Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 42 (2006) 1553–1566
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

An ego-centric citation analysis of the works of Michael O. Rabin based on multiple citation indexes

Judit Bar-Ilan *

Department of Information Science, Bar-Ilan University, Ramat Gan 52900, Israel

Received 16 March 2006; accepted 16 March 2006
Available online 18 May 2006

Abstract

The primary goal of this study was to carry out an ego-centric citation and reference analysis of the works of the mathematician and computer scientist, Michael O. Rabin. Until recently only a single citation database was available for such research – the ISI Citation Indexes. In this study we utilized and compared three major sources that provide citation data: the Web of Science, Google Scholar and Citeseer. Most cited works, citation identity, citation image makers and coauthors were identified. The citation image makers acquired through these sources differ considerably. Advantages and shortcomings of each of the tools are discussed in the context of computer science. A major issue in computer science is multiple manifestations of a work, i.e., its publication in several venues (technical reports, proceedings, journals, collections). The implications of multiple manifestations for citation analysis are discussed.

Keywords: Ego-centric citation analysis; Web of Science; Citeseer; Google Scholar; Multiple manifestations

1. Introduction

Being a former student of Michael O. Rabin, this paper grew out of a presentation I prepared for the Fifth Haifa Workshop on Interdisciplinary Applications of Graph theory, Combinatorics, and Algorithms in honor of Michael Rabin (http://cri.haifa.ac.il/index.html?http://cri.haifa.ac.il/events/2005/graph/graph.php). Michael Rabin is Professor of Computer Science at Harvard University and Professor Emeritus at the Hebrew University of Jerusalem. Michael Rabin, a mathematician and a computer scientist is a winner of numerous, prestigious awards including the Turing Award (for details see DEAS, 2005a). His Harvard research profile states: "his special interests are the applications of randomization in computations, cryptography, computer security, and parallel systems... and continues to work on creating efficient algorithms for problems in algebra, number theory, data structures, and combinatorics" (DEAS, 2005b). Randomization and non-determinism are the central themes of most of his work (see Wikipedia, 2005a, 2005b for definitions).

* Tel.: +972 523667326; fax: +972 3 5353937.
E-mail address: barilaj@mail.biu.ac.il

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2006.03.019

His career started out in algebra, mathematical logic and automata theory and over time his interests were extended to complexity theory, nondeterministic computations, probabilistic algorithms and cryptography. Michael Rabin's best-known works are in the areas of automata theory – "Finite automata and decision theory" (joint work with Dana Scott – the Turing Prize was awarded to them in appreciation of this work) and additional major contributions to automata theory, including probabilistic automata; primality testing (large primes are the basis for modern cryptography, Michael Rabin received the Kanellakis Prize for this achievement); complexity theory; contributions to cryptography (an encryption method provably equivalent to factorization, his "oblivious transfer" which was shown to be a cryptographic primitive and his more recent works related to hyper-encryption – joint work with Yonatan Aumann and Yan Zong Ding) and other well-known algorithms like fast, probabilistic pattern matching (joint work with Dick Karp), information dispersal and a randomized solution to the "Byzantine Generals" problem. This short overview is partially based on discussions with Michael Rabin (2005).

The notion of ego-centric citation analysis was introduced by Howard White (2000, 2001a). White defined the following constructs:

- Coauthors – authors (with multiplicity) who published with the specific author.
- Citation identity – authors (with multiplicity) who are cited by the specific author.
- Citation image makers – authors (with multiplicity) who cite works of the specific author.
- Citation image (White & McCain, 1998) – authors (with multiplicity) who have been co-cited with the specific author.

White (2000) carried out an ego-centric citation analysis of Eugene Garfield and created an egocentric citation map based on Eugene Garfield's citation image. In another paper (White, 2001a), eight information science authors were studied. There, in addition to the above-mentioned categories, White also looked at self-citations, unicitations (where a work has been cited only once by an author) and recitations, and computed the citations/cite ratios. Following Lawani (1982), he considered the synchronous and diachronous self-citation rates, where the synchronous rate is defined as the number of self-citations divided by the total number of references in the author's work, and diachronous rate is the number of self-citations divided by the total number of citations the author's works received. Cronin and Shaw (2002) using the methods introduced by White, created the citation identities and listed the citation image makers of three information scientists. They studied the changes occurring to the citation identities over the years, due to collocation, social ties and student–teacher relationships. One of the professors studied, Robert Kling, "migrated" to information science from computer science and management. This is clearly reflected in his citation image makers list, where 73% of the citations his works received were outside information science. They also compared the top citation identities with the top image makers lists, and found a relatively small number of names in the intersection of these sets.

White (2001b) discussed 15 kinds of personal author profiles that can be generated from bibliographic data available in databases. He called these profiles CAMEOs – characterizations automatically made and edited online. "Profiles of this nature personalize bibliometrics in a way rarely seen before, in that they center core-and-scatter distributions not on a principle, specialty, topic, or journal, but on a single author, construed as either an individual or an oeuvre" (White, 2001b, p. 608). The 15 profiles identified by White were: coauthor, journal, title, abstract, full-text, keyword, identifier, genre, publication year, descriptor, subject heading, classification code, citation identity, citation image-makers and citation image. White noted that whereas the first 12 profiles can be attained from a number of databases, the last three are uniquely available from the ISI Citation Databases. White compared the descriptors assigned to Allen Rothwarf's publications from two bibliographic databases, and found considerable differences in the indexing practices of the two databases.

The paper emphasized that coauthor, citation identity, citation image and image makers can be interpreted representing both subject matter and social ties. The author studied (the "focal author"), the authors in his/her citation identity and his/her citation image makers can be represented as nodes in a social network, where there is a directed edge from each image-maker to the focal author, and from the focal author to each author in his/her citation identity. The strength of the ties is a function of the number of times the author's works have been cited/referenced.

2. Methods

2.1. Data collection

The starting point for data collection was Michael Rabin's list of publications, available online (Rabin, 2004). Eighty-six publications are listed in that document. After searching in a wide range of bibliographic sources Web of Science (http://portal.isiknowledge.com/), ACM Digital Library (http://portal.acm.org/dl.cfm), Citeseer (http://citeseer.ist.psu.edu/), Google Scholar (http://scholar.google.com/), DBLP (http://www.informatik.uni-trier.de/~ley/db/), and AMS MathSciNet (http://www.ams.org/mathscinet) and references in the retrieved publications, the final list as of May 2005 was comprised of 113 items. Among the 113 items, there were 37 journal articles (33%), 46 papers in proceedings (41%), 15 technical reports (13%), 9 books/chapters in book (8%), 3 presentations, 1 notice, 1 report and 1 PhD thesis.

We were able to access all except three items. Some were available online (52 publications – free or through subscription), the rest were accessed through visits to the Mathematics and Computer Science Library of the Hebrew University of Jerusalem, through interlibrary loan and through the kind help of Professor Rabin. We listed the references appearing in all these publications, all the authors mentioned in the references were recorded. The 504 references of the "Computers at Risk" report prepared for the National Research Council prepared by the System Security Committee (Michael Rabin was a member of this committee) were excluded, and of course we have no information on the references in the three items we were not able to access (including Michael Rabin's PhD thesis, the results of which were published in two journal papers that were available to us).

White (2001b) stated that only the ISI databases are capable of listing a person's citation identity, citation image-makers and citation image. This was true at the time of his writing, and in the area of information science, however the bibliographic database scenery has changed considerably since. Citeseer is a free citation indexing system of computer science literature created by autonomous agents. It exits at least since 1998 (see Giles, Bollacker, & Lawrence, 1998) and indexes information science literature as well, but to a lesser extent. It indexed 723,140 documents as of May 10, 2005. It is a citation index, thus one can search for citations of an author or of a specific work. It not only provides bibliographic details of the citing source, but also provides a "snippet" that allows the reader to deduce the citation context. We searched Citeseer for citations of each of Michael Rabin's publication separately, where for each publication we chose a short characteristic phrase from the title of the publication (to "catch" misspelled references as well) combined with the authors name, e.g., *Program Transformations AND Rabin*, when searching for citations to "Efficient Program Transformations for Resilient Computation via Randomization" by Kedem, Palem, Rabin and Raghunathan.

Another new tool is Google Scholar (still in beta) that was launched in November 2004 (Sullivan, 2004). Google scholar indexes scholarly works from free and subscription based sources, and provides a list of items indexed by Google Scholar that cite the specific item. It also retrieves citations to items that are not indexed by it. We have no data on the number of publications indexed by Google Scholar. We searched for Michael Rabin's publications by submitting the author:"*Michael Rabin*" query to capture also publications where he is referenced without his middle initial. A semi-automated process was developed to download the data from Google Scholar, extract the necessary information and to present it in a columnar format for further analysis.

Naturally, we have also searched the ISI Citation Indexes (now owned by Thomson). The Web of Science interface was used. On May 11, 2005 the Web of Science contained 34,908,911 records from 1945 and on. Even though it only indexed 30 publications of Michael Rabin, it also lists citations to publications not indexed by the service (these citations are gathered from the reference lists of items indexed by the database). This is a crucial feature of all three citation indexing tools that were utilized in this study. The Web of Science offers to present the data in different formats, and we were able to find a format that allowed carrying out data analysis using Microsoft Excel. All the three citation databases were searched during April 2005.

Additional tools that provide partial citation data and are not included in the results are the ACM Digital Library (it only provides citations for items included in the library) and Scopus (http://www.scopus.com). At Scopus the user can choose a publication time span starting from 1960, however, at least for Michael Rabin, no documents were dated before 1972 were indexed, the citation counts are very low and there are no citation

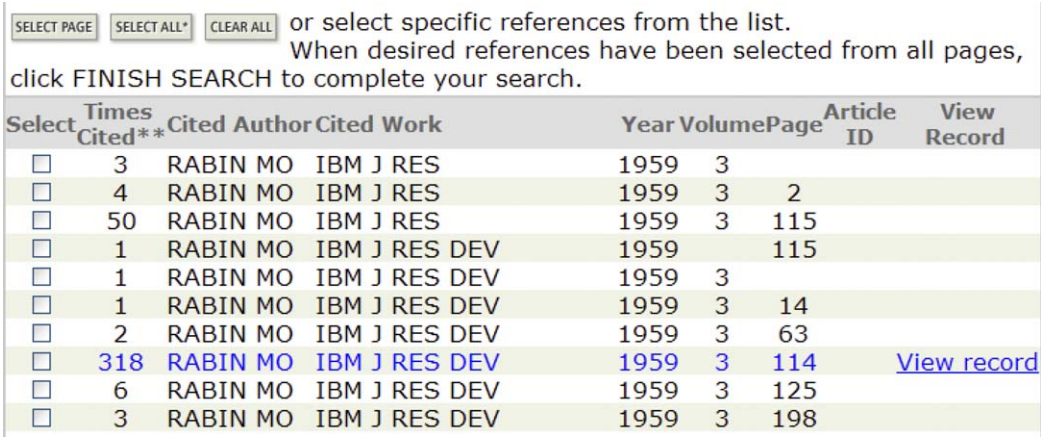
counts of items not indexed by Scopus. Scopus' current coverage is not extensive enough to be included in this study.

2.2. Data cleansing

Data collected from each source had to undergo some kind of editing/data cleansing. ISI only indexes "important" journals in each field, and does not cover proceedings. In computer science refereed proceedings are a major publication channel (often works published in proceedings only – for a discussion of this issue, see for example Goodrum, McKain, Lawrence, & Giles, 2001). An exception is Lecture Notes in Computer Science (LNCS, and its sub-series) – a series that publishes a large number of workshop and conference proceedings. However, the two major theoretical computer science conferences, the ACM STOC (Symposium on Theory of Computing) and the IEEE FOCS (Foundations of Computer Science) are not indexed by ISI, as all other proceedings published by ACM or IEEE. Citations to these publications are recorded, but there is no unified scheme to reference these sources, and the Web of Science only displays cited author, cited work – name of the publication, but not title of the specific article, year, volume and page (in a most recent change, the Web of Science does display the title of the article as well, but only for articles indexed by ISI). The number of characters allowed for cited work is limited, thus various abbreviations are used, and often it takes serious "detective work" to find out what item was cited. The fault is not always with the database, the citing authors often provide incorrect or partial information. Seemingly ISI makes no attempt to try to correct the data provided by the citers.

The ISI database is highly sensitive to errors in references. See for example Fig. 1. Probably most of the lines refer to the highly cited Rabin and Scott paper: Finite automata and their decision problems that appeared in 1959 in the IBM Journal of Research and Development, vol. 3, pp. 114–125. However, one cannot be sure, since Michael Rabin published another paper in the same volume of the journal together with W. W. Peterson (Rabin was second author on that paper) on pages 163–168. Thus, we may probably assume that the references to page 115 are to the Rabin & Scott paper, but what about the rest of the unassigned references? Especially "suspicious" are the two citations of a publication appearing on page 63 of the journal (note that the Peterson & Rabin paper starts on page 163). Of course, one should remember that Citation Indexes used to list the first author only, and citations were assigned only to the first author. ISI changed this policy a numbers of years ago, and it is not clear what is the situation with the backfiles is. Our experience with data related to Michael Rabin shows that assigning citations to second authors is not consistent, even for publications indexed by ISI, the second author does not always receive the citations.

In order to retrieve citations from the Web of Science, we searched for Rabin M as cited author in addition to Rabin MO (and discarded citations to other Rabin Ms). In addition, for every paper on which Michael Rabin was not the first author, we also looked for additional citations listed for the first author. Because


SELECT PAGE SELECT ALL* CLEAR ALL or select specific references from the list.
When desired references have been selected from all pages, click FINISH SEARCH to complete your search.

Select	Times Cited**	Cited Author	Cited Work	Year	Volume	Page	Article ID	View Record
☐	3	RABIN MO	IBM J RES	1959	3			
☐	4	RABIN MO	IBM J RES	1959	3	2		
☐	50	RABIN MO	IBM J RES	1959	3	115		
☐	1	RABIN MO	IBM J RES DEV	1959		115		
☐	1	RABIN MO	IBM J RES DEV	1959	3			
☐	1	RABIN MO	IBM J RES DEV	1959	3	14		
☐	2	RABIN MO	IBM J RES DEV	1959	3	63		
☐	318	RABIN MO	IBM J RES DEV	1959	3	114		View record
☐	6	RABIN MO	IBM J RES DEV	1959	3	125		
☐	3	RABIN MO	IBM J RES DEV	1959	3	198		

Fig. 1. A partial list of citations of Michael Rabin's works – The Web of Science.

of the above-mentioned problems, we were unable to assign to specific publications 52 citations (1.4%) out of the 3659 citations to works of Michael Rabin that were identified by ISI's Web of Knowledge.

Citeseer, a widely used bibliographic tool by computer scientists, turned out to be rather difficult to use. The site was overloaded and the service frequently timed out. There were often very large discrepancies between the number of citations reported and the number of citations actually displayed (may have been caused by overload, although we tried to retrieve results several times). Our analysis is based on the citation sources displayed (this is only about half of the number of citations reported). Sometimes the same citing source and citing context was displayed several times (over-counting citations) – this was usually caused by several identical (or near identical) versions of the same work collected by the autonomous agents from different URLs. Each citation was clicked on to retrieve complete information about the citing item. Quite often the title and author fields were not identified correctly. Although we have not recorded the identification errors, we felt that the error rate was rather high. Unlike the ISI Citation Indexes, Citeseer tries to normalize the references in order to group together slightly different references to the same work (see Goodrum et al., 2001), as can be seen from Fig. 2. In addition to this positive feature, Fig. 2 also illustrates problems with character recognition – it turns out the double *f* (*ff*) is especially problematic.

We complained about ISI's limited coverage, Citeseer, on the other hand, covers everything that seems to be a scholarly publication. These occasionally include summaries of lectures in courses. Theses and dissertations are routinely included – such publications usually have extensive literature reviews, thus raising the citation counts of a large number of publications, even if the citation context is only "perfunctory" (as opposed to "organic", see for example Baird & Oppenheim, 1994; Cronin, 1984; Liu, 1993 or White, 2001b). An additional problem encountered was chapters of books, with each chapter appearing as a separate document, but the bibliography of the whole book appears at the end of each chapter. Different versions of a book may show how the book evolves over time, but like the book chapters, these versions artificially increase citation counts.

We had fewer problems with Google Scholar, although Google Scholar also makes mistakes occasionally. Google Scholar bases a large portion of its data on information supplied by publishers, who provide free access to the abstracts of the articles, but require subscription or payment for accessing the full text. Thus, the bibliographic formats used are more predictable. Still, it was surprising to note the recurring mistakes with some of the publications of the IEEE. In all three citation sources, a few non-existing publications were cited (this of course is the citing authors' fault). Google Scholar lists at most four authors for an article. Where there were more than four authors, an attempt was made to identify the additional authors as well.

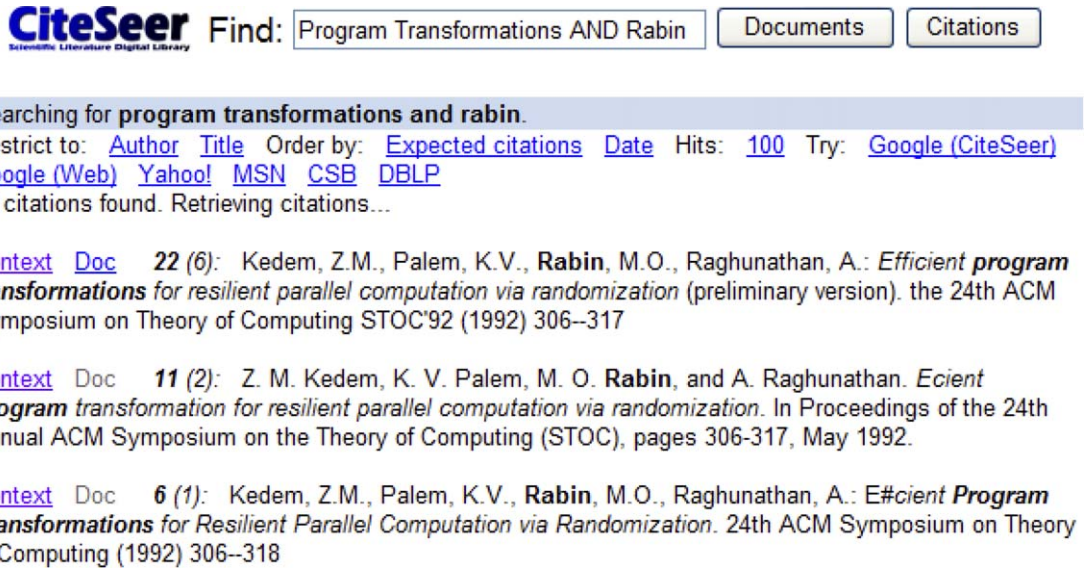


Fig. 2. Citeseer – normalization and character recognition.

3. Results and discussion

3.1. Most cited publications

Table 1 lists Michael Rabin's most cited publications according to the three different data sources. The table includes the 10 most highly ranked publications in each source. Note the differences in the rankings, especially the differences in rankings for the finite automata and the probabilistic automata paper. Citeseer relies on freely available online scholarly information mainly in computer science, and thus misses citations (even if the publications are freely available online) coming from other disciplines like mathematics and biology. Google Scholar also indexes only online information, but is not limited to a specific area and covers fee/subscription based sources as well. We were surprised to find back issues of a large number of journals on the Web, a welcome contribution of publishers and special projects (e.g., JSTOR or Project Euclid) to improve the effectiveness of information availability on the Web. Even with a relatively wide access to e-journals through the subscriptions of the MALMAD consortium of Israeli Universities and the additional online subscriptions of the Hebrew University and Bar-Ilan University, we were able to access online only 46% of the 113 publications of Michael Rabin – this probably reflects also on the percentage of citing documents available online. Thus, it seems that at this point of time when studying citation patterns of papers published before 1990, we still have to take into account sources available in print only.

Interesting to note that the paper ranked no. 3 by all three "databases" (Digitalized signatures) and the "oblivious transfer" paper (ranked 11, 5 and 4 by ISI, Google Scholar and Citeseer, respectively) are technical reports that have never appeared as a formal, peer reviewed publication. Still their value to computer science and cryptography are beyond dispute. Even more special is the case of the "oblivious transfer" paper: when collecting data for the presentation, I had to ask for Michael Rabin's help, and finally a copy of the report was mailed to me from his office at Harvard. Why was this report so inaccessible? Because it only exists as a handwritten manuscript! As Michael Rabin explained at the Haifa Conference, the report was typed on a stencil, and few printed copies were made, but then the stencil was lost, and no one ever retyped the manuscript. Most authors cite this report and make use of "oblivious transfer" without even knowing that it exists only in handwritten format (i.e., they have never seen an actual copy of it). After the Conference, Michael Rabin's daughter Tal, a computer scientist herself, submitted a scanned copy to the Cryptology Eprint Archive (http://eprint.iacr.org/2005/187), so that now the whole world can access a scanned copy of the handwritten original.

Michael Rabin's most cited paper according to the Web of Science is only ranked 4th and 7th by Google Scholar and Citeseer. A possible reason for the lower ranking at Google Scholar is that a considerable number of the citations covered by the Web of Science are not available online. Citeseer concentrates on publications in computer science, and this can explain the lower ranking of a mathematical paper. On the other hand, for the fourth item in Table 1, Google Scholar provides more citations than the Web of Science. This paper is a computer science paper, where the most of the papers are published in proceedings which are only very partially covered by ISI.

Returning to the last row of Table 1, we see that the top 10 most cited publications in each of the databases received more than 50% of the total citations of Michael Rabin's works. When viewing Table 1, Michael Rabin noted that the table contains his most important works. Fig. 3 depicts the graph of the cumulative citations, where the publications are arranged in decreasing order of citations received. The graphs for the other databases are very similar to the one appearing in Fig. 3.

3.2. Citation identity

The citation identity of an author is the list of authors cited by him/her (with multiplicity). The lists are usually long, so usually one concentrates only on the list of most often cited authors. However, before presenting this list, we must consider a crucial issue: what are the author's scientific works that constitute his/her oeuvre? This is especially a major issue in computer science, where an idea ("work" – under the FRBR terminology (IFLA, 1998)) may be published in several formats ("manifestations", IFLA, 1998): as a technical report, as an article in conference proceedings, as a journal article, as a chapter in a collection, etc.

Table 1
Most cited publications

Title	Source	Publication year	ISI-rank	ISI-citations	Google-rank	Google-citations	Citeseer-rank	Citeseer citations-identified
Finite automata and their decision problems	IBM Journal	1959	1	463	4	213	7	49
Decidability of second-order theories and automata on infinite trees	Trans. Amer. Math. Soc.	1969	2	358	2	365	1	117
Digitalized signatures and public-key functions as intractable as factorization	MIT/LCS/ TR-212	1979	3	244	3	298	3	82
Efficient Dispersal of Information for Security, Load Balancing, and Fault Tolerance	J. ACM	1989	4	176	1	371	2	112
Probabilistic Automata	Information and control	1963	5	173	9	127	6	52
Probabilistic algorithms	Algorithms and complexity	1976	6	172	8	140	15	29
Probabilistic algorithms for testing primality	J. of Number Theory	1980	7	153	7	173	9	41
Computable algebra, general theory and theory of computable fields	Trans. Amer. Math. Soc.	1960	8	125	22	54	25	16
Probabilistic algorithms in finite fields	SIAM J. on Computing	1980	9	111	13	102	10	40
Efficient randomized pattern-matching algorithms	IBM Jour. of Res. and Dev	1987	10	96	6	179	8	46
How to exchange secrets by oblivious transfer	Harvard TR-81	1981	11	95	5	204	4	65
Randomized Byzantine Generals	FOCS	1983	12	89	10	113	5	57
Total citations				3607		3880		1302
% Top-10 out of total				57.4%		53.4%		50.8%

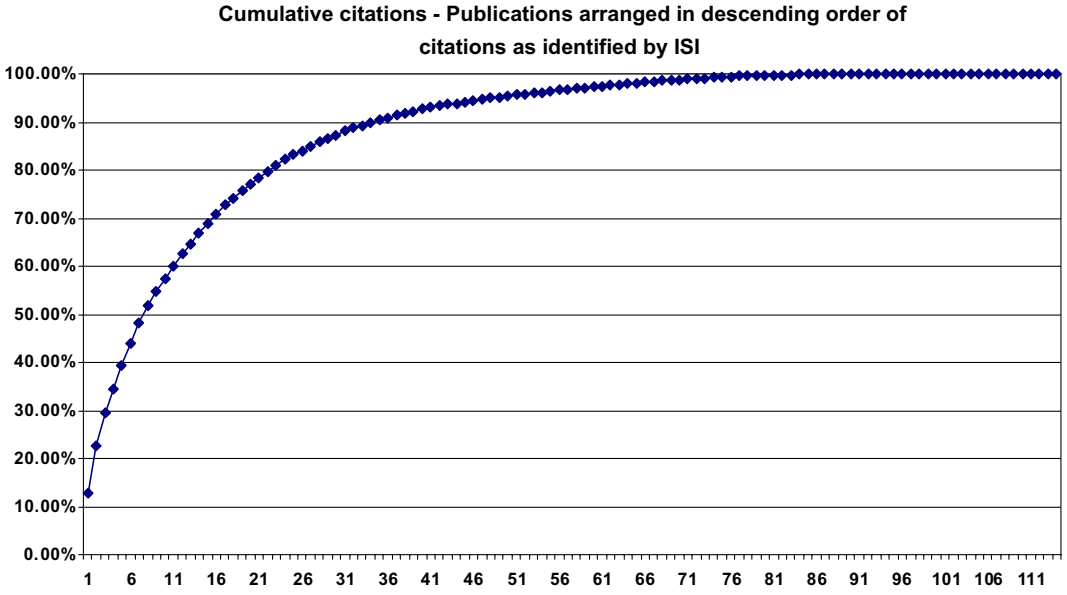


Fig. 3. Cumulative citations – ISI.

Two manifestations of the same work are not unusual in computer science, where refereed proceedings are the accepted modes of scholarly communication. The refereeing process and the publication delays for journal publications are relatively long in computer science, thus by the time the journal version appears the article is often outdated. Conference proceedings are refereed and proceedings papers are recognized publications, therefore quite often the authors do not even publish a journal version as well, however, sometimes because of academic pressure the work is published in a second format (as a journal paper). We have not investigated the percentage of works having a single vs. multiple manifestations in computer science, but we are confident that the size of neither group is negligible.

An additional complication arises when the journal version is not identical to the proceedings paper, it can be an extension of the proceedings paper, sometimes only in terms of literature review, but sometimes additional ideas and/or applications are introduced; sometimes the journal paper even has a different title and also a different set of authors. Thus, deciding what are the different manifestations of a work (“a distinct intellectual or artistic creation”, IFLA, 1998, p. 12) is not an easy task. It is possible that the use of the intermediate terminology “expression” is also in place (to indicate the content differences in the different versions).

Why do we raise this question here? As stated in Section 2.1, Michael Rabin’s list of publications included 86 items (Rabin, 2004), whereas we located 113 items authored by him. What is the reason for this huge difference? The list of publications is relatively updated, its date is September 2004, and the searches were carried out in April–May 2005. The main reason is that almost always (with only three exceptions), Michael Rabin lists only one manifestation of a work, out of the 28 additional items 14 were categorized as different manifestations of previously listed items. The other items not included in the list of publications were mainly presentations or invited lectures at conferences.

Multiple manifestations of a single work have two, slightly opposing effects. First they raise the citation counts of the cited items, since the reference lists are usually almost identical. On the other hand, they decrease the visibility of the work, because instead of a single manifestation receiving all the citations, the citations are spread between the different manifestations. Consider, for example, Jon Kleinbeng’s hubs and authorities paper – it first appeared at the SODA’98 conference and then it was extended and published in the Journal of the ACM in 1999. As of October 30, 2005, the Web of Science reports 227 citations to the journal version and 174 citations to the proceedings paper. Although it is theoretically possible that a paper cites both manifestations, we consider this highly unlikely. The journal version appeared within a year of the proceedings paper (6 years ago), but still the citations to the proceedings paper continue to cumulate.

When constructing Michael Rabin's citation identity we took into account all the items appearing in Michael Rabin's list of publications and the items discovered during our searches that were not additional manifestations of items already listed. Table 2 lists names of authors whose name appeared at list 10 times in Michael Rabin's citation identity. The whole list is comprised 620 authors, based on 1041 references to 1760 authors (authors counted with multiplicity). The references were counted in 100 publications, thus the average number of references is about 10, although there are considerable fluctuations, with the earlier, mathematical papers usually having only a few references and the newer computer science papers have many more references. The synchronous self-citation rate (Lawani, 1982) is 12.6%. Out of the 620 different authors cited, 296 (47.7%) were recitations, and 323 unicitations.

3.3. Coauthors

Michael Rabin had 46 coauthors so far in 52 coauthored publications (43 publications if only a single manifestation for each work is counted). Thus, the majority of his publications (independent on whether we count single or multiple manifestations) are singly authored. Table 3 displays the names of the authors who coauthored more than one paper with Michael Rabin, the counts are presented for both the single and the multi-manifestation cases.

3.4. Citation image makers

When constructing the list of citation image makers one has to take account not only the question of multiple manifestations, but also the citation index being used. The coverage, the accuracy and the selection policy of the citation index has direct influence on the results. Because of the large number of citations retrieved from each source (3607 from the Web of Science, 3880 from Google Scholar and 1302 from Citeseer) we were not able to eliminate multiple manifestations in the citing documents.

ISI has the most restrictive selection policy; it only indexes high quality journals. It provides a complete list of indexed publications (ISI, 2005). The result of such policy should be that at most one manifestation of a

Table 2
The most frequently occurring names in Michael Rabin's citation identity

Author	No. of times referenced
Rabin M.O.	131
Micali S.	41
Maurer U.M.	27
Aumann Y.	22
Goldwasser S.	22
Kedem Z.	20
Blum M.	19
Palem K.	19
Buchi J.R.	17
Shamir A.	16
Fischer M.J.	14
Tarski A.	14
Goldreich O.	13
Karp R.M.	12
Lynch N.	11
Naor M.	11
Raghunathan A.	11
Rivest R.L.	11
Bellare M.	10
Kleene S.C.	10
Rackoff C.	10
Spirakis P.G.	10
Wigderson A.	10

Table 3
Coauthors with more than one coauthored publication

Author	No. of papers – single manifestation	No. of papers – multi manifestations
Aumann Y.	7	9
Kedem Z.M.	3	3
Kushilevitz E.	3	4
Micali S.	3	3
Alon N.	2	2
Ding Y.Z.	2	2
Lehmann D.	2	1
Lipton R.	2	1
Mansour Y.	2	3
Palem K.V.	2	2
Tygar J.D.	2	2
Haastad J.	1	3
Sudan M.	1	3
Bender M.A.	1	2
Fischer M.	1	2
Halpern J.	1	2
Karp R.	1	2
Vazirani V.	1	2
Zuckerman	1	2

work is indexed. However, for some reason, ISI indexes the Lecture Notes in Computer Science Series, and thus even in the Citation Indexes multiple manifestations exists.

Citeseer has the least limited policy, according to Goodrum et al. (2001), Citeseer automatically categorizes a document (PDF or Postscript) as a research paper if it has a bibliography or reference section. Thus, the interpretation is very wide even if additional keywords are used to make sure that the document relates to computer science. It is supposed to detect duplicates. Indeed, during our searches we came across all sorts of documents: papers, drafts of papers, different versions of the same paper or chapter, technical reports, theses, dissertations, book chapters and even class summaries.

Google Scholar's indexing policy is as follows: it indexes "peer-reviewed papers, theses, books, abstracts and articles, from academic publishers, professional societies, preprint repositories, universities and other scholarly organizations" (Google Scholar, 2005). In practice it retrieves results also from personal sites.

The different indexing policies of the citation databases result in considerably different citation image makers. As stated before, we were not able to use Scopus for the current study, because it mainly covers publica-

Table 4
The top citation image makers according to the Web of Science

Author	No. of times citing Rabin's work
Rabin M.O.	44
Vardi M.Y.	38
Goldreich O.	33
Shelah S.	29
Micali S.	27
Kupferman O.	26
Pnueli A.	24
Niwinski D.	23
Ginsburg S.	22
Yung M.	22
Brassard G.	21
Courcelle B.	21
Reif J.H.	20
Stockmeyer L.	20

Table 5
The top citation image makers according to Google Scholar

Author	No. of times citing Rabin's work
Vardi M.Y.	64
Goldreich O.	52
Shvartsman A.A.	47
Micali S.	44
Rabin M.O.	42
Kedem Z.M.	41
Kupferman O.	41
Naor M.	38
Yee B.S.	36
Pnueli A.	31
Tygar J.D.	30
Daspupta P.	29
Bender M.A.	27
Crepeau C.	27
Maurer U.M.	26
Baratloo A.	25
Kushilevitz E.	24
Malewicz G.	23
Palem K.V.	23
Spirakis P.G.	23
Yung M.	23
Goldwasser S.	22
La Torre S.	22
Khoussainov B.	21
Aumann Y.	20
Blum M.	20
Lynch N.A.	20
Kanellakis P.C.	20
Rivest R.L.	20
Shoup V.	20
Russel A.	20
Walukiewicz Z.	20

Table 6
The top citation image makers according to Citeseer

Author	No. of times citing Rabin's work
Vardi M.Y.	30
Kupferman O.	21
Rabin M.O.	21
Bender M.A.	18
Comon H.	17
Kedem Z.	17
Malewicz G.	17
Naor M.	17
Yee B.S.	17
Goldreich O.	16
Micali S.	16
Tison S.	16
Tommasi M.	16
Dauchet M.	15
Gilleron R.	15
Lugiez D.	15
Ostrovsky R.	15
Shoup V.	15
Tygar J.D.	15

tions from 1994 and onwards, but in the future it will be very interesting to compare the list of citation image makers produced by Scopus to the lists produced by the other citation databases.

Tables 4, and 5 display the lists of citation image makers who cited Michael Rabin's works 20 times or more based on the Web of Science and Google Scholar. Since the number of citations identified through Citeseer was much smaller, 1302 versus 3607 and 3880 through the Web of Science and Google Scholar, respectively, in Table 5 one can view the list of citation image makers, acquired from Citeseer, who cited Michael Rabin's work 15 times of more is displayed.

Altogether 3059 different citing authors were identified through the Web of Science. Altogether 6513 names were extracted from the author lists of the 3607 publications citing works of Michael Rabin (an author can cite Rabin's works several times). The results in Table 5 are the counts after data cleansing, thus some minor mistakes can be expected.

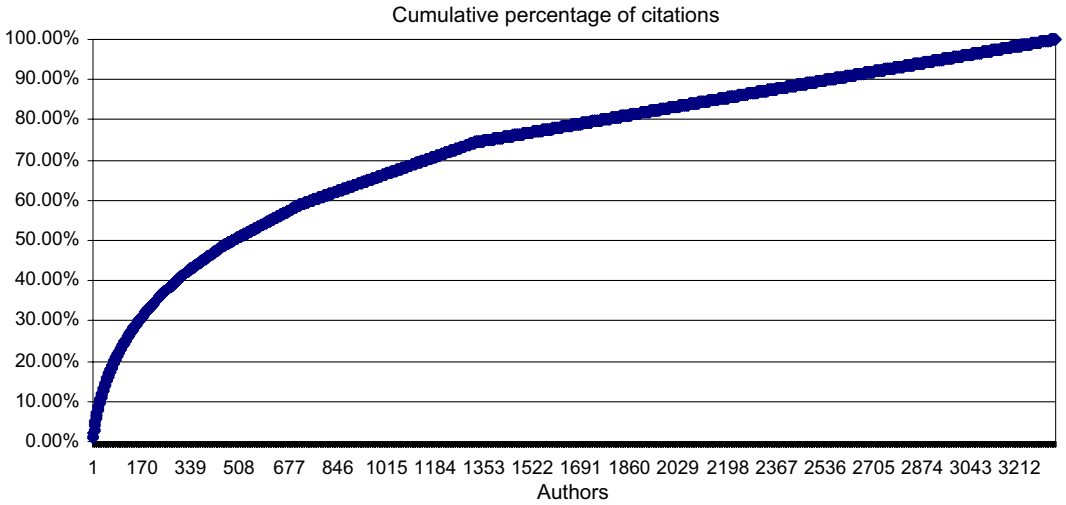


Fig. 4. Cumulative citations by authors based on data from Google Scholar.

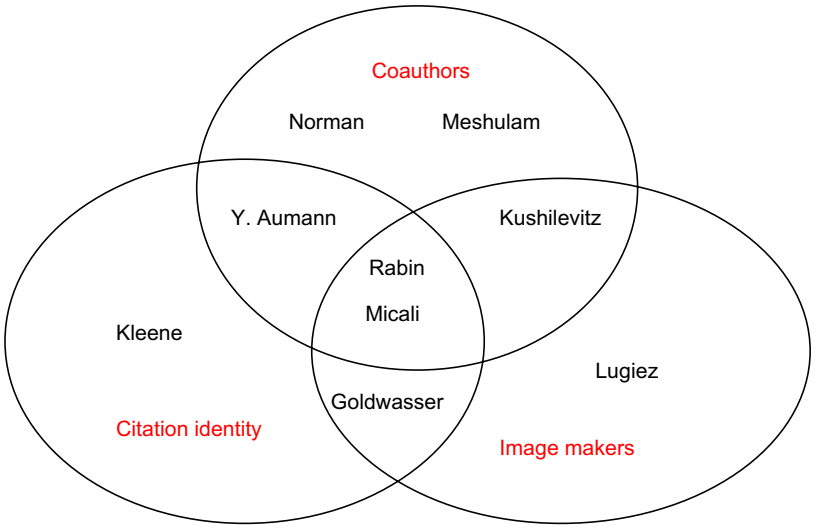


Fig. 5. A sketch of the intersections of the different sets of authors.

Through Google Scholar, 3335 different authors were identified. Altogether 7583 names were extracted from the author lists of the 3880 publications citing works of Michael Rabin (an author can cite Rabin's works several times). Here too, minor mistakes in the counts can be expected.

Citeseer retrieved the names of 1408 different citing authors. The author list consisted of 2886 names (with multiplicity) extracted from the 1302 citations. In Table 6, we present the numbers after data cleansing, thus some minor mistakes can be expected.

The graph of the cumulative citations by authors is considerable different form the graph of the cumulative citations by cited publications (Fig. 3). Fig. 4 is the graph of cumulative citations by authors, based on data retrieved from Google Scholar.

Only five names are common to all three lists of citation image makers (in alphabetical order): Oded Goldreich, Orna Kupferman, Silvio Micali, Michael Rabin and Moshe Vardi. Silvio Micali is a coauthor, and frequently cited author and a frequently citing author. Oded Goldreich is both frequently citing and frequently cited, whereas Yonatan Aumann is both a coauthor and a frequently cited author. The intersections between the different sets are visualized in the sketch appearing in Fig. 5. Since we listed only the most frequently occurring names in the all the lists, additional names are expected to appear in the different intersections.

4. Conclusion

In this study we have carried out an ego-centric citation analysis of Michael Rabin. We characterized his most cited works, his citation identity, coauthor list and citation image makers. Besides learning about his works, coauthors, references and citation, this "microscopic" analysis led to two major issues:

(1) Multiple manifestations of the same work (idea) have a crucial influence on citation analysis. Multiple manifestations are the norm in computer science. On the one hand multiple manifestations artificially raise citation counts, and on the other hand citations of an idea are distributed between a number of manifestations, thus decreasing the visibility of the specific work.
(2) We carried out an author citation study using several citation databases. The different collection and indexing policies of the different databases lead to considerably different results.

Based of this study, we raise the following questions: How are the science policy makers and promotion committees going to cope with both questions? How should citation studies handle multiple manifestations of the same work? When there are several citation databases available, which database should we use?

Acknowledgements

My special thanks to Michael Rabin for valuable discussions and insights on his works. Thanks to Martin Golumbic, Irith Ben-Arroyo Hartman and Seffi Naor, the organizers of the Fifth Haifa Workshop on Interdisciplinary Applications of Graph theory, Combinatorics, and Algorithms, for inviting all former students of Michael Rabin, and thus providing me the opportunity to prepare this work.

References

Baird, L. M., & Oppenheim, C. (1994). Do citations matter? *Journal of Information Science, 20*, 2–15.

Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication*. London: Taylor-Graham.

Cronin, B., & Shaw, D. (2002). Identity-creators and image-makers: Using citation analysis and thick description to put authors in their place. *Scientometrics, 54*(1), 31–49.

DEAS (2005a). *Michael O. Rabin – DEAS research profile*. Retrieved October 25, 2005. Available from http://www.deas.harvard.edu/ourfaculty/profile/Michael_Rabin.

DEAS (2005b). *Michael O. Rabin – Professional bio*. Retrieved October 25 2005. Available from http://www.deas.harvard.edu/directory/professionalbio/index.html?id=2542.

Giles, C. L., Bollacker, K. D., & Lawrence, S. (1998). Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM conference on digital libraries* (pp. 89–98). New York: ACM Press. Retrieved October 25, 2005. Available from http://clgiles.ist.psu.edu/papers/DL-1998-citeseer.pdf.

Goodrum, A. A., McKain, K. W., Lawrence, S., & Giles, C. L. (2001). Scholarly publishing in the Internet age: A citation analysis of computer science literature. *Information Processing and Management, 37*, 661–675.

Google Scholar (2005). About Google Scholar. Retrieved October 25, 2005. Available from http://scholar.google.com/scholar/about.html.

IFLA (1998). Functional requirements for bibliographic records. Final report. *UBCIM publications – New series, Vol. 19*. Retrieved October 25, 2005. Available from http://www.ifla.org/VII/s13/frbr/frbr.pdf.

ISI (2005). *Source index*. Retrieved October 25, 2005. Available from http://wos23.isiknowledge.com/searchaid/searchaid.cgi?

Lawani, S. M. (1982). On the heterogeneity and classification of author self-citations. *Journal of the American Society for Information Science, 33*, 281–284.

Liu, M. (1993). The complexities of citation practice: A review of citation studies. *Journal of Documentation, 49*, 370–408.

Rabin, M. O. (2004). *Michael O. Rabin*. Retrieved October 25 2005. Available from http://people.deas.harvard.edu/~rabin/cv9.04.pdf.

Rabin, M. O. (2005). *Personal communication*, May 2005.

Sullivan, D. (2004). Google Scholar offers access to academic information. In *Searchday*. Retrieved October 25, 2005. Available from http://searchenginewatch.com/searchday/article.php/3437471.

White, H. D. (2000). Toward ego-centered citation analysis. In B. Cronin & Atkins H. B. (Eds.), *The Web of knowledge, ASIST Monograph Series* (pp. 475–496).

White, H. D. (2001a). Authors as citers over time. *Journal of the American Society for Information Science and Technology, 52*(2), 87–108.

White, H. D. (2001b). Author-centered bibliometrics through CAMEOs: Characterizations automatically made and edited online. *Scientometrics, 51*(3), 601–637.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science, 49*(4), 327–355.

Wikipedia (2005a). *Nondeterministic algorithms*. Retrieved October 25, 2005. Available from http://en.wikipedia.org/wiki/Nondeterministic.

Wikipedia (2005b). *Randomized algorithm*. Retrieved October 25, 2005. Available from http://en.wikipedia.org/wiki/Randomized_algorithm.

Judit Bar-Ilan is a senior lecturer at the Department of Information Science of the Bar-Ilan University, Israel. She received her PhD in computer science from the Hebrew University of Jerusalem. Her areas of interest include: information retrieval, informetrics, the semantic Web, Internet research, information behavior and usability.