DOI:10.1145/1866739.1866751

Mark D. Ryan

Viewpoint
Cloud Computing Privacy Concerns on Our Doorstep

Privacy and confidentiality issues in cloud-based conference management systems reflect more universal themes.

CLOUD COMPUTING MEANS entrusting data to information systems that are managed by external parties on remote servers "in the cloud." Webmail and online documents (such as Google Docs) are well-known examples. Cloud computing raises privacy and confidentiality concerns because the service provider necessarily has access to all the data, and could accidentally or deliberately disclose it or use it for unauthorized purposes.

Conference management systems based on cloud computing represent an example of these problems within the academic research community. It is an interesting example, because it is small and specific, making it easier to explore the exact nature of the privacy problem and to think about solutions. This column describes the problem, highlights some of the possible undesirable consequences, and points out directions for addressing it.

Conference Management Systems

Most academic conferences are managed using software that allows the program committee (PC) members to browse papers and contribute reviews and discussion via the Web. In one arrangement, the conference chair downloads and hosts the appropriate server software, say HotCRP or iChair. The benefits of using such software are familiar:

- Distribution of papers to PC members is automated, and can take into account their preferences and conflicts of interest;
- The system organizes the collection and distribution of reviews and discussion, can rank papers according to scores, and send out reminder email, as well as email notifications of acceptance or rejection; and
- It can also produce a range of other reports, such as lists of sub-reviewers, acceptance statistics, and the conference program.

HotCRP and iChair require the conference chair to download and install software, and to host the Web server. Other systems such as EasyChair and EDAS work according to the cloud computing model: instead of installing and hosting the server, the conference chair simply creates the conference account "in the cloud." In addition to the benefits described previously, this model has extra conveniences:

- The whole business of managing the server (including backups and security) is done by someone else, and gains economy of scale;
- Accounts for authors and PC members exist already, and don't have to be managed on a per-conference basis;
- Data is stored indefinitely, and reviewers are spared the necessity of keeping copies of their own reviews; and

ILLUSTRATION BY GARY NEILL

- The system can help complete forms such as the PC member invitation form and the paper submission form by suggesting likely colleagues based on past collaboration history.

For these reasons, EasyChair and EDAS are an immense contribution to the academic community. According to its Web page, EasyChair hosted over 3,300 conferences in 2010. Because of its optimizations for multiconferences and multitrack conferences, it is mandated for conferences and workshops that participate in the Federated Logic Conference (FLoC), a huge multiconference that attracts approximately 1,000 paper submissions.

Data Privacy Concerns

Accidental or deliberate disclosure. A privacy concern with cloud-computing-based conference management systems such as EDAS and EasyChair arises because the system administrators are custodians of a huge quantity of data about the submission and reviewing behavior of thousands of researchers, aggregated across multiple conferences. This data could be deliberately or accidentally disclosed, with unwelcome consequences.

- Reviewer anonymity could be compromised, as well as the confidentiality of PC discussions.
- The acceptance success records could be identified, for individual researchers and groups, over a period of years; and
- The aggregated reviewing profile (fair/unfair, thorough/scant, harsh/undiscerning, prompt/late, and so forth) of researchers could be disclosed.

The data could be abused by hiring or promotions committees, funding and award committees, and more generally by researchers choosing collaborators and associates. The mere existence of the data makes the system administrators vulnerable to bribery, coercion, and/or cracking attempts. If the administrators are also researchers, the data potentially puts them in situations of conflict of interest.

The problem of data privacy in general is of course well known, but cloud computing magnifies it. Conference data is an example in our backyard. When conference organizers had to install the software from scratch, there was still a risk of breach of confidentiality, but the data was just about one conference. Cloud computing solutions allow data to be aggregated across thousands of conferences over decades, presenting tremendous opportunities for abuse if the data gets into the wrong hands.

The acceptance success records could be identified, for individual researchers and groups, over a period of years.

Beneficial data mining. In addition to the abuses of conference review data described here, there are some uses that might be considered beneficial. The data could be used to help detect or prevent fraud or other kinds of unwanted behavior, for example, by identifying:

- Researchers who systematically unfairly accept each other's papers, or rivals who systematically reject each other's papers, or reviewers who reject a paper and later submit to another conference a paper with similar ideas; and
- Undesirable submission patterns and behaviors by individual researchers (such as parallel or serial submissions of the same paper; repeated paper withdrawals after acceptance; and recurring content changes between submitted version and final version).

The data could also be used to understand and improve the way conferences are administered. ACM, for example, could use the data to construct quality metrics for its conferences, enabling it to profile the kinds of authors who submit, how much "new blood" is entering the community, and how that changes over different editions of the conference. This could help identify conferences that are emerging as dominant, or others that have outlived their usefulness.

The decisions about who is allowed to mine the data, and for what purposes, are difficult. Policies should be decided transparently and by consensus, rather than being left solely to the de facto data custodians.

Ways Forward

Policies and legislation. An obvious first step is to articulate clear policies that circumscribe the ways in which the data is used. For example, a simple policy might be that the data gathered during the administration of a conference should be used only for the management of that particular conference. Adherence to this policy would imply that the data is deleted after the conference, which is not done in the case of Easychair (I don't know if it is done for EDAS). Other policies might allow wider uses of the data. Debate within different academic communities can be expected to yield consensus about which practices are to be allowed in a discipline, and which ones not. For example, some communities may welcome plagiarism detection based on previously reviewed submissions, while others may consider it useless for their subject, or simply unnecessary.

Since its inception in 2002 and up to the time of writing, EasyChair has appeared not to have any privacy policy, or any statement about the purposes and possible uses of the data it stores. There is no privacy policy linked from its main page, and a search for "privacy policy" (or similar terms) restricted to the domain "easychair.org" does not yield any results. I have been told that new users are presented with a privacy statement at the time of first signing up to Easychair. I did not create a new account to test this; regardless, the privacy statement is not linked from anywhere or later findable via search. EDAS does have an easily accessed privacy policy, which (while not watertight) appears to comply with the "use only for this conference" principle.

Another direction would be to try to find alternative custodians for the data—custodians that are not themselves also researchers participating actively in conferences. The ACM or IEEE might be considered suitable, although they contribute to decisions about publications and appointments of staff and fellows. Professional data custodians such as Google might also be considered. It may be difficult to find an ideal custodian, especially if cost factors are taken into account.

In most countries, legislation exists to govern the protection of personal data. In the U.K., the Data Protection Act is based on eight principles, including the principle that personal data is obtained only for specified purposes and is not processed in a manner incompatible with the purposes; and the principle that the data is not kept longer than is necessary for the purposes. EasyChair is hosted in the U.K., but the lack of an accessible purpose statement or evidence of registration under the Act mean I was unable to determine whether it complies with the legislation. The Data Protection Directive of the European Union embodies similar principles; personal data can only be processed for specified purposes and may not be processed further in a way incompatible with those purposes.

Processing encrypted data in the cloud. Policies are a first step, but alone they are insufficient to prevent cloud service providers from abusing the data entrusted to them. Current research aims to develop technologies that can give users guarantees that the agreed policies are adhered to. The following descriptions of research directions are not exhaustive or complete.

Progress has been made in encryption systems that would allow users to upload encrypted data, and allow the service providers to perform computations and searches on the encrypted data without giving them the possibility of decrypting it. Although such encryption has been shown possible in principle, current techniques are very expensive in both computation and bandwidth, and show little sign of becoming practical. But the research is ongoing, and there are developments all the time.

Hardware-based security initiatives such as the Trusted Platform Module and Intel's Trusted Execution Technology are designed to allow a remote user to have confidence that data submitted to a platform is processed according to an agreed policy. These technologies could be leveraged to give privacy guarantees in cloud computing in general, and conference management software in particular. However, significant research will be needed before a usable system could be developed.

Certain cloud computing applications may be primarily storage applications, and might not require a great deal of processing to be performed on the server side. In that case, encrypting the data before sending it to the cloud may be realistic. It would require keys to be managed and shared among users in a practical and efficient way, and the necessary computations to be done in a browser plug-in. It is worthwhile to investigate whether this arrangement could work for conference management software.

Conclusion

Many people with whom I have discussed these issues have argued that the professional honor of data custodians (and PC chairs and PC members) is sufficient to guard against the threats I have described. Indeed, adherence by professionals to ethical behavior is essential to ensure all kinds of confidentiality. In practice, system administrators are able to read all the organization's email, and medical staff can browse celebrity health records; we trust our colleagues' sense of honor to ensure these bad things don't happen. But my standpoint is that we should still try to minimize the extent to which we rely on people's sense of good behavior. We are just at the beginning of the digital era, and many of the solutions we currently accept won't be considered adequate in the long term.

The issues raised about cloud-computing-based conference management systems are replicated in numerous other domains, across all sectors of industry and academia. The problem of accumulations of data on servers is very difficult to solve in any generality. The particular instance considered here is interesting because it may be small enough to be solvable, and it is also within the control of the academic community that will directly benefit—or suffer—according to the solution we adopt. C

Mark D. Ryan (M.D.Ryan@cs.bham.ac.uk) is Professor in Computer Security and EPSRC Leadership Fellow in the School of Computer Science at the University of Birmingham, U.K.

Many thanks to the *Communications* reviewers for interesting and constructive comments. I also benefited from discussions with many colleagues at Birmingham, and also in the wider academic research community. Thanks to Henning Schulzrinne, administrator of EDAS, for comments and clarifications. Drafts of this Viewpoint were sent to Andrei Voronkov, the Easychair administrator, but he did not respond.