Taking the Sting Out of Big Data

Dr Florian Kerschbaum

Scientia

TAKING THE STING OUT OF BIG DATA

To many people, the idea of Big Data is inseparable from the fear of a loss of privacy. Since many modern companies must collect vast amounts of personal information in order to operate effectively, this concern is now growing globally. **Dr Florian Kerschbaum** and his colleagues at the University of Waterloo aim to address the issues presented by this reality, through sophisticated new techniques that ensure all-encompassing privacy for a company's users. If the team's methods become more widely adopted, they could reassure many people that their sensitive information is being handled properly.

Data Flow

Many of the services we use in our everyday lives involve a flow of data from us, as users, to private companies, which typically store the information in their servers. Yet this flow is much more than a simple transaction: from the transfer of data through third parties, to the categorisation of the information it contains, the overall field of Big Data involves several distinct and important steps, before data can be effectively managed and manipulated by users.

Firstly, data is acquired as user information and is transferred to a server. Then it is prepared, as new data in the server is matched up with the user who provided it. At the management stage, server data is processed as it is subjected to operations by users; and through analysis and modelling, they can more easily access the information they need. Finally, use and inference involve categorising the data, making it easier for users to deal with. As a whole, this data pipeline runs so smoothly that we barely notice it. However, its operation still faces one particularly serious challenge.

Protection at Every Stage

In recent years, the issue of our right to privacy over the information that Big Data companies collect from us has gained widespread attention. It stems from the fact that much of this information is so personal and sensitive, that significant amounts of damage could be done if it fell into the wrong hands. Yet just as the data pipeline is intricately complex, the measures that would be required to ensure sufficient user privacy at each of the steps it involves are difficult to define.

In their research, Dr Florian Kerschbaum and his colleagues at the University of Waterloo have developed wide-ranging measures to combat this issue – each based around the concept that full user privacy is a fundamental requirement. In collaboration with SAP, a software company in Germany that relies on Big Data, and the Royal Bank of Canada, Canada's largest financial institution, the researchers have now applied their new techniques to several different situations, representing each stage of the data pipeline. As Dr Kerschbaum puts it, his work aims to 'take the sting' out of Big Data once and for all.

Acquisition

When companies collect information about their users, users can't necessarily trust the company to anonymise their data. To maintain privacy, therefore, the data must be transmitted in a distorted form, meaning that the company cannot read the value of a single user.

However, the company should be able to obtain aggregate data about a collection of its users. This can be achieved by a large distortion that cancels out over the aggregation, but if any user fails to transmit, the aggregate is unusable. Another way that this can be achieved is through a small distortion that diminishes over the aggregation, but many users are needed for an accurate aggregation.

Dr Kerschbaum and his team present a new computational technique based on statistics and cryptography, which can quickly and accurately gather private information through untrusted parties. These parties, which can be a few users or distributed servers, aggregate first and then apply a very small distortion. 'This involves collecting data from users with privacy and without a trusted collector,' Dr Kerschbaum summarises. 'Google and Apple collect user data with privacy, but their methods do not scale to lower than millions of users. We can do it for thousands of users.'

Preparation

Once data has been collected, another challenge arises: since multiple companies can possess information about single users, they often need to work together to identify matches in their records of individual people across multiple datasets. As they do this, two companies comparing records must use computer programs which can make these connections automatically, but this creates a significant privacy concern. If an individual has disclosed some information to one of the companies, but not the other, it would be a breach of their privacy for that data to be shared between both companies.

Again, there have been previous computational techniques that combat this issue, but they have typically come at the cost of reducing the accuracy and time efficiency of the matches. Now, Dr Kerschbaum and his colleagues have developed a new approach that offers full security for all parties involved, while maintaining high performance. 'This involves linking data across parties without disclosing it – which is done very often, but without data protection,' he explains. 'We hope to be able to scale it with data protection, and I am now thinking of founding a start-up on this idea.'

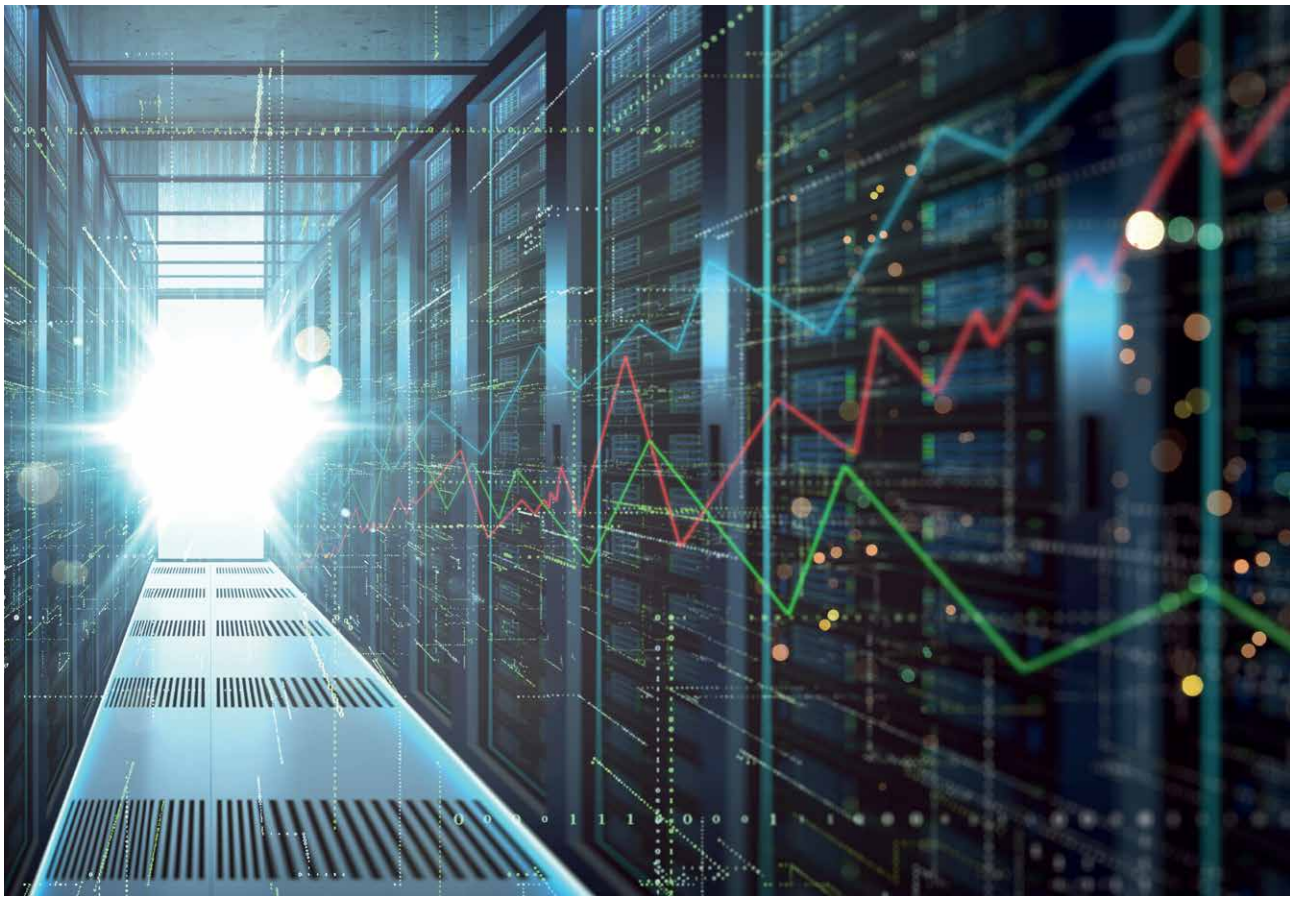
Management

When we use the services provided by a Big Data company, we are essentially commanding operations to be carried out on our own sensitive information while it is being held on their servers. However, as companies execute these operations for us, they must continually read and write new sensitive information, creating a further risk to user security. To alleviate this risk, companies must use databases that aren't able to 'see' the data they read and write. This is particularly important for cloud databases, in which users can store vast amounts of information on company servers in an encrypted form.

To achieve such a computationally demanding task, Dr Kerschbaum has developed an algorithm that is 'oblivious' to the data it operates on, which can account for realistic hardware and security capabilities of both users and companies. The structure of this algorithm makes it highly secure in various settings, while maintaining simplicity, operational efficiency, and high performance. 'Here, we focus on using secure hardware to process encrypted data in the cloud,' Dr Kerschbaum describes. 'This allows us to create a database that doesn't see what it is processing.'

Analysis and Modelling

Computers can exploit some useful techniques to help us analyse, organise, and retrieve our documents, which are key components of user-friendly interfaces. Sometimes, however, these same techniques can be used for more sinister purposes. There are many situations in which document writers might prefer to remain anonymous – from giving negative customer feedback, to criticising a restrictive government. Yet by



determining how frequently particular letter combinations are used in a document, and then checking these frequencies against previous reference texts, groups can often identify the authors of anonymous documents simply by the terms or phrases they use on a regular basis.

Dr Kerschbaum and his team have introduced an approach that artificially alters the terms used in original documents, altering their term frequencies, while retaining their meaning. This makes it far harder for hostile groups to use techniques to determine the authors of anonymous documents, without diminishing the accuracy of such otherwise useful identification tools. As Dr Kerschbaum summarises, 'this involves building a model of a textual database that protects private information, but also allows non-sensitive inferences, such as the identification of which news group a post belongs to.'

Use and Inference

When data has been gathered onto a server, it often needs to be categorised in order to make it easier to access and use. This can be done using tools named 'decision trees'. Based on the capabilities of artificial intelligence, these models are now used in applications ranging from spam filtering to healthcare. However, difficulties arise when servers use decision trees to classify user information, while maintaining their privacy. This requires the end results of decisions to be revealed to the user, while nothing is revealed to the server after the process has finished.

Currently, the existing techniques for doing this require many steps, making them computationally inefficient. Dr Kerschbaum's team has combated the issue by restructuring the decision tree so that it only shares the outcome of one branching point with the next point – enabling privacy to be maintained throughout the operation.

The team has now used the technique to categorise a large, real-world dataset – significantly reducing the computation time compared with previous approaches. 'The class of a data sample held at a client is inferred, using a decision tree held at a server,' Dr Kerschbaum explains. 'This is done without disclosing either the data sample to the server or the decision tree to the client using computation on encrypted data.'

Removing the Sting

Within the current technological landscape, it now seems increasingly inevitable that Big Data will play an ever-growing role in our lives going into the future. While many people are justifiably fearful over what this reality might entail, Dr Kerschbaum believes that his team's techniques could ensure that our personal information can never be compromised in such a world. By viewing privacy as a central requirement in every step of the data pipeline, the sting of Big Data could finally be removed, ensuring a more secure future for everyone.

Meet the researcher

Dr Florian Kerschbaum
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario
Canada

Dr Florian Kerschbaum completed his Doctorate in Computer Science at Karlsruhe Institute of Technology in 2010. He has worked as an Associate Professor at the at the University of Waterloo since 2017, has been Director of the Waterloo Cybersecurity and Privacy Institute since 2018 and was named the NSERC/RBC Chair in Data Security in 2019. Dr Kerschbaum's main research interests include data security and privacy in data management and machine learning. He also applies his findings to real-world systems, including databases and supply chain management systems. In 2019, he received the Outstanding Young Computer Science Researcher Award from CS-Can, and was also recognised as a Distinguished Scientist by the Association for Computing Machinery.

CONTACT

E: florian.kerschbaum@uwaterloo.ca
W: https://cs.uwaterloo.ca/~fkerschb/

FUNDING

NSERC
Royal Bank of Canada
SAP

FURTHER READING

B Khurram, F Kerschbaum, SFour: A Protocol for Cryptographically Secure Record Linkage at Scale, In The 36th International IEEE Conference on Data Engineering (ICDE), 2020, 277–288.

S Krastnikov, F Kerschbaum, D Stebila, Efficient Oblivious Database Joins, Proceedings of the VLDB Endowment (PVLDB), 2020.

J Böhler, F Kerschbaum, Secure Multi-party Computation of Differentially Private Median, In 29th USENIX Security Symposium (USENIX SECURITY), 2020.

B Weggenmann, F Kerschbaum, SynTF: Synthetic and Differentially Private Term Frequency Vectors for Privacy-preserving Text Mining, In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR), 2018, 305–314.

UNIVERSITY OF WATERLOO