Evolution of the CAiSE Author Community: A Social Network Analysis

Matthias Jarke, Manh Cuong Pham, and Ralf Klamma

Abstract The CAiSE community has always prided itself as more than just a normal conference – a successful social network with a very special culture. In this chapter, we apply formal social network analysis to study this community and its evolution of its first quarter-centennial of existence. Using a methodology and dataset developed for an analysis of Computer Science as a whole, we demonstrate the unusual positioning of CAiSE as a quasi-interdisciplinary conference between several sub-disciplines of Computer Science. We show that under an evolution model developed in our research CAiSE pursues a very successful and promising path, and we identify key topics and key players among the CAiSE authors. As the social network analysis focusses on formal aspects such as co-authorship and citations, we unfortunately must leave out one of the undoubtedly most critical success factors: the fun of being in the CAiSE community.

1 Introduction

The CAiSE community, as the community of other scientific conferences, can be considered as a community of practice (CoP) [13]. A community of practice is defined as "*a group of people who share a concern, a set of problem, or a passion about a topic, and who deepen their knowledge and expertise in this area by interacting on an ongoing basis*" [3]. CAiSE is a community of practice due to several aspects. First, members of CAiSE are working on a common research area, the Information Systems. Second, members are distributed across disciplines, which include information systems, database, requirement engineering, business process management, etc. Members are also distributed across organizations, cultures and

M. Jarke (✉) · M.C. Pham · R. Klamma
Information Systems and Databases, RWTH Aachen University, Ahornstr. 55, D52056, Aachen, Germany
e-mail: jarke@dbis.rwth-aachen.de; pham@dbis.rwth-aachen.de; klamma@dbis.rwth-aachen.de

J. Bubenko et al. (eds.), *Seminal Contributions to Information Systems Engineering*,
DOI 10.1007/978-3-642-36926-1_2,

geographical regions. Third, members communicate with each other via face-to-face conferences as well as technology-enhanced interaction. Finally, CAiSE attracts not only fundamental research, but also practical systems and architectures. That results in a very heterogeneous community where methods from different disciplines are used and practices are built on the basic and applied research.

In [12], we have developed a framework for analyzing the development of such scientific communities based on Social Network Analysis (SNA). The framework allows us to monitor the status of a community, qualify its development and compare its development pattern with other communities. It also enables the identification of key members and subgroups of the community. Different techniques are employed in this framework, including visualization, SNA ranking measures, and clustering techniques. Using the DBLP and CiteSeer databases as our data set, we applied this framework to the evolution of Computer Science as a whole. Moreover, we were able to show formally that a few leading computer science conferences are indeed equally important in terms of impact as the top journals in the field, which makes Computer Science quite different from many other disciplines where conference publications only play a marginal role.

In this chapter, we apply this framework to analyze the evolution of the CAiSE conference series. In particular, we are interested in the following questions:

- **Relationship with other communities**: what is the relationship between CAiSE and other communities in the field? What is the role of CAiSE to those communities?
- **Membership of CAiSE**: how do members come and stay in CAiSE? How is the community stabilized?
- **Connectivity**: how do members connect to each other? Does the connectivity grow over time? What is the pattern of the connections?
- **Topic analysis**: what topics are addressed by CAiSE community? How do topics connect to each other? who are the key researchers with the highest impact?

The rest of the chapter is organized as follows. In Sect. 2, we describe our analytical framework and the data we used in the analysis. Section 3 presents the results which aim to answer the above questions. The chapter finishes with a discussion and conclusion.

2 Methods and Data

Our general study of the evolution of digital libraries in general, and of computer science in particular [9] has resulted in a model to explain the community-building process, as well as the co-authoring and citation behavior in conferences and journals [12]. For example, a study of Technology Enhanced Learning research communities found interesting development patterns [11]. In this section, we describe this model and its underlying formal metrics as well as the data set we used for the analysis of CAiSE.

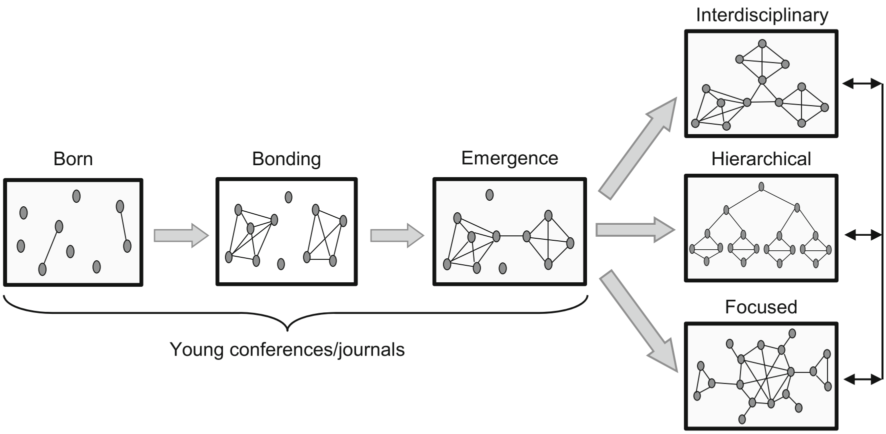

Fig. 1 The development model for scientific communities

Readers with a deeper methodological interest can also consult the Ph.D. thesis [10]. Moreover, an online version of the AERCS[1] analysis system by which the results in the thesis and in this paper were derived, is accessible for experimental use. AERCS does not just support the kind of long-term SNA we show in this paper, but also offers a component for mobile context-dependent advice to attendees of specific conferences.

2.1 The Development Model

Our basic evolution model depicted in Fig. 1 includes four stages: born, bonding, emergence, with the final stage being either interdisciplinary, hierarchical or focussed. Following earlier research in scientific community network analysis, the network employs two types of links: co-authorship and citation. The *co-authorship* subnet of a conference series consists of authors as nodes. There is an edge between two authors if they have co-authored at least one paper published in a conference event in that series. In the born phase, we typically find few connections between authors. After some events, author groups become apparent in the bonding phase. In the best case, they gradually integrated through joint publications from more than one group (emergence phase). Finally, successful conference series typically forms a network topology that features a strongly connected core group of authors that is connected to other smaller groups (*focused* topology). Alternatively, the co-authorship can develop into an interdisciplinary topology where several groups are connected via some gatekeepers, but where there is no core group. Or there

[1] http://bosch.informatik.rwth-aachen.de:5080/AERCS/

might emerge a hierarchical topology which exposes some "super gatekeepers" who connect a hierarchy of groups.

Time series analysis. To quantitatively characterize the development process of a community according to this development model, we apply time series analysis on the networks to reveal six parameters over time: densification law, clustering coefficient, maximum betweenness, largest connected component, diameter, and average path length. These parameters enable us to explain the community building process in Fig. 1. To interpret the shape of the community, one needs to use a combination of all of these parameters.

Formally, given the network $G = (V, E)$, where V is the set of vertices or nodes, and E is the set of edges, these network metrics are defined as follows:

- *Densification law*: [4] discovered that complex networks densify over time, with the number of edges growing super-linearly with the number of nodes, meaning that the average degree (i.e., number of edges) of the nodes is increasing. The densification follows a power-law pattern: $e(t) \propto n(t)^{\alpha}$, where $e(t)$ and $n(t)$ are the number of edges and nodes at time t, respectively, and α is an exponent that lies between 1 and 2 ($\alpha = 1$ corresponds to constant average degree over time, while $\alpha = 2$ corresponds to very dense graph where on average each node has edges to a constant fraction of all nodes). We use this exponent to differentiate the "speed" by which networks are densified.
- The *clustering coefficient* of a network [6] is defined as the total number of pairs of vertices that have a common neighbor and are themselves connected, divided by the total number of pairs of vertices that have a common neighbor:

$$C = \frac{3 \times \textit{ number of triangles in the graph}}{\textit{number of connected triples of vertices in the graph}} \tag{1}$$

  Intuitively, during the born phase, the clustering coefficient is low, since nodes are unconnected with each other. In the bonding phase, the clustering coefficient tends to increase quickly as nodes are clustered into very dense, yet unconnected components. When the unconnected components subsequently start to connect with each other, the clustering coefficient drops and stays relatively stable after some time.
- *Betweenness* measures the extent to which a particular node lies between the other nodes in the network:

$$B(u) \equiv \sum_{u \neq i \neq j} \frac{\sigma^{u}(i, j)}{\sigma(i, j)} \tag{2}$$

  where $\sigma(i, j)$ is the number of shortest-paths between nodes i and j, $\sigma^{u}(i, j)$ is the number of shortest-paths between i and j that pass through u. Nodes with high betweenness have more power to control the information flow in the network, and are normally the gatekeepers who connect several dense groups. For the network, the maximum betweenness of all authors is therefore a good indicator of whether there are strong gatekeepers within the network. Maximum betweenness increases when more components become connected (emergence stage) and continues to increase when the network develops toward a hierarchical or interdisciplinary topology. However, maximum betweenness will achieve a stable value when the network is at focused stage.

- *Largest connected component* (or giant component) measures the fraction of nodes that are connected with each other in the largest sub-network. As observed in Fig. 1, this fraction is small in the first two phases, and gradually increases as authors from different sub-networks connect with each other. It achieves a stable state when the fraction of nodes that connect to the largest component is equal to the fraction of new nodes that stay unconnected from the largest component.
- *Diameter* is the length of the greatest geodesic distance (i.e., the length of the longest shortest path) between any two nodes. Intuitively, in the beginning, the diameter is small, and then it increases. After some time, the diameter starts to shrink as new edges between existing nodes continue to be added. If the network develops toward a tree-like topology (hierarchical stage), the diameter will be larger than in the focused or interdisciplinary topologies.
- *Average path length* is the average length of all the shortest paths in the network. Clearly, during the first two phases, the average path length is small and increases when the network grows. In general, the average path length of a hierarchical network is larger than that of the other two topologies.

In summary, for the co-authorship network, the emergence of the giant component indicates the cohesiveness of collaboration within the community, while the betweenness shows the existence of gatekeepers and their importance. The clustering coefficient measures the extent to which the community is clustered into sub-communities. Other parameters such as diameter and average shortest path length, show whether the community is still developing or whether it is stable. For the citation network, combining these parameters helps to understand the interdisciplinarity of a conference.

2.2 Data: DBLP and CiteSeerX

The data set used in our study integrates the DBLP and CiteSeerX digital libraries. DBLP is a computer science bibliography, which also includes publications in interdisciplinary areas of computer science. We retrieved the publication lists of conferences from DBLP. However, DBLP does not record citations. Therefore, we used CiteSeerX to fill the citation list of publications in DBLP. DBLP data, as downloaded in July 2012, consists of 1,138,661 authors, 1,947,188 publications, 3,217 conference series and 1,193 journals. CiteSeerX data was downloaded in March, 2011, which includes 9,121,166 publications, 22,735,140 references and over 6 million author names. We combined DBLP and CiteSeerX using the canopy clustering technique [5]. Overall, the matching algorithm gave us 864,097 pairs of matched publications. From those data sets, we created the co-authorship and citation networks for our analysis. The co-authorship network is created based on DBLP data and the citation network is formulated by the combined DBLP and CitaSeerX data.

3 Development of CAiSE Community

In this section, we present the analytical results of CAiSE community, concerning the questions we posed in the introduction. We work inside out, starting with the positioning of CAiSE within Computer Science, then proceeding to the evolution pattern of CAiSE with respect to the development model of Sect. 2, and end with the internal structure of CAiSE concerning its main topics and its key players.

3.1 The Position of CAiSE in Computer Science

Our general study of the evolution of the Computer Science community [12] showed how the field has evolved a coherent giant component with clearly demarked subfields that have more or less strong citation interactions with each other (see Fig. 2); for example, Theoretical Computer Science interacts, albeit somewhat loosely, with almost all other areas. An extract from this map (see Fig. 3) shows that CAiSE can be seen as a kind of interdisciplinary gateway between neighboring research areas such as information systems, databases, software engineering, data mining and knowledge management, conceptual modeling, process modeling and world wide web. Overall, 237 conference and journals have at least 50 authors who also published in CAiSE. Table 1 lists the top 10 among them. Many other established conferences/journals also have common authors with CAiSE, such as SIGMOD Record (235 common authors), TKDE (212), TSE (191), ACM SIGMOD (183), CACM (156), VLDB Journal (127) and IJCAI (105). This demonstrates the diversity of CAiSE community membership, and its interdisciplinary nature.

The standing of CAiSE within the computer science community can also be assessed by ranking it in the citation network according to the centrality measures discussed in the previous section. The data set contains a total of 455 conferences in the fields of databases, data mining, and software engineering which are close to CAiSE in the graph. Among these 455 conferences, which include all the traditional top conferences of these fields, CAiSE is among the top 8 % in terms of PageRank [8] and the top 5 % in terms of authority [2], which is already quite good, but among the top 2 % in terms of betweenness. Thus, CAiSE is not just highly interdisciplinary but also an important bridge among the other fields and even a strong authority for its kinds of results.

Fig. 2 The map of Computer Science (giant component)

The interdisciplinarity of CAiSE can be shown in more detail by citation data, which indicate the knowledge exchange between publications. In detail, CAiSE publications have been cited by 472 conferences and journals. CAiSE publications have cited publications from 689 conferences and journals. Tables 2 and 3 list top conferences and journals who cited CAiSE or were cited by CAiSE. Note that the citation data we extract from CiteSeerX are incomplete and only cover conferences and journals indexed by DBLP. Therefore, the actual number of conferences and journals referenced to and cited by CAiSE is somewhat bigger.

To summarize, CAiSE is a very interdisciplinary conference, shown by the diversity in its membership as well as the citation data.

3.2 Evolution of Author Community Membership

The next question we want to address is how CAiSE community evolves over time. We apply the model described in Sect. 2 on the co-authorship network to analyze its development pattern. To make our analysis more meaningful, we compare the

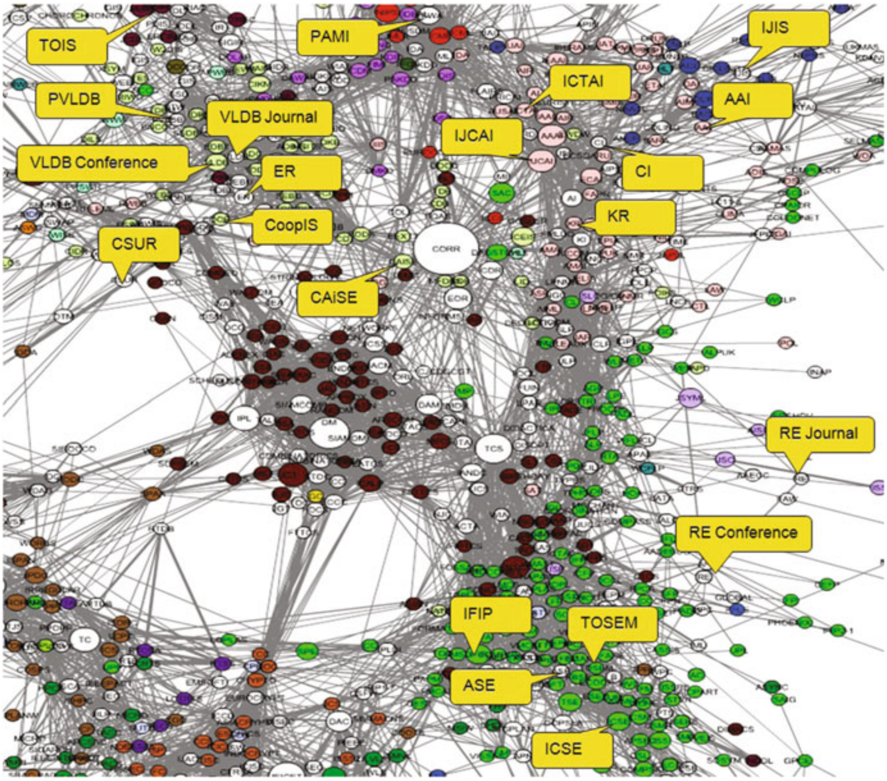

Fig. 3 The position of CAiSE in the map of Computer Science

Table 1 Top overlapping conferences/journals with CAiSE

	Name of conferences/journals	#Common authors
1	International Conference on Conceptual Modeling (ER)	624
2	OTM Conferences/Workshops	435
3	Information Systems	421
4	Data and Knowledge Engineering	358
5	Business Process Management	317
6	International Conference on Data Engineering (ICDE)	305
7	Very Large Data Bases (VLDB) Conference	290
8	International Conference on Information and Knowledge Management(CIKM)	281
9	International World Wide Web Conferences (WWW)	264
10	Information and Software Technology	256

evolution of CAiSE with that of the three well-known conferences at the top of Table 3: ER, VLDB, and ICDE. Some basic authorship data for these conferences is summarized in Table 4.

Table 2 Top conferences/journals who cited CAiSE

	Name of conferences/journals	#
1	International Conference on Conceptual Modeling (ER)	84
2	Business Process Management	52
3	Data and Knowledge Engineering	41
4	OTM Conferences / Workshops	33
5	Information Systems	28
6	Requirements Engineering	26
7	International Conference on Cooperative Information Systems (CoopIS)	22
8	Semantic Web	22
9	International Conference on Service Oriented Computing	22
10	Very Large Data Bases (VLDB) Conference	21

Table 3 Top conferences/journals cited by CAiSE

	Name of conferences/journals	#
1	Very Large Data Bases (VLDB) Conference	142
2	Communications of the ACM (CACM)	139
3	International Conference on Conceptual Modeling (ER)	132
4	IEEE Transactions on Software Engineering (TSE)	107
5	International Conference on Data Engineering (ICDE)	97
6	International Conference on Software Engineering (ICSE)	87
7	ACM SIGMOD Conference	87
8	Information Systems	82
9	Requirements Engineering	80
10	Data and Knowledge Engineering	78

Table 4 Data summary of ER, ICDE, VLDB and CAiSE conferences

Conference series	Events	#Authors	#Papers
International Conference on Conceptual Modeling (ER)	1979–2011	2,997	1,945
International Conference on Data Engineering (ICDE)	1984–2011	5,886	3,683
Very Large Data Bases (VLDB) Conference	1975–2010	3,660	2,397
Conference on Advanced Information Systems Engineering (CAiSE)	1990–2012	3,129	1,876

We begin with a simple analysis of the number of published papers over time. Figure 4a plots the absolute numbers of authors and papers of CAiSE over years. In general, the numbers of authors and papers increase over years, with a significant increase in 2002 and drop in 2007; the latter is obviously due to the decision of the steering committee at that time to reduce the acceptance rate sharply. A view into the individual proceedings shows that the number of submissions continued to increase, such that CAiSE nowadays attracts very high numbers of submissions despite acceptance rates that are among the toughest in the IS area.

Next, we study the distribution of authorship intensity, i.e. the number of CAiSE conferences authors have published in, and the number of papers they have written for CAiSE. Figure 4b plots this distribution in log-log axes. The distribution of

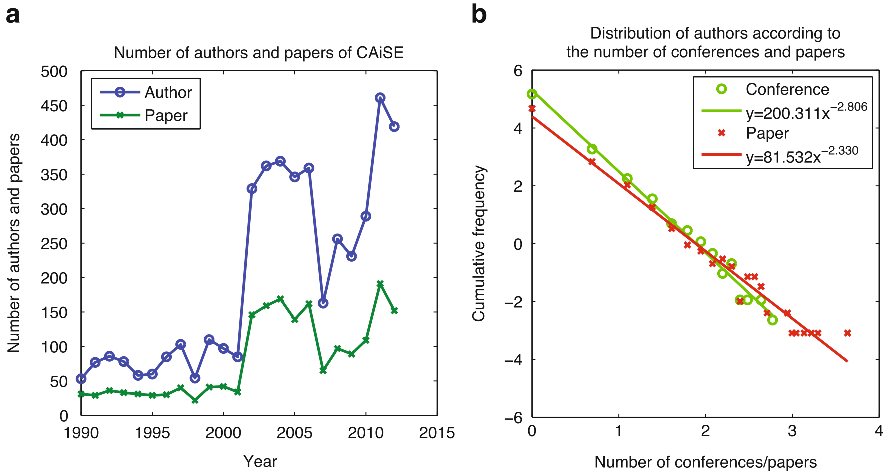

Fig. 4 Number and distribution of authors of CAiSE over years

authors according to the number of conferences and papers follows Power-law distributions with the exponent α equals to 2.806 and 2.33, respectively. This indicates that there is a "tail" of authors who significantly contribute to CAiSE despite appearing there only once. In detail, 79 % of authors contributed only to one conference, while 21 % contributed to at least 2 conferences and 94 authors (3 %) contributed to at least 5 conferences. In term of the number of papers, 76 % of authors contributed only 1 paper and 141 authors (about 4 %) contributed at least 5 papers to CAiSE.

To investigate the contributions of returning authors to CAiSE over the years, we calculated two measures in comparison to our three benchmark conferences: the rate of recurring authors and their publications over years. A paper is published by *recurring authors* if at least one of its authors has published in the previous conference. A high rate of recurring authors, together with a low rate of papers by recurring authors, indicates that recurring authors mainly collaborate with each other (one paper has more recurring authors). On the other hand, a high rate of recurring authors, together with a high rate of papers by recurring authors, indicates that recurring authors collaborate mainly with new authors, which contributes to community development. Those two measures allow us to assess one important principle to cultivate scientific communities [13]. On the one hand, a community needs to retain the authors in order to establish and keep the old ideas. On the other hand, it also needs to attract new authors who probably will bring new ideas.

In Fig. 5, we recognize that the basic trend during the early stage of all conferences is to retain authors. The frequency of papers by recurring authors also increased. In the first 11 years, CAiSE retained the authors at a lower rate (around 25 %) in comparison to VLDB, ER and ICDE. After that, CAiSE managed an author recurring at a similar rate as VLDB (around 38 %). Similar observation can be made for the papers by recurring author rate.

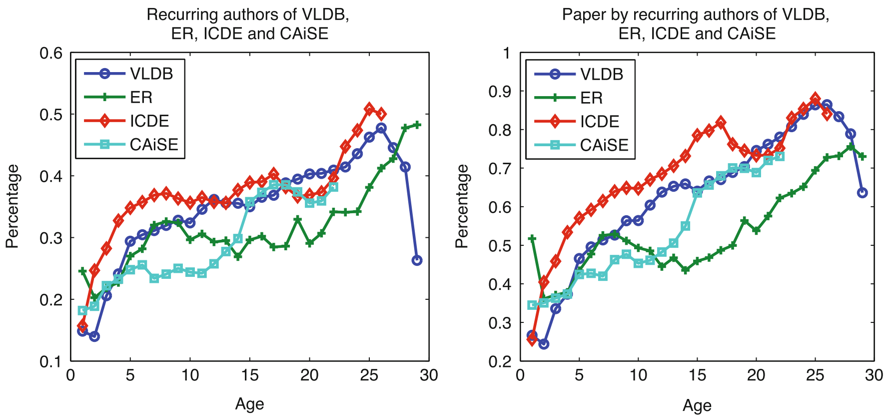

Fig. 5 Recurring authors and papers by recurring authors over years

In summary, CAiSE constantly developed in the last 24 years in term of authors and contribution intensity. There is set of authors who contribute continuously and greatly to CAiSE. Over time, CAiSE manages to not only retain authors who are working on the established ideas of the conference, but also to attract new authors who would bring fresh ideas to the community. A comparison of the returning rating of CAiSE authors and their contributions to other conferences shows that CAiSE now retains a healthy fraction of recurring authors in order to keep the community open.

3.3 The Evolution of Connectivity in CAiSE

Having looked at the phenomena of author activity at the individual level, we are now in the position to look at the question what this means for the shape of the CAiSE author community network as a whole. The basis for this are the co-authorship graphs, and the six network metrics we defined in Sect. 2.

The evolution of these six metrics for VLDB, ER, ICDE and CAiSE is shown in Fig. 6. VLDB, ICDE and CAiSE expose the same evolution pattern but with a slight delay for CAiSE. The maximum betweenness and largest connected component of the co-authorships of VLDB and ICDE started increasing after 10 years, while it took CAiSE 15 years. The ER conference faced an even bigger delay (22 years) which can perhaps be explained somewhat by the very late entry of US research into their community. Note that in our earlier studies we also found conferences where this has never happened, which was typically closely correlated with very low impact in terms of citation. In this sense, all four conferences can nowadays be considered successful. However, the decreasing parameters average shortest path length and diameter over long times (VLDB: 10 years, ICDE: 7 years, CAiSE: 5 years) suggest that these communities are now more stable while the ER

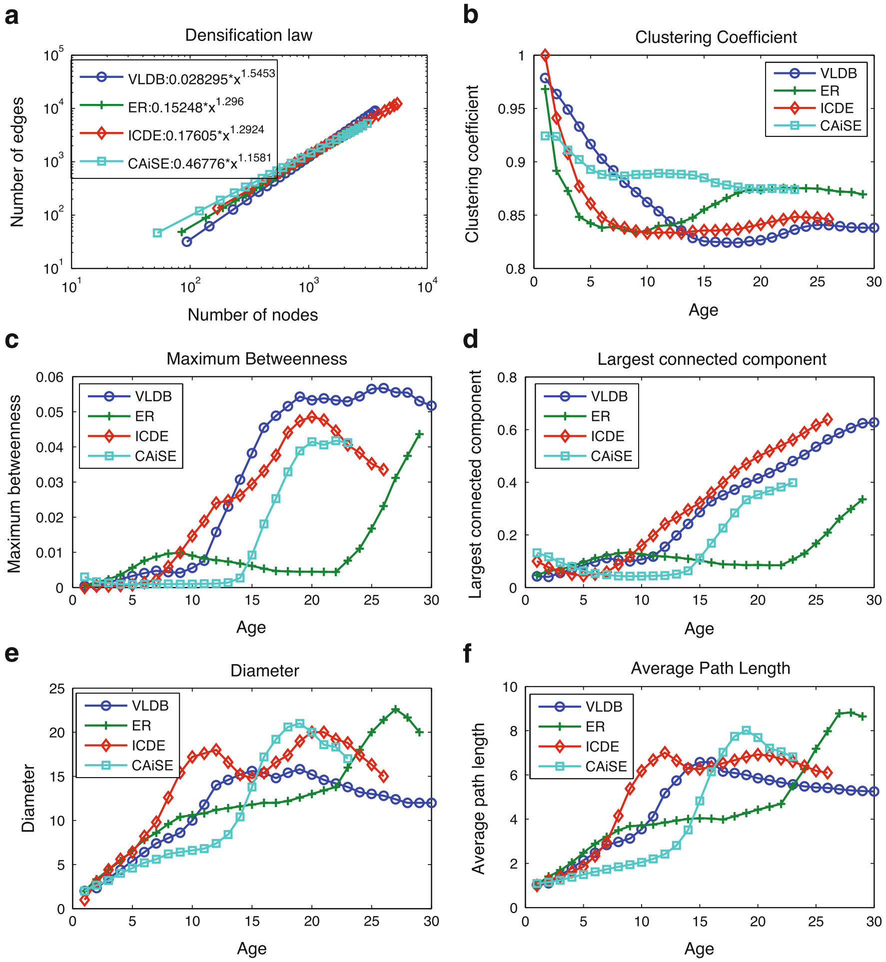

Fig. 6 Co-authorship network parameters of CAiSE and other conferences

community is still developing. Interestingly, the clustering coefficient of CAiSE and ER is higher than for VLDB and ICDE, indicating that CAiSE and ER are clustered in more sub-groups (with many disconnected components).

To summarize, the connectivity of CAiSE community has been increasing significantly over the last 12 years. The community is developing towards a well-connected and cohesive structure. Compared to other established conferences, we see that CAiSE is currently developing as fast as VLDB and ICDE. To illustrate where this might lead in the future, we compare in Figs. 7 and 8 snapshots of the co-authorship networks of CAiSE and VLDB for every fifth year of their respective histories (for CAiSE starting 1989, for VLDB 1975). Obviously, CAiSE and VLDB developed very similarly in the first 25 years, as both of them built their community

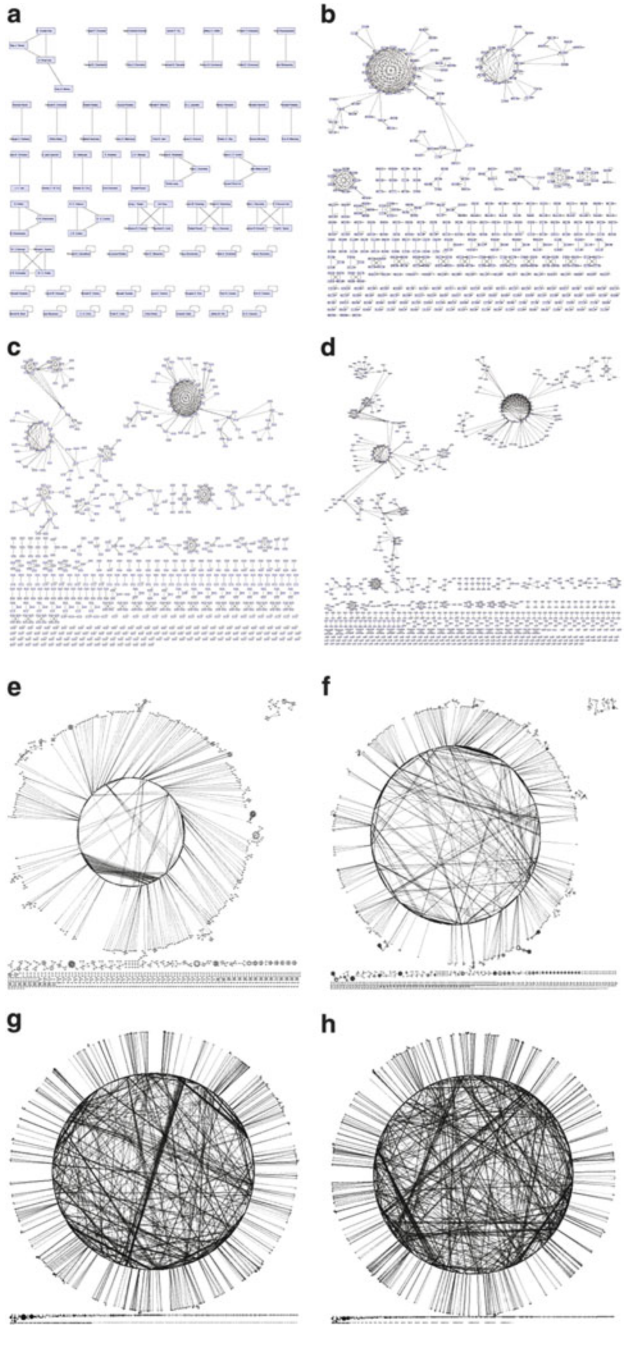

Fig. 7 Development of VLDB co-authorship network. (**a**) VLDB in 1975. (**b**) VLDB in 1980. (**c**) VLDB in 1985. (**d**) VLDB in 1990. (**e**) VLDB in 1995. (**f**) VLDB in 2000. (**g**) VLDB in 2005. (**h**) VLDB in 2010

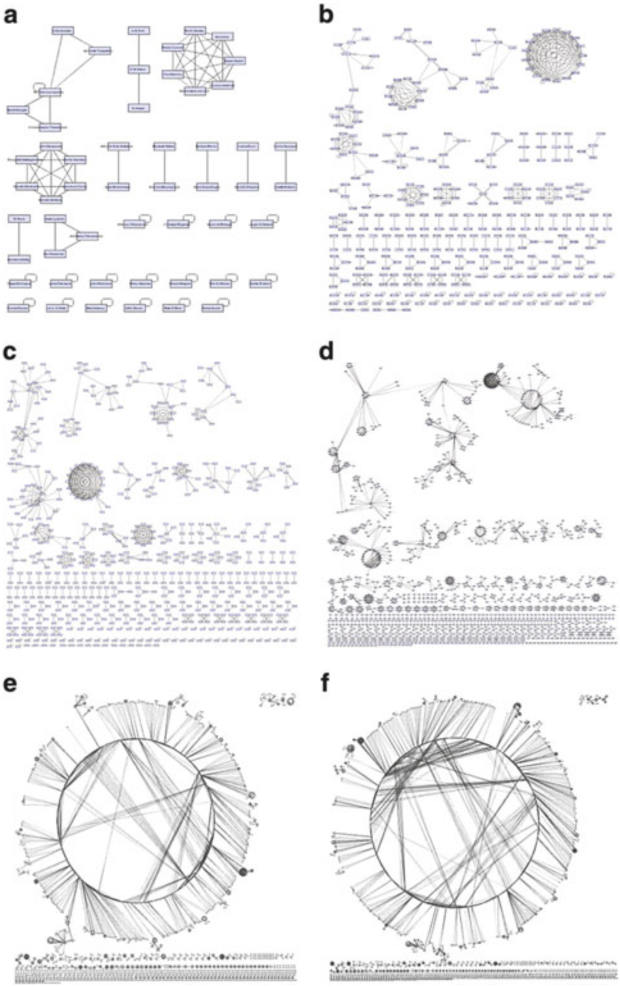

Fig. 8 Development of CAiSE co-authorship network. (**a**) CAiSE in 1990. (**b**) CAiSE in 1995. (**c**) CAiSE in 2000. (**d**) CAiSE in 2005. (**e**) CAiSE in 2010. (**f**) CAiSE in 2012

from a *born* to *bonding*, then *emergence* and finally *focused* topology. However, the last 10 years of VLDB (parts (g) and (h) in Fig. 7) also exhibit a possible danger; they show an ever denser giant component where new authors often can only enter by co-authoring with members of that component. It might then well happen that important new topics are not recognized early enough by the community, a danger that CAiSE has so far well managed to avoid.

The high betweenness centrality and big giant component of the CAiSE co-authorship network suggest that there is an increasing number of active members

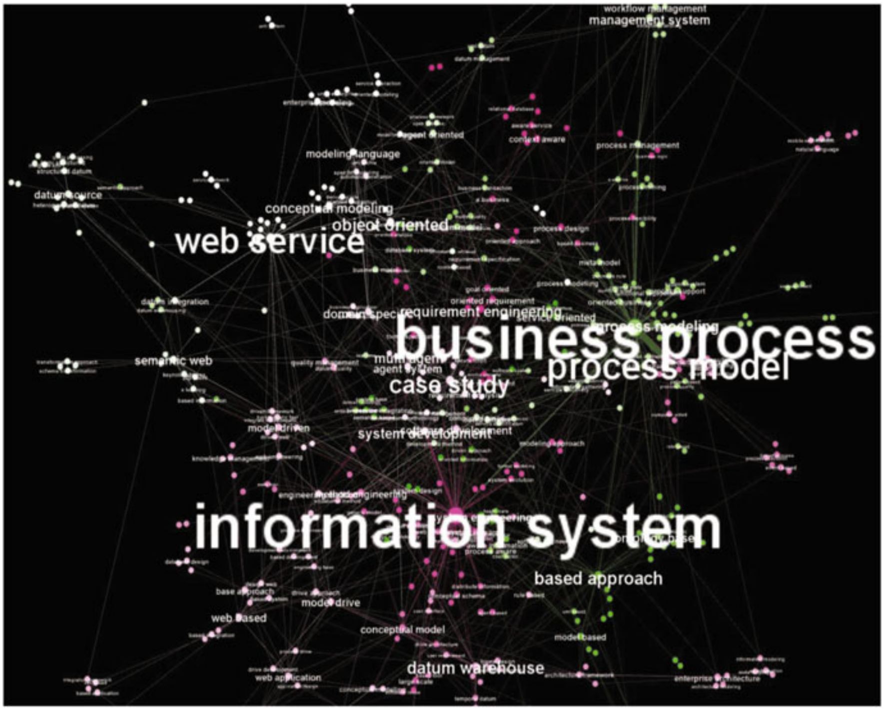

Fig. 9 The visualization of CAiSE keywords

who collaborate across sub-communities. These very active authors form the core of the community which we will identify in the next section.

3.4 Main Themes and Key Members of the CAiSE Author Community

It has been mentioned that CAiSE boasts a somewhat richer structure of sub-communities in its giant components than e.g. VLDB. This can likely be linked the diversity themes studied by these sub-communities. One of the simplest and most popular ways to study this phenomenon automatically is simply the generation of keyword networks based on the paper titles. There is a link between two keywords if they co-occur in at least one paper title. The connection is weighted by the number of co-occurences. We visualize this network using the ForceAtlas layout (see Fig. 9), where the size of nodes and labels denotes the PageRank score of keywords. Nodes are colored according to the clusters detected by a modularity-based clustering algorithm [7].

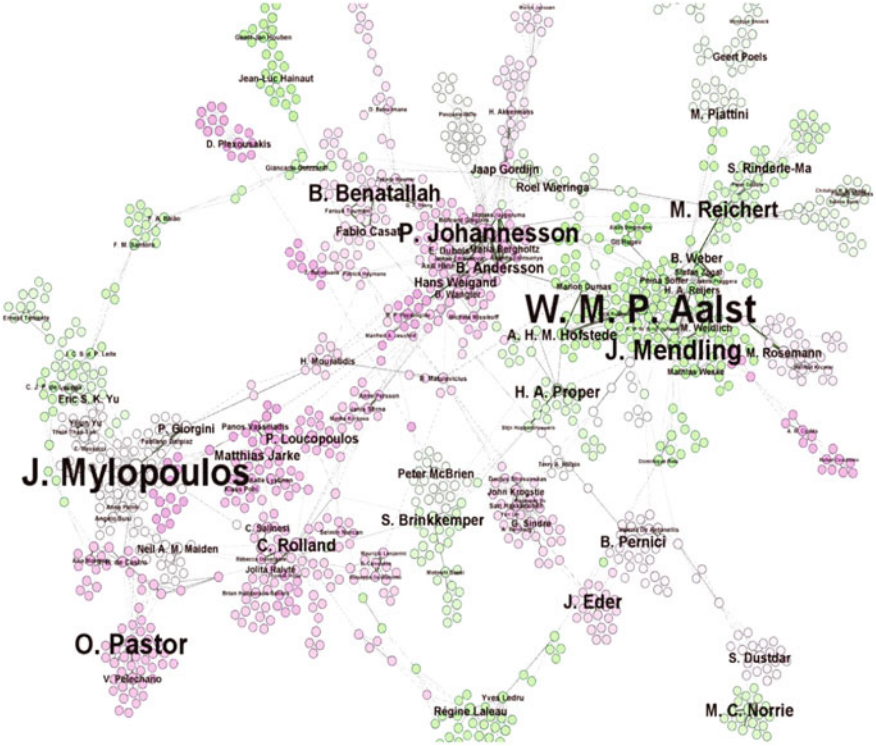

Fig. 10 Co-authorship network (giant component) of CAiSE (as of 2012)

With surprising clarity, the visualization shows two big clusters around the topics of Information System and Bussiness Process. For frequent attendees, it is probably obvious in hindsight that bringing these two themes – one more from Computer Science, the other more from MIS – together is one of the main attractions of CAiSE for members of both communities. Indeed, the other major keywords – Web Service, Process Model, Case Study, Data Warehouse, Object Oriented, Requirement Engineering, Multi Agent, Ontology Based, Conceptual Modeling, Management Systems and Semantic Web – are either closely related to one of these main topics or build some kind of bridge among them.

To investigate the key members in these sub-community and CAiSE as a whole, we applied two SNA ranking measures: the betweenness and the PageRank score of authors in the co-authorship network. The CAiSE co-authorship network in 2012 is given in Fig. 10, using again the ForceAtlas layout [1]. Nodes are colored according to their assignments to sub-communities detected by the modularity-based clustering algorithm [7]. The size of labels and nodes denotes the PageRank [8] score of authors. This visualization shows us the key members not only by their prestige (denoted by PageRank score), but also by their important position in the collaboration network.

Table 5 Top 20 authors by betweenness and PageRank in the co-author network

	Betweenness		PageRank	
1	John Mylopoulos	201,591	John Mylopoulos	0.0035
2	Wil M. P. van der Aalst	183,018	Wil M. P. van der Aalst	0.0034
3	Pericles Loucopoulos	148,551	Oscar Pastor	0.0024
4	Birger Andersson	147,759	Jan Mendling	0.0023
5	Raimundas Matulevicius	145,736	Paul Johannesson	0.0022
6	Benkt Wangler	142,569	Boualem Benatallah	0.0021
7	Marlon Dumas	128,972	Manfred Reichert	0.0020
8	Eric Dubois	97,785	Johann Eder	0.0018
9	Jan Mendling	87,883	Moira C. Norrie	0.0017
10	Paul Johannesson	80,494	Colette Rolland	0.0017
11	Arthur H. M. ter Hofstede	79,558	Henderik Alex Proper	0.0016
12	Haralambos Mouratidis	78,250	Barbara Pernici	0.0016
13	Colette Rolland	76,631	Sjaak Brinkkemper	0.0015
14	Paolo Giorgini	70,613	Birger Andersson	0.0015
15	Guttorm Sindre	69,171	Arthur H. M. ter Hofstede	0.0015
16	Boualem Benatallah	67,998	Schahram Dustdar	0.0015
17	Sjaak Brinkkemper	60,984	Mario Piattini	0.0014
18	Jaap Gordijn	60,481	Stefanie Rinderle-Ma	0.0014
19	Roel Wieringa	59,734	Matthias Jarke	0.0014
20	Manfred Reichert	55,062	Pericles Loucopoulos	0.0014

Interestingly, the topic structure is well reflected in this figure, as each of the themes has a clear "leader", John Mylopoulos in the case of Information Systems, and Wil van der Aalst for Business Process. In addition, the many smaller sub-groups are connected by a set of gatekeepers. For example, Paul Johannesson, Oscar Pastor, Colette Rolland, Jan Mendling, Manfred Reichert, Pericles Loucopoulo, Boualem Benatallah, Johann Eder and Barbara Pernici connect their own sub-groups with many other sub-groups. Those authors together form the core, which ensures the connectivity of the community as a whole. Moreover, it is interesting to note that former students and collaborators of the CAiSE founders and Advisory Committee members Janis Bubenko (most prominently Paul Johannesson), Arne Solvberg (e.g. John Krogstie and Peter McBrien), and more recently Colette Rolland still play important betweenness nodes linking the two main subfields. To provide a bit more detail, Table 5 gives the top 20 authors according to their betweenness and PageRank score in the co-authorship network.

Complementing the co-authorship network, there is of course the citation network which, however, extends far beyond the CAiSE community itself. Indeed, the selection of papers reproduced and commented in this volume was based on such an analysis, taking the ranking of numbers of citations as a starting point. We therefore do not discuss this aspect in this chapter. Suffice it to say that the truly outstanding h indexes of Wil van der Aalst (83) and John Mylopoulos (69) confirm impressively the exceptional role we also saw in the co-authorship network.

4 Conclusion

The Social Network Analysis of the CAiSE Conference author community shows a quite interesting strategic position within the Computer Science discipline, and – after a somewhat slow start – an impressive development towards a conference community that exhibits all the ingredients of success found in earlier success stories such as VLDB: a strong giant component of long-term collaborators with very high impact within and beyond the conference itself, combined with a topical openness and interdisciplinarity that promises sufficient openness for innovation. The long-term visionary but very open and friendly leadership of what is now called the senior Advisory Committee has certainly contributed to this success, as has the small "revolution" of a few junior key players around 2007 that made CAiSE one of the most strictly refereed conference in the field and thus – for the naive perhaps surprisingly – increased not just the prestige and quality, but also the quantity of submissions from several important collaborating sub-areas.

We hasten to admit that our choice of data sources implies some limitations of this study. First, both DBLP and CiteSeer show only author and published paper information about the conference; so our social network is limited to co-authorship, citation, and keywords. It leaves out the very important network of conference organizers but also conference attendees and authors of unsuccessful submissions. Second and perhaps more importantly, especially CiteSeer focuses on Computer Science only, so our analysis of the integrated data set cannot evaluate impact on or by related fields in other disciplines such as Management Information Systems.

Despite these limitations, regular CAiSE participants will find that many of their personal social experiences in CAiSE are reproduced fairly well by even by the co-author and citation analysis we employ. Perhaps – like the first author – they have also faced some interesting surprises in this paper which, however, can be well explained from their deep knowledge of the conference history on second thought. For ourselves, these limitations create the challenge to find and integrate data sources which are less narrow in their view of the IT field, yet – unlike much of the Web of Science – do include information about conferences and their impact. Especially the broad field of Management Information Systems seems in urgent need for such a study, as many of their representatives work in business schools where conference publications are not taken seriously at all.

Acknowledgements This work is supported by the DFG-funded excellence cluster UMIC, the B-IT Research School, and EU Integrated Project Layers.

References

1. Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*, pages 361–362.

2. Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, 46: 604–632.
3. Lave, J. and Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Learning in Doing. Cambridge University Press.
4. Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 177–187, New York, NY, USA. ACM.
5. McCallum, A., Nigam, K., and Ungar, L. H. (2000). Efficient clustering of high-dimensional data sets with application to reference matching. In *KDD '00: Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 169–178, New York, NY, USA. ACM.
6. Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proc.Natl.Acad.Sci.USA*, 98:404.
7. Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, 69:066133.
8. Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. Technical report;, Stanford University.
9. Pham, M. and Klamma, R. (2010). The structure of the computer science knowledge network. In *2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 17–24.
10. Pham, M. C. (2013). *Dynamic Social Network Analysis and Recommender Technologies in Scientific Communities: The Case of Computer Science*. PhD thesis, RWTH Aachen University, Aachen – Germany.
11. Pham, M. C., Derntl, M., Klamma, R., and Jarke, M. (2012). Development patterns of scientific communities in technology enhanced learning. *Educational Technology and Society*, 15(3):323–335.
12. Pham, M. C., Klamma, R., and Jarke, M. (2011). Development of computer science disciplines: a social network analysis approach. *Social Netw. Analys. Mining*, 1(4):321–340.
13. Wenger, E., McDermott, R., and Snyder, W. (2002). *Cultivating communities of practice: a guide to managing knowledge*. Harvard Business School Press.