# Rebooting the CS Publication Process

Dan S. Wallach, Rice University

August 12, 2010[*]

**Abstract**

Many computer science academics have been grousing about failures in our publication process. This paper catalogs many of the specific complaints that are raised and proposes some radical new solutions based on the assumption that, by eliminating physical paper entirely and going with a centralized system to manage papers, we can rethink the entire process: paper submission, revision and publication. Furthermore, having all of the metadata standardized and easily available, ranking algorithms can be easily conceived to aid in tenure cases and departmental rankings.

## 1 Introduction

The process of computer science academia is broken, and it's time to fix it. Consider the following list of problems that every practicing academic, either inside or outside CS has experienced at one or another point in their career:

**Acceptance rates.** The top conferences in our specialties, where publications can make or break a career, may publish 10% of the submitted papers. Submission rates have grown radically in the past decade with acceptance rates either flat or dropping, despite an increasing absolute number of papers accepted[1]. What happens to the other papers? Realistically, there are three categories. First, there are the "bubble" papers. If, for whatever reason, the conference were to double its acceptance rate, these papers would be published, but they were pushed out either because they were seen as too narrow or uninteresting, or they were considered to have some possibly reparable flaws (e.g., poor experimental design, incomplete proofs, or bad writing). Next are "second tier" papers that could well be publishable at area-specific workshops or less competitive conferences. Also in the "second tier" category would be "least publishable unit" (LPU) papers, where an author advances their own work by the smallest possible amount and the program committee wants more. Finally, there are "noncompetitive" papers, where the paper would have no chance at publication, ever, in any respectable venue.

**Overloaded reviewers.** As submission rates have gone up, program committees are struggling with either huge workloads per reviewer, or are adding PC members to the point where most members are completely disconnected from most papers' discussions. This appears to increase the degree of randomness in whether a paper gets in or is rejected.

[*] This manuscript was originally released June 14, 2010, and has been updated many times since then. To cite it, you might treat it as a technical report and note both the original and current date.

[1] Submission and acceptance rates for security conferences are tracked by Mihai Christodorescu (http://pages.cs.wisc.edu/~mihai/conference%20acceptance%20rates.html). Networking conference stats are tracked by Kevin Almeroth (http://www.cs.ucsb.edu/~almeroth/conf/stats/). Graphics conference stats are tracked by Robert Lindeman (http://web.cs.wpi.edu/~gogo/hive/AcceptanceRates/). Database conference stats are tracked by Peter Apers (http://wwwhome.cs.utwente.nl/~apers/rates.html)

**Resubmission.** What happens with all these rejections? Many are inevitably resubmitted. Major conferences try to coordinate their accept/reject announcement dates with subsequent conferences' submission dates, but these can still be quite tight. (Example: There was only one week, this year, between IEEE Security & Privacy's notification date and the USENIX Security Symposium's submission date.) This means substantially similar content is reviewed again and again, sometimes to the annoyance of members of both program committees.

**Journal latencies.** In many other fields, conferences only take short papers and the "real" work is submitted to journals. Journals offer the benefit of having the same set of reviewers through each phase of a paper's lifecycle. The reviewers can insist on improvements and can then agree that the authors satisfied their requirements. In computer science, most good work is never submitted to journals, and they get a large volume of noncompetitive submissions, consuming reviewing resources. Furthermore, a latency of one year from submission to publication is entirely normal, and it can be far worse.

**Scooped!** Very few things can cause more anger and frustration than seeing one's own ideas appear in print by somebody else. This can lead to unpleasant arguments and incriminations. Combine this with the acceptance rate issue, it's entirely possible to have been first to *submit* an idea but to end up late having *published* that idea.

**Costs.** Our library has been slowly dropping its journal subscriptions. Even without receiving printed paper journals, online access for our campus to major journals can cost thousands of dollars per year per journal title! Given that all of the relevant labor in writing and reviewing papers is "donated" by the authors and reviewers, why aren't these publications available free of cost? They can and should be.

**Inconsistent citations.** Because many papers are available online these days, academics will read them online and will then need to cite them. CiteSeer, DBLP, and other automated systems produce a variety of errors in their BibTEX references. Similarly, hand-written entries are prone to human error. This, in turn, complicates life for systems like Google Scholar, which attempt to capture citations from PDFs that it finds and use these to measure citation counts.

**Blinded submissions.** Some conferences are double-blind. Others are single-blind. Writing a double-blind paper makes it painful to do self-citation, yet writing unblinded papers allows the reviewers to bias their decisions in favor of the author ("I trust this group to be able to address my concerns.") or against ("Another crappy paper from this group."). Reviewers may also apply biases, conscious or otherwise, based on the author's apparent gender or ethnicity.

**Short incremental work.** Our current system of promotion and tenure strongly incentivizes authors to collect as many publications as possible, resulting in many different papers for any one given idea. Many current academics bemoan the good old days when you could have a group working on an involved, complex project, and have only one or a small number of groundbreaking papers.

**Plagiarism.** Lower-tier academics sometimes look to plagiarism as a career shortcut, submitting others' work to out-of-the-way conferences or journals where it is less likely to be noticed. The crime may not be discovered until well after it's been accomplished.

**Promotion and tenure.** Some universities understand that a top conference publication in CS carries the same prestige as a top journal publication in other areas of science and engineering, but it's a recurring battle, and it can have significant ramifications on promotion and tenure cases. Some universities go to other extremes and become obsessed with citation counts or *h*-indices, which possibly distort academic goals away from the (ideal?) goal of producing fewer, better papers. A related issue is that, as some conferences go to "paperless" publications, authors become concerned that the publications aren't "real" in the sense that they may not figure into official citation counts and may otherwise be discounted.

Given all these disparate concerns, can we evolve our way to a better structure for our academic processes? If evolution isn't an option, can we redesign our system from scratch in an effective manner? And if redesign is the answer, then how do we get there from here?

## 2 A Clean Slate Solution

This section presents the high-level design of a clean slate solution, called CSPub (clean-slate publication, or perhaps, ambitiously, computer science publication). CSPub is, at its core, a mashup of conference and journal submission and review management software, such as HotCRP [26], with technical report archiving services like arXiv and with bibliographic management and tracking and search services like DBLP, Google Scholar, and CiteULike (see Section 4 for extended discussion on how these systems work). In this section, we describe how CSPub would function and what benefits it could yield when widely adopted. The transition from our present system to the widespread adoption of CSPub is described in Section 3.

### 2.1 Tech reports on steroids

Today, computer scientists have three different mechanisms for disseminating technical reports: their own personal or laboratory home pages on the web, their departmental "official" technical report services (which may or may not be integrated with NCSTRL [32]), or on centralized services like arXiv. Efficient publication and dissemination is the first and most fundamental service that we'd have in CSPub. Ultimately, *every* paper published in our field can and should be available via this one mechanism, regardless of whether it's a "technical report," a "preprint," a "conference" paper, or a "journal" paper.

While the CSPub implementation could well be a distributed system, shared across many universities, perhaps using a system like LOCKSS [34] to guarantee durability in the face of hardware failures and even attempted censorship, it's easier to first consider the design of a centralized system, merging the functions we expect of searching services like Google Scholar, DBLP, or CiteSeer with the storage services we expect of arXiv or the IACR ePrint service. With such a service, every paper, old and new, could be posted in one place. There are really four classes of metadata that we want to have, distinct from the PDFs of the research papers: author, title, and relevant keywords; related home page, perhaps with additional source code or experimental data; citations to other work; and publication venue(s) and related awards.

The first two classes of metadata can and should be provided by the author when the paper is uploaded. Keywords can be a free list of whatever tokens the authors deem relevant as well as locations in relevant taxonomies such as the ACM's Computing Classification System[2]. Much like Twitter users associate "hash tags" with their tweets to indicate relevant topics of discussion, appropriate keywords will help draw attention to a given paper. Of course, some authors might choose to abuse this facility, but commercial search engines are already quite good at filtering out bogus pages. Similar mechanisms could apply here (see Section 2.2 for more on the challenges that this system would face).

[2] http://www.acm.org/about/class/

Citations to other work should ultimately be produced as a list of links to other papers within CSPub. CSPub can then trivially generate BibTEX entries with unique IDs as their handles. Furthermore, BibTEX could be modified to output a list of cited UIDs from a paper to feed into the metadata system. BibTEX could also make every citation in the PDF be a direct hyperlink to CSPub. This eliminates guesswork in the process of automatically extracting citations from PDF reports. (There are many possible ways of representing unique IDs for papers, including Digital Object Identifiers[3].) CrossTEX[4] would be a much better place to start than hacking BibTEX itself, as CrossTEX's searching facilities would potentially work nicely with the CSPub remote database.

Publications venues and awards (including "best paper" awards given by a conference or even "test of time" awards given in retrospect for papers that had a significant impact) could either be provided by the author or could instead by entered by the relevant professional society or conference steering committee. Such annotations can then help when somebody is searching CSPub for a paper to cite on a particular topic, and they may also contribute directly to how a paper is ranked (see below).

For works which only exist in the dusty shelves of a library, scanning them in and dealing with the necessary legal issues would be a challenge, but would be far from insurmountable. Some authors of classic papers have already redone them with modern typography and republished them on their web pages. See, for example, Saltzer and Schroeder's classic, "The Protection of Information in Computer Systems" [38]. Note how long the proper citation is and how many times the paper has been republished. This is exactly the sort of thing that CSPub would encourage, and CSPub should be able to handle this sort of thing more gracefully than our current techniques.

## 2.2 Newfound flexibility

If all of academic computer science's scholarship were available in CSPub, a variety of new features would become feasible and relevant. First and most obviously, search engines could efficiently compute simple citation counts and $h$-indices[5] as well as more sophisticated PageRank-like metrics [6, 36]. A publication that appears at a top conference might well automatically start out with a higher ranking than a small workshop paper. Of course, no two experts in search ranking will agree on the optimal metric for ordering search results to any given query, but there's no reason why we cannot offer searchers a variety of different ranking metrics. While Google keeps its metrics secret to dissuade "search engine optimization," CSPub's ranking function(s) should be public and well-documented, giving academics a clear understanding of their incentives to adjust their publishing behavior. (The related field of "bibliometrics" is discussed further in Section 4.3.)

Additionally, we can generalize the notion of a third-party adding metadata to a given paper. When Alice writes a paper and Bob cites it, that's one kind of metadata reference. When Alice submits it to a conference, the fact that it's in submission is metadata that Alice could choose apply in a public fashion, or it's something she could apply that would only be visible to the conference program committee. In a similar fashion, the conference reviewer's comments are still more metadata, applied by the conference to Alice's work, which she may choose then to keep private, to forward to another program committee, or even to make public to the world.

Of course, there are other relevant metadata worth tracking in CSPub. If Alice releases "version 2" of her paper, she could add bidirectional links, so "version 1" states that it has been superseded by "version 2", and "version 2" links to the earlier edition. This would allow ranking accumulated by earlier versions to

[3] http://www.doi.org/, see also http://en.wikipedia.org/wiki/Digital_object_identifier.

[4] http://www.cs.cornell.edu/people/egs/crosstex/

[5] See, e.g., http://en.wikipedia.org/wiki/H-index

be applied to later ones. Similarly, if Alice's paper has made a splash and she gets invited to give talks at a number of universities, those invited talks are, in effect, additional metadata links that are endorsed by the institution that invited Alice to give a talk. If a graduate seminar adds a paper to its reading list, that's still more relevant metadata.

Now that every paper spends its entire life in CSPub, papers that never get "formally" published in a peer-reviewed conference or journal can still acquire citations and grow in the rankings. Of course, if Alice simply pushes a paper onto the server and Bob is conducting a related work search, he is under no *obligation* to cite her report, but at least it will be easy for him to discover it and to determine if others think it's important. Furthermore, if there is ever a concern about whether Alice or Bob originated a seminal idea in the field, there will be no debate as to when each of their papers was first disseminated.

Consider also, for example, the PageRank citations [6, 36], above. I knew I wanted to cite Page and Brin for PageRank, but I didn't know exactly how they had published it. Via Google Scholar, I was quickly able to find their conference paper and subsequent tech report, but I invested far too much time trying to nail down exactly how to cite them. With CSPub, in conjunction with CrossTEX, I could instead cite something like `!author=brin-page:keyword=pagerank` and I could be confident that one or more proper papers from Page and Brin would be cited. If Brin and Page later wrote a journal paper on the topic, my own citation to them would automatically stay current. (If somebody clicked the link in my references section, they could be taken to a CSPub-generated web page with the most current query result.) Furthermore, a sophisticated ranking system could recognize this mode of citation such that Page and Brin would gain no ranking advantage in having multiple papers on the topic rather than one good one.[6]

One other natural benefit of centralizing papers in CSPub is that we can deal with authors who change names, as a result of marriage or divorce, and with different authors who happen to share common names. CSPub can maintain a database with unique IDs for all authors, allowing for proper disambiguation and tracking.

### 2.3 Problem solving

How well would such a system address the various concerns raised in Section 1?

The acceptance rate issue becomes less problematic. The top papers will still get in, as always. "Bubble" papers will, at the very least, get the proper priority date of their initial dissemination and will be able to start getting citations. More conferences might introduce an "accepted without presentation" category, allowing these papers to be recognized and to avoid the need for subsequent resubmission. Today, authors of lower-ranked papers must choose whether to edit and resubmit to a top conference, resubmit to a lower-ranked conference, or abandon a paper. In CSPub, these decisions can be delayed. If the paper turns out to be popular and starts gaining citations, then the authors will be motivated to update it and resubmit it. If the paper turns out to be a flop, authors can then abandon it and move on.

This might have the effect of reducing the reviewing load for conferences and encouraging fewer, deeper papers, because authors would be incentivized by the new ranking system to put out strong papers. Furthermore, "unpublished" papers, particularly popular ones, would be easy for program committees to discover on their own and "pull" into a conference. Second tier papers might thus get pulled into suitable venues without needing to be explicitly submitted and reviewed. A workshop organizer, for example, could troll through the system looking for relevant work and then invite the authors to present it. Likewise, noncompetitive papers are out there, still able to collect citations, without consuming reviewers' overcommitted

[6] One caveat: if I did scholarship on web page ranking, I might now be incentivized to change my name to "Dan Brin Page Wallach" or some other such monstrosity to attract these sorts of query citations to my own work. Presumably, the strongly negative reaction from the community would disincentivize me.

time. (An interesting question is whether a paper's history of having been submitted, but never having been accepted, should be public data. This might be a signal to a program committee that a paper is noncompetitive.) Of course, if a paper is *mistakenly* treated as noncompetitive, because the community is too blind to recognize the genius within, it will still be out there, published for all the world to see. Seven years later when the community recognizes their folly[7], there it will be, ready to pick up citations, and treated as the seminal manuscript for which it properly is.

Because this proposed system, at least so far as the metadata goes, treats a paper's acceptance and presentation in a conference as a similar event to an invited talk, strong papers that cross research areas can thus "appear" at multiple conferences without there being an issue of simultaneous submission. In fact, there's no reason an author couldn't submit a manuscript to multiple program committees simultaneously, simply by indicating the relevant submission states in the manuscript's metadata. At that point, the program committee chairs could make a decision of how to review the paper, whether it deserves to appear at both venues, and so forth.

When a paper is rejected, and the author receives feedback from the rejecting conference, that feedback would also be in this system, presumably (but not necessarily) private to the author. This creates the opportunity that the author could choose to make this feedback public to a subsequent program committee, along with a statement about how the previous committees' comments were addressed. This moves the treatment of the manuscript closer to the consistent handling available through the journal process, yet with the speed of the conference process. (See also the discussion of VLDB in Section 3.)

In CSPub, a journal is nothing more than an organization that adds metadata notations to papers in the system, with or without having a submission and review process. As such, anybody can start their own journal for almost no cost. Some journals would have calls for papers, as conferences do, and authors would indicate a submission in their metadata when posting a manuscript. Other "journals" would be nothing more than collections of thematically related papers, perhaps put together by graduate students as part of their related work searches. Of course, if a senior academic puts together a collection with a catchy title (e.g., "Alice's List of Seminal Papers in Blah-Blah Theory"), and Alice is a highly ranked professor, the collection would help increase the included manuscripts' rankings, both directly, due to Alice's strong personal ranking, and indirectly, by leading more academics to read and cite the papers on Alice's list.

**Ranking.** Once we have a notion of ranking papers, we can extend this to rankings for authors as well as rankings for whole departments. Of course, there will be many different ways to compute these rankings, and different ranking systems would be a subject of great debate. Consider, if a department's ranking is strictly based on the sum of its faculty rankings, then large departments will always outrank small ones, but if a department's ranking is based on a metric like the mean or median of individual rankings, then departments will be over-incentivized to hire senior faculty with large vitas rather than up-and-coming junior faculty. The selection of an ideal ranking function is clearly beyond the scope of this proposal, but it's important to note that CSPub could easily support many different ranking functions, publicly disclosed, allowing whoever wants a ranking to be able to compute one suited to their needs. For example, a prospective graduate student with a particular specialty in mind could rank departments on the strength of the subset of their papers meeting an appropriate search query. This would be radically superior to the one-size-fits-all U.S. News & World Report rankings which encourage departments to take otherwise undesirable steps to improve their USNWR rankings based on what they believe to be USNWR's ranking criteria. In fact, some academics have suggested boycotting USNWR [7, 8] and other such ranking systems, particularly due to their reliance on

[7] Thomas Edison, in a letter to William Emmet, congratulated him on an achievement, writing "The worst is to come, for it takes about seven years to convert the average man to the acceptance of a solved problem." [16]

"reputation" as reported by other academics. (Ranking and the related field of "bibliometrics" is discussed further in Section 4.3.)

**Plagiarism.** Accusations of plagiarism can normally be difficult to establish, but with CSPub, we can easily imagine plagiarism detection built into the submission process. By having a comprehensive library of prior work, a plagiarism detector can likely identify the source or sources of plagiarized material and point this out to the author at submission time. "If you wish to submit your paper, the authors of these papers will be notified." Algorithms for plagiarism detection are an area of active research and competition (see, e.g., the *3rd PAN Workshop and 1st Competition on Plagiarism Detection* [40]), and could easily be built into CSPub. While detecting plagiarism may well be feasible, dealing with it will be far more difficult, particularly when *historical* plagiarism is detected. Should any ranking given to plagiarists be forfeit to their victims? What if the manuscript in question only plagiarized its introductory material but has its own research results? Each of our professional societies have their own policies for such transgressions, which may need to be harmonized and centralized under CSPub. Any penalty assessed by CSPub would be amplified by its centrality to academic publishing.

**Promotion and tenure.** A system like CSPub could also have a huge impact on tenure cases. There could be a "tenure case report" feature that generates a standardized report, relative to one or more ranking metrics. The system could even automatically identify faculty in related areas with comparable personal and/or departmental ranking, giving the promotion and tenure committee a set of relatively objective facts and comparisons to use in making their decisions. Of course, our system of writing tenure letters will continue, but CSPub could help with the sometimes sensitive problem of identifying suitable senior reviewers as well as selecting suitable peers against whom the senior reviewers might be asked to make focused comparisons.

## 2.4 Unsolved problems

This proposal doesn't solve every problem in Section 1. This section discusses many of these issues.

**Bias.** CSPub doesn't solve the issue of double-blind versus single-blind submissions. (McKinley [35] discusses the literature, which finds that double-blind reviewing reduces incidence of gender bias and nepotism.) CSPub strongly prefers single-blind submissions, where a paper is always available to the public, with its metadata evolving as it's accepted by a conference, cited by other papers, and so forth. While the issue of gender bias could perhaps be muted by replacing first names with first initials, bias against ethnic background would be harder to hide since family names tend to signal ethnic or geographic information about the author. Similarly, bias in favor (or against) well-known authors would not be eliminated.

**Mutual admiration societies.** No academic would disagree that there are clusters of *other* academics who all love each other, cite each other, and whose output is completely irrelevant to academia outside of their clique, much less the real world. Any ranking function can count citations, but smart ranking functions will also be able to determine something of the size and vitality of any given clique. Furthermore, if a blockbuster result truly happened in an otherwise small clique, perhaps garnering external recognition for the author, this needs to be detected and worked into the rankings. Engineering the right sort of ranking function will likely be a work in progress even after CSPub has otherwise been widely deployed and accepted.

**Ownership.** While CSPub could certainly reduce costs for our community, it would smack into copyright ownership issues. While we, by definition, run our own professional societies (ACM, IEEE, etc.) and can vote to have them assign the necessary copyrights, we cannot necessarily do this for other publishers in our field. Of course, we all regularly ignore copyright assignments and post our papers on our personal home pages. Broadly, the notion of public availability of academic publications has been widely debated, often using the term "open access." Bachrach et al. [4] was an early proponent of giving non-exclusive licenses to traditional journals, while allowing academics to otherwise disseminate their works without interference. More recently, both MIT[8] and Harvard[9] have adopted policies wherein all faculty grant the university a non-exclusive license to disseminate their works, free of charge. If all major universities followed suit, this problem could well go away.

**Fewer/better publications.** I've hypothesized that CSPub could shift the incentive structure from producing lots of papers to producing a smaller number of "hit" papers, but that ultimately comes down to the ranking function, and as different organizations may apply different ranking functions, some may well still prefer paper volume. Furthermore, if other fields are sending out faculty applicants with 10-20 publications on their vita while CS sends out faculty applicants with a quarter of that ("but check out the ranking!"), this could make it more difficult for CS departments to compete against their peer departments for faculty slots.

**"Outside" works.** CS overlaps on its edges with a variety of other science and engineering disciplines which may or may not follow our lead. Furthermore, we often wish to cite material from outside of traditional academia, ranging from newspaper articles, web pages, and blogs, to judicial findings, books, personal correspondence, or other media. Even within CS, the proposed system here won't be adopted all at once. Consequently, we need a mechanism to consistently cite works that are "outside" the system such that, if and when they become available "inside", all the appropriate links are automatically updated. This leads to something of a security / anti-spam problem where somebody could well claim to be uploading a classic paper from the field while the actual content might be something radically different. A related problem is that there may be many variants of a popular paper, and we would like to unify the disparate citations, while again preventing spammers from trying to say they have a new version of a classic paper, when they in fact have nothing of the sort. While a variety of automated tasks could well deal with these concerns, manual intervention by trusted humans, working for our professional societies, will ultimately be necessary to maintain and police our intellectual commons.

**Metadata security policies.** In addition to attempts to upload fraudulent copies of historical work, there will be a variety of other security concerns where authors might wish to lie about a publication's status or awards given. If CSPub were to be implemented by a monolithic, centralized system, with a full-time staff to manage it, this would be all quite tractable, but from a reliability and availability perspective, it's almost certainly preferable to replicate the full system widely. At that point, additional techniques, perhaps a public key cryptographic infrastructure with professional societies delegating to conference steering committees, would be necessary to ensure the desired integrity and privacy policies. Needless to say, if and when a system like this were to be widely used, there would be strong incentives to game the system, so suitable protections would need to be engineered in from the beginning.

[8] http://info-libraries.mit.edu/scholarly/faculty-and-researchers/mit-faculty-open-access-policy/

[9] http://cyber.law.harvard.edu/node/3927

# 3 Rollout

Without a doubt, the biggest challenge of this proposal is turning it into a reality. Computer science sholarship is published under a variety of professional organizations including the ACM, IEEE, AAAI, USENIX, ISOC, IACR, and many more. It's enough of a challenge to imagine any *one* of these organizations moving to the wholesale adoption of a new publication model, much less all of them at once.

The only feasible path is for one organization to develop CSPub for itself and start using it one conference at a time. Initially, anybody could submit a paper, as in arXiv.org or the Crypto ePrint server, and for the pioneering conferences, this would be the exclusive mechanism for submitting a paper to be considered for inclusion. Citations to older papers, not included in the archive, would need to create metadata stubs for which the genuine authors could later insert their own papers. (This metadata could be seeded with collections that already track this, such as DBLP.) By making this mandatory, at least for the authors of the pioneer conferences, the system will be populated by those papers and will have its initial users.

Authors who presently serve PDFs of their papers from personal or lab home pages could migrate, one by one, to using CSPub instead. CSPub could also provide convenience functions, generating dynamic HTML that can be included in personal home pages, research group pages, and so forth. By providing such convenient services, academic authors may well upload all of their papers at once to take advantage of CSPub's features (and increase their work's visibility).

Inevitably, the switch will occur one research community at a time. First they would switch over their conferences, and next they would switch or reinvent their journals. If CSPub really takes off, it's easy to see the broader CS community jumping on the bandwagon, with the attendant growth in tools that understand the archive, and in the bureaucracy that manages the software and standardizes the ranking algorithms.

Of course, different research communities within CS have very different and sometimes experimental models for how they manage their papers. For example, the database community created a new journal, *Proceedings of the VLDB Endowment*[10], where authors may submit papers year-round and may prepare rebuttals to referee comments. Papers accepted to the journal also appear at a subsequent VLDB conference. Starting in 2011, the "journal" mechanism will be the exclusive way of publishing a VLDB paper. The International Conference on Logic Programming (ICLP) has similarly announced that accepted papers will appear in the journal *Theory and Practice of Logic Programing*[11]. CSPub could easily allow scholarship to follow any variation on the VLDB or ICLP models, with each community coming up with its own rules. At the end of the day, the only thing that changes in CSPub are the conditions under which a manuscript gains the metadata to indicate that it has been accepted for a conference or journal.

The largest cost concern would be the loss of revenue from the digital libraries hosted by our existing professional societies as well as the loss of income from research libraries' journal subscriptions. Of course, even today, virtually any current paper can be found on one of its co-authors' home pages. CSPub, by virtue of institutionalizing this practice, would require the ACM, IEEE, and so forth to redo their budgeting. Additional fees might need to come from registration fees at conferences, which might be balanced out by the fee savings that can come from eliminating printed proceedings. Also, if we save our institutional libraries from needing to spend large fees so we can read research papers, that money could be redirected in a variety of other ways, such as paying the labor costs of scanning and entering old works into CSPub.

If anything, the biggest concern will be "ownership." Should this service be "owned" by USENIX, ACM, and/or IEEE, all of whom have their own online services? Will Springer or Elsevier insist that they get a cut for works they've previously published on behalf of CS authors?

[10] http://www.vldb.org/pvldb/

[11] http://www.floc-conference.org/ICLP-cfp.html

# 4 Related work

## 4.1 Open access

Paul Ginsparg founded xxx.lanl.gov, an online pre-print server for physics manuscripts, in 1991. It has since expanded to other disciplines, has been renamed to arXiv.org, and was relocated to Cornell. The service is currently supported by the National Science Foundation, Cornell University, and a variety of other donors, with an annual budget of $400,000. (Wikipedia[12] has further details.) The cryptography community founded its own e-print server (eprint.iacr.org) in 2000. Interestingly, the crypto community has grappled with the issue of anonymous conference submissions with simultaneous non-anonymous pre-print submissions:

> It was ensured that authors are allowed to announce their results in public when they are in an anonymous refereeing process, that they can tell (and give away papers to) colleagues who work on similar matters and should know about an author's results. If an author announces a result widely, and committee members are on the distribution list, they should not be removed just because the paper is in submission. Authors are allowed to give talks on their papers and submit them to existing preprint servers, which will usually be announced widely. On the other hand, it is not intended that a submitter send letters to all the committee members saying who wrote which paper. Anonymous submission just means that papers are submitted without author's names and too obvious references.[13]

arXiv has a computer science section, the Computing Research Repository (CoRR). According to Joseph Halpern, who runs CoRR, CoRR has been growing at 35-40% per year and may experience 60% growth in 2010, which would yield 7000 submissions this year [21]. Halpern also notes that many conferences are considering using CoRR to host their archival publications. CoRR could well form the basis of a CSPub implementation.

Many conferences, workshops, and journals have gone to "paperless" submission, review, and publication. For example, consider the Web 2.0 Security & Privacy Workshop (W2SP) versus the Systematic Approaches to Digital Forensic Engineering Workshop (SADFE). Both workshops run alongside the IEEE Symposium on Security and Privacy. W2SP disseminates final papers exclusively online[14] while SADFE has printed proceedings. SADFE's 2010 registration fee for IEEE members was $295. W2SP's registration fee for IEEE members was $100 less. That savings resulted entirely from avoiding the publication costs. (W2SP's web server is hosted by a research server at Rice.) An example of a paperless journal is *Logical Methods in Computer Science*[15], published under the auspices of the International Federation of Computational Logic, with no cost to publishers or readers. Authors retain their copyright while agreeing to have their work distributed by the journal under a Creative Commons license. LMCS also distributes its publications through CoRR.

In the medical field, PubMed / MedLine[16] has become a one stop source for finding papers, but these ultimately just link back to the original publishers' web sites for full papers, which are not generally available without cost, although universities will pay for site licenses. Attempts to make all NIH-funded research available via a public, open-access system have met with resistance [41]. The White House's Office of Science and Technology Policy (OSTP) has solicited comments on how open access can and should occur [30].

[12] http://en.wikipedia.org/wiki/ArXiv

[13] See http://eprint.iacr.org/about.html

[14] http://w2spconf.com

[15] http://www.lmcs-online.org

[16] http://www.ncbi.nlm.nih.gov/pubmed

For legal documents, WestLaw and LexisNexis are comprehensive sources, but they charge significant fees. WestLaw has asserted copyright over their corrections and annotations to legal cases, including their numerical taxonomy and their internal page numbering, which were shot down in court [24]. Silversmith [39] summarizes the case and the issues, particularly the problem of uniformity among legal citations.

Ke-Sen Huang began collecting links to papers from SIGGRAPH and other graphics conferences, along with their related project pages[17], offering graphics researchers everything they might want from one web page. The ACM forced Huang to remove his page, which was only reinstantiated after strong protests from the community (see, e.g., [17, 22, 23]).

MIT's OpenCourseWare[18] and Rice's Connexions[19] have applied the open access idea to course textbooks and educational materials, allowing course instructors to easily mix and match for their own course needs. Presumably, as these materials are used and remixed, authorship can be tracked and CSPub ranking can be conferred upon both the texts and the authors.

## 4.2 Changing the rules

Some CS conferences have started experimenting with new models for how they accept and disseminate papers. For example, the ACM Conference on Electronic Commerce[20] accepts papers as normal but also accepts "working papers" under an interesting policy:

> To accommodate the publishing traditions of different fields, authors may instead submit working papers that are under review or nearly ready for journal review. These submissions will be subject to review and considered for presentation at the conference but not for publication in the proceedings. Abstracts of accepted working papers will be included in the proceedings and must be coupled with a URL that points to the full paper and that will be reliable for at least two years. Open access is preferred although the paper can be hosted by a publisher who takes copyright and limits access, as long as there is a link to the location.

Future publication models are a regular topic of discussion within the community. For example, the 2009 Conference on Neural Information Processing Systems had a panel[21] on exactly this topic. Langford [29] summarized the discussion, which covered many of the same issues that I discuss in this manuscript. For this panel, Lecun [31] suggested a system similar to CSPub, with papers going into a central archive, but with a more elaborate reviewing system, where reviews are public, are citable documents in their own right, and which would be subject to being voted up or down (akin to the discussion mechanism on Slashdot). Ghahramani [20] also proposed to adopt a model akin to the VLDB hybrid conference/journal system, alongside a community of reviewers who must submit some number of reviews per year in order to remain in good standing with the community.

STOC (The ACM Symposium on the Theory of Computing) is similarly having a debate on how it might change its rules, starting from a proposal by Lance Fortnow [19] to radically increase STOC's acceptance rate. After much debate, chronicled on a blog[22], STOC will this year accept 90 papers rather than 80.

Crowcroft et al. [11] offer a number of good ideas, including public reviews for papers (digitally signed, but using pseudonyms), allowing the community to determine not only the relative value of each paper,

[17] http://kesen.realtimerendering.com

[18] http://ocw.mit.edu

[19] http://cnx.org

[20] See., e.g., http://www.sigecom.org/ec09/papers.html

[21] http://nips.cc/Conferences/2009/PublicationModelsDebate

[22] http://futureofstoc.blogspot.com

as I discuss here, but also the relative value of each reviewer. In my own experience, I've observed that reviewers' quality of work depends on a variety of factors, such as how well they understand the paper they're reviewing, and furthermore how busy they happen to be at the time. A uniform reputation for a given reviewer seems like an unhelpful idea. I do like their idea of having a "best reviewer" award as part of a package of incentives that will help incentivize better reviews for papers.

Kelty et al. [25] discuss the bootstrapping process of changing from traditional peer review to an open publication model. They also address the concern of the increasing volume of publications in the world and consider alternative models for how manuscripts should be published and reviewed.

Vardi [43, 44] considers the issue that CS stands alone in using conferences rather than journals as our primary means of disseminating top research results. Fortnow [18] responded with a provocative title ("Time for computer science to grow up") and a critique of the failure of our conference system, suggesting radically higher paper acceptance rates and fewer meetings for conferences as one possible remedy.

Welsh [45], the incoming editor for ACM's *Transactions on Sensor Networks*, in an attempt to explicitly deal with the famously long lead times on journals, intends to make a formal fast-path linkage between top conferences in the area of sensor networks and the journal in the area as well as explicitly soliciting manuscripts that may be of interest to the community. His ideas could easily be applied to any other journal.

Berman and Schneider [5] discuss the overload of program committees in CS systems conferences and identify many of the same problems as others (lower review quality, incentives to produce incremental results, etc.). They suggest that the bottom 2/3 of submitted papers (in my lingo, the "second tier" and "noncompetitive" papers) should not get detailed reviews. They also suggest that authors be more mindful of not wasting the program committees' limited time, an idea also echoed by Lemire [33].

Further discussion on these topics can be found in Korth et al. [27], Wing [46], Dalal [13, 12] and Dupuis [14, 15]. Furthermore, the recent Workshop on Organizing Workshops Conferences and Symposia (WOWCS)[23] published a variety of relevant papers. Workshop attendees also created a wiki[24] to discuss additional steps and thoughts on the evolution of the process of publications.

The Liquid Publications project [9, 10] has more ambitious goals than CSPub, merging submission, review, and dissemination into a single web service. They further want to decompose traditional research papers into "scientific knowledge objects" that can be reused and remixed, much like Connexions allows for textbook chapters, but for all academic output. Arguably, such a radical change would be less amenable to incremental rollout and community acceptance than CSPub, which preserves the current conference publication model.

Mendeley[25] is a web and desktop tool that allows researchers to organize, share, and discover manuscripts. Many of its features are sympathetic with the goals of CSPub and would become much more valuable if CSPub were adopted.

### 4.3 Ranking and metrics

The topic of how academics, much less whole research universities, should be ranked is the subject of a vigorous and ongoing debate. *Nature* produced a special issue on this topic on June 16, 2010 (the same day as the initial draft of this manuscript). All the metrics-related articles are available online[26]. Notably, *Nature* polled 150 readers as well as provosts, department heads, and other research administrators at 30 major research institutions [1]. Three-quarters of those polled believed that metrics are being used in hiring

[23] http://www.usenix.org/events/wowcs08/tech/

[24] http://wiki.usenix.org/Main/Conference/CollectedWisdom

[25] http://www.mendeley.com

[26] http://www.nature.com/metrics/

decisions and promotion, and almost 70% believed that they are being used in tenure decisions and performance review. A majority (63%) were unhappy about the way in which some of these measures are used. Administrators countered that they don't particularly rely on metrics, favoring written letters. Nonetheless, they don't entirely discount them, either.

The field of "bibliometrics" has experienced an explosion of interest [42], and many of them have pursued PageRank-like metrics, notably including Radicchi et al. [37], who used the entire database of 407K articles published by the American Physical Society to measure per-author rankings as they evolve over time[27]. The science of bibliometrics is still a work in progress (see, e.g., Adler et al. [2], which has a good bibliography), and the whole concept has been subject to significant criticisms and may be too easy to game (see, e.g., Laloë and Mosseri [28] or Arnold [3]).

## Acknowledgements

This manuscript began through informal discussions with many of my peers. I would like to thank Drew Dean, Joseph Halpern, Mike Herf, Peter Honeyman, Carol Hutchins, Dave Johnson, Chris Kelty, Eric Rescorla, Moshe Vardi, Suresh Venkatasubramanian, and Ellie Young for their feedback and commentary on earlier drafts. Many of my colleagues have expressed skepticism at these ideas. Chris Jermaine notably had a negative reaction to the idea of creating a centralized ranking metric, feeling that this would only further push us away from the goal of publishing quality versus quantity. Arguably, a good metric could counteract this, but a bad metric might well exacerbate the problem.

I would also like to thank all the authors of the cited papers, all of which I was able to read online without having to pay anything, and thus making me more likely to cite them. I apologize to readers who cannot follow every link, some of which may have only been accessible to me because my employer pays for site licenses. When I could find freely downloadable copies, I preferred those URLs in my bibliography.

## References

[1] A. Abbott, D. Cyranoski, N. Jones, B. Maher, Q. Schiermeier, and R. Van Noorden. Metrics: Do metrics matter? *Nature*, 465:860–862, June 2010. http://www.nature.com/news/2010/100616/full/465860a.html.

[2] R. Adler, J. Ewing, and P. Taylor. Citation statistics. *Statistical Science*, 4(1):1–14, 2009. http://projecteuclid.org/euclid.ss/1255009002.

[3] D. N. Arnold. Integrity under attack: The state of scholarly publishing. *SIAM News, Talk of the Society*, Dec. 2009. http://www.siam.org/news/news.php?id=1663.

[4] S. Bachrach, R. S. Berry, M. Blume, T. von Foerster, A. Fowler, P. Ginsparg, S. Heller, N. Kestner, A. Odlyzko, A. Okerson, R. Wigington, and A. Moffat. Who should own scientific papers? *Science*, 281(5382), Sept. 1998. See also, http://www.dtc.umn.edu/~odlyzko/doc/paper.ownership.htm.

[5] K. Birman and F. B. Schneider. Program committee overload in systems. *Communications of the ACM*, 52(5):34–37, 2009. http://cacm.acm.org/magazines/2009/5/24644-program-committee-overload-in-systems/fulltext.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems*, pages 107–117, 1998. http://infolab.stanford.edu/~backrub/google.html.

[7] D. Butler. Academics strike back at spurious rankings. *Nature*, 447:514–514, May 2007. http://www.nature.com/nature/journal/v447/n7144/full/447514b.html.

[8] D. Butler. University rankings smarten up. *Nature*, 464:16–17, Mar. 2010. http://www.nature.com/news/2010/100303/full/464016a.html.

[27] http://www.physauthorsrank.org

[9] F. Casati, F. Giunchiglia, and M. Marchese. *Liquid Publications: Scentific Publications Meet the Web*. University of Trento, Oct. 2007. https://dev.liquidpub.org/svn/liquidpub/papers/deliverables/LiquidPub%20paper-latest.pdf.

[10] F. Casati, F. Giunchiglia, and M. Marchese. Publish and perish: why the current publication and review model is killing research and wasting your money. *ACM Ubiquity*, 8(3), Feb. 2007. http://www.acm.org/ubiquity/views/v8i03_fabio.html.

[11] J. Crowcroft, S. Keshav, and N. McKeown. Scaling the academic publication process to Internet scale. In *Workshop on Organizing Workshops, Conferences, and Symposia for Computer Science (WOWCS '08)*, San Francisco, CA, Apr. 2008. http://www.usenix.org/events/wowcs08/tech/full_papers/crowcroft/crowcrofthtml/, also reprinted in *CACM*, Vol. 52, No. 1, January 2009.

[12] A. C. Dalal. An interesting publishing model. *This is what a computer scientist looks like*, July 2010. http://acdalal.wordpress.com/2010/07/23/an-interesting-publishing-model/.

[13] A. C. Dalal. Publishing calculus. *This is what a computer scientist looks like*, July 2010. http://acdalal.wordpress.com/2010/07/15/publishing-calculus/.

[14] J. Dupuis. Time for computer science to grow up? *Confessions of a science librarian*, Sept. 2009. http://scienceblogs.com/confessions/2009/09/time_for_computer_science_to_g.php.

[15] J. Dupuis. Are computing journals too slow? *Confessions of a science librarian*, Mar. 2010. http://scienceblogs.com/confessions/2010/03/are_computing_journals.php.

[16] T. A. Edison. Letter to William Emmet, Nov. 1926. Reprinted in *Letters of Note*, http://www.lettersofnote.com/2010/02/worst-is-to-come.html.

[17] C. Ericson. ACM censors linking! *realtimecollisiondetection.net - the blog*, Nov. 2009. http://realtimecollisiondetection.net/blog/?p=101.

[18] L. Fortnow. Viewpoint: Time for computer science to grow up. *Communications of the ACM*, 52(8), Aug. 2009. http://cacm.acm.org/magazines/2009/8/34492-viewpoint-time-for-computer-science-to-grow-up/fulltext.

[19] L. Fortnow. Future of STOC proposal. *Future of STOC Blog*, June 2010. http://futureofstoc.blogspot.com/2010/06/future-of-stoc-proposal.html.

[20] Z. Ghahramani. A modest proposal. *Machine Learning (blog)*, Dec. 2009. http://hunch.net/?page_id=1115.

[21] J. Halpern. Private communication, July 2010.

[22] N. Hoffman. Ke-Sen Huang's paper pages are down, will soon go back up. *Real-Time Rendering Blog*, Nov. 2009. http://www.realtimerendering.com/blog/ke-sen-huangs-paper-pages-are-down-will-soon-go-back-up/.

[23] N. Hoffman. US gov requests feedback on open access - ACM gets it wrong (again). *Real-Time Rendering Blog*, Dec. 2009. http://www.realtimerendering.com/blog/us-gov-requests-feedback-on-open-access-acm-gets-it-wrong-again/.

[24] D. C. Johnston. West Publishing loses a decision on copyright. *New York Times*, May 1997. http://www.nytimes.com/1997/05/21/business/west-publishing-loses-a-decision-on-copyright.html.

[25] C. M. Kelty, C. S. Burrus, and R. G. Baraniuk. Peer review anew: Three principles and a case study in postpublication quality assurance. *Proceedings of the IEEE*, 96(6), June 2008. http://dsp.rice.edu/~richb/peer-review-anew-ProcIEEE-june08.pdf.

[26] E. Kohler. Hot crap! In *Workshop on Organizing Workshops, Conferences, and Symposia for Computer Science (WOWCS '08)*, San Francisco, CA, Apr. 2008. http://www.usenix.org/event/wowcs08/tech/full_papers/kohler/kohler_html/.

[27] H. F. Korth, P. A. Bernstein, M. Fernandez, L. Gruenwald, P. G. Kolaitis, K. McKinley, and T. Ozsu. Paper and proposal reviews: is the process flawed? *ACM SIGMOD Record*, 37(3):36–39, 2008. http://doi.acm.org/10.1145/1462571.1462581.

[28] F. Laloë and R. Mosseri. Bibliometric evaluation of individual researchers: not even right... not even wrong! *Europhysics News*, 40(5):26–29, 2009. http://dx.doi.org/10.1051/epn/2009704.

[29] J. Langford. Future publication models @ NIPS. *Machine Learning (blog)*, Dec. 2009. http://hunch.net/?p=1086.

[30] P. Larson. Public access policy update. *White House Office of Science and Technology Policy (OSTP) Blog*, Mar. 2010. http://www.whitehouse.gov/blog/2010/03/08/public-access-policy-update.

[31] Y. LeCun. A new publishing model in computer science, Dec. 2009. http://yann.lecun.com/ex/pamphlets/publishing-models.html.

[32] B. M. Leiner. The NCSTRL approach to open architecture for the confederated digital library. *D-Lib Magazine*, Dec. 1998. http://www.dlib.org/dlib/december98/leiner/12leiner.html.

[33] D. Lemire. Trading latency for quality in research. *Daniel Lemire's blog*, Feb. 2010. http://www.daniel-lemire.com/blog/archives/2010/02/08/trading-latency-for-quality-in-research/.

[34] P. Maniatis, M. Roussopoulos, T. Giuli, D. S. H. Rosenthal, M. Baker, and Y. Muliadi. LOCKSS: A peer-to-peer digital preservation system. *ACM Transactions on Computer Systems (TOCS)*, 23(1):2–50, Feb. 2005. See also, http://lockss.stanford.edu/lockss/Publications.

[35] K. S. McKinley. Editorial: Improving publication quality by reducing bias with double-blind reviewing and author response. *ACM SIGPLAN Notices*, 43(8):5–9, Aug. 2008. http://userweb.cs.utexas.edu/users/mckinley/notes/blind.html.

[36] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical Report 1999-66, Stanford InfoLab, Nov. 1999. http://ilpubs.stanford.edu:8090/422/.

[37] F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani. Diffusion of scientific credits and the ranking of scientists. *Phys. Rev. E*, 80(5):056103, Nov 2009. http://filrad.homelinux.org/Mypapers/pre_80_056103.pdf.

[38] J. H. Saltzer and M. D. Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, Sept. 1975. Reprinted in David D. Clark and David D. Redell, editors, *Protection of Information in Computer Systems*, IEEE 1975 CompCon tutorial, IEEE # 75CH1050-4. Also reprinted in Rein Turn, editor, *Advances in Computer System Security*, ArTech House, Dedham, MA, 1981, pages 105-135, ISBN 0-89006-096-7. Also reprinted in Marvin S. Levin, Steven B. Lipner, and Paul A. Karger, *Protecting Data & Information: A Workshop in Computer & Data Security*, Digital Equipment Corporation, 1982. Rendered as a web page in 1997 by Norman Hardy, http://web.mit.edu/Saltzer/www/publications/protection/.

[39] J. Silversmith. Universal citation: The fullest possible dissemination of judgments. *College Hill Internet Legal Practice Newsletter*, May 1997. http://www.thirdamendment.com/citation.html.

[40] B. Stein, P. Rosso, E. Stamatatos, M. Koppel, and E. Agirre, editors. *3rd PAN Workshop and 1st Competition on Plagiarism Detection*, Donostia-San Sebastian, Spain, Sept. 2009. http://www.uni-weimar.de/medien/webis/research/workshopseries/pan-09/.

[41] J. Timmer. Congress's copyright fight puts open access science in peril. *Ars Technica*, Sept. 2008. http://arstechnica.com/tech-policy/news/2008/09/open-access-science.ars.

[42] R. Van Noorden. Metrics: A profusion of measurements. *Nature*, 465:864–866, June 2010. http://www.nature.com/news/2010/100616/full/465864a.html.

[43] M. Y. Vardi. Conference vs. journals in computing research. *Communications of the ACM*, 52(5):5, May 2009. http://cacm.acm.org/magazines/2009/5/24632-conferences-vs-journals-in-computing-research/fulltext.

[44] M. Y. Vardi. Revisiting the publication culture in computing research. *Communications of the ACM*, 53(5), Mar. 2010. http://cacm.acm.org/magazines/2010/3/76297-revisiting-the-publication-culture-in-computing-research/fulltext.

[45] M. Welsh. Editor in chief. *Volatile and Decentralized (blog)*, June 2010. http://matt-welsh.blogspot.com/2010/06/editor-in-chief.html.

[46] J. Wing. Breaking the cycle. *Blog@CACM*, Aug. 2009. http://cacm.acm.org/blogs/blog-cacm/38402-breaking-the-cycle/fulltext, also reprinted in *CACM*, Vol. 52, No. 12, December 2009.