From the Editor in Chief...

Collective Wisdom
A Modest Proposal to Improve Peer Review, Part I

Fred Douglis • IBM T.J. Watson Research Center

After a few columns focusing at least in part on how scammers infest our inboxes and affect our Web rankings, I’d like to now look at something closer to home: academic publishing and peer review. I recognize that we have many types of IC readers: some participate in the “academic research” community by publishing peer-reviewed articles and serving on program committees; some publish more educational content in the form of columns, blogs, and so on; some are active in standards communities; and finally, some read IC for its content but don’t contribute material for publication, here or elsewhere.

Those who publish or review content will probably find that my comments in this column hit close to home, one way or another. Those who do neither might just skim this text to set the stage for a more Internet-centric discussion in the next issue.

I was prompted to write on this topic when a recent submission to IC turned out to be substantially similar to an earlier submission already under review with another IEEE publication. Misunderstanding the obligations of academic authorship, the authors disclosed the existence of the other submission only after that manuscript was accepted for publication; it took a few more weeks before those involved realized the extent of the overlap, at which time we had intended to accept the IC submission. Both publications ended up rejecting the manuscripts because of this violation.

The problem with the rejected submission made me realize two things. First, not all IC’s potential authors appreciate the rules for overlapping submissions, so elaborating on the guidelines could help avoid similar problems in the future. Second, we might have opportunities to improve the process and catch such overlap.

Submission Etiquette

Why do rules about simultaneous submission, overlapping content, and intellectual novelty exist? I can think of several reasons:

• Copyright. A publisher typically holds the copyright on material in its magazines, conference proceedings, and such. It can’t publish something to which someone else already holds copyright — publishing a significant piece of text without attributing it to the other publication is plagiarism and a copyright violation, even if the same person authored both works.

• Integrity. Even if content isn’t word-for-word identical, if it imparts the same basic information, it should cite the earlier work. This is true whether the earlier work is published, has been accepted for publication, or is currently under review. General guidelines exist for how much new material must appear in a manuscript to be published as a separate entity, as I’ll discuss later.

• Timeliness. Submitting papers simultaneously to multiple venues, then withdrawing them once a publication has accepted the submission, would let authors publish their papers a lot faster. However, this results in wasted effort and uncertainty for the other venues — especially if they’re close to accepting a paper, only to have it withdrawn. Generally, submitting a paper to a conference or periodical is an agreement that if the reviewers accept it, the authors will edit it and submit the final version for publication. Withdrawing an accepted paper breaks all assumptions about what will be published and when.

• Wasted resources. Reviewers are a valuable resource — they volunteer their time to help the peer-review process succeed. Having a reviewer evaluate a manuscript that will never be published wastes that time. In contrast, if an author submits to some venue and is rejected, he or she can improve the manuscript and submit it else-
where (that is, they haven’t wasted any effort, except to the extent that they tailor a submission to one venue and not another).

In our example, the authors took the online instructions at the **IC** submission site quite literally. It said that previously published content should be cited; because the other manuscript hadn’t yet been published, it wasn’t indicated. The site also said simultaneous submissions aren’t allowed. Here, the overlap was substantial but not complete. Providing the earlier submission and indicating the overlap would have let reviewers determine whether the later submission was a “new” publication. Although **IC** likely would have rejected the submission as being too similar, the first one, once accepted, would have been published. (And yes, I have requested that these instructions be clarified.)

If you plan to submit content (here or elsewhere) that’s related to earlier published or submitted work, I strongly recommend that you familiarize yourself with the appropriate guidelines. Both the IEEE (http://www.ieee.org/web/publications/rights/Section_822F.html) and the ACM (www.acm.org/pubs/plagiarism%20policy.html) offer help. Additionally, some academic publications cover the topic of self-plagiarism. Generally, the rule is, if it appears elsewhere, disclose it. Gray areas exist, of course — for instance, the related work sections of different papers on the same topic are likely to be fairly similar, and a copied sentence here or there won’t raise eyebrows for the same author the way it would if it were truly plagiarized from others.

In addition, some publishers offer specific guidelines on how much new material is required for them to republish material from an earlier publication in a “lesser” venue (the ACM requires 25 percent, for example). This means that authors can add to a conference paper to publish it in a magazine or journal, but are discouraged from submitting a conference publication to another conference, even with additional content. The same holds true for republishing one periodical’s content in another.

**Is It Time?**

Fortunately, in my experience, significant cases of self-plagiarism in computer science have been relatively rare, and plagiarizing other people’s work seems even rarer. Minor cases of self-plagiarism, such as including the same figures in different papers, are growing, and more and more venues exist in which we can publish academic work. When reviewers detect self-plagiarism early in the process, the consequences are minimal, and when detected after publication, the stigma is primarily on the authors (for instance, when the publisher must annotate the online copy of a paper to indicate the other work). But a very awkward window exists during which dropping a tainted paper has no effect on the authors (other than a rejected submission), yet the publication itself suddenly has one fewer paper. For a magazine with a specific page target each issue, such as **IC**, losing this article necessitates publishing other content in its place; worse, in the case of a special issue, this could result in too few “theme” articles appearing. Given these ramifications, should we have some sort of procedure to search for self-plagiarism more proactively?

In their article analyzing the types of self-plagiarism, Christian Collberg and Stephen Kobourov described a tool called SPlaT (for **Self-Plagiarism Tool**; http://splat.cs.arizona.edu/) that can crawl the Web for text published by the same authors and highlight possible self-plagiarism cases. Reviewers could use the tool to find self-plagiarism against previously published...
work, but it wouldn’t help with respect to simultaneous submissions unless those submissions are publicly available online.

Contrast this with tools that professors (and even high school teachers) use to detect student plagiarism. Many schools require their students to submit works electronically, both for comparison against other works and for storage in the corpus of materials used in later checks (http://tinyurl.com/yrej5k). However, some students have raised successful legal challenges to this requirement as a violation of their copyright; others have simply pushed back, resulting in relaxed requirements.

Would such a tool work for academic publishing? SPlaT effectively detects self-plagiarism against public documents; similar techniques could detect plagiarism of others’ work if it were a big enough issue. What about detecting overlapping submissions, given that they’re confidential until formally published? Presumably, a given organization such as the IEEE could detect two substantially similar submissions to its formal publications (magazines and transactions) easily enough, using a tool like SPlaT; in fact, the IEEE just announced that it will soon test a plagiarism-detection tool, which I expect would detect copied text from others’ work and an authors’ own published work, but not catch self-plagiarism of parallel submissions (http://tinyurl.com/yrej5k).

But it gets much harder when dealing with something submitted to conferences or different professional organizations.

I don’t have a solution here, only a challenge: I would like to see a system for detecting overlapping submissions without disclosing content. One obvious approach would be to submit manuscript signatures, rather than the content itself, using something such as Rabin-Karp fingerprints (http://en.wikipedia.org/wiki/Rabin-Karp_string_search_algorithm). These fingerprints, which Web search engines have used for several years to suppress similar pages from results, can efficiently hash sliding windows of content such that a few common fingerprints can indicate that two documents are very similar at a textual level. Researchers have used them to detect phrase-level similarity of Web pages (phrases from different pages strung together as “Web spam”) as well.3

However, the more detail we store about each manuscript in order to identify overlapping content, the more manuscript content is effectively revealed. Questions also arise about how to examine two suspected instances of overlap without violating submission confidentiality. (With periodicals, both can appoint a common reviewer, but conference program committees might be harder to manage.) Finally, issues exist with regard to managing the central repository to ensure that submissions, once rejected, are purged from the system.

What Are the Options?
One possible approach — legal rather than technical — would be to require authors to agree to manuscript submission, analysis by an independent party, and storage until and unless it’s formally rejected. In exchange, the repository owner would have to provide the same confidentiality guarantees that the organizations reviewing the manuscripts currently do. Would the academic publishing community agree to a system that many students have objected to so strenuously? To the extent that authors feel such a system presupposes guilt, it would be a tough sell. To the extent that authors feel it promotes academic integrity and would simply catch inadvertent self-plagiarism, it might be viable.

To provide a real-world analogy, I recently had a discussion about New
York state real estate rules. Apparently, home sellers must either submit a disclosure about various aspects of the home — and receive stiff penalties for any false statements — or pay buyers US$500. When I asked why anyone would buy a home without the disclosure, my colleague explained that virtually everyone pays the fee rather than risk issues with disclosure, so buyers don’t have a pool of homes with disclosures to choose from. I wonder if that model applies here: if an organization such as the IEEE started to require using a single repository for its conferences and periodicals, it would probably encompass enough publishing venues to enforce compliance in a way that a smaller organization might not.

What about making this system optional? Perhaps it would be sufficient to add a checkbox to let authors agree to submit their manuscript to the shared repository, permitting them to opt out. Some would opt in, some might opt out on general principles, and some might opt out because they have a legitimate fear of what such an analysis would find. The ones who opt out on general principles might be like the real estate sellers who won’t disclose information about their house just in case they got something wrong. Thus, in the event of a problem, the penalty for those who do allow this comparison should be small. As is the case today, the penalty would depend on the extent of the self-plagiarism (the IEEE guidelines give examples at www.ieee.org/web/publications/rights/ID_Plagiarism.html). I imagine that a publisher’s common response would be to require using a single repository for all manuscripts. One part of the solution to both the self-plagiarism issue and chronic reviewer misbehavior is Internet-based neutral, trusted agents, which I’ll delve into next time around.

Acknowledgments
I thank several people for contributing their thoughts to this discussion, as well as anecdotal evidence of the concerns I raised: Siobhán Clarke, Bob Filman, Stephen Kohourov, Doug Lea, Erich Nahum, Charles Petrie, Prabhakar Raghavan, Munindar Singh, Zhen Xiao, and Fan Ye. The opinions expressed in this column are my personal opinions. I speak neither for my employer nor for IEEE Internet Computing in this regard, and any errors or omissions are my own.

References