On 25 Years of CIAA Through the Lens of Data Science

Hermann Gruber[1] and Markus Holzer[2] and Christian Rauch[2]

[1] Knowledgepark GmbH, Leonrodstr. 68
80636 München, Germany
hermann.gruber@kpark.de
[2] Institut für Informatik, Universität Giessen
Arndtstr. 2, 35392 Giessen, Germany
{holzer,christian.rauch}@informatik.uni-giessen.de

Abstract. We investigate the structure of the co-authorship graph for the *Conference on Implementation and Application of Automata* (CIAA) with techniques from network sciences. This allows us to answer a broad variety of questions on collaboration patterns. Our findings are in line with (statistical) properties of other co-authorship networks from biology, physics and mathematics as conducted earlier by pioneers of network sciences.

1 Introduction

Shortly after the invitation of the second author to give an invited talk at the 26th Conference on Implementation and Application of Automata (CIAA), the idea grew to study collaboration patterns of the co-authorship network of this conference. As said in [15] "the structure of such networks turns out to reveal many interesting features of academic communities." Co-authorship networks and collaboration patterns thereof had been subject to scientific studies long before data science became a prominent subfield of artificial intelligence research, see, e.g., [4, 16]. Thus, besides the above mentioned interesting features of academic communities with such a study we familiarize ourselves with the techniques in data science and in particular in network sciences. Moreover, since the 25th jubilee of the CIAA conference passed due to the COVID-19 pandemic restrictions without further celebration, this paper may serve as a late birthday present to the whole community that is interested in implementation and application of automata.

Our study is based on collection and analysis of data gathered from open sources. The two main open sources we rely on are DBLP[3] (database systems and logic programming), which is the on-line reference for bibliographic information on major computer science publications, LNCS[4] (Lecture Notes in Computer Science), the prestigious conference proceedings series published by Springer,

[3] https://dblp.org

[4] https://www.springer.com/gp/computer-science/lncs

and the general website `https://www.informatik.uni-giessen.de/ciaa/` of the conference. The raw data from these sources were obtained during January to April 2022 and were pre- and post-processed with the help of the widely used Python[5] distribution "Anaconda,"[6] which includes a range of useful packages for scientific coding, such as `matplotlib`, `numpy`, `pandas`, etc.

Before we turn to the analysis of the co-authorship network we briefly give some history on the conference, which will obviously lack completeness. The CIAA conference actually started in 1996 as the "*Workshop on Implementation of Automata*" (WIA) in London, Ontario, Canada. The need for such a workshop was explained in [17] as follows:

> "Whence WIA? Why the need for a workshop of this type? As there are already many (perhaps too many) computer science conferences and workshops, any new meeting faces a rather stiff need to justify its existence. WIA came about primarily because there is no other good forum for systems that support symbolic computation with automata. [...] In addition [...] there is a vast amount of applied work, most of it undocumented, using automata for practic applications such as protocol analysis, IC design and testing, telephony, and other situations where automata software is useful.
> This is good and interesting work, and it needs a place to be exhibited and discussed. Existing journals and conferences, however, seem to have a difficult time in finding a place for what we do. Theoretical arenas sometimes treat this work as "mere" implementation, a simple working-out of the algorithms, theorems, and proofs that are the "real" contribution to the field. Systems-oriented venues, on the other hand, sometimes find this kind of work suspect because it appears to be aimed at theoreticians. It is tricky navigating between the Scylla of the too-abstract and the Charybdis of the too-practical."

At that time the general organization and orientation of WIA was governed by a Steering Committee (SC) composed of (in alphabetical order) Stuart Margolis, Denis Maurel, Derick Wood, and Sheng Yu. Sadly, both Derick Wood and Sheng Yu, our late lamented colleagues, passed away too early on October 4, 2010 and January 23, 2012, respectively. The first four workshops were held at London, Ontario, Canada (1996 and 1997), Rouen, France (1998), and Potsdam, Germany (1999). During the general WIA meeting in 1999 it was decided to rename the meeting to "*International Conference on Implementation and Application of Automata*" (CIAA) and to hold its first CIAA in London, Ontario, Canada, in the summer of 2000. There it was part of a tri-event conference together with the workshop on "*Descriptional Complexity of Automata, Grammars and Related Structures*" (DCAGRS) and a special day devoted to the 50th anniversary of automata theory, which was called "*A Half Century of Automata Theory.*" It is worth mentioning that CIAA is rarely co-located with

[5] `https://www.python.org`

[6] `https://www.anaconda.com`

other conferences. An exception was the Conference on "*Finite-State Methods and Natural Language Processing*" (FSMNLP) in 2011 in Rouen, France.

Already after around half a decade the conference became mature and started its way all around the globe: Pretoria, South Africa (2001), Tours, France (2002), Santa Barbara, California, USA (2003), Kingston, Ontario, Canada (2004), Nice, France (2005), Taipei, Taiwan (2006), Prague, Czech Republic (2007), San Francisco, California, USA (2008), Sydney, Australia (2009), Winnipeg, Manitoba, Canada (2010), Blois, France (2011), Porto, Portugal (2012), Halifax, Nova Scotia (2013), Giessen, Germany (2014), Umeå, Sweden (2015), Seoul, South Korea (2016), Marne-la-Vallée, France (2017), Charlottetown, Prince-Edward-Island (2018), Košice, Slovakia (2019), and Bremen, Germany (2021), which was held as a virtual event due to the COVID-19 pandemic. The 2020 conference, planned at Loughborough, United Kingdom, was canceled and thus was also a victim of the pandemic crisis. This year, CIAA takes place again in Rouen, France, where it was last held 22 years ago. A distribution of the locations w.r.t. the continents is depicted in Figure 1. There is a slight overhang on the number of locations for Europe (14) followed by North America (9). Then there is a large drop for Asia (2), Africa (1), and Australia/Oceania (1). South America and Antarctica have never been visited by CIAA, and for Antarctica, we personally think that there is no chance to organize it there. The current SC is encouraged to further globalize the conference and to fill the white or say gray spots on the continents' landscape.

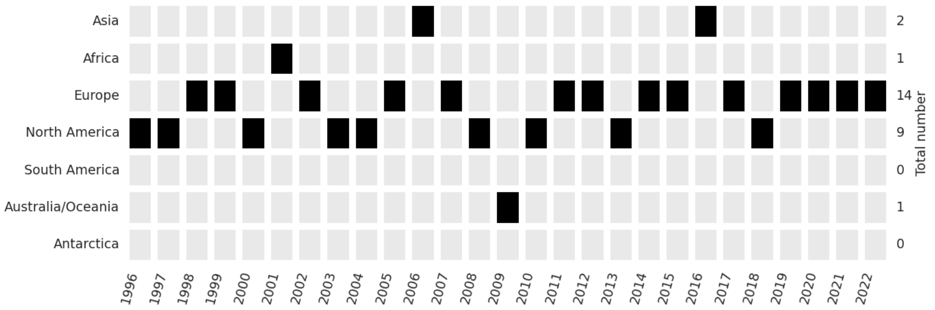


Fig. 1. CIAA destinations in relation to their continent locations.

Since the first WIA event in 1996, the proceedings appeared in the Springer LNCS series. This was not the case for the sister conferences "*Developments in Language Theory*" (DLT) and "*Descriptional Complexity of Formal Systems*" (DCFS), formerly known as "*Descriptional Complexity of Automata, Grammars and Related Structures*" (DCAGRS), that started slightly earlier than WIA. The authors of the best paper of the actual conference are awarded a monetary grant since 2004 (except for 2021). Until 2008, this was sponsored by the University of California at Santa Barbara and later by the conference itself. By acclamation the best paper award was subtitled "*Sheng Yu Award*" at the general CIAA meeting in 2012 and first awarded with this naming in 2014. So far, only five authors had the privilege of receiving the ”Best Paper Award" twice. These are (in alphabetical order) Janusz Brzozowski (2017 and 2018), Markus Holzer (2009 and 2015), Lisa Kaati (2006 and 2008), Lila Kari (2004 and 2018), and Mikhail V. Volkov (2007 and 2012). Since the renaming to CIAA in 2000, extended versions of selected papers from the proceedings of the conference series are usually retained for publication in special issues of either *International Journal of Foundations of Computer Science* (IJFCS) or *Theoretical Computer Science* (TCS), alternating each year.

The legacy of CIAA continues—the event is in its 26th edition and the expectations raised in [17] have been widely fulfilled:

> "Providing a forum for this work is a useful goal, and a sufficient one for WIA [CIAA]. But I think WIA [CIAA] is part of something more fundamental, and a process I want to encourage: the re-appraisal of the value of programming in computer science."

Nowadays the general organization and orientation of CIAA is directed by the SC members (in alphabetical order) Markus Holzer, Oscar H. Ibarra, Sylvain Lombardy, Nelma Moreira, Kai Salomaa, and Hsu-Chun Yen. Enough of the historical overview. Now let us concentrate on what can be deduced from the data that we extracted from the web.

The paper is organized as follows: In Section 2 we first briefly take a look on the topics of CIAA as communicated by the call for papers and the published papers. This will be a quick and shallow dive into natural language processing without to much details. Then in Section 3 the search for collaboration patterns is done in correspondence to previous systematic studies on co-authorship networks or more general on social real-world networks as conducted in [13, 14]. Finally, we conclude our tour through the world of data-science with some ideas for further investigations.

2 Conference *Versus* Paper Topics

The CIAA call for papers solicits research papers and demos on all aspects of implementation and application of automata and related structures, including theoretical aspects, as but not limited to:

- bioinformatics,
- complexity of automata operations,
- compilers,
- computer-aided verification,
- concurrency,
- data structure design for automata,
- data and image compression,
- design and architecture of automata software,
- digital libraries,
- DNA/molecular/membrane computing,
- document engineering,
- editors, environments,

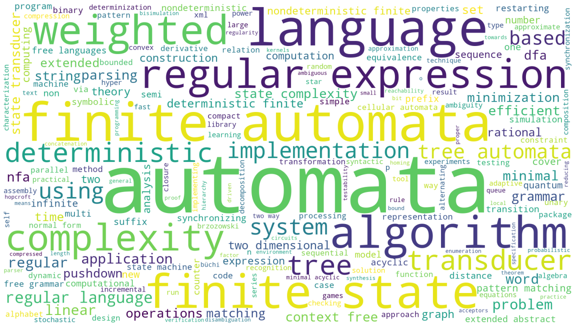
- experimental studies and practical experiences,
- industrial applications,
- natural language processing,
- networking,
- new algorithms for manipulating automata,
- object-oriented modeling,
- pattern-matching,
- quantum computing,
- speech and speaker recognition,
- structured and semi-structured documents,
- symbolic manipulation environments for automata,
- teaching,
- text processing,
- techniques for graphical display of automata,
- very large-scale integration (VLSI) research,
- viruses, related phenomena, and
- world-wide web (WWW).

How do the topics in the call for papers compare to the topics of the actual papers? To answer this question, we take a look at word clouds.

Word clouds have become a staple of data-visualization for analyzing texts. Usually words (unigrams) and bigrams, and the importance of each are shown with fontsize and/or color. Since the list of CIAA topics is condensed and limited one may consider all CIAA publications as a natural resource for natural language processing techniques. The decision to use only DBLP as data source considerably limits the analysis of the CIAA texts, because DBLP does not offer all relevant features of publications. For instance, the access to abstracts is not possible *via* DBLP. For such, information the relevant Springer websites have to be contacted. The only meaningful textual data DBLP provides is the title of a publication. With these titles, one can easily prepare a word cloud with the help of Python's `wordcloud` library. To this end the frequency of uni- and bigrams are determined. For a word the frequency is defined as the quotient of how often the word appears in the text and the number of all words of the text in question. Normalization is done by dividing with the maximal frequency. Usually preprocessing of the text incorporates removing of stopwords, such as, e.g., are, is, and, or, etc., stemming and lemmatization (word normalization). The word cloud obtained from the titles of all CIAA publications is depicted on the left of Figure 2, where only the removing of stopwords was applied. Words and bigrams related to automata and expressions attain high ranks. It is worth mentioning that the missing words "implementation" and "application" from the conference name CIAA appear on rank 12 and 22, respectively, with normalized frequencies 0.1215 and 0.0841, respectively.

3 Collaboration Patterns

In general the *co-authorship network* or *co-authorship graph*, for short, is an undirected graph built from a set of publications P restricted to a set of authors A from these publications with the following properties: (i) the set of nodes corresponds to the set of authors A and (ii) two authors are connected by an undirected edge if there is at least one publication in P jointly co-authored



word/bigram	frequency
automata	1.0000
finite automata	0.2850
language	0.2617
finite state	0.2196
algorithm	0.1963
weighted	0.1636
regular expression	0.1636
complexity	0.1402
transducer	0.1355
tree	0.1308

Fig. 2. (Left) Word cloud generated from all titles published at CIAA with standard stopwords and (right) the words and bigrams with the highest normalized frequencies.

by them. We call such a network a P-A co-authorship network. There are several ways to generalize co-authorship graphs, for instance, to introduce edge and node weights reflecting measures for collaboration (e.g., Newman's weighting scheme) and impact/productivity (e.g., h-index/number of papers), respectively. Note that co-authorship graphs are quite different from citation graphs. The latter is yet another important type of graph related to network sciences, but is not considered here.

We investigate (i) the publication venue co-authorship network of CIAA by using all publications of CIAA and hence all authors that ever published a paper at the conference (CIAA-CIAA co-authorship network) and (ii) the field co-authorship network, where all publications, not limited to the conference in question, of CIAA authors are used to construct the graph to be investigated (ALL-CIAA co-authorship network). For better comparability, we only take papers into account that appeared in 1996 or later when constructing the ALL-CIAA network. We think that the differentiation of these two graphs is important, because the conference only cannot describe the community behind CIAA completely. This may lead to different results of the analysis. As already mentioned earlier, for the analysis we decided to use only one data source, namely DBLP. This, in particular reduces the bias and simplifies identification problems such as, e.g., author identification, since we are acting within a closed world, namely DBLP. On the other hand, DBLP will not offer all relevant features of publications and authors as one would like to have. The raw data for the two networks contains lists of papers, including authors names and possibly other information such as title, pagination and so forth, but no information on abstract or affiliation of the authors, because these data are not communicated by DBLP. The construction of the co-authorship networks is straightforward by using Python's `networkx`[7] library and leads us to some basic results, which we report next.

[7] `https://networkx.org`

The findings on the basic results for our two co-authorship networks are summarized in Table 1. Let us comment on these numbers. The total number of papers is 688 respectively 38, 250. As a curious fact, for the CIAA-CIAA data set there are exactly two papers with the same title and authors, namely "Size Reduction of Multitape Automata" by Hellis Tamm, Matti Nykänen, and Esko Ukkonen that appeared in 2004 and 2005. Concerning the number of authors, as already said, the identification problem such as mentioned in [7] is not relevant to our study, thanks to the use of DBLP as the single source of truth for authors. DBLP does an excellent job in author name disambiguation, as reported in [9]. Author name disambiguation at DBLP is achieved by the combined effort of algorithms and humans, as described in [12]. For instance, the DBLP database identifies Kees Hemerik and C. Hemerik as the same person, while by relying on names only, one would rather count them as separate individuals. We are quite sure that the CIAA-CIAA data set is approximately correct w.r.t. the identification problem of authors. Hence, a bias from an incorrect identification is negligible for us. The average number of papers per author is 1.81 and the distribution of papers per author follows a power law. This was first observed by Lotka [10] and later confirmed by several studies, and is nowadays known as "Lokta's Law of Scientific Productivity"—see Figure 3.

	Co-authorship network	
	CIAA-CIAA	ALL-CIAA
total papers	688	38,250
total authors	839	839
mean papers per author	1.81	59.12
mean authors per paper	2.22	3.35
mean collaborators per author	2.57	43.80
size of giant component	192	696
as a percentage	22.76%	83.4%
2nd largest component	41	8
clustering coefficient	0.55	0.49
mean distance[8]	8.36	4.71
maximum distance[8]	22	12

Table 1. Summary of results of the analysis of the CIAA-CIAA and ALL-CIAA co-authorship network.

Simply speaking, if one plots two quantities against each other where both axes are logarithmically scaled (log-log scaled) and they show a linear relationship, this indicates that the two quantities have a power law distribution. Such a line can be described by $\ln f(x) = -\alpha \ln x + c$ and by taking exponentials we

[8] Since the CIAA-CIAA and ALL-CIAA co-authorship networks are not connected, the values are only computed for the largest connected component.

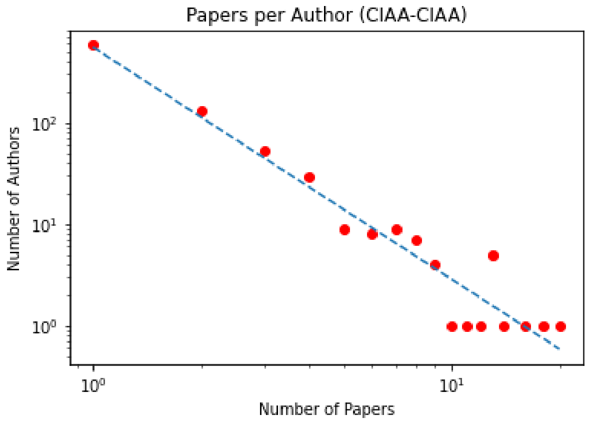


Fig. 3. Plot of the number of papers written by authors in the CIAA-CIAA co-authorship network. The plot is log-log scaled. The corresponding plot for the ALL-CIAA co-authorship network is similar but not shown due to space constraints.

end up with

$$f(x) = C \cdot x^{-\alpha},$$

where $C = e^c$. Distributions of this form are said to follow a *power law* and α is called the *exponent* of the power law. Observe, that a positive exponent α induces a negative slope on the straight line in the log-log plot. Mostly the constant C is not of particular interest. Power law distributions occur in an extraordinarily wide range of phenomena, e.g., [1, 3, 8, 10, 11, 18]. The distribution of papers per author follows a power law with exponent α approximately 2 in general [10] and we have $\alpha \approx 2.28$.

Now we turn to the ALL-CIAA co-authorship network. For many papers, DBLP identifies that an author name is shared by different authors, yet is not able to make an educated guess to which person the paper should be attributed. In that case, the author link in the DBLP record of the paper points to a disambiguation page, which lists the papers of all authors with that name. In our data set, we used the disambiguation page to serve as a list of papers by that author if DBLP cannot determine the author. In total, we have to deal with 11 disambiguation pages. While this number is quite modest compared to the total number of authors, these eleven pages list an amount of 2035 publications in total. So the average number of publications per disambiguation page is 185. Since those disambiguation pages list papers that are sometimes produced by many different actual persons, the disambiguation pages may introduce a sizable distortion in averages such as "papers per author" and "number of collaborators per author." For a moment, let us assume that each disambiguation page stands for an actual author who published only 1 paper overall - this will certainly underestimate the actual state of affairs. Then we have 839 authors and $(38250 - 2035 + 11)$ papers, which yields a figure of 43.12 papers per author on average, which is seizably lower but still in the same ballpark. So another explanation for the unusually high scores is in order. We will propose a hypothesis below, where we look at the top 10 in various different aspects.

In the first column of Table 2, we list the most frequent authors of both the CIAA-CIAA and ALL-CIAA co-authorship network. For the ALL-CIAA network, we observe that some CIAA authors are highly prolific writers, drawn from diverse fields in computer science: Robotics (Alois C. Knoll), artificial intelligence (Václav Snásel), string algorithms (Gonzalo Navarro), logic and verification (Moshe Y. Vardi), network coding (B. Sundar Rajan), to list the fields of the five most prolific authors. While only few of them regularly contribute to CIAA, this shows that the conference helps bringing the various applied fields of computer science together.

	number of papers	fractional no. of papers	number of co-workers
CIAA-CIAA	20 Martin Kutrib	9.08 Andreas Maletti	22 Jean-Marc Champarnaud
	18 Jean-Marc Champarnaud	9.00 Bruce W. Watson	16 Nelma Moreira
	16 Markus Holzer	8.20 Martin Kutrib	14 Kai Salomaa
	14 Kai Salomaa	7.58 Oscar H. Ibarra	13 Martin Kutrib
	13 Andreas Maletti	7.00 Markus Holzer	13 Borivoj Melichar
	13 Borivoj Melichar	6.87 Mehryar Mohri	13 Sheng Yu
	13 Mehryar Mohri	6.82 Jean-Marc Champarnaud	12 Rogério Reis
	13 Bruce W. Watson	6.23 Borivoj Melichar	11 Johanna Björklund
	13 Sheng Yu	6.00 Kai Salomaa	11 Markus Holzer
	12 Oscar H. Ibarra	5.75 Sheng Yu	11 Sylvain Lombardy
ALL-CIAA	695 Alois C. Knoll	269.8 Moshe Y. Vardi	1082 *Cheng Li*
	609 Václav Snásel	229.5 Gonzalo Navarro	875 Alois C. Knoll
	577 Gonzalo Navarro	222.0 B. Sundar Rajan	532 *Fei Xie*
	569 Moshe Y. Vardi	175.1 William I. Gasarch	516 *Bin Ma*
	501 B. Sundar Rajan	175.0 Alois C. Knoll	420 Václav Snásel
	475 Thomas A. Henzinger	167.6 Václav Snásel	374 *Xiaoyu Song*
	466 *Bin Ma*	162.6 Thomas A. Henzinger	365 Axel Legay
	438 Kim G. Larsen	152.2 Henning Fernau	359 *Yong Sun*
	438 Axel Legay	140.1 Jeffrey O. Shallit	341 Madhav V. Marathe
	384 *Cheng Li*	137.2 Andrzej Pelc	323 Kim G. Larsen

Table 2. The authors with the highest numbers of papers, fractional number of papers, and numbers of co-authors in the CIAA-CIAA and ALL-CIAA co-authorship network. Italicized items are disambiguation pages, i.e., possibly several actual authors.

Admittedly, among the top ten CIAA authors with most collaborators, five are actually DBLP disambiguation pages: Cheng Li, Fei Xie, Bin Ma, Xiaoyu Song and Yong Sun.[8] But as explained above, the amount of distortion due to fuzziness in the data is not too high. Together with the facts explained in the preceding paragraph, this may explain why the CIAA authors have, on average, very high scores regarding both research output and collaboration.

[8] The interested readers who is able to help with disambiguation is invited to suggest corrections to the DBLP team.

We thus identified two factors that may serve as a partial explanation for the very high scores in the ALL-CIAA network. Another factor is probably the way we construct the ALL-CIAA data set: we include the 839 CIAA authors and all their publications, but we exclude most of the co-authors that contributed to those publications. For comparison, the analysis carried out for computer science as a whole in [14] included all authors that were coauthors of at least one paper in the data set. For the community of the ACM SIGMOD conference, an analysis of the co-authorship graph was carried out in [13]. The network they construct is analogous to our CIAA-CIAA network, and there again, the considered set of authors is implied by the papers that were selected. In the ALL-CIAA network, we deliberately zoomed in on the set of CIAA authors, and as a consequence, the number of authors is much smaller than the number of publications we consider. Yet, as a *Gedankenexperiment*, let us extend the set of authors to all authors listed as co-author in papers of the ALL-CIAA network. Then we obtain a total of 24717 authors—and the average number of papers per author drops to 1.55. Then again, this figure appears too low—after all, we included only a fraction of the papers by those authors that collaborated with a CIAA author.

The alternative to counting the total number of papers is fractional number of papers. Each paper co-authored by a given author adds an amount of $\frac{1}{n}$ to the fractional number of papers instead of 1 as for number of papers, where n is the total number of authors on the paper. The rationale behind this choice is that in an ideal world, an authors collective equally divides the writing between all n authors who work on a paper. The fractional number of papers became famous among theoretical computer scientists by the author ranking to the "*International Colloquium on Automata, Languages, and Programming*" (ICALP) prepared by the late Manfred Kudlek and published in the EATCS Bulletin series. As expected, there is substantial overlap between the authors with a large number of papers and those with a large fractional number of papers.

By empirical results from the literature it is awaited that a power law also applies for the author per paper and collaborator per author. Both distributions are shown in Figure 4. The average number of authors per paper is 2.22, which is

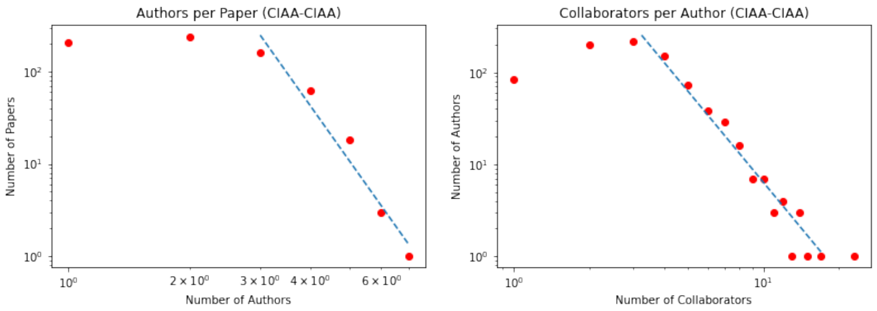


Fig. 4. Plots of the (i) number of authors per papers and (ii) number of collaborators per authors in the CIAA-CIAA co-authorship network. All plots are log-log scaled. The corresponding plots for the ALL-CIAA co-authorship network are similar.

in perfect fit with the average value 2.22 for computer science as a whole [14]. The largest number of authors on a single paper is 7 ("In Vitro Implementation of Finite-State Machines" by Max H. Garzon, Y. Gao, John A. Rose, R. C. Murphy, Russell J. Deaton, Donald R. Franceschetti, and Stanley Edward Stevens Jr.). In the CIAA-CIAA network, the mean on number of collaborators is 2.55 and this somewhat less than 3.59 for computer science as report in [14]. The distributions of the number of collaborators are depicted in Figure 4. The third column of Table 2 shows the authors of the CIAA-CIAA and ALL-CIAA co-authorship network with the largest numbers of collaborators.

In the CIAA-CIAA network, it is remarkable that although Jean-Marc Champarnaud is already retired and has published his last CIAA paper in 2012, i.e., a decade ago, he still has the highest number of 22 co-workers. His co-workers are (in alphabetical order) Philippe Andary, Pascal Caron, Fabien Coulon, Tibor Csáki, Jean-Philippe Dubernard, Gérard Duchamp, Jason Eisner, Jacques Farré, Marianne Flouret, Tamás Gaál, Franck Guingne, Hadrien Jeanne, André Kempe, Éric Laugerotte, Jean-Francis Michon, Ludovic Mignot, Florent Nicart, Faissal Ouardi, Thomas Paranthoën, Jean-Luc Ponty, Are Uppman, and Djelloul Ziadi.

Next let us come to more graph theoretical properties and measures that are relevant in the network analysis community. The obvious measures of a graph are the number of nodes n and the number of (undirected) edges m. Both measures give rise to the density, which is defined as $d = 2m/(n(n-1))$. The CIAA-CIAA co-authorship network with $n = 839$ nodes and $m = 1077$ edges has density $d = 0.0031$. The values for the ALL-CIAA co-authorship network are $n = 839$, $m = 2150$, and thus $d = 0.0061$. Hence both networks are sparsely connected. Density is a measure in the theory of graphs with limited meaning for real-world networks. Real-world networks are unlike random graph or regular lattices, and, as empirical observation suggests, they are more like small-worlds [2]. Networks of this kind are characterized by at least two main features:

1. The diameter of the network grows logarithmically in the size of the network like in random graphs and
2. the network is highly clustered as it happens in lattices.

By the first property any two nodes can be reached from each other (if they are in the same connected component) using only a few number of steps, even if the network is large. The second trait induces that any two neighbors of a given node have a large probability of being themselves neighbors. In other words the network has the tendency to form tightly connected neighborhoods.

Both studied co-authorship networks are disconnected. The CIAA-CIAA co-authorship network contains 219 connected components, and the giant component is built by 192 nodes, which is approximately 22.76% of the whole graph. The ALL-CIAA co-authorship network contains 73 connected components, and the giant component has 696 nodes (83.35%).

Further basic concepts of graph theory are the diameter and the clustering coefficient. The diameter and the clustering coefficient can be found in Table 1 and they are defined as follows: the *diameter* is the maximum eccentricity of the nodes of a graph G. Here the *eccentricity of a node* v of G is the maximum distance from a given node v to all other nodes in the graph G. The *periphery* this is the set of nodes whose eccentricity is equal to the diameter. For the CIAA-CIAA co-authorship network the periphery is the set that contains Mohamed Faouzi Atig and Antonio Restivo, while for the ALL-CIAA co-authorship network the members of the periphery are Juan Otero Pombo, Leandro Rodríguez Liñares, Gloria Andrade, Niels Bjørn Bugge Grathwohl, Ulrik Terp Rasmussen, Lasse Nielsen, and Kenny Zhuo Ming Lu. The diameter of the giant component in CIAA-CIAA network and the ALL-CIAA network is 22 and 12, respectively. The *clustering coefficient* C for a graph G with vertex set V is the average

$$C = \frac{1}{n}\sum_{v\in V} c_v,$$

where n is the number of nodes of G and c_v is defined as the fraction of possible triangles through that node that exist,

$$c_v = \frac{2T(v)}{\mathit{deg}(v)(\mathit{deg}(v)-1)},$$

where $T(v)$ is the number of triangles through node v and $\mathit{deg}(v)$ is the degree of the node v. The clustering coefficient of the CIAA-CIAA network and the ALL-CIAA network is 0.55 and 0.49, respectively. The obtained values are in correspondence to previous empirical results for diameters and clustering coefficients obtained from real-world co-authorship networks [14].

In order to identify the most influential individuals in (small-world) networks one may take a closer look on the measure of betweenness. Loosely speaking betweenness is an indicator who bridges the flow of information between most others. In the literature one can find several competing definitions of betweenness, see, e.g., [5], which cover different aspects of being important. In our analysis we rely on the following definition: the *betweenness*, or *betweenness centrality*, of a node v in the graph G with vertex set V is the sum of the fraction of all-pairs shortest paths that pass through v, namely

$$c_B(v) = \sum_{s,t\in V} \frac{\sigma(s,t\mid v)}{\sigma(s,t)},$$

where V is the set of nodes, $\sigma(s,t)$ is the number of shortest (s,t)-paths, and the value $\sigma(s,t\mid v)$ is the number of those paths passing through some node v other than s or t. If $s=t$, then $\sigma(s,t)=1$, and if $v\in\{s,t\}$, then $\sigma(s,t\mid v)=0$. The first column of Table 3 summarize our findings on betweenness.

Now let us come to the strength of collaboration. Cooperation in co-authorship networks is measured in several different ways in the literature [19, Chapter 5]. We will only consider two measures that can be seen as counterparts to the number of papers and the number of fractional papers that are assigned to the authors (nodes of the graph). The easiest way is to assign a weight to a pair of co-authors, which is an edge in the co-authorship graph, is to use the number of

	betweenness ($\times 10^{-2}$)		collaboration weight (straight)		collaboration weight (Newman)	
CIAA-CIAA	2.98	Stavros Konstantinidis	9	Nelma Moreira/Rogério Reis	6.00	Markus Holzer/Martin Kutrib
	2.91	Lila Kari	8	Markus Holzer/Martin Kutrib	4.25	Cyril Allauzen/Mehryar Mohri
	2.51	Galina Jirásková	7	Martin Kutrib/Andreas Malcher	4.25	Martin Kutrib/Andreas Malcher
	2.48	Juraj Sebej	7	Sylvain Lombardy/Jacques Sakarovitch	4.17	Sylvain Lombardy/Jacques Sakarovitch
	2.45	Kai Salomaa	7	Kai Salomaa/Sheng Yu	3.75	Martin Kutrib/Matthias Wendlandt
	2.40	Markus Holzer	6	Cyril Allauzen/Mehryar Mohri	3.75	Nelma Moreira/Rogério Reis
	2.33	Yo-Sub Han	6	Martin Kutrib/Matthias Wendlandt	3.67	Kai Salomaa/Sheng Yu
	2.30	Michal Hospodár	5	Jean-Marc Champarnaud/Djelloul Ziadi	3.50	Yo-Sub Han/Sang-Ki Ko
	2.07	Derick Wood	4	Cyrill Allauzen/Michael Riley	2.83	Jean-Marc Champarnaud/Djelloul Ziadi
	1.81	Jean-Luc Ponty	4	Jurek Czyzowicz/Wojciech Fraczak	2.75	Cyril Allauzen/Michael Riley
ALL-CIAA	6.49	Bruce W. Watson	223	Luiza de Macedo Mourelle/Nadia Nedjah	151.87	Luiza de Macedo Mourelle/Nadia Nedjah
	5.36	Andreas Maletti	184	Shunsuke Inenaga/Masayuki Takeda	69.33	Martin Kutrib/Andreas Malcher
	5.14	Moshe Y. Vardi	181	Hideo Bannai/Shunsuke Inenaga	63.12	*Sanjay Jain*/Frank Stephan
	4.79	Axel Legay	167	Bin Ma/*Haizhou Li*	59.90	Luca Aceto/Anna Ingólfsdóttir
	4.61	Juhani Karhumäki	166	Hideo Bannai/Masayuki Takeda	59.76	*Krishnendu Chatterjee*/Thomas A. Henzinger
	4.13	Markus Holzer	149	Luca Aceto/Anna Ingólfsdóttir	55.98	Markus Holzer/ Martin Kutrib
	4.07	Sheng Yu	141	*Ajith Abraham*/Václav Snásel	51.50	Shmuel Tomi Klein/Dana Shapira
	4.02	Jean-Marc Champarnaud	136	*Pavel Krömer*/Václac Snásel	51.50	Shunsuke Inenaga/Masayuki Takeda
	3.93	Jeffrey O. Shallit	132	*Jan Platos*/Václav Snásel	50.43	*Pavel Krömer*/Václav Snásel
	3.82	Sebastian Maneth	132	*Sanjay Jain*/Frank Stephan	49.18	Bin Ma/*Haizhou Li*

Table 3. The authors with the highest betweenness, the strongest straight collaboration weight, and the strongest Newman's collaboration weight in the two co-authorship networks. Non-CIAA-authors are italicized.

commonly co-authored papers. This measure is called the *straight collaboration weight*. A more complex measure also takes other co-authors into account—we refer to this measure as *Newman's collaboration weight*. For a pair of co-authors it is defined as the sum over all co-authored papers of $1/(n-1)$, where n is the number of collaborators of the paper under consideration in the summing. The idea behind the choice of the value $1/(n-1)$ is that the researchers divide their time equally between the $n-1$ co-authors. Observe, that Newman's collaboration weight does not take into account the actual order in which the names appear in a publication. This is a reasonable assumption for computer science publications, since there are only around 130 CIAA publications that don't list the authors lexicographically. This is approximately 18.90 percent. The obtained results for both co-authorship networks are depicted in in the second and third column of Table 3.

An important issue for real-world networks is the identification and extraction of meaningful communities in order to better understand complex networks. Common to all community definitions [19] is the idea that a community is a group of densely interconnected nodes that are only sparsely connected with the rest of the network. On the variety of algorithms, the community detection algorithm from [6] based on a measure called modularity performed best, in the sense that the identified communities fit very well with the given data. Figure 5 illustrates the result of this algorithm running on the giant component of the CIAA-CIAA co-authorship network. Overall 12 communities are detected, of sizes (in decreasing order) 28 (blue), 22 (green), 22 (red), $19, 19, 16, 13, 12, 12, 11, 10$, and 8. There are 72 nodes contained in the three largest communities. This is 37.5%, and thus more than a third, of the giant component of the CIAA-CIAA co-authorship network. Take a closer look at the largest community . There is an

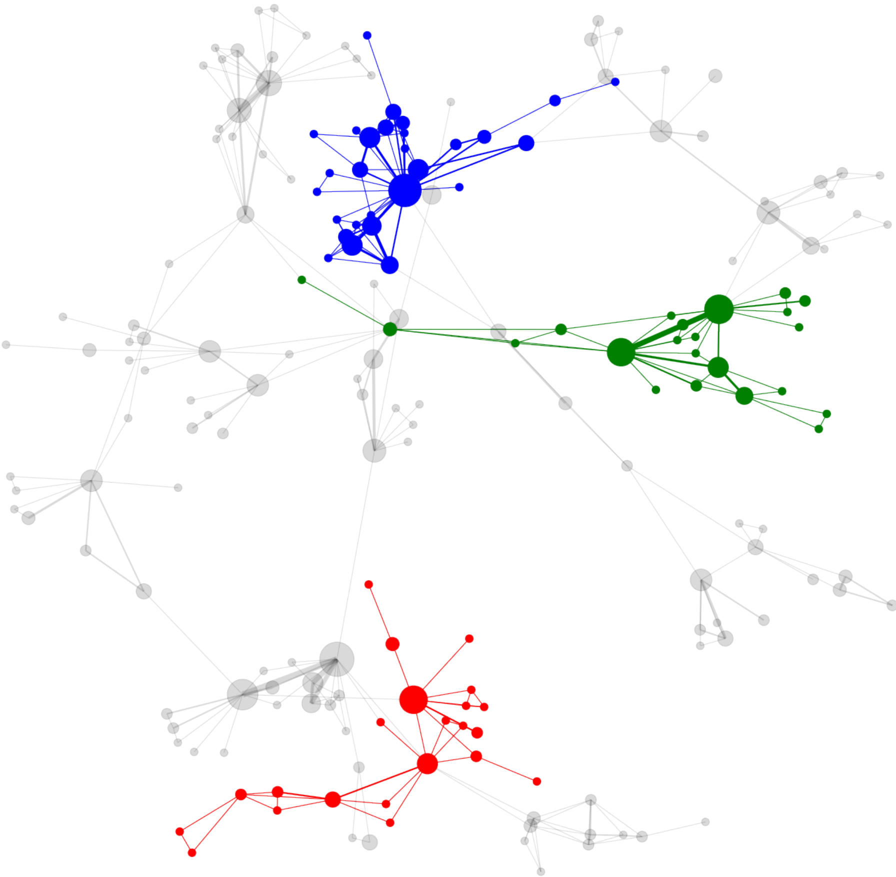


Fig. 5. The giant component of the CIAA-CIAA co-authorship network, which contains 192 authors; names are not shown in order to keep the drawing readable. There are 647 more authors in smaller components. Application of the Clauset-Newman-Moore community structure algorithm produces 12 communities, where the three size-largest ones (top, middle, bottom) are shown by colors (blue, green, red). Node size corresponds to the number of papers and edge width to collaboration weight.

eye-catching node with high degree. An educated guess is that this node stands for Jean-Marc Champarnaud. The analysis confirms this—the authors that from the largest community are Jean-Marc Champarnaud, his 22 collaborators already mentioned earlier, except Jacques Farreé, and (in alphabetical order) Houda Abbad, Samira Attou, Christof Baeijs, Dominique Geniet, Gaëlle Largeteau, and Clément Miklarz.

The presented results can be seen as a starting point for more complex analyses, including, e.g., analysis of the growth of the co-authorship network over time, analysis of the citation network, text analytics and natural language processing (NLP) to cluster research texts, etc. Let us close with congratulations to CIAA and all the best for the coming 25 years.

References

1. Adamic, L.A., Huberman, B.A.: The nature of markets in the world wide web. Quarterly Journal of Electronic Commerce **1**, 512 (2000)
2. Amaral, L.A.N., Scala, A., Barthélémy, M., Stanley, H.E.: Classes of small-world networks. Proceedings of the National Academy of Sciences **97**(21), 11149–11152 (2000)
3. Auerbach, F.: Das Gesetz der Bevölkerungskonzentration. Petermanns Geographische Mitteilungen **59**, 74–76 (1913)
4. Barabási, A.L., Pósfai, M.: Network Science. Cambridge University Press, Cambridge, UK (2016)
5. Brandes, U.: On variants of shortest-path betweenness centrality and their generic computation. Social Networks **30**(2), 136–145 (2008)
6. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. Physical Review E **70**(6), 066111 (2004)
7. Grossman, J.W., Ion, P.D.F.: On a portion of the well-known collaboration graph. Congressus Numerantium **108**, 129–132 (1995)
8. Gutenberg, B., Richter, R.F.: Frequency of earth-quakes in california. Bulletin of the Seismological Society of America **34**, 185–188 (1944)
9. Kim, J.: Evaluating author name disambiguation for digital libraries: a case of DBLP. Scientometrics **116**(3), 1867–1886 (2018)
10. Lotka, A.J.: The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences **16**(12), 317–324 (1926)
11. Lu, E.T., Hamilton, R.J.: Avalanches of the distribution of solar flares. Astrophysical Journal **380**, 89–92 (1991)
12. M.-Ch.Müller, Reitz, F., Roy, N.: Data sets for author name disambiguation: an empirical analysis and a new resource. Scientometrics **111**(3), 1467–1500 (2017)
13. Nascimento, M.A., Sander, J., Pound, J.: Analysis of SIGMOD's co-authorship graph. SIGMOD Record **32**(3), 8–10 (2003)
14. Newman, M.E.J.: The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences **98**, 404–409 (2001)
15. Newman, M.E.J.: Coauthorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Sciences **101**(suppl_1), 5200–5205 (2004)
16. Newman, M.E.J.: Network: An Introduction. Cambridge University Press, Cambridge, UK (2010)
17. Raymond, D.R.: WIA and the practice of theory in computer science. In: Raymond, D.R., Wood, D., Yu, S. (eds.) Automata Implementation, First International Workshop on Implementing Automata, WIA '96, London, Ontario, Canada, August 29-31, 1996, Revised Papers. LNCS, vol. 1260, pp. 1–5. Springer (1996)
18. de S. Price, D.J.: Networks of scientific papers. Science **149**, 510–515 (1965)
19. Savić, M., Ivanović, M., Jain, L.C.: Complex Networks in Software, Knowledge, and Social Systems. Springer, Cham, Switzerland (2019)