ADCS Reaches Adulthood:

An Analysis of the Conference and its Community over the last Eighteen Years

Bevan Koopman[1,2], Guido Zuccon[2], Lance De Vine[1], Aneesha Bakharia[1],
Peter Bruza[1], Laurianne Sitbon[1], Andrew Gibson[1]

[1]Faculty of Science & Technology, Queensland University of Technology, Brisbane, Australia
[2]Australian e-Health Research Centre, CSIRO, Brisbane, Australia

ABSTRACT

How influential is the Australian Document Computing Symposium (ADCS)? What do ADCS articles speak about and who cites them? Who is the ADCS community and how has it evolved?

This paper considers eighteen years of ADCS, investigating both the conference and its community. A content analysis of the proceedings uncovers the diversity of topics covered in ADCS and how these have changed over the years. Citation analysis reveals the impact of the papers. The number of authors and where they originate from reveal who has contributed to the conference. Finally, we generate co-author networks which reveal the collaborations within the community. These networks show how clusters of researchers form, the effect geographic location has on collaboration, and how these have evolved over time.

Categories and Subject Descriptors

A.m [**Miscellaneous**]; H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

Measurement

1. INTRODUCTION

The first Australian Document Computing Symposium (ADCS'96) was held in March, 1996. It was organised by Ron Sacks-Davis (RMIT) with Justin Zobel (RMIT) as the founding programme chair. This inaugural meeting set in train a series of eighteen successive symposia. ADCS'96 featured a very well attended industry day, as well as a memorable keynote delivered by Prof. Mary O'Kane, who at that time held a senior position in the ARC and eventually became NSW Chief Scientist. Symposium delegates included many people from government agencies who were working with documents, particularly from Tasmania. This emphasis on documents was by no means an accident and the name of the symposium reflected the goal to focus on various forms of computation over documents. Branding the symposium in this way allowed both NLP and IR to be covered, including additional topics. That is, at its inception ADCS aimed to be a broad church, not just an Australian IR conference. The irony here is that the lions share of the pioneers behind the conference were in fact IR researchers active in the SIGIR community.

The roots of ADCS trace back to the "First Australia-Japan Joint Symposium on Natural Language Processing", which was held November, 1989 at RMIT/Uni. Melbourne. Although the original name of ADCS suggested an Australian focus, significant contributions have been made by overseas researchers. For example, an Asian connection was already present at the second symposium, when Tengku Sembok served on the small programme committee of seven members under the affiliation UMK (Malaysia). Other early contributions to the internationalisation of ADCS came from Mun Kew Leong (from Singapore) who had been a regular reviewer pre-2000 ADCS, and from articles by Charlie Clarke (Uni. Waterloo, Canada) and L. E. Hodge (Cardiff University, Wales). As we entered the new millennium, ADCS became the "Australasian Document Computing Symposium" to provide a more international positioning.

Long running symposia are not possible without a supportive community and that very community infuses a culture into the conference. The culture of the ADCS symposia is reflected, in part, by a low budget (and at times disdainful) approach to organisation, an informal atmosphere, and the intention to be as supportive as possible to research students. Over the years, numerous students have presented at ADCS. As the symposia have for many years taken place in early December, this has allowed researchers to gain feedback on preliminary work which may eventually be submitted to SIGIR or CIKM the following year. The ADCS culture is not insular. There were ongoing discussions at various stages about how best to colocate the event with like conferences. A successful solution was found in the form of the ongoing mutually supportive joint arrangement with the Australasian Language Technology Association (ALTA) Workshop.

ADCS 2013 is the eighteenth instalment of the symposium. Like a child who reaches adulthood, perhaps it is time to look back over its childhood. Who has been involved? What have they been doing and how has that evolved, and importantly, has what they have been doing mattered? In an attempt to unveil how the community and the research top-

ADCS'13, December 05 – 06 2013, Brisbane, Queensland.
Copyright 2013 ACM 978-1-4503-2524-0/13/12 ...$15.00.

ics changed over time, as well as what the impact of ADCS is, in this paper we answer a number of questions about the conference and its community. To do so, we search for ADCS articles on the Web and extract the associated metadata (e.g. authors, affiliations, number of citations, etc.). Different aspects of the conference and its community are discussed through a quantitative and qualitative analysis of the acquired data.

2. ACQUISITION OF ADCS ARTICLES AND CITATION INFORMATION

Two sources of ADCS data were used for this study. First, the ADCS proceedings, including PDFs of the articles published in ADCS, and secondly ADCS articles indexed by Google Scholar, including the number of associated citations.

2.1 Obtaining past ADCS proceedings

Prior to 2012, when ADCS proceedings were first included in the ACM Digital Library, ADCS articles were uploaded to the ADCS website hosting that year's conference. Therefore, to obtain the PDF proceedings we visited each of the past ADCS websites and downloaded the published articles. The practice of copying the previous year's website and adapting it to the next year proved beneficial as most sites conformed to a very similar format, making crawling the proceedings easier. However, the websites for 1996 to 2002 and for 2005 were no longer available. To obtain the articles for these years we used the Wayback Machine web archive.[1] Where the proceedings could not be obtained using the Wayback Machine, we obtained the paper proceedings, scanned them and applied Optical Character Recognition (1997, 1998, 1999, 2000). Using these methods we were able to obtain PDF versions of the proceedings for all years except 1996, 2001 and 2003.[2] For these years, either the proceedings were never uploaded to the conference website or the site administrator expressly disabled crawling of the site using the `robots.txt`, the Robots Exclusion Protocol, and therefore no Wayback Machine archive existed.

The proceedings were used to analyse the various themes, topics and author affiliations, and how these have changed over time.

2.2 Obtaining ADCS articles in Google Scholar

To obtain the BibTeX entries and citation information for ADCS articles, Google Scholar was used to search for indexed ADCS articles[3]. This was performed by searching Google Scholar's "published in" field using the three queries: *Australian Document Computing Symposium*, *Australasian Document Computing Symposium* and *ADCS*. For each article returned by Scholar, the BibTeX entry was automatically downloaded and the number of citations to that article recorded. Each article was then manually reviewed to ensure it was an article published in ADCS. Using this method, a total of 191 ADCS articles were obtained from Google Scholar; out of a total of 268 articles published in ADCS (1996–2012). Figure 1 summarises the number of articles published in the ADCS proceedings over the seventeen years of the symposium, and contrasts this to the number of articles that were retrievable using Google Scholar. The figure shows that the number of articles published varied between years (mean 15.8 articles/year, stddev. 4.1). The least articles published was 10 articles in 1997, while the most was 27 articles in 2002. Generally, it was easier to obtain articles from later years using Google Scholar. The articles not available in Scholar were either never uploaded online or did not report the venue as published in ADCS.

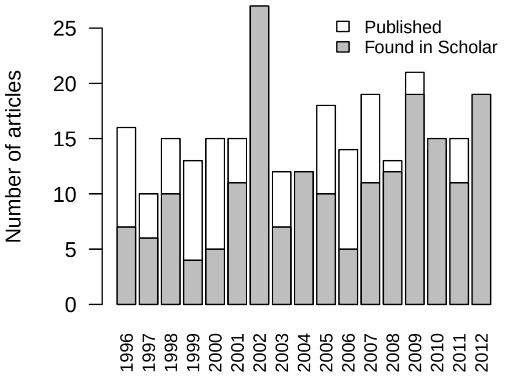


Figure 1: Yearly coverage of ADCS articles in Google Scholar.

The articles from Google Scholar were used to analyse the ADCS community, including co-authorship and how this evolved over time. Additionally, the citation counts were used to assess the impact of ADCS articles.

3. ANALYSIS OF ADCS: THE CONFERENCE

3.1 How influential is ADCS?

To study the influence of ADCS on the research community, we extracted the number of citations received by each article published in ADCS and indexed by Google Scholar; these are reported in Figure 2. The figure exhibits a power law distribution: a few articles that are highly cited, and a large number of articles that have received little or no citation. However, articles with no citation are not the majority: of the 191 acquired articles, about 150 have been cited at least once. The top 10 cited articles all have at least 18 citations, with the most cited article receiving 44 citations.

To further study the influence of ADCS articles on the international research community, we examined specifically which venues are citing the 20 most cited ADCS articles. Table 1 reports the venues most frequently citing ADCS articles, along with the number of citations.

The results show that citations to ADCS articles largely originate from IR venues, reflecting a sustained and unsurprising focus on this area. Included in the list of venues are those covering foundational work on IR, such as the *Modern Information Retrieval* book, *Foundations and Trends in Information Retrieval* and *ACM Computing Surveys*. Citation by articles in foundational venues showed that some ADCS articles cover research foundational to IR.

Although IR venues cite ADCS articles the most, there is

[1] The Wayback Machine is a digital archive of the World Wide Web; crawls of the web are performed periodically. Users can enter a URL and view past version of webpages at time-points. See http://archive.org.

[2] Although the full-text PDFs could not be obtained, we did acquire the list of papers for these years.

[3] Note that an article may appear as a record in Scholar but without the corresponding PDF file.

Venue	#
Conference on Research & Development on Information Retrieval (SIGIR)	9
Conference on Information & Knowledge Management (CIKM)	7
Information Retrieval (JIR)	7
ADCS	6
Journal of the American Society for Information Science and Technology (JASIST)	5
World Wide Web Conference (WWW)	5
European Conference in Information Retrieval (ECIR)	5
Transaction on Information System (TOIS)	4
Modern Information Retrieval Book [1]	4
International Journal of Human Computer Studies	4

Table 1: Top 10 venues most frequently citing ADCS articles.

a notable diversity of other venues, reflecting the diversity of topics covered in ADCS. For example, ADCS articles receive citations from works in Human Computer Interaction, represented by venues such as *SIGCHI*, *Advances in Human Computer Interaction*, *The Ergonomics Open Journal*; Biomedical Informatics, including venues such as *BMC Bioinformatics*, *Journal of Biomedical Semantics* and *PLoS ONE*; Cognitive Science, including venues such as *Behaviour Research Methods* and the *International Journal of Knowledge and Learning*; Artificial Intelligence, represented by venues such as *Journal of Artificial Intelligence Research*, *IJCAI*; Machine Learning and Pattern Recognition, covered by venues such as *ICML*, *Pattern Analysis and Machine Intelligence* and Pattern Recognition (ICPR); and Data & Knowledge Engineering, including venues such as *SIGKDD*, *SIGMOD* and *Knowledge Engineering Review.*

There are few citations to ADCS articles from NLP venues; this is surprising given the NLP and IR focus at the conception of ADCS and the co-location with the NLP conference, ALTA. Regarding within community citations, there are a sufficient number to show that there is evolution and interaction within the community, but too few to reflect a community with an insular focus.

In summary, the number of citations and the quality of venues citing ADCS articles reflect that ADCS has had a small but significant focus on both Information Retrieval and a number of other fields.

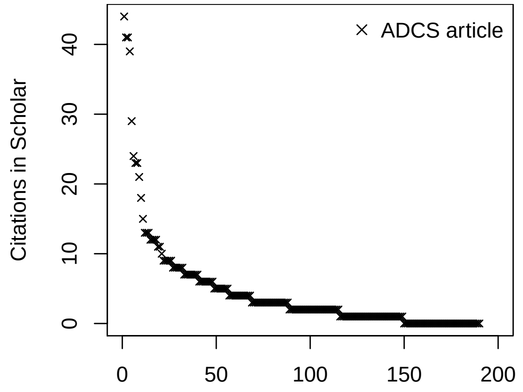


Figure 2: ADCS articles ordered by citation count.

3.2 What are the most influential ADCS articles?

A small number of ADCS articles are highly cited. The top 10 most cited articles are reported in Table 2, along with the total number of citations and the number of citations per year. If the citation count is used as a measure of influence, then these highly cited articles can be considered as the most influential. We observe a diversity of topics covered in the top 10 influential ADCS articles. The previous section showed that ADCS articles are most likely to be cited by IR related venues. Here instead, many of the top 10 influential articles are not about search related topics; for example, there are articles on NLP (#3) and Document Classification (#4, #10). Recall that the scope of ADCS topics relates to document computing in genreal, not just search, and this aim is reflected in the most influential articles over the years.

3.3 What are the topics of ADCS?

Figure 3 reports the word cloud obtained from the titles of ADCS articles (only those appearing in Scholar). The word cloud was produced using Wordle[4] and by applying a standard English stop list.

To further identify the most salient topics of the ADCS proceedings, we used a topic clustering algorithm based on Non-Negative Matrix Factorisation [4], which is used to decompose the initial article-term co-occurrence matrix into a product of two matrices highlighting associations between latent topics and terms, and latent topics and articles. The initial matrix was initialised by applying Singular Value Decomposition to the matrix reporting the number of occurrences of each term (words stemmed with the Porter Stemmer algorithm) in each article [2]. A projected gradient descent was used to estimate the product.

Figure 4 summarises the 19 topics identified from the automatic analysis of available ADCS articles (in PDF); topic labels are assigned manually by interpreting the related top terms (omitted for space restrictions). The figure also reports the spread of articles over the conference years. The results highlight the consistent focus over time of ADCS articles on issues related to document management, document structure, IR evaluation and medical IR. It also shows the emergence of recent interests in cognitive aspects of search, mobile and user studies.

[4] http://www.wordle.net/

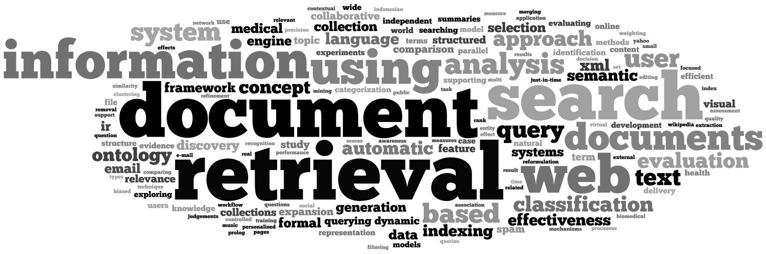


Figure 3: Word cloud generated from the titles of ADCS articles.

#	Author Surnames	Article	Year	Cit.	Cit./Y
1	Dennis, McArthur, Bruza	Searching the World Wide Web Made Easy? The Cognitive Load Imposed by Query Refinement Mechanisms	1998	44	3.14
2	Upstill, Craswell, Hawking	Predicting fame and fortune: Pagerank or indegree	2003	41	4.56
3	Fuller, Zobel	Conflation-based comparison of stemming algorithms	1998	41	2.93
4	Smith	Machine mapping of document collections: the leximancer system	2000	39	3.25
5	Zobel	Collection selection via lexicon inspection	1997	29	1.93
6	Billerbeck, Zobel	Document expansion versus query expansion for ad-hoc retrieval	2005	24	3.43
7	O'Keefe, Koprinska	Feature selection and weighting methods in sentiment analysis	2009	23	7.67
8	Vercoustre, Dell'Oro, Hills	Reuse of information through virtual documents	1997	23	1.53
9	D'Souza, Zobel, Thom	Is CORI Effective for Collection Selection? An Exploration of Parameters, Queries, and Data.	2004	21	2.63
10	Crawford, Koprinska, Patrick	Phrases and Feature Selection in E-Mail Classification.	2004	18	2.25

Table 2: The 10 most cited ADCS articles in Google Scholar (1997–2012).

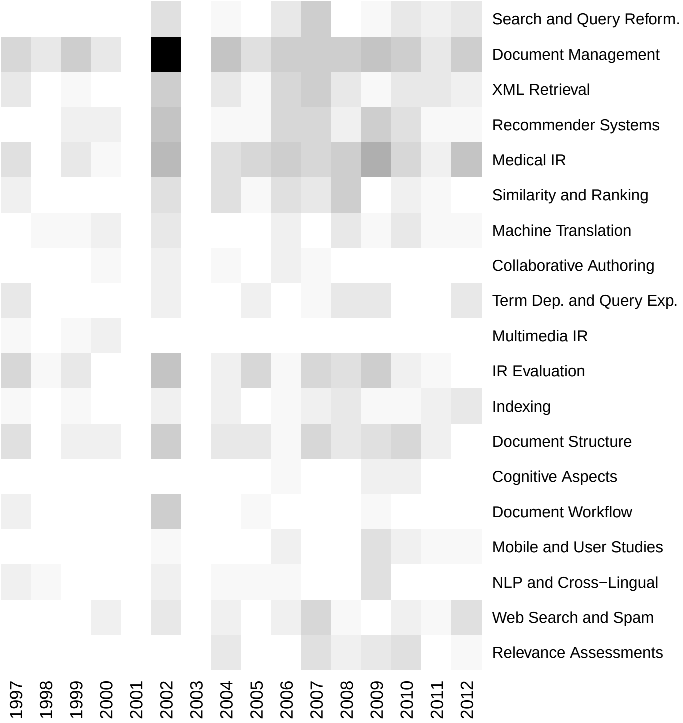


Figure 4: Research topics ADCS articles belong to, per year of publication. In the image, darker cells indicate that more articles pertaining to the corresponding topic have been published in ADCS on the specific year.

4. ANALYSIS OF ADCS: THE COMMUNITY

4.1 How has the community grown and where do ADCS authors come from?

To answer this question, we manually recorded the number of authors and their country of affiliation. The number of contributing authors has fluctuated over time (mean 41 authors/year, stddev. 12), as shown in Figure 5. A large number of authors often occurs when the conference is hosted in Sydney (2002, 2005, 2009). The first hosting of the conference outside of Australia (Dunedin, 2012) also saw above average number of authors. Although numbers have fluctuated, recent years have shown a steady growth in the number of authors.

ADCS was born as an Australian conference but quickly broadened focus, as reflected in the name change in 1999. Although the majority of authors still originate from Australian institutions, there has also been a substantial number of overseas contributions (mean 6.6, stddev. 3.5). Contributing authors by region are as follows: Australia 425, New Zealand 40, Europe 26, North America 15 and Asia 12. Figure 5 also shows the portion of international contributions. ADCS has consistently attracted international authors over the years (a notable exception being 2002). In recent years (2009 onwards), there has been a consistent growth in the number of New Zealand authors, while international contributions have remained steady.

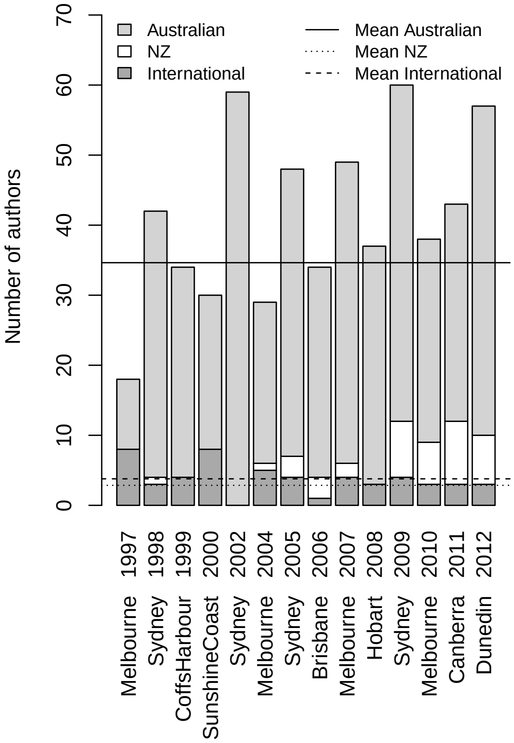


Figure 5: Number of contributing authors by region.

4.2 How large is the ADCS community? How is it connected?

To further analyse the ADCS community, we considered the co-authorship network. The network was built by analysing the BibTeX data referring to ADCS articles crawled through Google Scholar. The data was manually cleansed by conflating different references to the same author to a common author name (e.g. P. Bruza and P. D. Bruza refer to the same author). As mentioned previously, this data does not cover the totality of articles published in ADCS, yet it does contain more than 70% of the total articles.

Figure 6 shows the co-authorship network constructed from articles published until 2012. The network contains 266 nodes/authors and 441 edges co-authorship, and thus each author has co-authored with an average of about 1.6 authors.

The network is composed of a main backbone component and several islands. The network shows marked clusters of co-authors based principally on geographical location and institution. These clusters are then connected to form the main backbone through edges between a few main authors. The main backbone features clusters from three major geographic regions; Melbourne (centred around Moffat, Scholer, Thom, Turpin, Zobel), Dunedin (Trotman), Brisbane (Bruza, Geva), and authors from Canberra (Hawking, Thomas). Indeed, geographical location seems to be the main driver for collaboration, with topic of research being a secondary driver.

The network exhibits a number of disconnected islands. The largest island revolves around J. Kay (Sydney), who published on personalisation and recommender systems with local collaborators in the early years of ADCS. A second large island revolves around T. Baldwin and represents part of the NLP and health informatics contribution to ADCS, driven also by the collocation of the symposium with the ALTA workshop.

4.3 How has the ADCS community evolved over time, and did ADCS seed new collaborations?

To analyse the growth of research collaborations in the community, measured by co-authorship, we consider the evolution of the co-authorship network over time. Figure 7 shows that the community has grown over time, and highlights the creation of the main backbone representing the core of the community, containing authors providing sustained contributions. In the first years of ADCS (1996-2003), many new collaborations were created along with new authors joining the community. While new authors kept joining the community in the period 2003-2009, new collaborations grew at a slower rate. This observation is supported by Figure 8(a) which shows the edge to node ratio (co-authorship to author ratio). In the last three years (see Figure 6) new authors have continued to join the community, but more significant is the growth in collaborations: this is once again confirmed by the growth in edge to node ratio (Figure 8(a)).

The Trotman Effect

The most notable phenomena that occurred over the years in the ADCS's co-authorship network was the formation and growth of the main backbone component. Historically, the main backbone started emerging with collaborations between Wu, Paris and Lu, which allowed for connections between authors around the Hawking and Paris clusters. However, the main propellent for the creation of the large backbone that spanned heterogeneous research topics, connecting geographically distant researchers, was Andrew Trotman. It is through the collaborations Trotman-Jones and Trotman-Geva (2007-2009) that the main backbone has largely emerged within the network, thus increasing the diameter[5] of the ADCS network from 7 edges in 2006 to 12 in 2009.

ADCS is not a Small World (Yet!)

It has been noted that co-authorship networks in the scientific community often form a small world network [5]. A small world network is characterised by a linear relationship between the growth of the average path length and the growth of the log of the number of nodes. This is not the case for ADCS: Figure 8(b) reports the growth of the average path length[6] over time, which does not grow linearly with the log of the number of nodes. This means that authors tend to create new collaborations with authors within their cluster, often characterised by their geographic location. As observed above, there are only a few authors that span clusters to form the main backbone.

Figure 8(b) does however highlight that the growth in average path length has slowed, meaning that although new authors join the community and collaborations are created, these do not contribute much in increasing the distance between any two authors. This is because in the period 2009-2012, there have been collaborations that connected two authors who were previously far apart. This phenomenon is exemplified by the collaboration between P. Vines and X. Zhang that has reduced the number of edges to be traversed to go from M. Wu to P. Vines from 5 to 2. The creation of future collaborations between authors currently distant in the co-authorship network would quickly transform the ADCS community in a small world network (at least its main connected component, i.e. the backbone).

Figure 8(b) also shows a marked change in the trend of clustering coefficient, which represents how 'dense' the co-authorship network is. The density of collaborations in the ADCS community has decreased over time, as authors belonging to new groups joined the community (i.e., adding weakly connected leaves to the network). However, in recent years (2009-2012) the clustering coefficient has increased because new collaborations have been created between authors that shared a common co-author (thus closing co-authorship triangles). This is the case for the Brisbane cluster (Bruza, Geva) where a denser web of co-authorship has arisen.

5. DISCUSSION AND CONCLUSION

The analysis of a scientific conference and its community does not simply amount to self-celebratory folklore, but often unveils its impact on the research field and offers points of reflection for future growth. A similar study of the SIGIR community has for example unveiled how articles appearing in that forum have changed in topic and language over time [3]. A recent investigation of the CLEF community has shown the impact of that evaluation forum on the whole

[5] We refer to the diameter of the network as the average of the diameters of each connected component.

[6] We refer to the average path length of the network as the mean of the average path lengths of the single components forming the network.

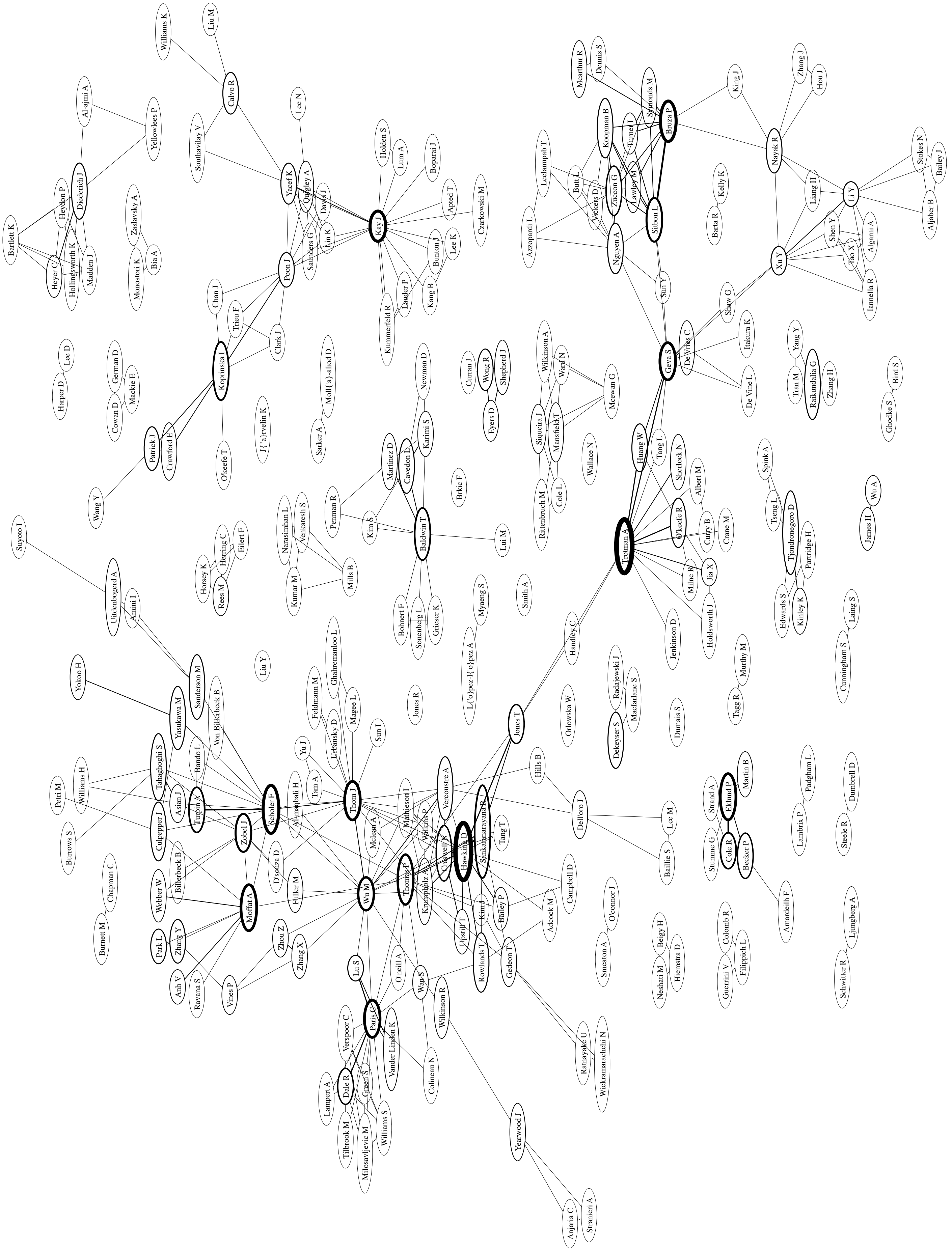
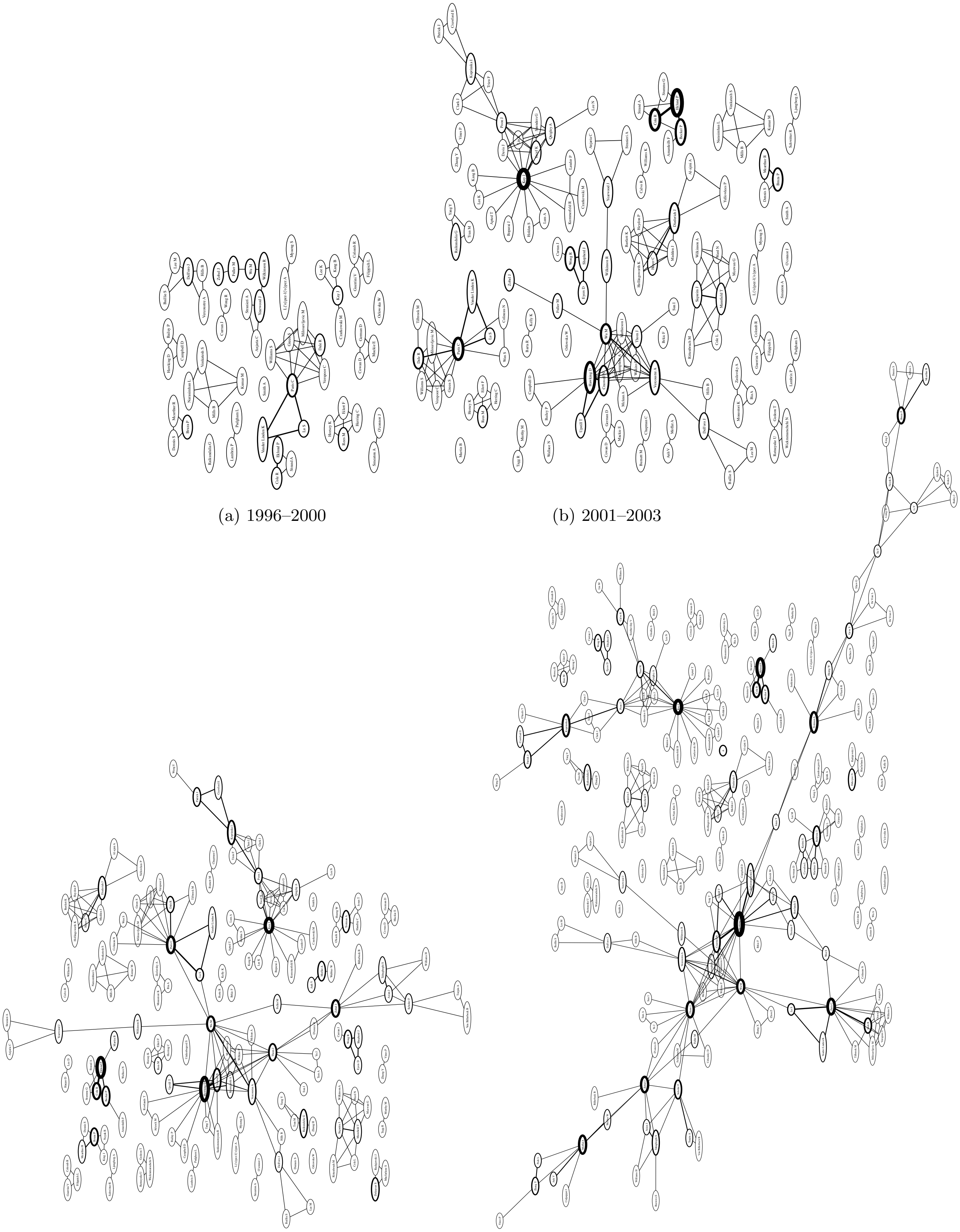


Figure 6: Co-authorship network: nodes are authors, edges represent co-authored articles.



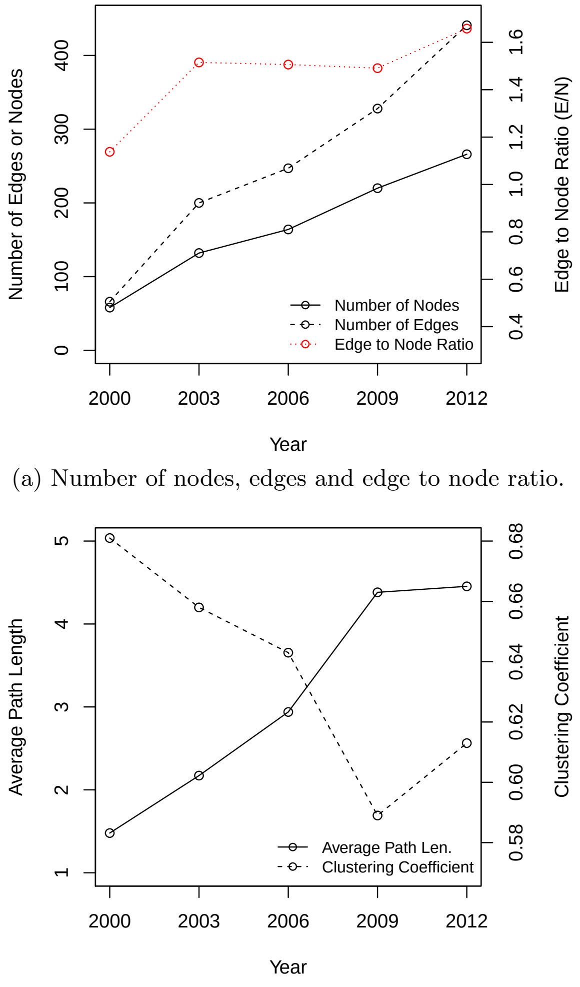
(a) 1996–2000

(b) 2001–2003

(c) 2004–2006

(d) 2007–2009

Figure 7: Evolution of the co-authorship network over time; each sub-figure represents the cumulative co-authorship connections.



(a) Number of nodes, edges and edge to node ratio.

(b) Average path length and clustering coefficient.

Figure 8: Network measures for ADCS co-authorship network (2012).

research field [6]. The analysis presented in this paper has revealed a number of interesting aspects of ADCS:

- Not all ADCS articles can be retrieved using the popular Google Scholar tool. This is not the case just for old articles: even some of the articles published in 2011 are not available in Scholar.

- ADCS articles have had an impact on the larger IR community, with works cited in top conferences and journals, as well as in surveys that cover foundational IR topics and methods. Interestingly, a number of articles are also cited outside of the IR community, with work from Human Computer Interaction, Biomedical Informatics, Cognitive Science, Artificial Intelligence and Machine Learning, drawing from research published in ADCS.

- Even if ADCS was born as a national conference, and later renamed to address the growing international contributions, the symposium still enjoys significant contributions from Europe, North America and Asia.

- Analysis of the co-authorship networks has shown that geographic location is a strong driver for collaboration; however there are a few key authors that connect geographically disparate clusters (dubbed the "Trotman effect").

- The topic analysis of published articles has shown that although ADCS has a strong IR focus, the conference has continued to cover topics related to document computing more generally.

Besides the ADCS specific analysis, this paper makes general contributions in the application of content based analysis techniques to scientific articles, including thematic and temporal analysis, as well as the analysis of co-authorship networks, revealing collaboration within, and evolution of, a scientific community over time.

A number of resources generated in this study are provided online, including: past ADCS proceedings (PDF and plain text), topic content-analysis (including topic-keyword mappings), Scholar citation counts and high definition co-author network (PDF and Graphviz); these are available at http://github.com/ielab/adcs_adulthood.

ACKNOWLEDGEMENTS

The authors thank Graham Dawson (QUT Library) and ADCS pioneers and stalwarts: Ross Wilkinson, Alistair Moffat, David Hawking, Paul Thomas and Justin Zobel, for access to their historical reflections.

6. REFERENCES

[1] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[2] C. Boutsidis and E. Gallopoulos. SVD based Initialization: A Head Start for Nonnegative Matrix Factorization. *Pattern Recognition*, 41(4):1350–1362, 2008.

[3] D. Hiemstra, C. Hauff, F. de Jong, and W. Kraaij. SIGIR's 30th Anniversary: An Analysis of Trends in IR Research and the Topology of its Community. In *ACM SIGIR Forum*, pages 18–24, Amsterdam, The Netherlands, 2007. ACM.

[4] C.-J. Lin. Projected Gradient Methods for Nonnegative Matrix Factorization. *Neural computation*, 19(10):2756–2779, 2007.

[5] M. E. Newman. The Structure of Scientific Collaboration Networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[6] T. Tsikrika, B. Larsen, H. Müller, S. Endrullis, and E. Rahm. The Scholarly Impact of CLEF (2000–2009). In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 1–12. Springer, 2013.