Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing

C. Lee Giles* and Isaac G. Councill

School of Information Sciences and Technology, Pennsylvania State University, 311 IST Building, University Park, PA 16802

Communicated by James N. Gray, Microsoft Corporation, San Francisco, CA, November 2, 2004 (received for review July 7, 2004)

Acknowledgments in research publications, like citations, indicate influential contributions to scientific work. However, acknowledgments are different from citations; whereas citations are formal expressions of debt, acknowledgments are arguably more personal, singular, or private expressions of appreciation and contribution. Furthermore, many sources of research funding expect researchers to acknowledge any support that contributed to the published work. Just as citation indexing proved to be an important tool for evaluating research contributions, we argue that acknowledgments can be considered as a metric parallel to citations in the academic audit process. We have developed automated methods for acknowledgment extraction and analysis and show that combining acknowledgment analysis with citation indexing yields a measurable impact of the efficacy of various individuals as well as government, corporate, and university sponsors of scientific work.

acknowledgment analysis | information extraction | machine learning

Since the introduction of the Science Citation Index (1), researchers, funding agents, promotion and tenure committees, and others have used citation index measures to ascertain the quantity and quality of the impact of articles and authors as well as to explore the topical and social structure of scientific communities (2). However, citations alone can fall short of describing the full network of influence underlying primary scientific communication. In addition to referencing published material, many researchers choose to document their appreciation of important contributions through acknowledgments. Acknowledgments may be made for a number of reasons but often imply significant intellectual debt. Just as citation indexing proved to be an important tool for evaluating research contributions, acknowledgments can be considered a metric parallel to citations in the academic audit process (3). Whereas citations are formal expressions of debt, acknowledgments are arguably more personal, singular, or private expressions of appreciation and contribution. We have developed automated intelligent methods for acknowledgment extraction and analysis and show that analysis of acknowledgments uncovers important trends not only in reference to individual researchers but also regarding institutional and agency sponsors of scientific work.

Acknowledgments embody a wide range of relationships among people, agencies, institutions, and research. Classification schemes (4) have been proposed for six categories of acknowledgment: (*i*) moral support; (*ii*) financial support; (*iii*) editorial support; (*iv*) presentational support (e.g., presenting a paper at a conference); (*v*) instrumental/technical support; and (*vi*) conceptual support, or peer interactive communication (PIC). Of all of the categories, PIC has been considered the most important for identifying intellectual debt (5); some researchers have considered acknowledgments of PIC to be at least as valuable as citations (3, 6).

In addition to analyzing PIC, we show that analysis of "financial support" and "instrumental/technical support" acknowledgments give insights into other trends in scientific communities. For example, acknowledgments of financial support may be used to measure the relative impact of funding agencies and corporate sponsors on scientific research (7–9). Acknowledgments of instrumental/technical support may be useful for analyzing indirect contributions of research laboratories and universities to research activities. In short, acknowledgments can help us to better understand the context of scientific research.

Despite their promise as an analytic tool, acknowledgments have remained a largely untapped resource. Presumably, the reason that acknowledgments are not currently included in major scientific indices has to do with cost. Until recently, two models for dealing with the cost of data extraction have been proposed for citations: a centralized model in which an organization pays employees for manual indexing and offers the results as a service [this model is used by the Institute for Scientific Information (ISI), although ISI does not index acknowledgments], and a distributed model that would shift the labor of citation indexing to authors (10). Recently, an approach similar to Cameron's was proposed that would require authors to provide tagged descriptions of the contributions of all intellectual contributors, including those warranting acknowledgment (11). Although distributed models promise to reduce the cost of indexing while increasing coverage, such systems have not been realized.

Autonomous citation indexing (ACI) has recently emerged as an alternative for the creation of citation indices (12, 13). Through ACI the cost associated with manual information extraction is eliminated with manual intervention replaced by parsing algorithms that automatically create citation indices. Because neither the centralized nor distributed models of citation indexing have yet been successfully applied to acknowledgment indexing, we look to ACI as a framework for mining acknowledgment information. To this end, we have created an information extraction algorithm to automatically extract acknowledgments from research publications.

We use the CiteSeer digital library (http://citeseer.ist.psu.edu), created in 1998 as a prototype to demonstrate ACI, as both data source and deployment architecture for our algorithm. At the time of this study the CiteSeer archive contained cached copies of over 425,000 unique computer science research papers harvested from the web and submitted by users. To explore the viability of using the CiteSeer archive as a sample of computer science publications, we have cross-referenced the archive with the Digital Bibliography and Library Project (DBLP; http://dblp.uni-trier.de), a database of bibliography information for 438 journals and 2,373 proceedings in the field of computer science. The DBLP contained 500,464 records at the time of this study, in comparison with the 141,345 records in the Association for Computing Machinery (ACM) digital library and the 825,826

COMPUTER SCIENCES

Abbreviations: DBLP, Digital Bibliography and Library Project; PIC, peer interactive communication; SVM, support vector machine.

*To whom correspondence should be addressed. E-mail: giles@ist.psu.edu.

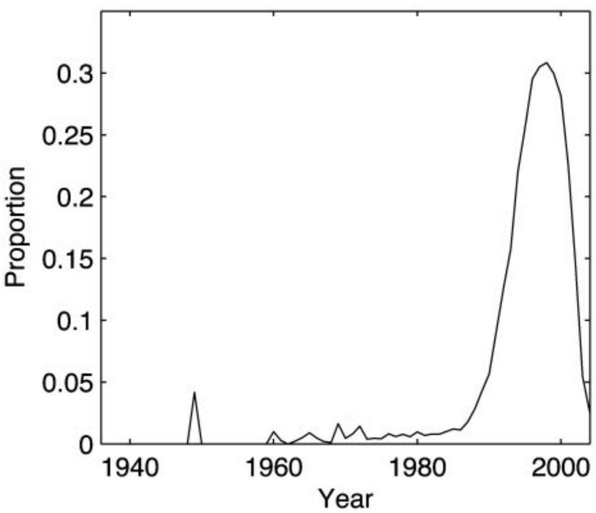

Fig. 1. The proportion of all documents indexed by DBLP that are contained in CiteSeer by year.

records contained by the more comprehensive ACM Guide. The DBLP contains records for a significant portion of the ACM digital library: complete data for 29 of 41 ACM journals (70.7%) and 117 of 209 ACM proceedings (56.0%).

By using exact title match, we obtained a lower bound estimate of the proportion of documents indexed by DBLP contained in CiteSeer. It was found that there are at least 86,467 documents overlapping between CiteSeer and the DBLP, comprising 20.2% of CiteSeer's total archive and 17.3% of the DBLP archive. The DBLP indexes publications from as early as 1936; however, CiteSeer contains mostly documents from the 1990s to the present (see Fig. 1). Given the bias of our sample, we restrict our analyses to the time period from 1990 to 2004. Fig. 2 shows the proportion of all DBLP journals and proceedings contained in CiteSeer from 1990 to 2004. We observe imbalanced coverage of CiteSeer for publication venues in the DBLP, indicating bias in our document sample. Not all venues are represented equally, indicating that computer science subcommunities may also have disproportionate representation. This bias complicates the comparisons of entities through either citation counts or acknowledgment counts. However, we believe that our comparison of CiteSeer with the DBLP shows that our collection is large and diverse enough to generate interesting analyses. The bias in our results could be alleviated in future studies either by using complete archives of publication venues or by restricting our analyses to documents within particular subcommunities of computer science.

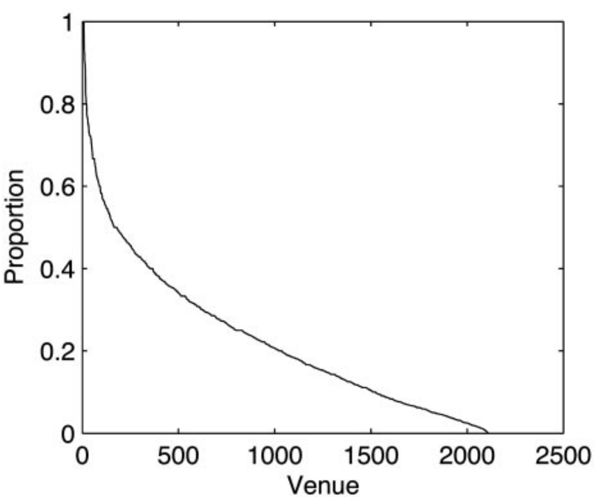

Fig. 2. The proportion of all publication venues in DBLP contained by CiteSeer, where the venues are ordered by the amount of coverage received in CiteSeer.

We extracted acknowledgments from 335,000 unique documents from CiteSeer and have analyzed the results for the top acknowledged funding agencies, corporations, universities, and individuals.

Automatic Acknowledgment Extraction and Indexing

The problem of extracting acknowledgments from research articles can be viewed as a specific case of automatic document metadata extraction. Several approaches have been proposed for automatic metadata extraction, with the most common tools including regular expressions, rule-based parsers, and machine learning algorithms. Regular expressions and rule-based parsers are easily implemented and can perform acceptably well if data are well behaved. Machine learning techniques are generally more robust and easily adaptable to new data. Machine learning methods used for information extraction include inductive logic programming, grammar induction, symbolic learning, hidden Markov models, and support vector machines (SVMs). Because of recent success using SVMs for learning in high-dimensional feature spaces (14, 15), SVMs are becoming increasingly popular tools for classification. Recent work has shown it possible to recast the problem of information extraction as a classification task (16), and SVMs have been proven to be effective for chunk identification and named entity extraction (17–20).

While highly effective at metadata extraction, much recent work using machine learning for information extraction (17, 21) exploits the semistructured format of document headers for chunk identification and classification. The problem of acknowledgment extraction involves the identification of chunks of a single class found most often within free text. We have found that regular expressions work acceptably well for identifying the names of acknowledged entities within identifiable acknowledgment passages.

The first step in extracting acknowledgments is extracting text that is likely to contain acknowledgments. We have two techniques for achieving this based on whether acknowledgment passages are labeled or unlabeled. Most acknowledgments in research papers are found in clearly identifiable acknowledgment sections within documents. Acknowledgment sections are easily identified using regular expressions by searching for lines containing only the word "acknowledgment" in various forms and extracting all of the following text until the next section header. However, acknowledgment passages may also be found in unmarked sections, within the document header, or within footnotes. These acknowledgment passages are typically found at the beginning of documents (before the abstract or introduction, or on the first page) and at the end (before the references or first appendix). To identify these passages, we extract roughly the first page of the document and the last page before the reference section or the first appendix, whichever comes first. We then classify the lines of extracted text by using a SVM to identify those lines containing acknowledgments.

Our SVM line classifier may produce errors of recall for multiline acknowledgment passages. For example, a footnote may contain patterns that indicate an acknowledgment in the first line but the second line may only contain names of acknowledged entities with no other context. Our SVM would produce a false negative on the second line in this example. To make matters worse, the misclassified line may contain only partial names (for example, only "Giles" from the complete phrase "C. Lee Giles"), producing errors of precision. We mitigate these problems by merging positively classified lines with surrounding lines of negatively classified text. The context merging technique improves line classification recall by 17.34% and produces an 8.70% precision improvement for subsequently extracted entity names.

Text passages extracted by using the above methods are parsed by using a regular expression to extract the names of acknowledged entities. Finally, name variants are merged to account for different ways of referring to entities. For example, our algorithm identifies "National Science Foundation" and "NSF" as references to the same entity. We achieve this through two methods. First, full names and acronyms that are adjacent to each other in acknowledgment passages, and ordered name acronym, i.e., "National Science Foundation" and "NSF," are compared. If it is found that the acronym letters match the first letters of all words in the expanded name, the two name variants are identified as referring to the same entity for all occurrences. A weakness of this method is that not all name variants are merged. This is particularly true for individuals. For example, a person might be acknowledged with or without a middle initial in different acknowledgment passages, resulting in the identification of two individuals where there is only one. Full entity name disambiguation is not a trivial task (22) and can be a topic for future work.

Through rigorous testing involving 1,800 manually labeled documents we have shown our algorithm to achieve 78.45% precision and 89.55% recall, reflecting an intentional bias toward recall.

Acknowledgment Metrics for Funding Agents, Companies, Educational Institutions, and Individuals

We have applied our acknowledgment extraction algorithm to 335,000 unique research documents within the CiteSeer computer science archive. Of these documents, 188,052 were found to contain acknowledgments (≈56% of our papers). This result is consistent with a previous study of acknowledgments in information science journals (23). The names of acknowledged entities were automatically extracted and linked to the source articles for analysis.

Initial analyses revealed that the distribution of acknowledgments to named entities (e.g., "National Science Foundation" or "John Smith") within the CiteSeer archive follows a power law such that only a few entities are named very frequently and a great many entities are named only rarely (see Fig. 3). The power law trend in acknowledgments has been reported in a study involving manual extraction of acknowledgments from research papers within information science and sociology journals (3, 24). An analysis of the Institute for Scientific Information (ISI) data set (25) has shown that citations also follow a power curve. The ISI study shows an exponent of approximately −0.5 for the distribution of citations, comparable to our finding that CiteSeer's citation distribution follows an exponent of −0.55. Our acknowledgment data fits a power law with an exponent of −0.65, a significantly steeper slope than that exhibited by citations. We explain this by noting a high proportion of acknowledgments given to a relatively small and static list of funding agencies. These agencies fund work in many subcommunities within computer science. In contrast, we expect but have not shown that a greater number of research papers will be found within the top echelons of cited work and that citations will be shared among many classic papers according to particular scientific communities.

In addition to acknowledgment frequency, acknowledgment results were coupled with data from CiteSeer's citation index to measure the collective impact of acknowledging articles. CiteSeer maintains a graph of all citations made within the document collection, such that it is possible to retrieve the number of times each document is cited by other documents within the collection.

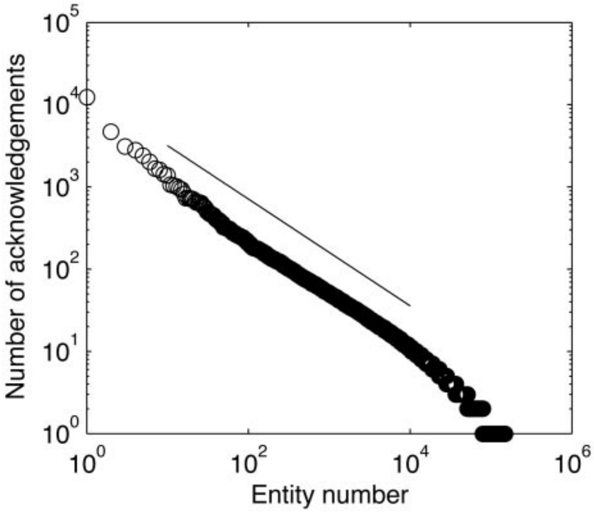

Fig. 3. The distribution of acknowledgments in the CiteSeer document collection follows a power law with the exponent −0.65. A line with −0.65 slope is drawn for reference.

For each acknowledged entity, we calculate the number of acknowledgments made to the entity and the total number of citations made to those articles acknowledging the entity. Additionally, we calculate the ratio of the total number of citations over the number of acknowledgments for each entity, which we define as the C/A metric. In this manner, we measure the relative impact of documents acknowledging each entity as well as the average impact.

The most acknowledged entities were manually reviewed to split the results for the most acknowledged entities into four categories: funding agencies, corporations, universities, and individuals. We assume that acknowledgments to funding agencies and companies represent acknowledgments of financial support and that acknowledgments to individuals represent PIC. Although it is unreasonable to suggest that all acknowledgments to individuals in our data represent PIC, a manual review of 100 randomly sampled acknowledgments to each of the 15 most acknowledged individuals verified this assumption. We have verified our assumption regarding the type of acknowledgments received by funding agencies, corporations, and universities through similar analyses. To extend our analyses to less-acknowledged entities within our data, it will be necessary to develop automatic means of classifying both entity types and the context of acknowledgments (from Cronin's typology). We are currently exploring solutions to this problem through a combination of lookup tables and machine learning techniques.

We believe our acknowledgment counts generate a fairly complete picture of the informal influence funding agencies and individuals have had within our document collection. However, accounting for the influence of companies and educational institutions is not so easily achieved through acknowledgments. Specifically, it is not common to explicitly acknowledge one's home institution for supporting published work. More complete analyses could be generated by taking into account author affiliation data found in document headers and treating statements of affiliation as de facto acknowledgments.

The 15 most acknowledged entities in each category are presented in Table 1. The results show significant variation not only in total acknowledgments received by entities but also in the average citations to acknowledging articles. For example, the most acknowledged entity (the National Science Foundation) received 2.6 times the total acknowledgments of the next most acknowledged entity (the Defense Advanced Research Projects

COMPUTER SCIENCES

Table 1. The 15 most acknowledged entities in four categories: funding agencies, companies, educational institutions, and individuals

Name	No. of acknowledgments	Total citations	C/A metric
Funding agencies			
National Science Foundation	12,287	144,643	11.77
Defense Advanced Research Projects Agency	4,712	80,659	17.12
Office of Naval Research	3,080	48,873	15.87
Deutsche Forschungsgemeinschaft	2,780	9,782	3.52
National Aeronautics and Space Administration	2,408	21,242	8.82
Engineering and Physical Sciences Research Council	2,007	16,582	8.26
Air Force Office of Scientific Research	1,657	16,850	10.17
National Sciences and Engineering Research Council of Canada	1,422	12,050	8.47
Department of Energy	1,054	5,562	5.28
Australian Research Council	1,010	5,464	5.41
European Union Information Technologies Program	825	9,594	11.63
National Institutes of Health	709	7,279	10.27
Army Research Office	666	7,709	11.58
Netherlands Organization for Scientific Research	646	2,843	4.4
Science and Engineering Research Council	489	6,976	14.27
Companies			
International Business Machines	1,380	23,948	17.35
Intel Corporation	962	14,441	15.01
Digital Equipment Corporation	831	16,390	19.72
Hewlett-Packard	735	11,186	15.22
Sun Microsystems	651	12,042	18.5
Microsoft Corporation	368	6,061	16.47
Silicon Graphics, Inc	279	3,898	13.97
Xerox Corporation	265	4,309	16.26
Siemens Corporation	241	8,395	34.83
Bellcore	192	2,393	12.46
Nippon Electric Company	164	942	5.74
SRI International	163	1,450	8.9
AT&T Bell Labs	146	1,549	10.61
Apple Computer	135	3,159	23.4
Motorola	122	1,352	11.08
Educational institutions			
Carnegie Mellon University	640	10,840	16.94
Massachusetts Institute of Technology	500	10,509	21.02
California Institute of Technology	464	4,170	8.99
Santa Fe Institute	368	3,387	9.2
French National Institute for Research in Computer Science	321	3,399	10.59
Stanford University	314	3,693	11.76
University of California at Berkeley	306	10,439	34.11
National Center for Supercomputing Applications	261	4,777	18.3
International Computer Science Institute	180	2,078	11.54
Cornell University	180	1,656	9.2
University of Illinois at Urbana–Champaign	177	5,304	29.97
USC Information Sciences Institute	176	3,283	18.65
University of California, Los Angeles	176	2,003	11.38
McGill University	152	3,001	19.74
Swedish Institute for Computer Science	134	2,017	15.05
Individuals			
Olivier Danvy	268	8,000	29.85
Oded Goldreich	259	4,615	17.82
Luca Cardelli	247	10,846	43.91
Tom Mitchell	226	5,494	24.31
Martin Abadi	222	9,647	43.46
Phil Wadler	181	7,252	40.07
Moshe Vardi	180	6,094	33.86
Peter Lee	167	8,941	53.54
Avi Wigderson	160	2,901	18.13
Matthias Felleisen	154	4,705	30.55
Benjamin Pierce	152	4,641	30.53
Noga Alon	152	2,388	15.71
John Ousterhout	152	6,369	41.9
Frank Pfenning	148	2,049	13.84
Andrew Appel	144	7,630	52.99

Table 2. Number of citations to the most acknowledged individuals

Author	Acknowledgments	Citations
Olivier Danvy	268	847
Oded Goldreich	259	3,277
Luca Cardelli	247	3,847
Tom Mitchell	226	3,336
Martin Abadi	222	3,507
Phil Wadler	181	3,780
Moshe Vardi	180	3,786
Peter Lee	167	1,790
Avi Wigderson	160	2,566
Matthias Felleisen	154	1,622
Benjamin Pierce	152	1,484
Noga Alon	152	2,640
John Ousterhout	152	3,693
Frank Pfenning	148	1,639
Andrew Appel	144	2,064

Agency, DARPA), but the NSF-supported articles received only 1.8 times the total number of citations of DARPA and <0.7 times the mean citations of DARPA. Likewise, the Army Research Office (ARO) has been acknowledged only 63% as much as the Department of Energy (DOE), but ARO-supported work has 37% more total citations, indicating that the ARO has had more impact within our document collection despite being less frequently acknowledged.

For the most acknowledged companies, we see companies that are or were known for their support of research. For education institutions, we see the expected collection of research-oriented universities, although some of a much larger number of acknowledgments than others. In the category of individual researchers, the results show that some individuals have received more acknowledgments than some popular corporate sponsors of research and well respected educational institutions. For example, within our document collection only seven educational institutions and seven companies have been acknowledged more frequently than Olivier Danvy. We interpret these results to indicate a large degree of intellectual debt to individuals documented through the mechanism of acknowledgment. However, it should be noted that among the top 15 acknowledged entities in all categories, funding agencies received more acknowledgments than any other category by an order of magnitude. Counting author affiliations as acknowledgments may reveal that companies and educational institutions have impacted scientific work on a scale similar to funding agencies.

Table 2 shows that the number of citations made to the most acknowledged individuals does not correlate well with the number of acknowledgments to those individuals. This is consistent with previous studies of acknowledgment trends (3). We have cross-referenced our acknowledgment data with author names in the CiteSeer database and found that 9,474 of the 10,000 most cited author names are acknowledged. Using this sample, we found that there is a 0.3406 correlation coefficient between the number of acknowledgments and the number of citations received by authors. Although anonymous entities received more acknowledgments than any single entity (12,228), such acknowledgments are excluded from our analyses.

A temporal analysis of the 10 most acknowledged entities indicates distinct patterns of acknowledgment over time. Although most of the top acknowledged entities exhibit a stable proportion of acknowledgments each year, it can be seen from Fig. 4 that both the German Science Foundation (Deutsche Forschungsgemeinschaft) and the United Kingdom Engineering and Physical Sciences Research Council (EPSRC) display a steady upward trend in the proportion of acknowledgments received each year during the 1990s, whereas the Office of Naval Research and IBM slowly become overshadowed by other entities over the decade.

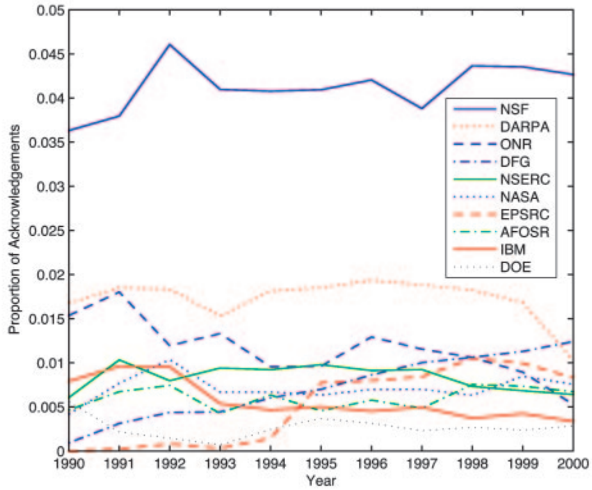

Fig. 4. The proportion of documents per year acknowledging the 10 most acknowledged entities in the CiteSeer document collection.

Shown in Table 3 is another measure of impact: the most acknowledged entities collected for the 100 most cited papers in the CiteSeer database. We found that 429 acknowledgments were made within this document sample, averaging 4.29 acknowledgments per paper. Not surprisingly, funding agents known for funding research are at the top, but a university and two companies round out the top 10. The National Science Foundation was acknowledged by an impressive 26% of the top 100 papers. When acknowledgments for the 100 most cited papers are ranked by entity type, as shown in Table 4, individuals have more acknowledgments. This is natural because individuals who can contribute to scientific work greatly outnumber institutional contributors.

Impact of Acknowledged Entities

The results obtained from our acknowledgment extraction algorithm have shown it to be a viable tool for automatically creating initial analyses of the relative impacts of acknowledged entities in document collections. We believe that our technique is general enough that it can readily be applied to digitized collections of research publications other than CiteSeer. We have presented the most acknowledged entities within the CiteSeer document collection with two distinct measures: the

Table 3. Most acknowledged entities in the 100 most cited papers in the CiteSeer database

Entity	No. of acknowledgments
NSF	26
DARPA	19
DOE	9
ONR	6
IBM	5
AFOSR	5
NASA	4
AT&T Bell Labs	3
MIT	3
NSERC	3

COMPUTER SCIENCES

Table 4. Number of acknowledgments by entity type for 100 most cited papers in the CiteSeer database

Entity type	No. of acknowledgments
Funding agency	91
Educational institution	19
Company	21
Individual	298

number of acknowledgments received and the mean citations of the acknowledging papers. We take the raw number of acknowledgments to measure the breadth of contributions entities have made to the research community. For funding agencies and corporate sponsors this may correlate with the amount of funding contributed to research. For individuals, the number of acknowledgments may indicate the extent to which acknowledged persons influence other researchers through informal channels of communication. The distribution of acknowledgments within our document collection follows the distribution found through prior studies of information science and sociology publications; thus, our results may indicate trends across disciplines.

Our results show that individual scientists may be more widely acknowledged than popular corporate funding sources. Additionally, our work supports prior studies showing that acknowledgment trends for individuals do not correlate well with citation trends, perhaps indicating a need to reward highly acknowledged researchers with the deserved recognition of significant intellectual debt.

By cross-referencing the number of acknowledgments made to entities with the number of citations made to the acknowledging papers, we can measure the average impact of the research influenced by an entity. This is particularly interesting for analyzing the relative impacts of funding agencies and companies who invest in research. Through impact measures it will be possible to compare the effectiveness of funding programs and to calculate the return on investments in terms of the average research impact per dollar spent. It should be noted, however, that the average citations to all funded works should not be used to measure the efficacy of funding agencies directly because some funding programs may realize their impact, in part, by providing educational opportunities to young scientists rather than funding the "best" work in the field. It should be possible to provide a more detailed level of analysis in the future by capturing grant numbers and titles during the acknowledgment extraction process. Further work could explore temporal, national, and international trends in acknowledgments. For most funded research, acknowledgments to the appropriate funding agency are requested. Combined with access to all published documents and other measures such as funding levels, we speculate that these measures could be used to evaluate the efficacy of funding agencies and programs both at the national and international level.

We thank Steve Lawrence, David Mudgett, and Frank Ritter for useful discussions and the comments provided by anonymous reviewers. This work was partially supported by National Science Foundation Grants 0330783 and 0121679 and by Microsoft Research.

1. Garfield, E. (1964) *Science* **144,** 649–654.
2. Shiffrin, R. M. & Börner, K. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5183–5185.
3. Cronin, B., McKenzie, G., Rubio L. & Weaver-Wozniak, S. (1993) *J. Am. Soc. Inf. Sci.* **44,** 406–412.
4. Cronin, B., Shaw, D. & La Barre, K. (2003) *J. Am. Soc. Inf. Sci. Technol.* **54,** 855–871.
5. McCain, K. W. (1991) *Sci. Technol. Hum. Val.* **16,** 491–516.
6. Edge, D. (1979) *Hist. Sci.* **17,** 102–134.
7. Cronin, B. & Shaw, D. (1999) *J. Doc.* **55,** 404–408.
8. Henderson, C., Howard, L. & Wilkinson, G. (2003) *Br. J. Psychiatry* **183,** 273–275.
9. Jeschin, D., Lewison, G. & Anderson, J. (1995) in *Proceedings of the 5th Biennial Conference of the International Society for Scientometrics and Informetrics*, eds. Koenig, M. & Bookstein, A. (Learned Information, Medford, NJ), pp. 235–244.
10. Cameron, R. D. (1997) *First Monday* **2,** www.firstmonday.org.
11. Davenport, E. & Cronin, B. (2001) *J. Am. Soc. Inf. Sci. Technol.* **52,** 770.
12. Giles, C. L., Bollacker, K. & Lawrence, S. (1998) in *ACM Conference on Digital Libraries*, eds. Witten, I., Akseyn, R. & Shipman, F. M. (ACM Press, New York), pp. 89–98.
13. Lawrence, S., Giles, C. L. & Bollacker, K. (1999) *IEEE Computer* **32,** 67–71.
14. Joachims, T. (1998) in *Advances in Kernel Methods: Support Vector Machines*, eds. Schölkopf, B., Burges, C. & Smola, A. (MIT Press, Cambridge, MA), pp. 169–184.
15. Dumais, S., Platt, J., Heckerman, D. & Sahami, M. (1998) in *Proceedings of the 7th International Conference on Information and Knowledge Management*, eds. Makki, K. & Bourganim, L. (ACM Press, New York), pp 148–155.
16. Chieu, H. L. & Ng, H. T. (2002) in *Proceedings of the 18th National Conference on Artificial Intelligence*, eds. Dechter, R., Kearns, M. & Sutton, R. (AAAI Press, Menlo Park, CA), pp 786–791.
17. Han, H., Giles, C. L., Manavoglu, E. & Zha, H. (2003) in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, eds. Delcambre, L., Henry, G. & Marshall, C. C. (IEEE Computer Soc., Washington, DC), pp. 37–48.
18. McNamee, P. & Mayfield, J. (2002) in *Proceedings of the 6th Conference on Natural Language Learning*, eds. Roth, D. & van den Bosch, A. (Association for Computational Linguistics, New Brunswick, NJ), pp. 183–186.
19. Kudoh, T. & Matsumoto, Y. (2000) in *Proceedings of the 4th Conference on Natural Language Learning*, eds. Cardie, C., Daelemans, W., Nedellec, C. & Sang, T. K. (Association for Computational Linguistics, New Brunswick, NJ), pp. 142–144.
20. Takeuchi, K. & Collier, N. (2002) in *Proceedings of the 6th Conference on Natural Language Learning*, eds. Roth, D. & van den Bosch, A. (Association for Computational Linguistics, New Brunswick, NJ), pp. 119–125.
21. Seymore, K., McCallum, A. & Rosenfeld, R. (1999) in *Proceedings of the AAAI 99 Workshop on Machine Learning for Information Extraction*, ed. Califf, M. E. (AAAI Press, Menlo Park, CA), pp 37–42.
22. Han, H., Giles, C. L., Zha, H., Li, C. & Tsioutsiouliklis, K. (2004) in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, eds. Chen, H., Wactlar, H. & Chen, C.-C. (ACM Press, New York), pp. 296–305.
23. Cronin, B. (2001) *J. Doc.* **57,** 427–433.
24. Davis, C. H. & Cronin, B. (1993) *J. Am. Soc. Inf. Sci.* **44,** 590–592.
25. Redner, S. (1998) *Eur. Phys. J.* **4,** 131–134.