PROFESSOR MURAT KANTARCIOGLU

Securing 'big' data

Professor Murat Kantarcioglu is immersed in studies on the secure manipulation of large data. Here, he talks about the importance of his work in protecting the privacy of individuals

To begin, could you describe your current research and what you hope to achieve? From what context did this project emerge?

I believe that capturing, storing and mining 'big data' may create significant value in many industries, ranging from healthcare to location-based services. However, several important issues need to be addressed to capture the full potential of big data. I believe security and privacy, as well as incentives to share data, are critical. To address these challenges, we have been carrying out research for several years on creating technologies that can efficiently extract useful information from any data without sacrificing privacy or security. Specifically, we have obtained some highly innovative results on ensuring security and privacy while sharing, linking, mining and storing large quantities of data. We also proposed novel interdisciplinary techniques gleaned from data mining, cryptography and game theory, to address big data challenges and have developed several software tools which are being utilised in practical applications. In addition, we made significant contributions to developing privacy-preserving techniques and algorithms for distributed data mining, record linkage and genomic data storage.

What experiences led you to become interested in this field of study?

I have always found the issues surrounding security and privacy in managing data very interesting and, subsequently, this – and the surrounding research challenges – were the focus of my PhD studies. This interest has continued to grow and I have worked extensively on questions concerning data security and privacy.

Why have organisations become increasingly dependent on integrating secure methods of sharing private data to execute critical tasks?

The ability to communicate and share data has many benefits as well as obvious security risks, and the idea of an omniscient data source carries great value. For example, in the US Food and Drug Administration's (FDA) Mini-Sentinel Program, healthcare organisations led by the FDA implemented a distributed data analytics protocol to answer questions related to drug effectiveness. Similarly, different pharmaceutical companies are sharing clinical trial data to reduce drug development costs for diseases like Alzheimer's.

On the other hand, an omniscient data source is open to misuse, such as sensitive data disclosure. To prevent this, there has been a recent surge in laws mandating the protection of confidential data. However, this protection comes at a real cost, through added security expenditure and penalties, costs associated with disclosure and the barriers that security measures sometimes present to those attempting to productively use the information. For example, CardSystems was terminated by Visa and American Express after having credit card information stolen. This indicates the need for securely sharing and mining distributed data without disclosing anything other than the necessary results.

Where do you see your research progressing over the coming years? Do you have any other plans in the pipeline?

We are currently working on many different projects. For example, in our recently funded National Science Foundation project, we are working on addressing various security and privacy issues in managing provenance information. In addition, as a part of a National Institutes of Health project, we are collaborating with colleagues from Vanderbilt Medical School to build a framework for understanding and managing risks in disseminating genomic data.

Another area that we are very interested in is developing techniques for securely managing data that is stored in public clouds. In this Air Force-funded project, we are using a novel combination of access control, query optimisation and encryption to offer practical solutions for cloud data security solutions.

Finally, we are very interested in using data analytics for detecting adversarial activities. As a part of a recent Army Research Office grant, we are applying game theoretic ideas for building novel data mining algorithms that can have long-term success in detecting malicious events that are important in many different application domains: such as fraud detection, spam detection and homeland security.

If you were to single out your greatest achievement and sole challenge in high-assurance data, what would they be?

We are absolutely delighted that our proposed techniques can potentially achieve two orders of magnitude improvement in terms of performance for typical privacy-preserving distributed analytics tasks. On the other hand, packaging all those novel techniques into easy-to-use software that can be adopted by practitioners is quite a challenge.

Privacy-preserving distributed data analytics

Ongoing studies at the **University of Texas at Dallas** are facilitating novel computational methods for reconciling organisations' needs to analyse large amounts of data with the importance of protecting the privacy of the individual

AS COMPUTER AND web technologies continue to become more powerful, and an increasing amount of potentially compromising data is held about individuals by government and business, the need for secure sharing of private data is more critical than ever. Although techniques exist for providing privacy-preserving data exchange analytics, each presents a unique set of strengths and weaknesses, and none has emerged as a silver bullet. In response, a series of studies at the University of Texas at Dallas, led by Professor Murat Kantarcioglu, is attempting to find a solution and create an efficient, secure and cost-effective means of linking, mining and storing large amounts of data quickly and easily.

CURRENT METHODS

In recent years, several approaches for secure data use have come into circulation. These have revolved around either a secure multi-party computation (SMC) technique, or one based on sanitisation/anonymisation of data. For a party attempting to analyse data, the former relies on learning only the final result of their query; after which no further information can be inferred about the original data other than what can be inferred based on their own input and this final result.

Kantarcioglu clarifies with a simple analogy: "In famous Yao's millionaire problem, two millionaires, Alice (A) and Bob (B), want to learn who is richer without disclosing their actual wealth to each other". Translated to SMC, the goal is to solve this inequality without compromising either A or B's true value. In this method, privacy protection guarantees are provided through the use of cryptographic tools but remain computationally expensive procedures for many applications. Whilst they enable trade-offs between privacy protection and efficiency by allowing users to select their desired size of cryptographic key, they lack provisions for a compromise between privacy protection and accuracy. They are also ill-equipped for up-scaling for use with very large amounts of data.

Sanitisation/anonymisation of data techniques, on the other hand, enables an organisation to reveal privacy sensitive data by distorting it. Proponents of this approach include Census Bureaus, which reveal sanitised versions of private information on demographics, as Kantarcioglu elaborates: "Over the years, a plethora of methods have been proposed to perturb and sanitise data to protect individual privacy; in these methods, higher levels of protection typically translate into further deviation from original data and, consequently, produce less accurate results".

Neither method is fault free. Medical researchers who need to integrate datasets from two separate hospitals, to provide just one example, would require a number of computationally-expensive operations with existing SMC, while a sanitised dataset – which the hospital could safely disclose only after removing certain identifiers – would give incomplete or inaccurate results.

A SYNERGISTIC MODEL

In order to combat these downfalls, Kantarcioglu and his colleagues are developing an approach which integrates the positive aspects of each model. By looking at sanitised datasets, it is possible for users to see that certain records do no match. For example, if two patient records have different age values, it is clear that they cannot belong to the same person. Using this as a starting point, Kantarcioglu has developed a novel technique which can be used to record pairs of records which could potentially match and thus belong to one person. The results demonstrated by his collaborative 2012 study showed that this assimilation could reduce the running time of the process to just 1 per cent of traditional SMC techniques, whilst maintaining accuracy.

> Kantarcioglu's approaches could be adopted in areas such as homeland security and intelligence, enabling large amounts of data to be analysed without sacrificing security and privacy

The researchers have principally been looking at computation methods which address capture, linkage, mining and storage requirements of big data. So far they have been successful in: developing a technique for linking and mining data from two separate sources, whilst managing to maintain the privacy requirements of each; delivering an initial prototype to more

PRIVACY-PRESERVING DISTRIBUTED DATA ANALYTICS

OBJECTIVES

Develop efficient technologies that can efficiently extract useful information from distributed data sources without sacrificing privacy or security.

KEY COLLABORATORS

Elisa Bertino, Purdue University, USA

Ali Inan, Isik University, Turkey

FUNDING

National Science Foundation (NSF) Grant Career award no's CNS-0845803, CNS-0964350 and CNS-1016343

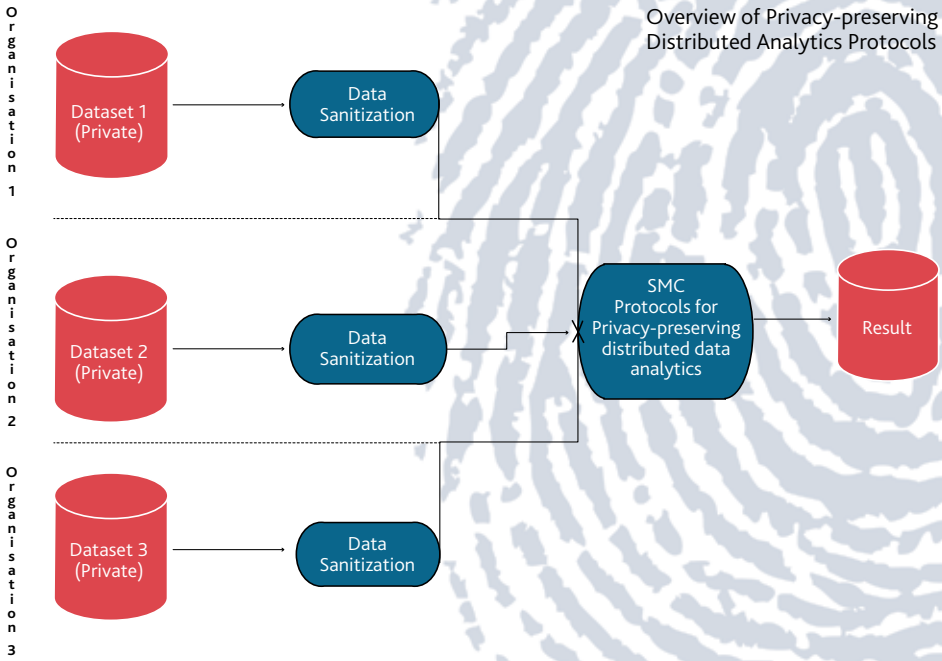
CONTACT

Professor Murat Kantarcioglu
Principal Investigator

Computer Science
University of Texas at Dallas
2601 North Floyd Road
Richardson
Texas
75083
USA

T +1 972 883 6616
E muratk@utdallas.edu

MURAT KANTARCIOGLU is an Associate Professor of Computer Science and Director of the UTD Data Security and Privacy Lab at the University of Texas at Dallas. He is a recipient of the NSF CAREER award and the Purdue University Center for Education and Research in Information Assurance and Security (CERIAS) Diamond Award for Academic excellence. He has published over 100 papers related to data security and privacy in peer reviewed journals and conferences. His research has been supported by grants from the National Science Foundation, Air Force Office of Scientific Research, Office of Naval Research, National Security Agency, Army Research Office and National Institutes of Health. His work has been featured by the Boston Globe, ABC News and other media outlets, and he has received two best paper awards.



Overview of Privacy-preserving Distributed Analytics Protocols

accurately and safely query encrypted data within databases; and exploring query optimisation issues with regards to encrypted data and the associated risks of data being disclosed to undesirable parties, with a system which operates using existing cryptographic hardware.

Kantarcioglu believes that the techniques he and fellow researchers have been developing could decrease the cost of carrying out data analytics. "Our work could potentially have a direct economic impact, opening the way for new applications in, for example, eHealth and eGovernment applications, that are at present considered unfeasible due to the lack of necessary privacy preserving solutions for large datasets," he enthuses.

CLOUD COMPUTING

Additional studies by the Kantarcioglu lab have sought to find more secure methods for storing data in the cloud, something which has grown in popularity enormously over the last few years. Because of advantages such as its speed, convenience and reliability, there are currently huge amounts of data stored in the cloud by a wide range of businesses and other organisations. However, there are ongoing concerns about the privacy and security of this method of data storage and, in order to reduce some of these issues, data is often outsourced in an encrypted form. This technique protects data from being accessed illegally, but has the disadvantage of muddying some basic functions such as the ability to carry out accurate searches. Several potential approaches for searching encrypted data have been suggested, but very few are able to carry out similarity matching efficiently for big data, instead offering the facility to execute exact query matching. This is a major problem for many real-world applications, as users increasingly need a flexible approach to search through data held in the cloud and perform similarity testing for large amounts of data.

Kantarcioglu proposes a more efficient system that enables nearest neighbour search over encrypted data. As part of their study into this resolution, the researchers demonstrated the system's efficiency on a real dataset, suggesting its widespread use could be established for a range of applications, whilst ensuring the protection of sensitive data.

LOOKING AHEAD

Kantarcioglu describes that the approach his lab has been advocating is orthogonal to new SMC or sanitisation techniques, and are therefore in a position to incorporate efforts in either method to improve functioning. Making its potential application even more diverse and widespread, he elucidates that, "the ultimate goal of our project has always been to preserve the individual's privacy, rather than preventing the disclosure of any sensitive information related to the data holder".

In addition to health research, Kantarcioglu's approaches could be adopted in areas such as homeland security and intelligence, enabling large amounts of data to be analysed without sacrificing security and privacy. As methods of data storage change and evolve, and the amounts being stored continue to grow, advances such as those made by Kantarcioglu are essential in order to guarantee the protection and privacy of the individual.

UTD

UT DALLAS
The University of Texas at Dallas