Affiliation analysis of database publications

David Aumüller, Erhard Rahm
University of Leipzig, Leipzig, Germany
{aumueller, rahm}@informatik.uni-leipzig.de

ABSTRACT

We analyze the author affiliations of database publications to determine the main institutions contributing research results in our field. We consider the publications of the last decade (2000—2009) that appeared in the top conferences SIGMOD and VLDB and in the VLDBJ and TODS journals. We determine the top affiliations in terms of number of papers and aggregate the numbers at the levels of entire countries and continents. Further, we analyze to which degree authors from different affiliations and countries cooperate on jointly authored papers, and study the development over time. We also consider the number and size of affiliations of different countries.

1. INTRODUCTION

Previous bibliographic studies of computer science and database publications mainly focused on the number of papers and citations per author or per venue as well as co-authorship relations [4, 6-9]. However, there has been very little analysis of the affiliation of authors to determine where research results are produced. Commercial bibliography services such as Elsevier Scopus and Thomson Web of Science provide some affiliation information but are still mainly limited to journals. By missing most conferences they do not sufficiently cover the computer science research literature. In [8] the affiliations of database publications have already been evaluated but based on a largely manual effort. The study only considered first author affiliations of publications with more than 20 citations at the time of the evaluation.

In this paper we present a more comprehensive affiliation analysis (considering all authors) based on a largely automatic determination of author affiliations. Determining the affiliation information automatically is quite challenging and requires a substantial effort for information extraction, data cleaning and matching heterogeneous representations of the same affiliations. We mainly extracted the affiliation data from bibliographic portal web sites, such as ACM Digital Library and SpringerLink; in some cases we had to extract the information from lists of accepted papers or directly from the fulltext documents. For conferences, we also determined the type of paper (research, industry, demo) which required the integration of further information such as the tables of contents. The collected affiliation strings are highly heterogeneous (frequent use of acronyms and abbreviations, etc.), often inconsistent and partially incomplete (e.g., "Microsoft Research" without city information). Consider for instance the two strings "MIT" and "Department of Mechanical Engineering, Massachusetts Inst. of Technology, MA 02139, Cambridge, USA" referring to the same institution (neglecting department). The pursued approach for entity recognition and affiliation matching (utilizing existing web search engines) is described in [1]. We extract institution and location information from the affiliation strings but ignore departmental information as this information is unstable over time and not always given. Our affiliation information is thus at the level of institution and city from where it can be aggregated at coarser geographic levels such as country and continent.

Our affiliation analysis focuses on database publications of two top database conferences (ACM SIGMOD, VLDB) and two top journals (ACM TODS, VLDB Journal) over ten years (2000 until 2009). These venues are known to be highly selective and of high quality so that (frequent) publications in these venues can be viewed as a quality indicator not only for authors but also their institutes. In this initial study, we will analyze the number of papers of different affiliations and their countries and continents, and study the development over time. We also analyze to which degree authors from different affiliations and countries cooperate on jointly authored papers. Furthermore, we evaluate the number and size of affiliations of different countries. Due to space constraints, we leave an affiliation-based citation analysis for future work. At *dbs.uni-leipzig.de/affiliations* we set up a website to browse the papers.

In the next section, we provide some base statistics on the considered publications. In section 3 we study author affiliations at the levels of continents and countries while section 4 focuses on the top affiliations and the most prolific authors within.

2. BASE DATA

Table 1 provides some base statistics for the considered papers in the four venues (TODS, VLDBJ, SIGMOD, VLDB). It shows the number of papers that appeared in the two journals and the two conference series. The conference papers are further discriminated by track into research, industrial, and demo papers. In the lower part of the table we differentiate the paper counts by first and second five year spans; we will use these two time intervals to illustrate some temporal trends.

In total we analyze the author affiliations for more than 2,700 papers: over 1,900 research papers and more than 800 demo and industrial papers. Slightly more than a quarter of the research publications appeared in the two journals (per year about 50 on average vs. 140 research papers in the two conferences). The number of papers per year almost doubled during the decade (188 in 2001 vs. 352 in 2009); about 60% of the papers appeared in the second half of the decade.

Year	pubs	Jrn.	Conf.	*Conf. track: r, i, d*			res.
2000	188	26	162	*95*	*31*	*36*	121
2001	203	35	168	*103*	*28*	*37*	138
2002	212	32	180	*111*	*32*	*37*	143
2003	225	35	190	*128*	*17*	*45*	163
2004	292	42	250	*150*	*43*	*57*	192
2005	295	54	241	*150*	*38*	*53*	204
2006	276	56	220	*141*	*26*	*53*	197
2007	318	52	266	*175*	*27*	*64*	227
2008	360	91	269	*179*	*30*	*60*	270
2009	352	83	269	*171*	*39*	*59*	254
1^{st} 5y	1,120	170	950	*587*	*151*	*212*	757
2^{nd} 5y	1,596	331	1,265	*816*	*160*	*289*	1,147
Decade	2,716	501	2,215	*1,403*	*311*	*501*	1,904

Table 1: Base data per year and 5-year periods

Most publications have more than one author so that it is of interest to what degree authors of different affiliations and countries publish together. Fig. 1 provides some base information in this respect by illustrating the relative shares of publications with specific numbers of authors, affiliations, and countries. We observe that less than 5% of all demo and research papers and 25% of the industrial papers are written by a single author; the majority of research publications share two to four authors. While the majority of industrial and demo papers originate from a single affiliation, there are slightly more research papers from two or more affiliations than from a single institution. Almost 25% of the research papers have authors from two or more countries. Table 2 lists the average number of authors, affiliations, countries, and continents per paper. Interestingly, research papers involve on average more affiliations and countries per paper despite a lower number of authors than industrial and demo papers.

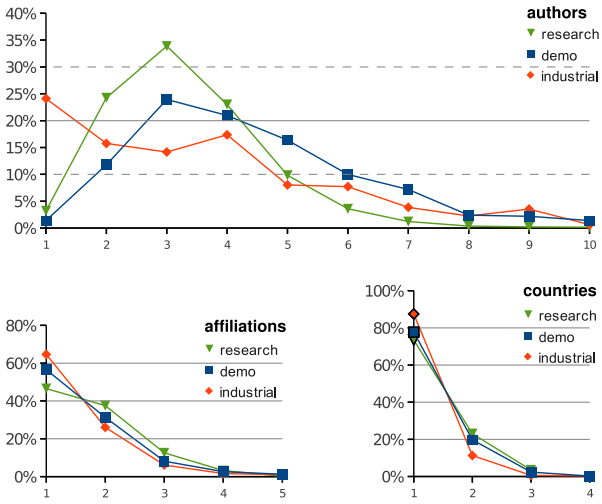

Figure 1: Frequency distribution of publications with distinct number of authors, affiliations, or countries

Entity/track	research	industrial	demo
Author	3.34	3.68	4.58
Affiliation	1.73	1.51	1.62
Country	1.30	1.15	1.25
Continent	1.21	1.08	1.15

Table 2: Average participants per paper by track

Figure 2: Geographic distribution of affiliations publishing in the top database venues in the last decade

The map in Fig. 2 pinpoints places of the world with papers in our collection. Larger bubbles denote more papers, visualizing the concentration of larger paper counts to only few hubs (US West and East Coast/Great Lakes, Central Europe, Hong Kong, and Singapore).

3. CONTINENTS AND COUNTRIES

We first analyze which continents and countries published most papers in the considered top venues and how the productivity developed during the decade. The reported number of papers per entity, e.g. author, affiliation (institution, city), country or continent, is derived by crediting each entity once when at least one author of the paper is affiliated with it. Thus, the sum of papers can be larger than the sum of unique papers due to multiple authors. In the following tables 3 to 7 we list this total number of papers as *pubs*. We also report fractional counts (*frac)* where each author (and her affiliation, country or continent) is credited only the n-th part of a paper in case of n authors [2, 5]. This is a simple approach to account for cooperative efforts since the fractional counts sum up to the total number of papers. Furthermore, as rough indicators for the degree of cooperation, the average number of contributing entities (authors, affiliations, countries, continents) per paper are shown in the tables. We also list the number of affiliations (institutions at the city level) that contributed publications.

3.1 Continents

We aggregate author affiliations into countries as well as into the continents North America (N.A.), Europe, and Asia. We further aggregate Africa, Oceania, and South America into the Southern Hemisphere (S.H.) as only few papers originate from this region. We observe from Table 3 that by far most papers originate from affiliations in North America (USA and Canada). Almost three quarters of the research papers as well as the industry/demo papers have at least one author from this area. Europe contributes the second most papers followed by Asia. The number of contributing affiliations is somewhat differently distributed since less than 50% (283 of 623) are located in North America. The number of affiliations is relatively high for Europe as we will cover further when discussing the country and affiliation statistics. The level of cooperations (average number of continents per paper) is higher for continents with fewer papers indicating that they depend most on cooperations with affiliations from other continents to publish in the top venues.

Cont.	affil	pubs	*frac*	res, ind, dem	C./p (r i d)
N.A.	283	**1,982**	*1,766*	1,396, 258, 328	1.3, 1.1, 1.2
Europe	217	**642**	*504*	436, 46, 160	1.4, 1.3, 1.3
Asia	95	**513**	*309*	407, 29, 77	1.5, 1.4, 1.4
S.H.	28	**79**	*51*	63, 3, 13	1.8, 2.0, 1.8

Table 3: Publication counts per continent

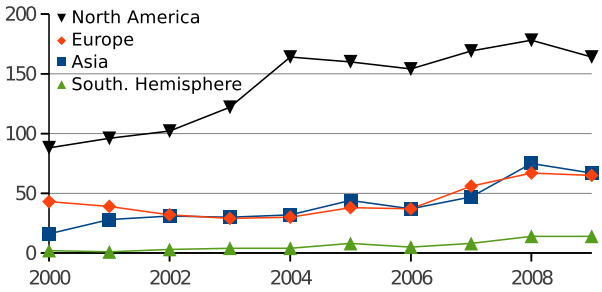

Figure 3: Research paper trends by continent

Fig. 3 illustrates how the number of research papers per continent developed during the decade. While the number of papers increased for all continents during the decade, the strongest increases are observed for Asia and North America. Asia caught up with Europe in the last years and the lead of North America has even increased during the decade. The trends for industrial and demo papers are similar, albeit Europe has retained a lead over Asia so far.

3.2 Countries

Table 4 lists top countries contributing most publications in the three categories research, industry, and demo. We observe that USA leads by far for all three paper categories. Germany and Canada are runners up for industrial and demo papers, and also among the top countries for research papers. The fractional paper counts lead to a largely similar ordering of countries but can be better used to determine the relative paper shares per country (since the fractional counts sum up to the total number of papers). For instance, the fractional numbers show that US authors contribute about 60%, 75% and 53% of all research, industry and demo papers, respectively. This underlines that the US dominance is especially pronounced for industrial papers as the major DBMS vendors are from the US. The table also differentiates the number of papers in the first and second half of the decade illustrating some interesting trends. Regarding research publications, China and Singapore more than tripled their number of papers in the second half of the decade. For the whole decade, this helped China and Singapore to contribute the second and fifth most research papers from all countries. The dominance of US institutions has slightly reduced since the share of research papers with an US-based co-author changed from 72% in the first to 67% (770/1147) in the second half of the decade. Also, the UK and Australia have achieved significant increases in the second half of the decade, positioning them among top ten.

country	affils	pubs	*frac*	cntr/p	1^{st}	2^{nd}
USA	179	**1,315**	*1,131*	1.32	545	770
China	24	**176**	*124*	1.73	37	139
Canada	15	**160**	*93*	1.89	61	99
Germany	50	**147**	*109*	1.53	69	78
Singapore	4	**102**	*64*	1.92	24	78
Italy	23	**57**	*36*	1.70	22	35
France	24	**56**	*37*	1.79	32	24
India	14	**56**	*38*	1.63	30	26
UK	10	**46**	*29*	1.83	10	36
Australia	14	**46**	*31*	1.76	5	41
USA	115	**247**	*231*	1.15	118	129
Germany	22	**23**	*17*	1.48	11	12
Canada	10	**23**	*16*	1.61	14	9
India	11	**13**	*9*	1.62	7	6
South Korea	6	**6**	*5*	1.00	1	5
USA	116	**305**	*267*	1.28	124	181
Germany	35	**73**	*58*	1.38	28	45
Canada	11	**45**	*27*	1.80	20	25
China	19	**32**	*23*	1.78	12	20
Italy	16	**28**	*20*	1.63	13	15

Table 4: Countries by research, industrial, demo

The average number of countries per paper in Table 4 is especially high for Canada and Singapore, denoting an over-average amount of cooperations with authors from one or more other countries. The high degree of cooperation also led to significantly reduced fractional paper counts for these countries. To gain additional insight, we illustrate in Fig. 4 the number of intra- and cross-country papers for the top five countries. It shows the amount of papers attributed to a single country and between countries, giving overall as well as research, demo, and industrial counts. We observe that cross country collaboration mostly takes place in connection with the USA, especially regarding neighboring Canada, where nearly as many research papers were co-authored with US-based authors as without. Singapore authors co-published to a similar degree with authors from USA and China. Both Germany and China published most papers with co-authors from their own country and co-published a similar number of papers with colleagues from US institutions. While we cannot derive a strong connection between the degree of collaboration and productivity (similar as in [3]), it seems that countries like Singapore and Canada having fewer affiliations than China and Germany were able to substantially benefit from the higher degree of international collaboration.

As shown in Table 4 there are significant differences in the number of contributing affiliations per country even for countries with a comparable number of papers. For instance, Germany has many more research affiliations (50) than China, Canada, and especially Singapore (4). To further analyze this ob-

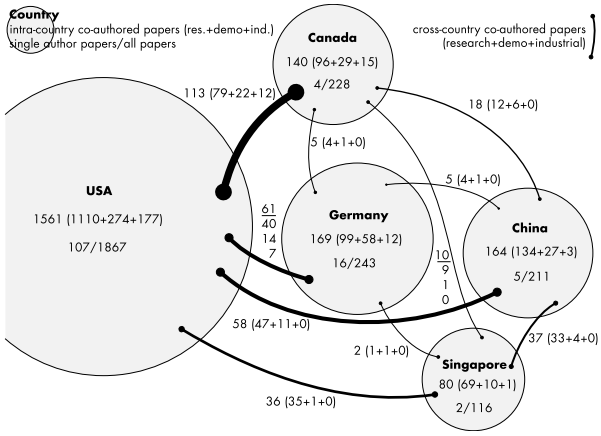

Figure 4: Intra- and cross-country co-operations

servation we also determine the number of authors per affiliation (within the considered ten years) as an indicator of the affiliation size. In Fig. 5 we illustrate for every country its number of affiliations and the average affiliation size; the bubble size indicates the number of papers of the country. We observe that from the five leading countries Germany has the smallest average number of authors per affiliation (6) while Singapore has the largest average group size of 22; the average size of US institutions is twice as high as for Germany. We conclude that large teams with many authors are generally favorable to achieve a high number of papers in the considered quality venues. This will also be confirmed in the next section indicating that the size of the top affiliations is significantly above the country averages.

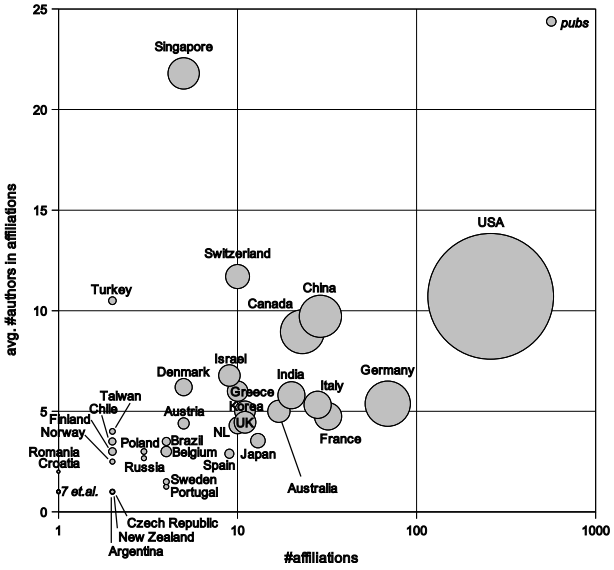

Figure 5: Country publications by no. of affiliations vs. average no. of authors within affiliations

4. AFFILIATIONS AND AUTHORS

4.1 Institutions

Table 5 lists the affiliations contributing most publications in the research, industry, and demo categories. We consider all departments of an institution or company located in the same city as one affiliation resulting in several affiliations for companies such as IBM or Microsoft. The table shows that three US companies with large research departments contribute most research papers in the last decade: IBM San Jose, Microsoft Redmond, and AT&T Florham Park. IBM and Microsoft are also top contributors for industrial and demo papers while DBMS vendor Oracle is only prominently visible for industrial papers. Microsoft Research nearly doubled the number of research papers in the second half of the decade and outnumbered its competitors in this time period.

Regarding research publications from academia we observe that universities from Singapore and Hong Kong have achieved similarly high publication counts as the traditionally strong US universities Stanford, Wisconsin, and Berkeley. While Berkeley and especially Stanford had declining publication counts in the second half of the decade, the National University of Singapore (NUS) and Hong Kong UST had strong increases. Further affiliations with a strongly growing number of research papers include the Canadian universities of Toronto and Waterloo, and the Chinese University of Hong Kong. When aggregating the paper counts per institution across all locations, most research papers come from IBM followed by Microsoft (196 and 128 papers). Aggregating at the city level (across all institutions) reveals that most research publications (133) originate from Hong Kong.

The last column in Table 5 indicates the number of authors contributing to the affiliations' publications. In general the teams of the listed affiliations are very large. With the exception of AT&T, the top ten research affiliations have teams of about 50 or more active authors; the by far largest number of authors (126) comes from IBM San Jose. AT&T has the highest average number of institutions per paper indicating a strong degree of collaboration with other affiliations. Such intensive collaborations seem to facilitate a high number of publications with a moderate number of local authors.

Despite the high number of research papers from Europe (cf. Table 3), no European institution is among the top 20 listed in Table 5. A likely reason for this is the low average affiliation size that we observe for most European countries in Fig. 5. Table 6 shows which European institutions contributed most research papers. We observe that only two institutions had more than 20 active authors so that – with the exceptions of ETH Zurich and INRIA, the team sizes of the leading European database affiliations are substantially below the ones of the globally leading affiliations.

Institution	pubs	*frac*	inst/p	1^{st}	2^{nd}	aut
IBM, San Jose	115	*73*	2.04	49	66	126
Microsoft, Redmond	110	*67*	1.97	39	71	52
AT&T, Florham Park	87	*43*	2.44	49	38	35
Natl. Univ. of Singapore	83	*55*	1.98	23	60	74
Univ. of Wisconsin, Madison	81	*55*	1.88	38	43	71
Stanford University	81	*52*	1.80	54	27	66
Hong Kong Univ. of Science and Technology	79	*46*	2.29	27	52	49
Univ. of California, Berkeley	67	*45*	1.85	37	30	65
Univ. of Illinois, Urbana-Champaign	65	*43*	1.98	25	40	49
Univ. of Maryland, College Park	58	*41*	1.81	30	28	49
Univ. of Toronto	58	*32*	2.29	15	43	37
Univ. of Washington Seattle	57	*41*	1.82	30	27	38
Bell, Murray Hill	55	*32*	2.13	41	14	35
Univ. of Michigan Ann Arbor	49	*31*	2.02	22	27	32
CMU, Pittsburgh	47	*28*	2.19	23	24	48
Purdue Univ., West Lafayette	41	*26*	2.02	15	26	39
Chinese Univ. of Hong Kong	41	*23*	2.32	6	35	32
IBM, Yorktown Heights	39	*19*	2.21	15	24	31
Univ. of Waterloo	37	*22*	1.89	9	28	26
UC Riverside	37	*19*	2.38	13	24	31
Oracle, Redwood Shores	42	*33*	1.67	13	29	88
Microsoft, Redmond	35	*31*	1.54	16	19	83
IBM, San Jose	24	*20*	1.58	16	8	64
Oracle, Nashua	16	*11*	1.88	7	9	33
IBM, Toronto	16	*9*	1.94	11	5	26
IBM, San Jose	29	*21*	2.07	10	19	73
Microsoft, Redmond	24	*17*	2.00	5	19	58
Univ. of Toronto	20	*13*	1.90	8	12	25
AT&T	19	*8*	2.58	11	8	18
WPI, Worcester	17	*14*	1.35	9	8	42

Table 5: Institutions by research, industrial, demo

4.2 Authors

As authors drive the productivity of their affiliation, we finally investigate which authors contributed most research, industrial, and demo papers (Table 7). Some authors published their papers for up to four different affiliations. Regarding research publications, the two most prolific authors come from industry labs (Microsoft, AT&T). The other authors in the re-

Institution	pubs	*frac*	inst/p	1^{st}	2^{nd}	aut
ETH Zurich	29	*23*	1.52	11	18	39
Aalborg University	26	*14*	1.96	11	15	18
Univ. of Edinburgh	25	*15*	2.16	5	20	16
INRIA Le Chesnay	23	*12*	2.35	20	3	33
Univ. of Athens	22	*12*	2.32	6	16	17
MPI Saarbrücken	17	*12*	1.71	2	15	9
CWI Amsterdam	14	*12*	1.43	5	9	15
Univ. of Munich	12	*9*	1.67	10	2	20

Table 6: Top European research institutions

search top ten are mostly from American and Asian universities. Some leading affiliations (IBM San Jose, Stanford) have no author in the top ten for research papers. The authors with most demos are primarily from universities that apparently emphasize building of prototypes. There are significant differences in the average number of authors per paper. For the most prolific authors of research papers, this value is generally higher than for all research papers (average 3.43, Table 2) and the highest value is noted for the top listed researcher D. Srivastava. On the other hand, S. Chaudhuri has a relatively low (and below average) number of co-authors indicating that a very high productivity can also be achieved with a moderate level of cooperation.

5. CONCLUSIONS

We analyzed the author affiliations of database publications that appeared in four top venues in the last decade. We observed that most papers originate from US institutions and that industry labs run by IBM, Microsoft, and AT&T are most prolific. Asian institutions have achieved a comparable research output as European institutions. Top universities from Singapore and Hong Kong have reached a similar number of publications as the traditionally strong US universities Stanford, Wisconsin, and Berkeley.

Almost all research papers are co-authored; half of them involve at least two affiliations and almost a quarter two or more countries. The most frequent cross-national co-authorships occur between USA and Canada; Singapore has frequent cooperations with USA and China. The most prolific affiliations have relatively large teams; a high degree of collaboration also tends to improve the publication counts in the considered top venues. Europe hosts many but mostly small affiliations that do not yet achieve the paper counts of the top affiliations world-wide.

In future work, we plan to evaluate additional aspects such as affiliation-specific citation counts.

Author *affiliations*	pubs	*frac*	aut/p	1^{st}	2^{nd}
D Srivastava *AT&T*	39	*10*	4.31	20	19
S Chaudhuri *Microsoft Redmond*	38	*13*	3.00	19	19
N Koudas *AT&T, UW Seattle, Univ. Toronto*	36	*10*	3.81	11	25
MN Garofalakis *Bell, UC Berkeley, Yahoo, TU Crete*	35	*12*	3.29	19	16
Y Tao *HKUST, CityU of HK, CMU Pittsburgh, CUHK, Univ. of HK*	34	*12*	3.38	13	21
J Han *SFU Vancouver, U of I at Urbana-Champaign*	34	*10*	3.74	13	21
HV Jagadish *UMich Ann Arbor, AT&T*	33	*11*	3.61	15	18
D Papadias *HKUST*	32	*10*	3.41	16	16
R Ramakrishnan *UW-Madison, Yahoo*	32	*9*	4.09	6	26
BC Ooi *Natl. Univ. of Singapore*	32	*8*	3.97	11	21
MJ Carey *Propel Software San Jose, BEA San Jose, UC Irvine*	9	*3*	6.67	4	5
C Galindo-Legaria *Microsoft Redmond*	8	*3*	3.50	5	3
M Poess *Oracle Redwood Shores*	7	*3*	2.57	3	4
E Rundensteiner *WPI Worcester*	17	*4*	5.18	9	8
G Weikum *MPI Saarbr., Univ. Saarbrücken*	12	*3*	4.67	4	8
J Pei *SFU Vancouver, SUNY Buffalo*	11	*3*	4.73	7	4

Table 7: Authors by research, industrial, demo

6. REFERENCES

[1] Aumüller, D., Rahm, E.: Web-based Affiliation Matching. *Information Quality (ICIQ 09),* 2009

[2] Egghe, L., et al.: Methods for accrediting publications to authors or countries: Consequences for evaluation studies. *JASIST*, 2000

[3] Egghe, L., et al.: Collaboration and productivity: an investigation in Scientometrics and in a university repository. *Collnet, scientometrics and information management,* 2008

[4] Franceschet, M.: A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. *Scientometrics*, 2010

[5] Hagen, N.T.: Harmonic publication and citation counting: sharing authorship credit equitably – not equally, geometrically or arithmetically. *Scientometrics*, 84, 2010

[6] Nascimento, M., et al.: Analysis of SIGMOD's co-authorship graph. *Sigmod Record*, 2003

[7] Rahm, E.: Comparing the scientific impact of conference and journal publications in computer science. *Information Services and Use* 28 (2), 2008

[8] Rahm, E., Thor, A.: Citation analysis of database publications. *Sigmod Record*, 2005

[9] Sidiropoulos, A., et al.: Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics* 2007