Sci Eng Ethics (2009) 15:467–489
DOI 10.1007/s11948-009-9119-4

ORIGINAL PAPER

Programmers, Professors, and Parasites: Credit and Co-Authorship in Computer Science

Justin Solomon

Received: 21 September 2008 / Accepted: 27 January 2009 / Published online: 27 February 2009

Abstract This article presents an in-depth analysis of past and present publishing practices in academic computer science to suggest the establishment of a more consistent publishing standard. Historical precedent for academic publishing in computer science is established through the study of anecdotes as well as statistics collected from databases of published computer science papers. After examining these facts alongside information about analogous publishing situations and standards in other scientific fields, the article concludes with a list of basic principles that should be adopted in any computer science publishing standard. These principles would contribute to the reliability and scientific nature of academic publications in computer science and would allow for more straightforward discourse in future publications.

Keywords Co-authorship · Computer science research · Publishing

In November 2002, a team of computer scientists, engineers, and other researchers from IBM and the Lawrence Livermore National Laboratory presented a conference paper announcing the development of a record-breaking supercomputer. This supercomputer, dubbed the BlueGene/L, would sport a "target peak processing power" of 360 trillion floating-point operations every second [1], enough to simulate the complexity of a mouse's brain [2]. While the potential construction of the BlueGene/L was a major development for the field of supercomputing, the paper announcing its structure unwittingly suggested new industry standards for sharing co-authorship credit in the field of computer science. The paper spanned twenty-two pages of conference proceedings, sixteen of which contained written text. The authors of the paper, on the other hand, spanned both coasts of the United States,

J. Solomon (✉)
Department of Computer Science, Stanford University, PO Box 13092, Stanford, CA 94309, USA
e-mail: justin.solomon@stanford.edu

An Overview of the BlueGene/L Supercomputer

The BlueGene/L Team
IBM and Lawrence Livermore National Laboratory[1]

NR Adiga, G Almasi, GS Almasi, Y Aridor, R Barik, D Beece, R Bellofatto, G Bhanot, R Bickford, M Blumrich, AA Bright, J Brunheroto, C Caşcaval, J Castaños, W Chan, L Ceze, P Coteus, S Chatterjee, D Chen, G Chiu, TM Cipolla, P Crumley, KM Desai, A Deutsch, T Domany, MB Dombrowa, W Donath, M Eleftheriou, C Erway, J Esch, B Fitch, J Gagliano, A Gara, R Garg, R Germain, ME Giampapa, B Gopalsamy, J Gunnels, M Gupta, F Gustavson, S Hall, RA Haring, D Heidel, P Heidelberger, LM Herger, D Hoenicke, RD Jackson, T Jamal-Eddine, GV Kopcsay, E Krevat, MP Kurhekar, AP Lanzetta, D Lieber, LK Liu, M Lu, M Mendell, A Misra, Y Moatti, L Mok, JE Moreira, BJ Nathanson, M Newton, M Ohmacht, A Oliner, V Pandit, RB Pudota, R Rand, R Regan, B Rubin, A Ruehli, S Rus, RK Sahoo, A Sanomiya, E Schenfeld, M Sharma, E Shmueli, S Singh, P Song, V Srinivasan, BD Steinmacher-Burow, K Strauss, C Surovic, R Swetz, T Takken, RB Tremaine, M Tsao, AR Umamaheshwaran, P Verma, P Vranas, TJC Ward, M Wazlowski
IBM Research

W Barrett, C Engel, B Drehmel, B Hilgart, D Hill, F Kasemkhani, D Krolak, CT Li, T Liebsch, J Marcella, A Muff, A Okomo, M Rouse, A Schram, M Tubbs, G Ulsh, C Wait, J Wittrup
IBM Rochester

M Bae (IBM Server Group), K Dockser (IBM Microelectronics)

L Kissel, MK Seager, JS Vetter, K Yates
Lawrence Livermore National Laboratory

Fig. 1 The author list from the BlueGene/L paper [1]

five research labs, and all but four letters of the alphabet. In total, 115 researchers were listed as "authors" of the paper (see Fig. 1)—enough so that had the work been divided evenly each would have contributed about six lines of writing.

As the sheer number of authors on the BlueGene/L paper suggests, the simplest definition of an author by the *Oxford English Dictionary* as "[o]ne who sets forth written statements" no longer holds, if it ever did, for computer science research [3]. Instead, authorship on a computer science paper indicates not that the author made concrete written contributions to the paper itself but rather that he or she played a substantial role in the development of a larger project. As would be expected, the issue of authorship is a crucial one in computer science since authorship can be the key to promotion, tenure, and prestige for researchers advancing through the ranks of academia. A computer scientist whose name appears on a conference presentation slide or published paper has indicated a certain degree of involvement with "publishable" research and has proven that his or her work can withstand the pressure of peer review by a community of academics, engineers, and practitioners. Furthermore, in computer science, publication can be an initial stepping stone toward obtaining a patent for marketable products that stem from research results. Thus, it comes as no surprise that computer scientists compete for positions on lists of authors for published research papers. This "publish-or-perish" attitude, which is well-known to have existed in the natural and social sciences for some time [4], has driven computer science researchers to become more prolific writers as well as competitors for increasingly exclusive spots in academic journals and conferences.

This pressure to publish has led to several practical and ethical problems concerning the assignment of academic credit in computer science research reports. When submitting to academic journals, computer science researchers must decide who to list as an author, who to mention in a footnote or "acknowledgements" section, how to order names, and which author should be contacted for further information. They also must decide whether to list interns or less experienced researchers on the publication to help them establish a place in the research community and whether senior researchers or professors *emeriti* should be granted honorary authorship despite being only nominally associated with a research project. These decisions are especially important in computer science research, as many developments in computer science can be placed directly on the market with little to no modification. The degree of academic credit that a researcher receives thus can affect not only his or her academic career but also the possibility of entering into business given a lucrative development or product.

If a consistent publishing standard were put into place, these questions could be answered satisfactorily. No such standard, however, has been released by either of the principal organizations governing computer science research and publication, the Association for Computing Machinery and the Institute of Electrical and Electronics Engineers [5, 6]. Instead, these organizations have practiced a policy of "salutary neglect" in which there exists an unwritten, informal, and oft-disregarded understanding that authorship of a computer science paper reflects some sort of considerable contribution to the development of the project described by the paper. Consequently, papers published in academic computer science journals exhibit inconsistent patterns in co-authorship and author credit. These patterns represent a hodgepodge of citation procedures from more established fields, particular research groups' policies, and other informal standards. Using these inconsistent methods, it is difficult if not impossible to produce a list of contributors that satisfies every participant in a research project. For this reason, after publication, the only way to discern researchers' respective levels of contribution is through guesswork or retrospection on the part of the researchers themselves; these methods cannot determine easily a concrete list of contributions that would be accepted by all authors. Such irregularity has led to confusion over the nature of particular individuals' contributions to research projects, squabbles over credit, and even legal action in some cases.

Clearly, computer science needs some sort of consistent publishing standard governing attribution of credit in papers, presentations, and other venues for the dissemination of findings. The design of such a standard should rely upon current practices, ethical concerns, legal issues, and successful practices in other fields to produce a reasonable set of rules that gives fair credit to all researchers involved in a particular project. The potential implementation of this policy would make scholarly work in computer science easier to interpret and more reliable, helping the field establish itself as a rigorous "science" in which authors take full responsibility for their research. Additionally, such attention to author credit and responsibility would help computer science differentiate itself from computer engineering and related fields, in which a publication is more likely to represent the end of a line of inquiry than a step toward solving an open problem.

Past and Present Practices

Few studies have examined the evolution of credit in computer science research publications as computer science grew from a small subfield of math or electrical engineering to a field unto itself. Given that computer science has developed more rapidly than its peers in the natural sciences, however, the possibility exists that trends in academic credit and co-authorship in computer science indicate the formation of a unique system that differs significantly from those in other fields. After all, in contrast with protocols for credit in more traditional areas of study, computer science policy as it exists today is mostly the result of no more than forty years of development. Fortunately, the DBLP Computer Science Bibliography database makes its listings of over 8,85,000 computer science papers available for download and subsequent analysis [7]. These entries span the history of computer science, going as far back as Church and Turing's theoretical work in the 1930s on the concept of computability, allowing for large-scale analysis of computer science publishing trends. In all, the DBLP represents about 30% of all computer science literature from a representative set of subfields, journals, and conferences [8].

Statistical analysis of publication records from the DBLP[1] reveals that while certain trends within computer science follow more global trends in scientific publishing, others are unique to computer science itself. As expected, computer science research starting in the 1980s has experienced a steep increase in the number of papers with multiple authors (see Fig. 2). This trend is logical given that computer science expanded rapidly in the 1980s due to the invention and eventual ubiquity of the personal computer. As the field became more popular throughout the decade, research groups in computer science grew as well, leading to papers with larger author lists. Furthermore, a similar rise in multiple authorship has been documented extensively within other sciences through the "scientometric" analysis of various databases of scientific publications [9]. After all, although the phrase "publish or perish" may have evolved in the 1940s or earlier, the past thirty years have shown the largest amplification of publishing pressure for academic researchers [10]. In fact, the 1993 Ig Nobel prize for "improbable research" in literature was awarded to "E. Topol, R. Califf, F. Van de Werf, P.W. Armstrong, and their 972 coauthors, for publishing a medical research paper which has one hundred times as many authors as pages," providing an extreme instance of bloat in publication lists for medical papers [11].

In general, team sizes of two to three members have become the most prominent in computer science (see Fig. 3). Since current publishing standards are so vague, however, the lack of a method for discerning between contributing authors, assistants, and honorary authors makes it impossible to tell whether these small

[1] Analyses of the DBLP and NRC *Research-Doctorate Programs in the United States* data were carried out by the author. Programs in C++ were devised for parsing and analyzing the data; for example, the NRC data analysis program is shown in the Appendix. Figures 2–6 were produced using the output of these programs, exported to a spreadsheet application. Occasionally it was not possible to parse the data correctly (due to incorrect formatting or other inconsistencies); these situations were documented (see the Appendix for an example). Instances of this problem were relatively small and should not affect the trends observed in this study.

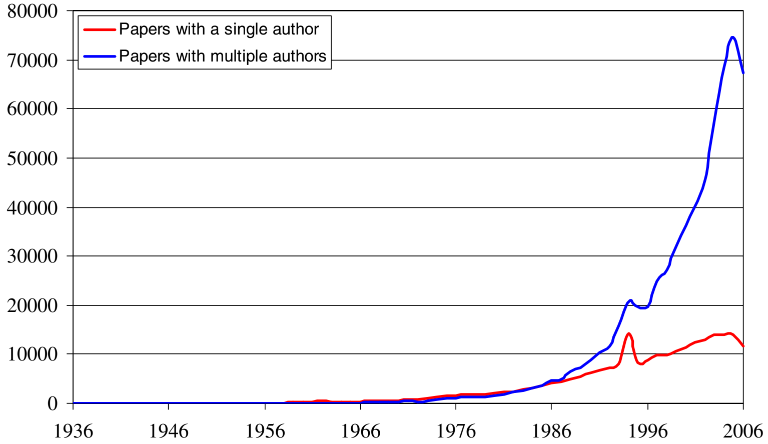


Fig. 2 Trends in single and multiple authorship in computer science papers (data gathered from the *DBLP Computer Science Bibliography*)

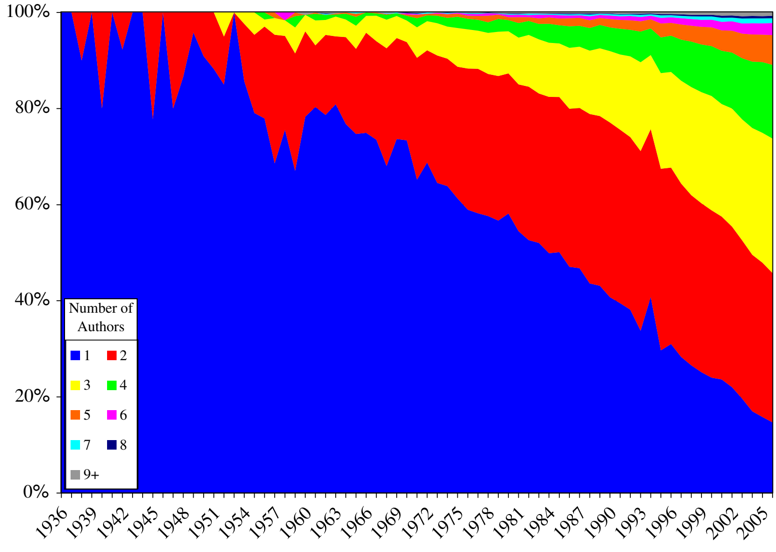


Fig. 3 Trends in single and multiple authorship in computer science papers by percent (data gathered from the *DBLP Computer Science Bibliography*)

teams consist primarily of one main author and one to two assistants, one main author and his or her advisors, groups of equally-contributing members, or some other combination of various members in the research process. Future research could determine the nature of these relationships through the use of survey data, although secondary contributors to research projects could overestimate their involvement, making the data difficult to interpret. Regardless, the proportion of papers with team sizes larger than two to three members has grown significantly as

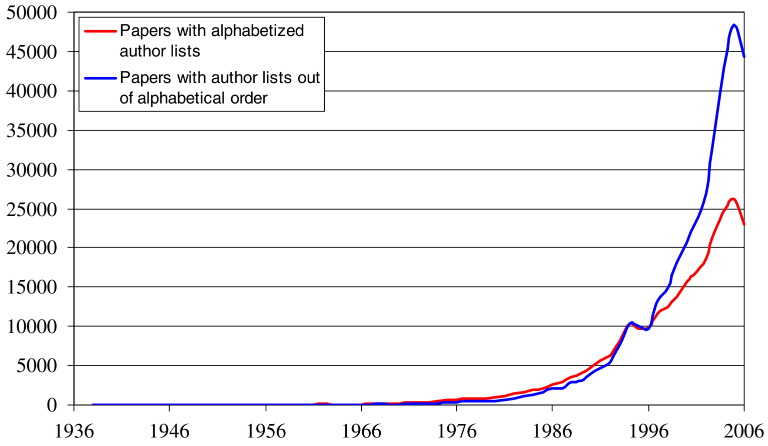


Fig. 4 Trends in author list alphabetization in computer science papers (data gathered from the *DBLP Computer Science Bibliography*)

well, making a standard governing credit for publications by larger-sized teams necessary. Such a standard at least would have made the author list on the BlueGene/L paper much easier to navigate.

Interestingly, the proportion of co-authored computer science papers with author lists in alphabetical order by last name has decreased significantly within the last decade (see Fig. 4). In particular, the ratio of alphabetized to non-alphabetized author lists began shrinking rapidly in 1996 from approximately 1:1 to 1:2 within the span of ten years. This pattern is not common to all research areas; in fact, alphabetization rates in "top tier" economics journals increased between 1978 and 2000 [12]. Thus, any explanation for this type of trend must be discipline specific and possibly even sub-discipline specific. For instance, alphabetized author lists have become more common in agricultural economics but less common within other subfields of economics [13]. Furthermore, basic explanations for such a trend involving standards released or academic institutions changing their policies in 1996 are not apparent. Neither the ACM nor the IEEE, the two organizations that govern most publications in computer science, released any sort of instructions concerning author lists for their publications in 1996 or any other nearby years [5, 6]. In addition, there is little evidence showing that any of the principal organizations actively pursuing computer science research at that time put out a similar standard.

On a larger scale, concern about the order in which authors appear on papers has existed in the sciences ever since co-authorship became the norm. Carl Djerassi, a Stanford chemistry professor known for synthesizing the first oral contraceptive, acknowledges this concern in his novel *Cantor's Dilemma*, which describes the efforts of a young female researcher attempting to establish a career in the sciences:

> When I was a senior at Brown—and a very ambitious one, almost unpleasantly so—I paid very much attention to where my name would ultimately appear. Of course, I'd never published a paper; I hadn't even decided where to go to

> graduate school. To my father's shock, I announced one day that I would change my name from Jean Yardley to Jean Ardley. Just like that! […] I went to the courthouse and did it legally. I told the judge, "It's best to be first, it's been true since prehistoric times." [14]

Statistical evidence suggests that alphabetical order had little effect on a scientist's career near the time the novel was written [15]. Still, Jean Ardley's attitude, whether or not it reflected Djerassi's personal experience, certainly reflects a widespread concern over academic credit in collaborative works. This concern has led most areas of research with established publishing standards or precedent, notably excluding mathematics, to encourage journals to list authors in order of contribution rather than by last name.

Several factors may have brought about the trend toward non-alphabetized author lists in computer science as opposed to other fields. Mark Mandelbaum, director of the Office of Publications for the Association for Computing Machinery (ACM), suggests that the trend may involve the sharp increase in computer science research conferences that occurred in the mid-1990s. In this case, the rise in alphabetization may be due to particular conferences' policies or the nature of teams submitting to the conferences. Conference presenters also could feel the need to list assistants who helped prepare presentations or demonstrations. Then, author lists would go out of alphabetical order if the original authors want to subordinate the amount of credit these assistants would receive. Note that most if not all ACM conferences and journals, however, "accept the order of the listed authors" as it was received on the original manuscript, implying that the trend toward non-alphabetized author lists would be the result of decisions made by individual teams of researchers rather than official policy changes.[2] In general, most potential explanations involving the conferences or policies of the ACM and IEEE represent gradual policy changes or small-scale decisions. Since these explanations do not justify the suddenness of the trend away from alphabetization, it remains possible that outside circumstances affected publication trends in computer science.

Other potential explanations for the change in alphabetization rates involve larger assessments of academia as a whole. For instance, in 1995 the National Research Council (NRC) published *Research-Doctorate Programs in the United States: Continuity and Change* [16], which ranked graduate computer science (and other) programs using statistical methods. The statistics analyzed included the number of publications by researchers at the various institutions and the number of citations those publications received. The NRC also distributed data accompanying the study concerning the nature, frequency, and authors of the publications used to rate the various institutions [17]. Analysis of this dataset, however, reveals that the study did not consider all publications equal in the determination of "scholarly quality." Specifically, while there is a clear positive correlation between "scholarly quality" scores for computer science and the number of primary-authored publications

[2] Personal communication.

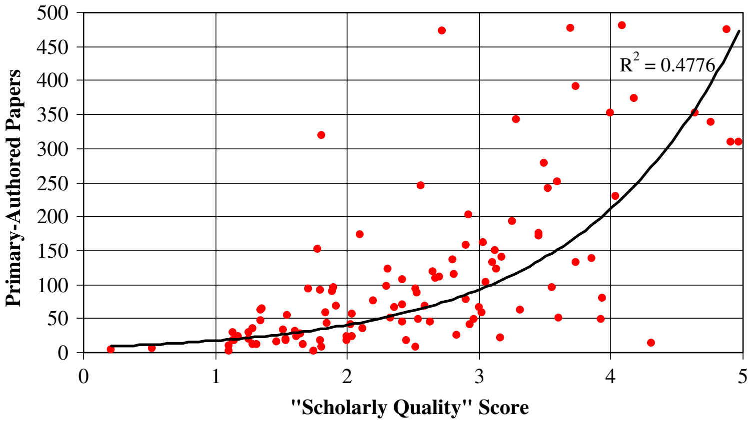


Fig. 5 NRC "Scholarly Quality" scores versus number of primary-authored papers (data gathered from *Research-Doctorate Programs in the United States: Data Set*)

within particular computer science departments ($R^2 = 0.4776$),[3] there is little to no correlation between this score and the number of secondary-authored publications published by faculty in the department ($R^2 = 0.0489$) (see Figs. 5 and 6). Although the specific ranking formula is unclear in the *Research-Doctorate* publication, according to Charlotte Kuh, staff officer for the Assessment of Research Doctorate Programs and Deputy Executive Director of the Policy and Global Affairs Division of the National Academies, the 1995 study did indeed give some credit to the schools of the first *two* authors of each publication[4]; evidently, this credit was not sufficient to seriously affect most departments' respective scores.

Regardless of how the "scholarly quality" scores were computed, these patterns indicate that authorship in computer science was not sufficient to estimate the strength of a particular department. Even if the NRC study did not consider primary versus secondary authorship statistics in producing their final rankings, the ordered list of departments clearly honored those schools whose professors were concerned about author order in their academic publications. The placement of the study in 1995, near the time when computer science research papers moved toward non-alphabetized author lists, may indicate one of several facts. For instance, it may be the case that the NRC study itself inspired professors at various research institutions to reevaluate their publishing policies and obtain more credit for their work. This possibility is unlikely as it would have to involve a conscientious effort on the part of several professors and their research associates. Still, since the NRC assessment is often considered "the gold standard for anyone […] seeking a national,

[3] The R^2 value is the "coefficient of determination" for a statistical fit line. R^2 values close to 1 represent ideal fit lines, while $R^2 \approx 0$ implies little to no correlation between a fit curve and the data. To produce these values, optimal least-squared fit curves were chosen from standard models (exponential, logarithmic, linear) for statistical variation. The curves are imposed on Figs. 5 and 6 for inspection. Here we see that primary authorship and computer science "scholarly quality" are related by a fit line with sufficiently high R^2 value to indicate some type of correlation, while the relationship between secondary authorship and "scholarly quality" is insubstantial.

[4] Personal communication.

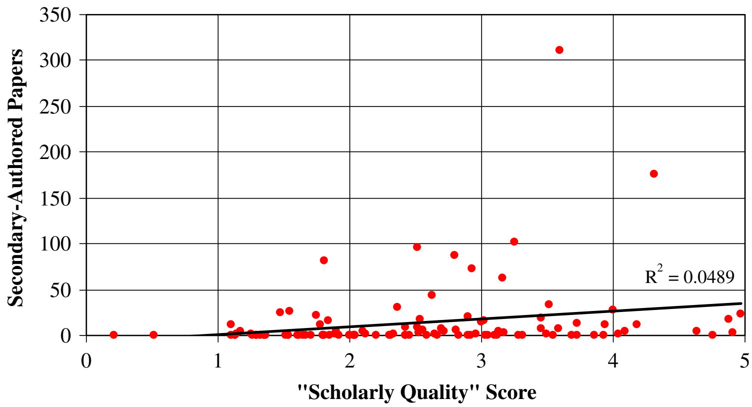


Fig. 6 NRC "Scholarly Quality" scores versus number of secondary-authored papers (data gathered from *Research-Doctorate Programs in the United States: Data Set*)

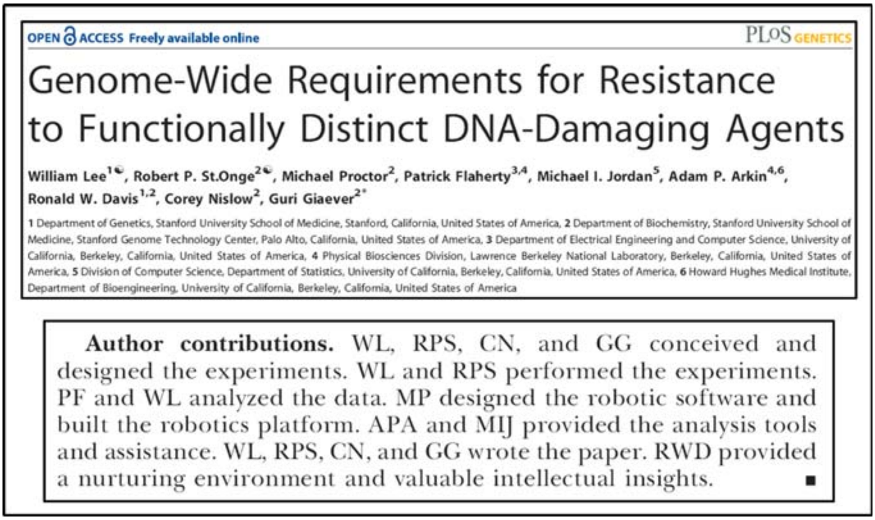
standardized way of measuring the quality of graduate programs in dozens of disciplines" [18], it could be the case that school admissions departments or faculty supervisors encouraged professors to pay more attention to their publication practices to raise their school's ranking. It is more likely, however, that the study indicates changing viewpoints on the necessity of author order to gain publication prestige. Whereas the sheer number of publications by a particular professor may have been sufficient to judge the quality of his or her work in earlier decades, by the time the 1995 study was completed, professors were judging each other's contributions to published works rather than their volume of output.

Without a doubt, the publishing situation in computer science is ripe for change. As co-authorship becomes more common and team sizes grow, publishing houses and organizations of computer scientists no longer can take a *laissez-faire* attitude toward establishing policies for assigning academic credit, determining authorship, and ordering authors. As they stand now, publishing practices not only are inconsistent with each other but also are changing over time, as indicated by the increasing rates of co-authorship and decreasing rates of alphabetization. This mutability partially invalidates any study, such as that by the NRC, evaluating research productivity based on citation or prolificacy, since it becomes difficult to normalize for changing publication conditions. For instance, any statistics involving author order or only honoring primary authorship are invalidated if the percent of papers with alphabetized author lists changes over time. For a similar reason, current publishing practices make it difficult to discern particular researchers' contributions to a project, because author order has little consistent meaning and most papers employ no other means of separating the involvement of each author.

Current Possibilities for Publishing

The design of a successful policy for assigning credit in academic computer science work requires the consideration of several somewhat disjointed factors. From an academic perspective, the policy must allow for the acknowledgement of all parties who were involved significantly in a research project to indicate who should be contacted about possible extensions or questions. From a business or patent law perspective, the policy must preserve patentability for the main authors or principal researchers. From an ethical perspective, the policy must honor those researchers whose effort brought about the main developments in the project, rather than those who contributed only monetary means or who gain authorship positions based on past reputations in their respective fields. The consideration of these broad criteria will lead to an acceptable and realizable standard for academic credit in computer science.

Several of the most important concerns in designing a standard for assigning academic credit in computer science are related directly to similar considerations in other fields. Most prominently, standards for assigning academic credit must devise a system by which the amount of work contributed by each team member can be judged. On the one hand, many "co-authorship" situations amount to more one-sided relationships, in which a researcher includes his or her superiors or assistants as co-authors on academic publications. In this case, some publications attempt to separate the contributions of the various authors by making a note clarifying the specific contributions of each author; this way, those authors whose names appear simply for providing a "nurturing environment" can be separated from those who made more substantial contributions to the research (see Fig. 7). On the other hand, some co-author groups represent truly symbiotic relationships. In their book *(First Person)^2: A Study of Co-Authoring in the Academy*, Kami Day and Michele Eodice describe their personal experiences in completing a research project exploring collaboration and co-authorship as fully collaborative partners themselves:


OPEN ACCESS Freely available online PLoS GENETICS

Genome-Wide Requirements for Resistance to Functionally Distinct DNA-Damaging Agents

William Lee[1], Robert P. St.Onge[2], Michael Proctor[2], Patrick Flaherty[3,4], Michael I. Jordan[5], Adam P. Arkin[4,6], Ronald W. Davis[1,2], Corey Nislow[2], Guri Giaever[2*]

1 Department of Genetics, Stanford University School of Medicine, Stanford, California, United States of America, 2 Department of Biochemistry, Stanford University School of Medicine, Stanford Genome Technology Center, Palo Alto, California, United States of America, 3 Department of Electrical Engineering and Computer Science, University of California, Berkeley, California, United States of America, 4 Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, 5 Division of Computer Science, Department of Statistics, University of California, Berkeley, California, United States of America, 6 Howard Hughes Medical Institute, Department of Bioengineering, University of California, Berkeley, California, United States of America

Author contributions. WL, RPS, CN, and GG conceived and designed the experiments. WL and RPS performed the experiments. PF and WL analyzed the data. MP designed the robotic software and built the robotics platform. APA and MIJ provided the analysis tools and assistance. WL, RPS, CN, and GG wrote the paper. RWD provided a nurturing environment and valuable intellectual insights. ■

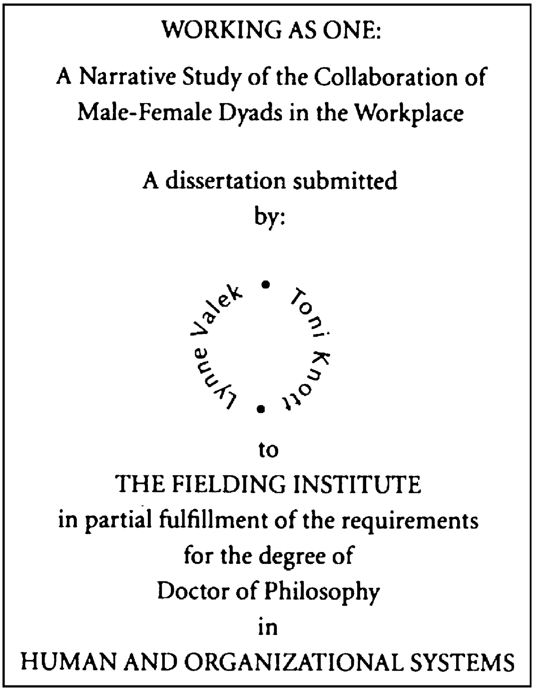
Fig. 7 The title and "author contributions" section of a genetics paper [32]

> We co-wrote the literature review, a chapter on collaborative dissertations, and part of the design and methodology for both studies; and we became co-researchers in each other's projects—Kami team-taught with Michele during the classroom study, and Michele took part in the interviews for Kami's study. Because we live together, proximity allowed us to participate jointly in all aspects of analyzing our data. We transcribed side by side, listening from time to time to each other's tapes to provide a second interpretation of what we were hearing and to check for accuracy; after one of us had coded a section of transcript, the other often coded it again to test and expand our understanding. And we talked—as we worked, as we cooked, as we ate, as we drove, as we walked [19].

Clearly, Day and Eodice deserved equal standing in any publication resulting from their work. Unfortunately, most published papers imply some sort of hierarchy between the various authors. For instance, even if Day and Eodice had published several short papers each representing the "least publishable unit" of research and alternated between primary authorship [20], the final list of authors on each paper could be interpreted as indicating each author's individual contributions to the overall study.

Despite their insistence on balanced collaboration throughout the research process, even Day and Eodice fail to suggest any completely acceptable publishing practices for expressing close co-authorship relationships, although they do present some creative possibilities. For instance, they point to one co-authored thesis in which the two authors wrote their names in a circle to avoid giving one or the other preferential treatment in the author list (see Fig. 8). Even this solution, however, fails to give equal credit to both authors. First, reading the "circular author identifier" from left to right still gives Valek higher standing than Knott. Also, many publishing firms may find it unprofessional, difficult to read, or space-consuming to write the author list in such a style. From the standpoint of future research, other papers looking to cite this one, as well as databases listing paper titles and authors, will have to give one author preference over the other in storing records or creating bibliographies. Thus, the problem of assigning *equal* credit can be just as difficult if not more difficult than the problem of *differentiating* between authors.

The question of how authors are listed is doubly important when we consider that the list of authors not only assigns credit but also responsibility. Anybody whose name appears on the author list for a publication or presentation must agree to take responsibility for the work presented therein. For instance, guides for academic authors suggest that several academic presses require authors of papers or books to sign contracts making authors "responsible for reviewing the editing, getting permissions, indexing, and so forth" [21]. In signing such a contract, the co-authors obligate themselves to complete a certain amount of work surrounding the publication itself as opposed to the academic work that went into its conception. More importantly, submissions for publications indicate, formally or informally, to publishers and editors that each of the coauthors agree to the statement, "This is my work and to the best of my knowledge it is correct" [22]. When this basic statement breaks down, it becomes nearly impossible to judge who should be responsible for


WORKING AS ONE:

A Narrative Study of the Collaboration of
Male-Female Dyads in the Workplace

A dissertation submitted
by:

Lynne Valek • Toni Knott •

to
THE FIELDING INSTITUTE
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in
HUMAN AND ORGANIZATIONAL SYSTEMS

Fig. 8 A creative solution to the credit problem [19]

publishing fraud or mistakes. The peers of Dr. John Darsee, who was caught in 1981 for fabricating medical data for a Harvard heart study, form one example of this effect. When Darsee's questionable practices came to light, his former colleagues at Emory claimed to have "no responsibility at all for what happened" despite the fact that their names appeared as co-authors on some of Darsee's publications [23]. As Marcel C. Lafollette [22] puts it in his book *Stealing into Print: Fraud, Plagiarism, and Misconduct in Scientific Publishing*, "When suspicions of wrongdoing are raised, […] coauthors tend to disappear." This disappearing act clearly indicates unhealthy co-author relationships in which the parties did not take equal responsibility for the published work.

Darsee's co-authors also indicate a different type of publication fraud to be avoided in designing a credit policy for any type of work: the inclusion of authors who contributed little to no work toward the published results, as well as the exclusion of authors who completed significant research. Rennie and Flanagin suggest in an American Medical Association paper that there are three general types of questionable authorship caused by publication pressure in academia. These classes are personified by three figures: the "guest," whose name appears for honorary rather than intellectual reasons on a list of authors; the "ghost," who writes papers that are attributed to more well-known scientists; and the "grafter," who appears at the end of a list of authors for making negligible contributions to a project. Rennie and Flanagin acknowledge that research institutions "rely on publications as the coins academics must use to get through the tollgates on their way to academic promotion," providing a believable motivation for the appearance of these characters [24]. While the authors' explanation of the causes of intellectual parasitism is plausible, additional statistics or analysis of promotion policies at research institutions is needed to prove a relationship between publication and success in academia. Fortunately, other researchers have presented surveys of scientists in various fields to examine this relationship. Birnholtz presents a comprehensive study of interviews of researchers in the High Energy Physics (HEP) community; in doing so, he found an interviewee who had "several publications written in Russian, a language he cannot speak or read," and others whose names were on lists of authors that were several hundred lines long [25]. This situation may be parallel to the situation in computer science, although HEP projects involve work by large-scale teams, while computer science projects usually involve smaller groups of researchers.

While most types of publication fraud remain unpunished unless explicitly revealed, the parasitic publishing relationships suggested by Rennie and Flanagin actually can backfire for the wrongfully-listed authors. Rennie and Flanagin mention that in some cases "researchers did not desire authorship so much when it meant being publicly acknowledged as 'al'" [24]. In this case, "grafters" can put their integrity into question, since their peers may recognize a pattern of academic parasitism rather than original work if their name is consistently last on author lists. For instance, in a letter to the MIT community defending the rejection of tenure for a biological engineering professor, reviewers suggested that, "Only three of the six publications list [the professor] as the first or corresponding author […], the status most highly valued for promotion decisions" [26]. Here, the review board acknowledges that the "publish or perish" model is insufficient for evaluating professors since they may have committed authorship abuse. Instead, the board relied not only on publication numbers but also on position on author lists for a more accurate portrayal of a researcher's involvement in a particular project. A more explicit publishing standard would allow for such review boards to have a better idea of professors' contributions to projects, thus avoiding the debate over what constitutes "significant" involvement.

Other factors in designing the academic credit standard must be specific to computer science itself. As mentioned earlier, patentability and other legal concerns may affect the optimal distribution of credit. While many members of a research project should receive credit for their work, this credit should not preclude the principal researchers' right to obtain a patent on any novel, marketable products that come about as a result of the research. Whereas researchers in the natural sciences may need to find a way to fairly credit research assistants, computer science researchers must concern themselves with crediting programmers or interns involved only in implementing aspects of a research project rather than inventing new components. From the publisher's standpoint, the fast pace of computer science research does not allow for extensive background checking on the part of the publisher. For example, in describing the review process for the major SIGGRAPH computer graphics conference, Jim Kajiya states, "In 10 weeks, SIGGRAPH can do what other major publications take 10 months to do. In a fast-moving field like computer graphics, this is crucial" [27]. Thus, by the time a paper reaches the main part of the review process, its list and ordering of authors should be somehow verified for reliability.

Parallel Practices

Although an exhaustive analysis of parallel publishing practices in other sciences and academic fields deserves a completely separate study, Table 1 lists sample publishing policies from a wide variety of organizations. The principal missions and publishing policies of these organizations contrast with each other because they reflect different interests in the publishing process. For instance, the American Psychological Association (APA) presents a comprehensive set of guidelines governing the ethics and mechanics of publishing psychology papers. Since the APA [28] is an organization principally composed of scientists and practitioners in psychology, the APA guidelines represent a practical and professional interest in making publication procedures specific and easy to follow. Contrastingly, the "Guidelines on Good Publication Practice" released by the Committee on Publication Ethics (COPE) represent the broader views of a group bound by ideological rather than professional common grounds. Even though COPE's members primarily are from the medical field, their attempts to make the guidelines applicable to a wide variety of fields have led to a broader, philosophical approach to publishing guidelines. Since the guidelines are much more general than the specific details presented in the APA manual, COPE is able to cover the basics of research, from data analysis and authorship to media relations, in five pages. The remaining standards compared in Table 1 represent interests involved in the publishing process itself, rather than the research or authorship of academic papers. Elsevier represents the viewpoint of a publishing house, whose business is solely in the publication of academic journals and the management of authors and editors. Elsevier's minimal "Ethical Guidelines for Journal Publication" [29] protect Elsevier's legal interests in publishing and maintain only basic ethical behavior on the part of its authors. Finally, the International Committee of Medical Journal Editors (ICJME) [30] standards represent the point of view of scientific journal editors. Since the editors serve as bridges between researchers or authors and publishing houses, the ICJME standards are simple and clear, establishing a protocol by which authors can communicate their concerns about credit with publishers. Thus, this broad set of standards includes the points of view of authors, special interest groups, publishing houses, and editors.

As would be expected, for the most part these policy statements are very similar. After all, few ethical or professional groups would be willing to endorse openly policies of academic fraud or misplaced credit. At the same time, subtle differences between the various standards mark possible policy choices for a publication standard in computer science. For instance, COPE's guidelines include a "dealing with misconduct" section that outlines broad punishments for plagiarism and other types of publication fraud, while most other organizations leave the formulation of suitable punishments up to journal editors. Another concern would be the specificity of the publication standard. While the APA guidelines span 439 pages, covering

Table 1 Publication policies in other fields

	American Psychological Association	Committee on Publication Ethics	Elsevier	International Committee of Medical Journal Editors
Definition of "author"	"Authorship is reserved for persons who receive primary credit and hold primary responsibility for a published work. Authorship encompasses, therefore, not only those who do the actual writing but also those who have made substantial scientific contributions to a study." (350)	"There is no universally agreed definition of authorship, although attempts have been made. As a minimum, authors should take responsibility for a particular section of the study." (sec. 3)	"Authorship should be limited to those who have made a significant contribution to the conception, design, execution, or interpretation of the reported study. All those who have made significant contributions should be listed as co-authors."	"Authorship credit should be based on (1) substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; (2) drafting the article or revising it critically for important intellectual content; and (3) final approval of the version to be published. Authors should meet conditions 1, 2, and 3." (II.A.1)
People without publication rights	"Principal authorship and other publication credits accurately reflect the relative scientific or professional contributions of the individuals involved, regardless of their relative status. Mere possession of an institutional position, such as Department Chair, does not justify authorship credit. Minor contributions to the research or to the writing for publications are appropriately acknowledged, such as in footnotes or in an introductory statement." (395–396)	"If there is no task that can reasonably be attributed to a particular individual, then that individual should not be credited with authorship." (3.1)	"The corresponding author should ensure that all appropriate co-authors and no inappropriate co-authors are included on the paper, and that all co-authors have seen and approved the final version of the paper and have agreed to its submission for publication."	"Acquisition of funding, collection of data, or general supervision of the research group, alone, does not justify authorship." (II.A.1)

Table 1 continued

	American Psychological Association	Committee on Publication Ethics	Elsevier	International Committee of Medical Journal Editors
Responsibility of authors	"Psychologists take responsibility and credit, including authorship credit, only for work they have actually performed or to which they have contributed." (395)	"All authors must take public responsibility for the content of their paper. The multidisciplinary nature of much research can make this difficult, but this can be resolved by the disclosure of individual contributions." (3.4)	"Authors of reports of original research should present an accurate account of the work performed as well as an objective discussion of its significance. Underlying data should be represented accurately in the paper. A paper should contain sufficient detail and references to permit others to replicate the work. Fraudulent or knowingly inaccurate statements constitute unethical behavior and are unacceptable."	"Each author should have participated sufficiently in the work to take public responsibility for appropriate portions of the content." (II.A.1)
Author order	"Authors are responsible for determining authorship and for specifying the order in which two or more authors' names appear in the byline. The general rule is that the name of the principal contributor should appear first, with subsequent names in order of decreasing contribution. If authors played equal roles in the research and publication of their study, they may wish to note this in the second paragraph of the author note" (351)	N/A	N/A	"The order of authorship on the byline should be a joint decision of the co-authors. Authors should be prepared to explain the order in which authors are listed." (II.A.1)

Table 1 continued

	American Psychological Association	Committee on Publication Ethics	Elsevier	International Committee of Medical Journal Editors
Acknowledgments	"Lesser contributions, which do not constitute authorship, may be acknowledged in a note. These contributions may include such supportive functions as designing or building the apparatus, suggesting or advising about the statistical analysis, collecting or entering data, modifying or structuring a computer program, and recruiting participants or obtaining animals. Conducting routine observations or diagnoses for use in studies does not constitute authorship. Combinations of these (and other) tasks, however, may justify authorship." (350)	"To avoid disputes over attribution of academic credit, it is helpful to decide early on in the planning of a research project who will be credited as authors, as contributors, and who will be acknowledged." (3.2)	"Where there are others who have participated in certain substantive aspects of the research project, they should be acknowledged or listed as contributors."	"All contributors who do not meet the criteria for authorship should be listed in an acknowledgments section." (II.A.2)

specific typographic issues and authorship situations, the Elsevier guidelines provide a bare-bones outline that is sufficient to avoid legal action against the publishing company. Although the computer science guidelines would need to err on the side of specificity to ensure consistency between publications, they also should be short enough that a researcher looking to adopt the policy could read them in a reasonable amount of time. Related to this issue is that of the formality of the guidelines designed for computer science. The guidelines, as in the COPE standard, can be informal about the particular definition of terms such as "authorship," allowing for a more flexible but less controlled interpretation, or very formal, as in the ICJME standard, to avoid any ambiguity.

The Future of Publishing in Computer Science

The establishment of any academic credit policy in computer science will require the agreement of the major computer science research organizations, academic institutions, and individual researchers. Given the above analysis of various research organizations' publishing policies, the current publishing situation in computer science, and successful practices in other fields, however, the following basic set of principles are proposed for any new policy:

I. Authorship credit should be distributed *only* to those researchers directly involved with the paper or project in question. Researchers with indirect or minimal involvement may be mentioned in an additional "acknowledgements" section if necessary. All contributors should appear on a paper; "ghost writing" is an invalid way even for a busy researcher to produce publications.
II. All authors should be paired with short descriptions of their contributions to the project. These descriptions need not be on the title page but should be apparent for anybody seeking further information about the research presented. This principle extends to the acknowledgments list. In general, any individuals or organizations mentioned by the paper should be identified to avoid "honorary" authorship and make explicit the division of work leading to the final results.
III. The list of authors should be divided by level of contribution. Within each division, authors should be ordered by the amount they contributed to the particular paper in question. Truly equal co-authorship relationships should be marked as such, with none of the authors identified as a "corresponding" author. The lack of a single corresponding author can be addressed by creating a simple email alias that contacts all of the principal authors simultaneously. Those researchers who would be considered "inventors" should be marked as such for the purposes of verifying future patent applications.
IV. Upon publication, authors should be required to sign that the work in the paper is at least partially their own and that no other authors should be given credit.
V. Any and all decisions involving authorship should involve the mutual consent of all authors, which should be established via individual contact.
VI. Any discovered cases of authorship fraud should be dealt with in much the same way as data fabrication. Once they are caught, authors should be

required explain their incorrect practices in a published statement and rectify any disadvantages suffered by parties not receiving appropriate credit.

Of course, a final, more official standard would have to be more careful in defining particular terms, explaining procedures for verifying authorship, and outlining punishments for non-compliance. This standard would have to acknowledge and ideally align with ethical standards already put into place by particular institutions or organizations; in cases where the new policy and that already in place for some organization are not reconcilable, papers affected by some compromise should be noted as such. The definition of standards for measuring various authors' contributions to a research project would be particularly important; the requirement that all authors agree on the final listing, however, will help make this decision consistent. Additionally, procedures would need to be put into place for mediating authorship disputes without discouraging junior authors from confronting their superiors.

Even so, the implementation of these six basic principles would lead to a much more consistent and interpretable publishing landscape for computer science. The principles take every precaution to avoid situations in which authors receive insufficient or excessive credit relative to their work on the project presented in a published paper. By requiring that all authors agree on the order and division of their names and that the contributions of each author be outlined specifically, fraudulent authors will be discouraged from wrongfully adding their names to papers. For example, the BlueGene/L paper would be sorted by contribution rather than research lab, giving the leaders and developers of the project they credit they deserve. Authors would not be punished for having last names later in the alphabet, and when an author whose name is toward the beginning of the alphabet appears at the top of an author list, there would be less doubt regarding his or her contribution. The more comprehensive system in which authors describe their contributions also would allow editors to act as final gatekeepers, calling into question author lists if they are presented in an unreasonable fashion. If discrepancies or disagreements arise regarding authorship credit, requirements put in place for greater accountability would ensure that the issues are addressed before publication and that all authors consent to the final arrangements. Additionally, such increased accountability would make it easier to address accusations of authorship fraud, since it would be not only a violation of abstract ethical standards but also a breach of contract.

As computer science continues to develop rapidly from a small subfield of engineering into a complete science unto itself, the need for a consistent and usable publishing standard governing the assignment academic credit to all members of a research project will become more and more urgent. As it stands now, computer science publication represents a collection of informal, changing standards that make it difficult to judge specific individuals' contributions or find the responsible team member for academic fraud. This confusing situation makes it nearly impossible to honor individuals' respective contributions to a research project, allowing senior researchers to overshadow less experienced authors and making processes such as applying for patents needlessly convoluted. The implementation of a consistent and clear policy for academic credit will help computer science become a more unified discipline, preparing it to enter the ranks of the “traditional” sciences as a rigorous and respectable field with a sustainable and fair publishing practice.

Acknowledgements Special thanks to James Robert Wood, Stanford Department of English and Program in Writing and Rhetoric, for his advice throughout the writing process and revisions of the final paper. Additional thanks to Prof. Clifford Nass (Stanford Department of Communications), Prof. Andrea Lunsford (Stanford Department of English), and my family for their support.

Appendix: C++ Code Used to Parse NRC Study Data

The following short program was used to parse the data accompanying the National Research Council’s study, *Research-Doctorate Programs in the United States*. It is included to make transparent the methods used for data analysis and to enable easier analysis of the NRC data in future studies.[5]

```
#include <iostream>
#include <fstream>
#include <vector>
#include <string>
#include <conio.h>
#include <set>
using namespace std;
int main() {
 //Citation data
 ifstream infile(“PUB_CIT.dat”,ios::binary|ios::in);
 //List of faculty
 ifstream faclist(“FACLIST.dat”,ios::binary|ios::in);
 //Output file
 ofstream outfile(“pub_cit_analysis2.txt”);
 set<string> validnames;//set of names of CS researchers
 string curline;
 int b = 0, cs=0;//b = number of invalid lines; cs = number of CS professors
 while (getline(faclist,curline), !faclist.fail()) {
  //according to NRC standard, all lines in FACLIST.dat should have 63
characters
  if (curline.length() < 63) {
   b++;//invalid line
   continue;
```

[5] Incidentally, the NRC data concerning primary versus secondary authorship apparently is somewhat flawed. Certain schools with considerable numbers of publications are identified as having *no* secondary-authored publications, which is unlikely given that any team with multiple researchers from the same school would have to have one or more secondary authors since only one person can appear first on an author list. Furthermore, not all lines in the data file have the right number of characters to agree with the description of the data format.

```
  }
  string facname = curline.substr(0,5);//Name of faculty; NRC 5-character code
  string progcode = curline.substr(60,2);//Program/department code
  if (progcode == "26") {//code 26 = computer science
   validnames.insert(facname);
   cs++;
  }
 }
 cout ≪ b ≪ "bad lines.\n";
 cout ≪ cs ≪ "CS profs.\n";
 //all schools indexed by 3-digit code, so vectors have 1000 elements to cover all codes
 vector<int> primary_authorship(1000);//number of primary-authored papers per school
 vector<int> secondary_authorship(1000);//number of secondary-authored papers per school
 vector<int> single(1000);//number of singly-authored papers per school
 vector<int> multiple(1000);//number of multiple-authored papers per school
 vector<int> total(1000);//total number of papers per school
 string empty="";
 int numlines = 0;//number of lines parsed
 int numbad = 0;//number of bad lines
 while (getline(infile,curline), !infile.fail()) {//for each publication
  numlines++;//update status
  if (curline.length() < 98) {//invalid line
   numbad++;
   continue;
  }
if (!validnames.count(curline.substr(0,5)))//not a CS publication
   continue;
  char t = curline[2];
  if (t == 'A' || t == 'O' || t == 'S' || t == 'Y')
   continue;//only accept proceedings, journals
  int numAuthors = curline[19]-'0';//number of authors (between 0 and 9)
  if (numAuthors == 1) single[c]++;
  else if (numAuthors > 1) multiple[c]++;
  if (curline[31] == 'P') primary_authorship[c]++;//'P' indicates primary authorship
  else if (curline[31] == 'S') secondary_authorship[c]++;
  total[c]++;
 }
 for (int i = 0; i < 1000; i++)
  if (total[i]) {//if school published in CS, output data
   outfile ≪ i ≪ '';
   outfile ≪ total[i] ≪ '';
   outfile ≪ single[i] ≪ ''≪ multiple[i] ≪ '';
```

```
      outfile ≪ primary_authorship[i] ≪ ''≪ secondary_authorship[i] ≪ '';
      outfile ≪ (double)primary_authorship[i]/(primary_authorship[i] +
       secondary_authorship[i]);
      outfile ≪ endl;
    }
  cout ≪ "Data processing is done.\n";
  cout ≪ numlines ≪ "lines processed.\n";
  cout ≪ numbad ≪ "bad lines.\n";
  getch();
}
```

References

1. Adiga, N. R., et al. (2002). An overview of the BlueGene/L Supercomputer. In *Proceedings of the 2002 ACM/IEEE Conference on Supercomputing*, November 16–22, 2002, 1–22.
2. Frye, J., Ananthanarayanan, R., & Modha, D. S. Towards Real-Time, Mouse-Scale Cortical Simulations. IBM Research Report RJ10404 (A0702-001). Retrieved February 7, 2007, from http://www.modha.org/papers/rj10404.pdf.
3. Cliff, P. (1989). *The Oxford English Dictionary*, (2nd ed.) OED Online. Retrieved May 26, 1989, from http://dictionary.oed.com/cgi/entry/50015051. Accessed 2007.
4. Monastersky, R. (2005). The number that's devouring science. *The Chronicle of Higher Education, 14*, A12.
5. Association for Computing Machinery. (2007). *ACM*. Retrieved May 7, 2007, from http://www.acm.org.
6. Institute of Electrical and Electronics Engineers. *IEEE*. Retrieved May 7, 2007, from http://www.ieee.org.
7. Computer Science Bibliography. Michael Ley, maintainer. Retrieved May, 7, 2007, from http://www.informatik.uni-trier.de/~ley/db/.
8. Petricek, V., et al. (1994). Modeling the author bias between two on-line computer science citation databases. Special interest tracks and posters, *The 14th International World Wide Web Conference*, May 10–14, 2005, 1062–1063.
9. Glänzel, W. Coauthorship patterns and trends in the sciences (1980–1998): A bibliometric study with implications for database indexing and search strategies. *Library Trends, 50*(3), 461–474.
10. Garfield, E. (1996). What is the primordial reference for the phrase 'publish or perish'? *Scientist (Philadelphia, PA), 10*(12), 10–11.
11. Abrahams, M. (2002). *The Ig Nobel prizes: The annals of improbable research*. New York: Dutton.
12. Joseph, K., Laband, D., & Patil, V. (2005). Author order and research quality. *Southern Economic Journal, 71*(3), 545–555.
13. Laband, D., & Tollison, R. (2006). Alphabetized coauthorship. *Applied Economics, 38*(14), 1649–1653. doi:10.1080/00036840500427007.
14. Djerassi, C. (1989). *Cantor's dilemma*. New York: Doubleday.
15. Rudd, E. (1977). The effect of alphabetical order of author listing on the careers of scientists. *Social Studies of Science, 7*(2), 268–269. doi:10.1177/030631277700700208.
16. Goldberger, M., Maher, B., & Flattau, P. E. (Eds.). (1995). *Research-doctorate programs in the United States: Continuity and change*. Washington: National Academy Press.
17. *Research-doctorate programs in the United States. Data set.* (1995). CD-ROM. Washington: National Academies Press.
18. Lederman, D. Rating doctoral programs. *Inside Higher Ed*, 23 Nov. 2005. Retrieved June 7, 2007, from http://insidehighered.com/news/2005/11/23/graduate.
19. Day, K., & Eodice, M. (2001). *(First Person)^2: A study of co-authoring in the academy*. Logan: Utah State University Press.

20. Macrina, F. L. (2005). *Scientific integrity* (3rd ed.). Washington: ASM Press.
21. Luey, B. (2002). *Handbook for academic authors* (4th ed.). Cambridge: Cambridge University Press.
22. La Follette, M. C. (1992). *Stealing into print: Fraud, plagiarism, and misconduct in scientific publishing*. Berkeley: University of California Press.
23. Broad, W. J. (1983). Notorious Darsee case shakes assumptions about science. *New York Times*, 14 June, C2.
24. Rennie, D., & Flanagin, A. (1994). Authorship! Authorship!: Guests, ghosts, grafters, and the two-sided coin. *Journal of the American Medical Association, 271*(6), 469–471. doi:10.1001/jama.271.6.469.
25. Birnholtz, J. (2006). What does it mean to be an author? The intersection of credit, contribution, and collaboration in science. *Journal of the American Society for Information Science and Technology*, 57(13), 1758–1770.
26. Belcher, A., et al. (2007). Letter from members of the biological engineering division faculty: Statement of facts in regard to the James Sherley tenure case. *MIT Faculty Newsletter, 19*(6), 13–15.
27. Kajiya, J. How To Get Your SIGGRAPH Paper Rejected. 29 August 2006. Retrieved March 28, 2007, from http://www.siggraph.org/publications/instructions/rejected.
28. American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.
29. Ethical Guidelines for Journal Publication. *Elsevier*. Retrieved May 25, 2007, from http://www.elsevier.com/wps/find/intro.cws_home/ethical_guidelines. .
30. International Committee of Medical Journal Editors. *Uniform requirements for manuscripts submitted to biomedical journals: Writing and editing for biomedical publication*. February 2006. Retrieved May 26, 2007, from http://www.icmje.org/icmje.pdf.
31. Committee on Publication Ethics. Guidelines on Good Publication Practice. Retrieved May 28, 2007, from http://www.publicationethics.org.uk/guidelines.
32. Lee, W., et al. (2005). Genome-wide requirements for resistance to functionally distinct DNA-damaging agents. *PLOS Genetics, 1*(2), 235–246. doi:10.1371/journal.pgen.0010024.