2016)
14. Agarwal, S., Mittal, N., Katyal, R., Sureka, A., Correa, D.: Women in computer science research: what is the bibliography data telling us? ACM SIGCAS Computers and Society **46**(1) (2016) 7–19
15. Sonnert, G., Holton, G.: Who Succeeds in Science?: The Gender Dimension. Rutgers University Press (1995)
16. Shneiderman, B.: Tree visualization with tree-maps: 2-d space-filling approach. ACM Trans. Graph. **11**(1) (January 1992) 92–99
17. Asahi, T., Turo, D., Shneiderman, B.: Visual decision-making: Using treemaps for the analytic hierarchy process. In: Conference Companion on Human Factors in Computing Systems. CHI '95, New York, NY, USA, ACM (1995) 405–406
18. Henwood, F., Miller, N., Senker, P., Wyatt, S.: Technology and In/equality: Questioning the Information Society. Taylor & Francis (2002)
19. Foundation, N.S.: Women, minorities, and persons with disabilities in science and engineering. field of degree: Women. https://www.nsf.gov/statistics/2017/nsf17310/digest/fod-women/computer-sciences.cfm [Online; accessed January 24, 2018].
20. Abbate, J.: Recoding Gender: Women's Changing Participation in Computing. History of Computing. MIT Press (2012)