Gerome Miklau Speaks Out
on His SIGMOD Distinguished Dissertation Award, How Great It Is to Be a Professor, and More

by Marianne Winslett

http://www.cs.umass.edu/~miklau/

I would like to take a moment to thank the many people who have helped me devise interview questions over the years. Often people propose questions under a promise of anonymity, so I will not name the individuals who have suggested questions---but you know who you are! Without you, these columns would not be possible. Thank you for your many excellent suggestions over the years.

Welcome to this installment of ACM SIGMOD Record's series of interviews with distinguished members of the database community. I'm Marianne Winslett, and today we are at the SIGMOD 2006 conference in Chicago, Illinois. I have here with me Gerome Miklau, who is an assistant professor of Computer Science at the University of Massachusetts at Amherst. Gerome is the recipient of the 2006 ACM SIGMOD Dissertation Award, for his dissertation entitled Confidentiality and Integrity in Distributed Data Exchange. *His PhD is from the University of Washington, where his advisor was Dan Suciu. So, Gerome, welcome!*

I'm delighted to say that Gerome is the first *recipient of the SIGMOD Dissertation Award. Probably many of our readers haven't heard about this new award, which will be given every year to recognize excellent research in a database dissertation that was submitted during the previous year to a computer science department anywhere in the world. Our runners-up for the award this year were Marcelo Arenas from the Pontificia Universidad Católica de Chile, and Yanlei Diao, who is also at the University of Massachusetts at Amherst.*

So, Gerome, what is the thesis of your thesis?

My thesis is a response to the fact that, as a result of successes in the database community, there has been an explosion in the collection of data, and an explosion in the exchange and sharing of data. As a result, there are new challenges in balancing the need to share and exchange data and the need to protect sensitive data.

What do you see as the threats?

The general high level threat is that sensitive data will be misused. That misuse can cause harm, either by violating our personal privacy, by disclosing facts about us that we don't want others to know; or through incorrect data introduced into records about us, such as medical records and credit reports.

How does your thesis address those issues?

The goal of my thesis is to provide some theoretical techniques, and some practical tools. On the theoretical side, I address the challenge of understanding information disclosure in databases. The challenge here is to understand, when we decide to share data by publishing a view, what exactly that view may reveal about other facts that we would like to protect. We recognized that there was a lack of good definitions for understanding that disclosure precisely. We proposed a new definition, and analyzed it theoretically, providing complexity results and decision procedures and so forth. So that is the theoretical side of the work.

The more practical side of the work has to do with exchanging data, beyond a trusted domain. In a trusted domain, you can use your own systems to negotiate access to data. When you publish data beyond that trusted domain, you can't rely on others to respect the policies that you want respected. So, we rely on cryptography to enforce access control. We devised a framework for publishing data and passively enforcing access control with cryptography.

Is it a solution based on keys that you give to the authorized people?

Yes, that's right. The goal is to move from a model where you would publish one version of the data for each of the authorized recipients (of which there could be hundreds or thousands), and to instead construct a single version of the data, protected and partially encrypted in a somewhat complicated way, and publish that single version to a web server. The recipients can all take that data, but require appropriate cryptographic keys to process the data.

Are you using some sort of group cryptography scheme? How does the crypto part work?

It is fairly straightforward. There is some connection to secret sharing schemes in cryptography. I think the main insight is the view; we work with XML data, we view the data as a tree, and we annotate the tree with key expressions. So access to the data must be from traversing the tree, and you must satisfy the key expressions by possessing the appropriate keys. There is a precise semantics to access in this model, which makes it easy to analyze and easy to work with.

Do the keys intuitively express group membership, or some quality of the recipient, or are they related to the recipient's identity?

They are related to the recipient's identity and they are derived from the access control policy. The way you would construct this partially encrypted data is first to state an access control policy over the XML data in a declarative way. Then in the processing phase, the keys are actually generated automatically. An aspect of the keys corresponds to the identity of the recipients, and another aspect corresponds to the positions in the data that they have access to.

Does the policy list the identities of the authorized people, or their properties or characteristics?

Identities, but it could also support properties.

What do you see as the major open problems in the area where you were working?

There was one really challenging part of this problem that it was not possible for me to address in my work. The challenge was to construct a proof that the encrypted data produced by our process has had the encryption functions applied correctly and that the only information an attacker could get is the information implied by our precise access semantics, which is more at a logical level. This kind of problem has been addressed by others in other settings. Martin Abadi and Bogdan Warinschi wrote a follow-on paper that appeared in PODS last year and proved the security of our constructions. I view that as an important piece of follow-on work. It was really beyond my scope, as I am not a cryptographer.

Where do you see this work being applied?

With regard to the disclosure work, our results provide a theoretical ideal of security. We have done some follow-on work to make these notions more practical in nature. Some recent work that appeared at this conference (SIGMOD 2006) has improved our complexity bounds for special cases. That improves the practicality of our work as well.

I think that there is a great opportunity to make progress and achieve this balance of sharing data and also protecting it. We need good measures to help us understand what is disclosed when you publish an anonymized data set, and that is something that has trailed behind techniques for *producing* anonymized data sets. In addition, current techniques do not handle many common data sets. There are pressing real-world problems in this area. One example is the case of networked data sets. As you may have heard, recently the National Security Agency has collected and analyzed records of phone calls made by US residents. Apparently the NSA does not anonymize those records, but they probably should. If they did, they would have a massive graph of relationships that they would like to study for patterns. So an important open question is, can you publish a graph like that about entities and their relationships and provide a resistance to re-identification of individuals. There are some connections to the techniques of k-anonymity, but k-anonymity is intended for tables that represent entities, not relationships between entities.

You are the only assistant professor that I have ever interviewed, so you have a unique perspective to offer our readers. What do you know now that you wish you had known as a graduate student?

During my first year as an assistant professor I've confirmed that academia is great. You can never be sure until you get there! I am not sure what I would have done differently had I known that earlier, but I think I would have been able to act with more certainty. Maybe I would have finished my PhD sooner, had I known that this job would suit me to the degree it does, how exciting it is to work with students and come up with new ideas, and to teach.

If you magically had enough extra time to do one additional thing at work that you are not doing now, what would it be?

I have a small stack of books on my desk about various aspects of privacy in society, written by non-technical people. These books talk about the legal and philosophical foundations of privacy, as well as practicalities about how data is collected and who legally owns data. I feel strongly that it would benefit

my research if I had time to read more of these books. Understanding the societal background and history of privacy would provide inspiration for my research.

If you could change one thing about yourself as a computer science researcher, what would it be?

A stronger background in statistics would benefit me, I think. My background is more in logic and discrete math, and certainly there seems to be a trend towards the increased importance of statistical modeling in computer science.

Thank you very much for talking with me today.

Thank you.