MIT Big Data Initiative at CSAIL Member Workshop #2: Big Data Privacy *Exploring the Future Role of Technology in Protecting Privacy* June 19, 2013	bigdata @CSAIL

WORKSHOP REPORT

BIG DATA AND PRIVACY

"One thing should be clear, even though we live in a world in which we share personal information more freely than in the past, we must reject the conclusion that privacy is an outmoded value. It has been at the heart of our democracy from its inception, and we need it now more than ever"
- President Barak Obama

"Technology in this world is moving faster than government or law can keep up. It's moving faster I would argue than you can keep up. You should be asking the question of what are your rights and who owns your data."
- Gus Hunt, CTO CIA, speaking at GigaOm 2013

Big Data promises a better world. A world where data will be used to make better decisions, from how we invest money to how we manage our healthcare to how we educate our children and manage our cities and resources. These changes are enabled by a proliferation of new technologies and tools that have the ability to measure, monitor, record, combine and query all kinds of data about us and the world around us -- but how will that data get used and for what purpose? Who owns the data? How do we assure accountability for misuse?

Just as Big Data lays out many promises, it lays out many questions and challenges when it comes to privacy. We must think carefully about the role of technology and how we design and engineer next generation systems to appropriately protect and manage privacy, in particular within the context of how policy and laws are developed to protect personal privacy. Decisions about how to address privacy in big data systems will impact almost everyone as we push to make more data open and available inside organizations and publicly. Governments around the world are pushing themselves and private companies to make data transparent and accessible. Some of this will be personal data. We will need new tools and technologies for analysis, for anonymizing data, for running queries over encrypted data, for auditing and tracking information, and for managing and sharing our own personal data in the future. Because issues of data privacy will be relevant across so many aspects of our life, including banking, insurance, medical, public health, government, etc, we believe it is important to collectively address major challenges managing data privacy in a big data world.

The goal of this workshop was to bring together a select group of thought leaders, from academia, industry and government, to focus on the future of Big Data and some of the unique issues and challenges around data privacy. Our aim is to think longer term (5 years +) and better understand and help define the role of technology in protecting and managing privacy particularly when large and diverse data sets are collected and combined. We will use the workshop to collectively articulate major challenges and begin to layout a roadmap for future research and technology needs.

Workshop Organizers: Daniel Weitzner, Sam Madden, Elizabeth Bruce, CSAIL, MIT

We gratefully acknowledge all of our contributors, including speakers, panelists, writers, reviewers and breakout groups.

This workshop was supported by the MIT Big Data Initiative at CSAIL and by a Grant from The Alfred P. Sloan Foundation.

bigdata @CSAIL

ALFRED P. SLOAN FOUNDATION

Table of Contents

Workshop Summary **4**

Introduction and Workshop Objectives **7**
Daniel Weitzner, Sam Madden, Elizabeth Bruce - CSAIL, MIT 7

Defining "Privacy" in a Big Data World **7**
David Vladeck - Georgetown University Law Center 7

Session I: Understanding User Perspectives - the Value of Data and the Issues of Privacy **11**
Big Data, Systemic Risk, and Privacy-Preserving Risk Measurement **11**
Andrew Lo - Sloan School of Management, MIT 11
Big Data and Privacy **13**
Maritza Johnson - Technical Privacy Manager, Facebook 13
Big Data Privacy **14**
Robert Zandoli - SVP and Global Chief Information Security Officer, AIG 14
Big Data: New Oil of the Internet **15**
Alex (Sandy) Pentland - Media Lab, MIT 15

Session II: Approaches to Managing Data Privacy – Systems, Tools, and Theory **18**
Differential Privacy **18**
Kobbi Nissim - Visiting Scholar, Harvard CRCS and Faculty at Ben-Gurion University* 18
No Free Lunch and the Pufferfish Approach to Privacy **22**
Ashwin Machanavajjhala - Duke University 22
Accountable Systems **24**
Lalana Kagal - Distributed Information Group, CSAIL, MIT 24
De-identification Methods – Anonymization of Patient Spatial Data **26**
Shannon Wieland - Visiting Scholar, MIT 26
Encrypted Query Processing **28**
Raluca Ada Popa - CSAIL, MIT 28
Encrypted Databases – Private Information Retrieval Using Secure Hardware **29**
Srini Devadas - CSAIL, MIT 29

APPENDIX: Breakout Groups ..... Error! Bookmark not defined.

Workshop Summary

Recent interest in "Big Data" arises from the convergence of several major technology trends, including the pervasiveness of data (the digitization of information via web, social media, medical records, etc); the availability of inexpensive sensors, including location based devices; and the decreasing cost of computation and data storage. The ability to collect, analyze, and report on massive amounts of data is now possible for institutions of all sizes. This leads us to new challenges related to managing privacy.

The goal of this Workshop was to invite industry, government and academic leaders to discuss these complex challenges and potential solutions. In BigData@CSAIL, we were motivated to host this Workshop because Big Data has such potential to improve lives, but is also fraught with ethical, societal, and technical issues related to mitigating the privacy risks of data.

Consider three Big Data use cases:

1) **Social Sciences:** Social media services, like Facebook, Google, Twitter, now generate data sets on a global scale that provide new means for understanding trends, sentiment, and opinion across large populations of society, and for organizing groups with related interests. On the other hand, these services record minute details of individual lives that have the potential to be misused in a variety of ways.

2) **Public Health:** Big Data can play a role in helping predict epidemics, and, with early detection the opportunity to minimize impact, whether its the flu, outbreaks of cholera, or other potentially devastating diseases. Despite these advantages, much of this data is sensitive personal information that needs to be protected.

3) **Financial Markets:** Economic data can be used to better understand economic risks at a national level which will guide policy makers and regulators, and help to better manage systemic risk in the financial sector.

At the workshop, a number of general privacy risks in applications like these were discussed, including:

1) **Indiscriminate collection of data.** Big Data encourages the indiscriminate collection and over collection of data. Data now holds the promise of discovery, the opportunity to develop new products and services, yet this runs counter to the first rule of good data hygiene: do not collect and hold on to data that is not needed, especially personal information. Do we throw out this rule in the age of Big Data and embrace ubiquitous data collection? Or do we risk limiting the potential of big data by limiting the data we collect?

2) **Discrimination by algorithm.** Data may be used to make determinations about individuals as if correlation were a reasonable proxy for causation.

3) **Obtaining informed consent.** Today, data is collected about individuals, from the medical treatment they receive, to goods and services they buy, to websites they search, phone calls they

make, etc. However, much of this data is collected without consumer's consent, and rarely do consumers have the opportunity to consider and consent to the aggregation of this data, or consent to secondary uses that are not even contemplated at the time consent is sought.

4) **Allowing people to manage and control their own data.** Often people want to share personal information with certain groups of people, friends and/or family meaning privacy controls are not simply binary (public vs. private). As the amount of personal data we generate grows, how do we effectively design mechanisms that enable users to have more fine-grained control of what they share, when, and with whom? How design for transparency and control in the context of multiple data sharing relationships?

5) **De-identification alone is not a solution.** There are many examples where efforts to render Big Data safe by deleting identifying information can be undermined by techniques permitting re-identification of the data, often through linking to external data sets..

6) **Data breaches may lead to catastrophic harm.** A malicious hack or inadvertent mistake -- could cause incalculable harm, especially as the scale of data collection and aggregation increase.

New technologies will enable better ways to manage, protect and secure data in the future, from new de-identification techniques to advanced encryption methods. But there is no silver bullet that solves all privacy challenges: each method has its advantages and disadvantages. Additionally, different applications require different levels of privacy, depending on the type of data and how it is used. In this Workshop, we discussed a number of methods for managing privacy including:

>> Open Personal Data Store (open PDS) [Pentland]
>> Secure Multi-Party Computation (SMPC) [Lo]
>> Differential Privacy [Nissim]
>> Pufferfish Approach [Machanavajjhala]
>> Accountable Systems [Kagal]
>> De-identification: the LP method [Wieland]
>> CryptdB: Encypted Query Processing [Popa]
>> Private Information Retrieval (PIR) using Secure Hardware [Devadas]

Brief introductions to each of these approaches are included in this report. Solutions to managing Privacy are not purely technical, as Big Data raises serious legal, policy and ethical questions that remain unanswered. We must develop the right mix of technology and public policy approaches. Existing privacy policy frameworks are based on basic principles around preventing unfair or deceptive acts and practices that need to be extended to Big Data.

During the Workshop some suggestions were made for best practices inside organizations including:

1) Create new roles inside organizations responsible for assessing privacy risks
2) During initial phase of a big data project, do a privacy risk assessment
3) Privacy by design -- provide transparency to users from the beginning; be clear about the purpose of data collection and potential for future use
4) Do regular assessments of the data storage environment and maintain knowledge of usage patterns
5) Keep security controls close to the data and use encryption for all types of personal data (both static and in transit).

Next Steps

As a follow on to this Workshop, we intend to form a Big Data Privacy Working Group where we can continue this dialogue, to explore and learn about different technology solutions to managing privacy, to understand policy requirements, and frame future research needs.

Introduction and Workshop Objectives

Daniel Weitzner, Sam Madden, Elizabeth Bruce - CSAIL, MIT

The mission of the MIT Big Data Initiative at CSAIL is to collaborate with industry and government to identify and develop new technologies that are needed to solve the next generation of data challenges; this will involve the development of scalable and reusable software platforms, algorithms, interfaces, and visualizations that are designed to deal with data that is high-volume, high-rate, or high-complexity, or some combination of these characteristics.

In this second workshop in the Big Data series, the focus is Privacy in the age of Big Data; the goal was to bring thought leaders from academia, business, and government together to explore some of the unique issues around privacy in a Big Data environment. One of the challenges lies in promoting a better understanding of the role of technology in protecting and managing privacy; the work done within the initiative may help to define that role, particularly in cases where large and diverse data sets are collected and combined for business and government uses.

Defining "Privacy" in a Big Data World

David Vladeck - Georgetown University Law Center

Professor Vladeck served as former Director of the US Federal Trade Commission, Bureau of Consumer Protection

The Federal Trade Commission's (FTC) mandate from Congress is to prevent unfair or deceptive acts and practices in or affecting interstate commerce. It does not have jurisdiction over insurers or depository institutions, but most other business entities are under FTC jurisdiction. In recent years, the FTC has been involved extensively in issues relating to privacy. The FTC has taken action against Google, Facebook, My Space, and others, alleging that these companies deceived consumers by making commitments to limit data sharing practices and then made changes to those practices that resulted in sharing sensitive personal information with third parties without consumer knowledge and consent. As a result, each of those companies is now under a 20-year consent decree with the FTC, which requires the creation of elaborate privacy procedures and biannual audits by outsiders that are carefully reviewed by the FTC.

The agency also has authority over unfair practices, which are defined by statute as practices which harm or threaten to harm consumers, that consumers cannot reasonably avoid, and where the harm outweighs the benefits. The FTC has used this authority to bring action against companies that do not provide reasonable security measures to protect sensitive private information. The FTC also enforces the Fair Credit Reporting Act, which was enacted when Congress took notice that growing databases in private hands could be used in ways that were invisible to consumers, but could cause them harm. The Act sets out rules for companies that use information relating to creditworthiness, insurance underwriting, and suitability for employment. Historically the FTC has had great interest in the collection and use of data and has been looking at the challenges posed by big data for quite awhile. The Fair Credit Reporting Act (FCRA) is in fact an example of 'big data' oriented consumer protection before it's time. Those credit bureaus were the 'big data' repositories of their day. The rules set out in the FCRA find a good balance

between enabling intensive analysis of large amounts of data while at the same time assuring that consumers are not unfairly discriminated against in the process.

Big Data Risks and Challenges

The FTC's role in Big Data is to make sure consumer privacy is safeguarded. Given the risks posed by the greater and greater volume of data being aggregated, there are some serious risks that should be addressed. It is useful to catalog the most serious risks and evaluate how those risks can be mitigated or avoided.

One of the most serious risks posed by big data is that it explicitly encourages the indiscriminate collection and over collection of data. Data is now seen as the raw material of innovation and holds the promise of discovery, development of new products and services, and the expansion of knowledge. However, the most basic risk is that the drive to collect and amalgamate data runs directly counter to the first commandment of good data hygiene: "Thou shall not collect and hold on to data that is not needed, especially personal information."

Increasingly big data is being used to make determinations about individuals as if correlation were a reasonable proxy for causation. I am worried that consumers will suffer what it calls "discrimination by algorithm", or a kind of data determinism, because correlation will lead individuals to be categorized, not merely because of their own actions, but because of general trends that will be seen as sufficiently robust to draw conclusions about their individual behavior, often with no process for mitigation if the conclusion is wrong.

Big data is assembled from little data and becomes big only when sufficient amounts of little data are compiled into some kind of database. This little data often reflects the activities of identifiable human beings. The medical treatment they receive, the goods and services that they buy, the websites they search, the phone calls they make, their whereabouts constantly tracked by their smart phones and the geo location features of the apps they use, their interactions with family members, their political activities, their sexual preferences, and on and on, that is the little data that now comprises big data databases. Today, much of this data is collected without consumer's consent, which is why the FTC has advocated so strongly in favor of the creation of a robust do not track system. But even when consent is obtained at the front end, consumers rarely if ever are given the opportunity to consider and consent to the aggregation of this data, let alone consent to secondary uses that are not contemplated at the time consent is sought.

Add to that the fact that combining data sets can be messy because data needs to be processed and integrated as it is combined; along the way, the quality of the information may be compromised. Predictions based on imperfect data will be imperfect, and companies using big data need to be clear about the limitations of big data. If the purpose of collecting big data is to discover knowledge, to discern general trends, or to deal with subjects that have nothing to do with individuals, (e.g. predicting the weather, safeguarding networks, making systems work better) then maybe some of these concerns don't apply. But to the extent that big data may be used to make determinations that relate to individuals, these concerns apply with full force.

For example, is it alright for consumers to be offered less favorable credit, not because of their financial condition, but because of their behavior, or the behavior imputed by them to a computer program on social networking sites, or because of their browsing behavior, and that data somehow suggests that they may not be perfect credit risks?

One standard response to this dystopian vision is that de-identification will protect individuals and will make it difficult for big data to be used to make reductive decisions about individuals. As it stands now, in converting little data into big data, attention is ordinarily paid to stripping out names, social security numbers, and other unique identifiers, thereby rendering the data, at least for the time being, anonymous. But de-identification is not a wholly satisfactory answer. Some data, particularly data used in medical research, needs to be re-identifiable. But now there is a technical arms race being waged over anonymization and de-identification. Professor Paul Ohm, a fellow colleague at the FTC, wrote a famous Law Review article in 2009 called "The Promises of Privacy Responding to the Surprising Failure of Anonymization," emphasizing the fact that efforts to render big data safe by deleting identifying information can be undermined by techniques permitting re-identification of the data, often through the introduction of more data. The inability to ensure that big data is not subject to re-identification demands a response.

The FTC, in its March 2012 Privacy Report ("Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers"), the FTC recognizes the possibility of re-identification and calls on the technology community to provide better tools for de-identification. The report urges companies to take several actions: 1) take reasonable technological measures to ensure that the data is de-identified; 2) consider the deletion and modification of data fields and the addition of sufficient noise to make re-identification more difficult; 3) make a public commitment to maintaining the data in a de-identified fashion; and 4) if the company makes the data available to other companies, whether those may be service providers, affiliates, or other third parties, the company should contractually prohibit such entities from making an effort to re-identify the data and to provide for harsh sanctions in the event that the entity does so.

Another serious concern about big data is the looming possibility of a data breach, either as the result of a malicious hack, an inadvertent mistake, or an attack by so-called hack activists. No matter how a breach is affected, the harm to the public could be incalculable.

In my view, from a legal perspective, the aggregation of big data is no different than the impoundment of water. We talk about data flows -- so water makes a good analogy. If one puts water in a glass and the glass breaks, the damage caused by small quantities of water is generally manageable. If I cause that damage—suppose I drop a glass and water stains the floor—the legal question is: was I negligent? Did I act unreasonably? If not, I am not liable, despite the fact that I caused the stain by dropping the glass. On the other hand, if one builds a dam to form a reservoir to collect large amounts of water, the damage can be catastrophic. The town adjacent could be flooded. People could die. Ever since the Industrial Revolution, the courts have said that the potential for businesses to cause massive harm justifies what lawyers call a "strict liability" legal regime. That is, the business will be held liable even if it acted reasonably, even if it met contemporary standards.

The justifications for a tougher legal standard are two-fold. One is fairness. That is, given the magnitude of the potential injury, the burden of remediation has to be on the party who caused the risk and not the unfortunate person who suffered the harm. The second, is that the law wants to provide strong incentives to take precautions that may go beyond mere reasonableness. When the risks of catastrophic injury are considerable the incentives to avoid that risk must be strong.

Take a moment and consider the consequences of a successful malicious hack on a bank, on an insurer, a healthcare provider, or even, an educational institution like MIT. The keeper of a big data could, if that data is out there, contribute to devastating consequences to people's financial and personal life. One of the reasons I talk about legal regimes is because we need to drive home

to companies that maintain big data that they must be good stewards of that data and must use all available means to safeguard it.

Big data raises serious legal questions that remain unanswered. The question of big data is tied inexorably to how, not if, and by whom, the collection, storage, and use of big data will be regulated. Who regulates big data? Entities that store and use big data will have to be accountable for their stewardship of that data. This regulatory issue is particularly complex because big data generally resides in the Cloud, with massive amount of data in servers that could be housed almost anywhere; the Cloud does not respect political borders. In many jurisdictions, including the United States, there is a push to insist that sensitive personal data be stored with servers within the US's physical borders. The debate on this is animated by very different conceptions of privacy protection in dominant economic markets, and how this debate over regulation is resolved will have considerable impact on the future of big data.

Addressing these Big Data Challenges - Some Suggestions

These suggestions are drawn from the principles of fair information practices:

1) **First and foremost, when it is time to create or add to a database engage in a privacy risk assessment.** Ask basic questions. Is the data being used to discern general trends or to answer questions that do not relate to individuals? If so, and the data doesn't contain sensitive personal information then the risks to individuals are likely minor. On the other hand, if the data is being used to recognize correlations and then make determinations that may affect individuals, consideration should be given to a range of factors including are the security measures appropriate, given the volume and sensitivity of the data? What would be the impact of a data breach? Should additional steps be taken to identify the data? Does the quality of the data match the sensitivity of the analytic inquiry? Is the analytic inquiry sufficiently fine-tuned to draw the correlations sought with reasonable confidence? Could the contemplated uses of the data cause harm to individuals, ranging from financial loss or damage to livelihood, to damage to reputation, to unjustified discrimination, to the revelation of a potentially embarrassing fact, and can these risks be avoided or mitigated?

2) **Consent is a significant issue and where possible we should work with those engaged in the initial collection of information and urge them to get consent and to be candid when the data may be used down the road for secondary purposes.** We need to start designing data acquisition systems that design privacy in from the start and provide transparency from the very first interaction. Purpose specification is also relevant; consumer and individuals should be told in as much detail as possible the purpose for the initial data collection and what the potentiality is for future use.

3) **Create a new role inside organizations to review big data analytical processes and assess privacy risks.** This is an idea pulled from Ken Cuckier and Victor Mayer-Schonberger new book ("*Big Data, A Revolution That Will Transform How We Live, Work, and Think*."). The suggestion is that businesses create a new position called an "algorithmist", whose job it would be to review all big data analytical processes impartially to support the organization in making appropriate decisions about whether and how to use and apply those processes. This would help to provide meaningful oversight over the aggregation and use of big data. An algorithmist would have a thorough understanding of the methods of acquiring data, processing it, integrating it, and expertise in the analytic techniques applied to the data and he or she would assess whether the use of big data is appropriate in particular cases.

A key dimension to this concept is that someone has to be minding the legal and ethical factors that go into the use of big data; there are a lot of legal questions embedded in the use of big data. The FTC's enforcement record shows that some of the uses of big data violate existing laws. There are also normative and ethical questions that should be examined and so it is important we carefully define an algorithmist's role and catalog the competencies that would be needed in order to perform this job effectively.

How can we make sure that we harness the power of big data effectively without sacrificing personal privacy completely?

Session I: Understanding User Perspectives - the Value of Data and the Issues of Privacy

This session offered perspectives on the value of Big Data and examined the challenges that organizations face in addressing privacy in various domains.

Big Data, Systemic Risk, and Privacy-Preserving Risk Measurement

Andrew Lo - Sloan School of Management, MIT

In the financial industry privacy is a tremendously hotly contested issue. The problem is that in the financial system, where we don't use patents to protect our intellectual property, we use trade secrecy, we equate data privacy with profitability. This is "big data versus big dollars".

Today governments use various economic data measures, which help guide efforts to understand economic risk at the national level, with a focus on rendering the risks more transparent to policy makers, regulators, and market participants. These measures include for example, rates of inflation, unemployment, GDP growth, non-farm payroll figures, housing starts, the Federal Reserve's balance sheet figures, and other macro indicators, which are calculated by governments in the U.S. and around the world.

In contrast, in the private financial sector we have very little data with which to calculate economic measures at the national level. Let's consider Hedge Funds. The hedge fund industry, as you may recall, was the industry that got us into a fair bit of trouble in 1998 and in 2007 hedge funds played a significant role in causing some of the early defaults of banks and financial institutions. What data do we have on Hedge Funds? The answer is, not much. This sector of the financial industry is well-known for keeping its strategies, performance, and risk management practices secret; competition is fierce and firms are not required by law to share much data with government regulators or, implicitly, their competitors.

Yet the amount of assets under management at the top Hedge Funds have grown tremendously over the past decade. If we look at the Top 25 hedge funds, ranked by their assets under management (AUM), we find that the two largest, Bridgewater Associates and JP Morgan Asset Management, each top $50 billion; the smallest of these firms, ESL Investments, holds assets of $14 billion, all figures as of February 2011. The actions of one firm at this scale, should it face a rapid downturn, could have a significant impact on the financial markets as a whole.

While we can come up with various indirect measures of the amount of risk that these financial institutions pose these methods are limited -- we don't have the hard data that we need.

> *"Can you imagine the National Weather Service or the US Geological Survey or the Census Bureau doing their jobs without the data? Imagine trying to forecast hurricanes without actually having meteorological data at your fingertips? We need data to measure risks in the financial system, and today, we simply don't have the data to do that".*

The problem is that in the financial sector, where we don't use patents to protect our intellectual property, we use trade secrecy, we equate data privacy with profitability. When considering the issue of privacy as opposed to risk transparency, we find that indirect measures are suggestive, but not conclusive; we need concrete data to measure systemic risk. And we need risk transparency if we're going to prevent the kind of fiasco that has occurred over the last five years in the financial system. Is there are compromise to be made between transparency and privacy? The emphatic answer is "Yes."

One solution, a technology solution, comes out of the computer science literature—in fact comes out of a number of pioneers that did their work, early days, here at MIT at CSAIL -- Secure Multi-Party Computation (SMPC).

With Secure Multi-Party Computation, transparency and privacy needs can both be met: the individual data is kept private, the encryption algorithms are "collusion robust" allowing aggregate statistics to be computed, such as means, variances, correlations, percentiles, Herfindahl indexes, VaR, CoVar, MES, etc.

Recent research we've completed [ref: Abbe, Khandani, and Lo (2012)] takes these well known algorithms and applies them to financial data to be able to encrypt the data in such a way so that no one can reverse engineer individual firm's data, yet allows us to calculate average statistics, like the average leverage, the average risk exposure, and the average concentration across different industries. Focusing on computation of aggregates provides meaningful information and there is no need to expose individual firms raw data.

> *"When it comes to [big data and data privacy], technology can play a critical role in letting us have our cake and eat it too. But de-identification techniques are not nearly enough. We need to be able to ensure de-identification and the way to do that is using encryption."*

In closing, the global financial system has become much more important and complex in recent decades. The measurement of risk and other factors is a first step in addressing that complexity – the goal is to seek out accurate feedback mechanisms which provide information about the health of the system as a whole. The ability to collect data and compute risk analytics are critical to this endeavor. However, privacy, in this case for certain types of financial firms, notably hedge funds, is necessary to support innovation. Using privacy-preserving techniques, such as Secure Multi-Party Computation, to calculate systemic risk measures can resolve this conflict between the need to know and the need of individual firms to protect their trade secrets.

"Privacy-Preserving Methods for Sharing Financial Risk Exposures" Abbe, Khandani, Lo (2012)

Big Data and Privacy

Maritza Johnson - Technical Privacy Manager, Facebook

People are sharing details about their lives via social media services like Facebook and Twitter at enormous rates. As more of the world interacts online with their friends and with public content, the value of social network data continues to grow. That being said, as people continue to share and the world becomes more open and connected, users want to control how and what they are sharing. We as researchers are in a great position to design tools that empower users to manage and control their own data.

Facebook has an incredible opportunity to examine the meaning and value of social network data through understanding how people interact with each other and how information is shared. Today, there are approximately 1.15 billion monthly active users on Facebook, who spend an average of 20 billion minutes on Facebook each day. These users share about 2.5 billion photos each month and approximately 3.5 billion pieces of other content each week. Social networks, like Facebook, allow businesses, social scientists and technologists to measure human behavior and interactions with unprecedented scale, scope, and precision.

Some say that Facebook users are contributing to a sociologist's dream data set. Our data science team is continually conducting research that showcases the value of social network data. For example, researchers recently measured overall user sentiment around different time periods by creating a happiness index across geographic boundaries. Another idea would be to examine national trends and count the number of posts, which include mentions of phrases like "laid off" as a way to gauge unemployment on a state-by-state basis.

Looking at the strength and expansion of social networks, Facebook researchers can plot the degrees of separation (or "hop distances") between people, similar to the research done several decades ago by Stanley Milgram and his concept of "six degrees of separation". Today, using Facebook data we can quickly show that we're more connected than we may have thought: the average number of hops between two people worldwide is about five, and in the United States the average is about four.

For individuals who are curious about learning more about their habits and the characteristics of their friend network, Wolfram Alpha has leveraged Facebook's API to enable people to study their own Facebook data using various visualization techniques.

At Facebook, we understand that the information our users put on the site is personal, which is why we have worked hard over the years to give users controls over what information is shared and who it is shared with. Research shows that slightly more than 20% of men and slightly less than 20% of women maintain their Facebook presence in a completely public manner; the vast majority seeks to maintain a limited audience. One question we often face is how to empower people so they can easily reach their audience while ensuring that they maintain an appropriate level of control. As technology continues to evolve, how do we design privacy controls that satisfy these goals? In some cases, online services have taken the path of providing one option: share publicly or with your friend network. This choice is binary, and is in some ways easier to manage. Social sites like Twitter and Tumblr are good examples of this approach; however, sites like Facebook and Google Plus need more fine-grained controls. At Facebook we allow people to

control every single piece of content they add to the site. In addition, we have developed a series of inline controls whereby people can select an audience for their content when they are in the act of sharing.

Users are making use of additional control features to curate their online presence by untagging photos or deleting comments. To assist users in curating their content, we introduced the activity log feature that allows them to see the privacy settings for each picture, post, or other content that they have added. This feature increases transparency and is great for fine-grained control.

Now, think about the different information you have shared with social networks or web services and ask yourself – what tools can we design where you can understand what you're sharing and with who are sharing it with? That is an open research question.

To summarize, "How do we design for transparency and control in the context of other data-sharing relationships?" and "How do we evaluate effectiveness?" We need to focus on evaluating effectiveness in the form of user studies, to be sure that users are informed, and that transparency and control mechanisms truly work.

Big Data Privacy

Robert Zandoli - SVP and Global Chief Information Security Officer, AIG

AIG has 88 million customers and we must ensure that as we bring Big Data into AIG that we protect it. Information security officers focus on understanding the value of the data and knowing the owners of the data. The keynote speaker at the RSA conference this year was Art Coviello, Executive Chairman at RSA, who said as our data services increase, so do our vulnerabilities. This absolutely underscores what I see as a major challenge, protecting our data.

Big Data in the public and private sector is not an entirely new phenomenon; what is new and different now is the pervasiveness of the data, the vast expansion of sources, particularly of external, sensitve data, and the rapidly decreasing cost of data storage. Given these enhanced technical and affordability factors, the ability to collect, analyze, and report on massive amounts of data is now possible for institutions of all sizes.

In this environment, the importance of the Big Data Lifecycle is central to responsible implementations and participants in this arena must bear in mind the following factors:

In this environment, the importance of the Big Data Lifecycle is central to responsible implementations and participants in this arena must bear in mind the following factors:

- **Risks** – Data in transit, at rest, and in use must be treated with equal or greater care than it would be in a production environment.
- **Laws and Regulations** – Compliance models are a primary consideration and firms and governments must respect and adhere to global privacy standards.
- **Use and Misuse** – Organizations must ensure that the data is used in a meaningful and ethical manner and that access to sensitive data of any type must be logged and monitored to preserve trust with both existing and potential customers.

As organizations devise new ways to store, process, and analyze large amounts of data, they will face a greater array of threats to their efforts to protect the information i.e., to ensure the confidentiality, integrity, and availability of that data.

These threats come in at least three forms:

- **Insights** – Threats stemming from the desire of outside parties to take advantage of the intellectual property assets within organizations.
- **Behavior** – Challenges in monitoring and profiling employee, customer, and adversary behavior.
- **Scale** – The pervasive use of social media, mobiles, cloud computing, and othe new technologies is altering the security landscape significantly.

There are pros and cons to the current technological scenario. New forms of information analysis will help people to become more productive and knowledgable about the world, but there are significant challenges to be met in understanding the tools and using them in an effective and ethical manner, at indiviudal, corporate, and governmental levels. Furthermore, the rapid expansion in the ability to collect, store, and move data runs in tandem with the opportunity for "the breach of the century."

Understanding the Big Data lifecycle is important the

- **Collection** – What kind of data is being collected? Is it reliable and secure?
- **Storage** – How is the data being stored? Where and with what type of protection?
- **Uses/Users** – How is the data being used and by whom?
- **Transfer** – How is the data being moved? Where is it going and is the transfer being done securely?
- **Destruction** – What are the data retention cycles? Who decides when to destroy the data and how will the destruction take place?

Recommendations for responsible data maintenance include: an assessment of the data storage environment and knowledge of usage patterns, keeping the controls close to the data, encryption for all data (both static and in transit) and incorporating logging functions into the data cluster.

In summary: Security should be a central consideration, not an afterthought. Information theft is a greater risk in the Big Data environment. There must be an appropriate balance between the ideal level of security and the processing and analysis of large datasets. Human factors are a key element in the security context; staff should understand the nature of the data, the policies governing its protection and use, and those who will access it must be trustworthy.

Big Data security is a matter of prudent risk and protections must be devised that are appropriate to the level of risk faced by the organization.

Big Data: New Oil of the Internet

Alex (Sandy) Pentland - Media Lab, MIT

I'd like to start off talking about big data and public good, by looking at several field projects we have done lately, where the focus shifts from privacy per se to the value of the data, who owns the data and who captures it.

In the D4D ("Data for Development") Challenge [ref: http://www.d4d.orange.com/home], a project completed with the support of the United Nations and the World Economic Forum, Orange contributed, and made public, all of the data they had collected in the Ivory Coast. Research teams from eight-six universities around the world collaborated on projects using that data to improve living conditions through intelligent data acquisition and analysis. Data was massaged algorithmically in order to make it difficult to re-identify and, along with the technical protection, there were contractual agreements with project participants that they would not attempt to re-identify the data and would only use it for the specific purpose that had been suggested originally.

Results for various project teams included a 10% reduction in commuting time in a large city - because the researchers analyzed how people were traveling, they were able to rationalize the bus lines. Similarly, another group found a 20% reduction in the impact of malaria through better management practices, and an HIV program was also transformed by this sort of data analysis. Other project focused on ability to create real-time census data, which provided a platform for the analysis of poverty and ethnic divisions in this country, which were previously unavailable due to the civil war in this region.

In making the data available, users had to agree to a legal contract that said they would not attempt to re-identify individuals and the data could only be used for the specific purposes of the competition. I want to emphasize that our solution to data sharing, was both technical (aggregate anonymization algorithms) and legal.

I lead a group at the World Economic Forum focusing on Personal Data and what we've realized is that there is enormous value in data for society - both for individuals and for companies. Rather than trying to lock everything down, we want data to be very liquid, allowing it to flow. There is a risk reward - risk being damage to an individual, reward being societal good such as saving lives.

In a report, "Personal Data: Emergence of a New Asset Class," prepared for the World Economic Forum, I proposed a "New Deal on Data" framework with a vision of ownership rights, personal data stores, and peer-to-peer contract law. The report findings helped to shape the EU Human Rights on Data document and the US Consumer Privacy Bill of Rights. Among the proposals in the report was the notion that there could be a combination of informed consent and contract law that allowed for auditing of data about oneself. The personal data would also have meta data that would accompany the personal data, showing provenance, permissions, context, and ownership. At MIT, a open source version of such a scheme has been created in the form of Open Personal Data Store (openPDS) [ref: http://openpds.media.mit.edu/]

OpenPDS features contractual agreements built into code, forming "trust networks." It also provides for Secure Distributed Identity through OpenID Connect, an initiative that is now supported by the MIT Kerberos Consortium. The system allows for auditability; users can verify what data is read, which data can leave the system, and how it is being used. It is built on a distributed computing platform and leverages the data from multiple PDSs in a privacy-preserving fashion. openPDS focuses on answers, not raw data, and preserves the safety required for location and other high-dimensional data.

> *"One of the key things in debates about data sharing is that it is rare when you actually have to share data. What you have to share is certified answers about data. This is what banks do; they don't tell you about all of the accounts, customers, and investments that*

> *they have, they just certify that they have enough money to fulfill the transfer that is being proposed."*

One final field project we are working on that collects data from individuals, is the Mobile Territorial Lab. The goals are to understand the needs and behavior of users. The project provides individuals with a mobile phone equipped with a sensing middleware, which collects the data for analysis. The first community for deployment was young families with newborn babies. Short-term outcomes include developing and testing a new model of data ownership, understanding the dynamics of peoples' needs, and understanding the nature of peoples' interactions in freshly-generated social networks.

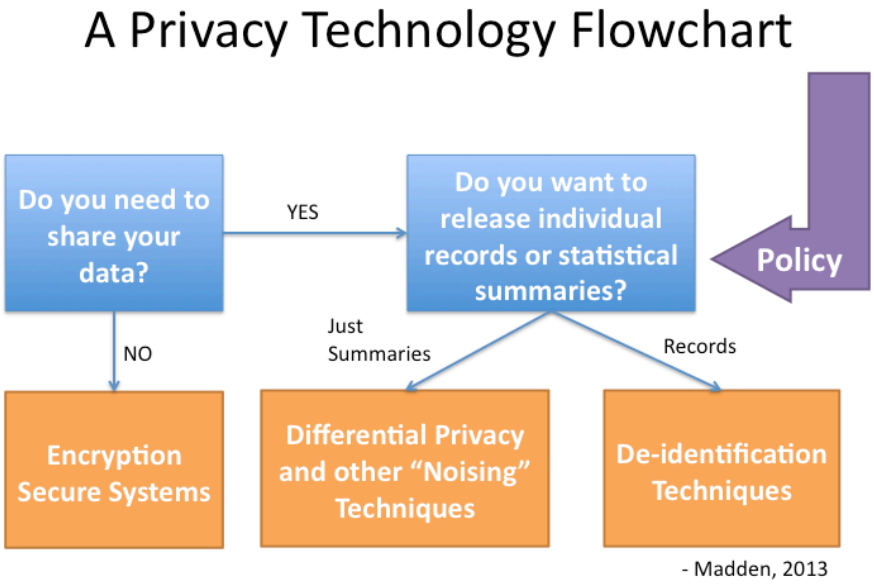
Finally, as the Grandfather of Google Glass (my students are technical lead for the project, based on work done in my lab during the 1990's), I want to point out the problems that such new devices will bring:

> *"One of the biggest debates around Google Glass concerns privacy, because when you can take pictures of people without them knowing, just by walking by them, and you combine that with facial recognition, it gets complicated…and, as there are more sensors, more cameras, more RFIDs, more GPS units deployed in the world, these issues are going to become more acute."*

We struggled with these questions while prototyping the first of these devices during the 1990's, and these struggles highlighted questions for further research, which include how to obtain informed consent and how to structure permssions in this environment.

Session II: Approaches to Managing Data Privacy – Systems, Tools, and Theory

This session covered different techniques for managing data privacy. The discussion focused on the latest thinking and research aimed at privacy issues surrounding mining, querying, integrating, and sharing large data sets, with the goals of assessing the applicability of different techniques and framing future research directions.



What we mean by "privacy" falls along a spectrum and we are seeing changing norms about what is acceptable. We recognize that technology is not going to "solve" privacy, but technology can increase our ability to protect and privately share data. In parallel, laws and regulations are also needed as a necessary safeguard.

Differential Privacy

Kobbi Nissim - Visiting Scholar, Harvard CRCS* and Faculty at Ben-Gurion University

* Center for Research on Computation and Society

I think there is a great promise for the marriage of Big Data and Differential Privacy:

- Big data brings with it a promise for research and society, but often the data contains detailed sensitive information about individuals, making privacy a real issue.
- Heuristic privacy protection techniques were designed for an information regime very different from todays. Many failures were demonstrated in the last decade.
- Differential privacy provides provable guarantee for individuals and provides good utility on larger datasets.

Privacy: Our goal is, given a data set of sensitive individual information (e.g., health records, census data, or financial data), to compute and release functions of the data without

compromising individual privacy. I will assume that there is a trusted entity – a server or an agency (this assumption can be sometime relaxed using tools from cryptography). I'll call this trusted entity a *data creator*. The data curator collects information about individuals, and then computes functions of this information or answers queries about it. The curator may try and use "traditional" techniques for privacy protection. But, these should be used very cautiously, as most of "traditional" techniques are heuristics designed in the days before massive data collection was possible and before the Internet was present to make massive datasets accessible. Unfortunately, we now have over a decade of demonstrated failures with such techniques.

One example of a traditional privacy protection technique is *anonymization*. It entails an anonymized database, resulting by the removal of identifying information by the trusted curator. Often, the removed identifying information is replaced with random identifiers. This concept of anonymization is use widely. It is hard-wired into practices, regulations, and thought about privacy, but in reality this technique has produced a series of failures that were pronounced in both academic and public literature.

One of the main points of failure is the inability to cope with external information – Sweeney demonstrated that it is often not hard to reveal hidden identities (this is called re-identification) by linking an external data source with the anonymized dataset (Sweeney, 2000).

The failure of anonymization techniques to cope with this availability of external information poses a big challenge – One may try to design an anonymization scheme that is safe relative to the kind of external information that exists at the time of the design, but how will it cope with new information that may be made accessible subsequently? The lesson is: we need to cope with general or arbitrary external information.

Differential Privacy: *Differential privacy* (Dwork, McSherry, Nissim, Smith 2006) presents a formal "crypto-inspired" approach to privacy; the goal of this line of work is to protect privacy of data sets against all "feasible" attackers, which may have access to arbitrary external information. Differential privacy captures a strong and intuitive notion of individual privacy, it has a strong mathematical foundation, and yet it enables the extraction of useful information from a data set. Intuitively, differential privacy works by the introduction of carefully crafted noise by the data curator: large enough to hide the individual contributions to the data set, yet small enough to allow for useful analyses. The promise for big data is that these analyses will improve with the data size.

Unlike many older attempts at defining individual privacy, differential privacy does not simply restrict the outcome of a privacy preserving algorithm (as is the case with anonymization where the outcome should not include any identifying information). Instead, differential privacy puts a restriction on the input-output behavior of the algorithm, namely, that for all datasets that differ on the entry of just one individual, the outcome distribution of the algorithm should remain (almost) the same. The difference between the two distributions is bounded by a function of the privacy parameter ε (the smaller, the better; typically, ε is a small constant, e.g., $\varepsilon = 0.1$ or $\varepsilon = 0.01$).

Formally:

> *Definition:* An algorithm A mapping data sets to outcomes preserves ε -differrential privacy if for all data sets D_1, D_2 which differ on the input pertaining to just a single individual, and for all subsets S of possible outcomes of A :
>
> $$Pr[A(D_1 \in S] \leq e^{\varepsilon} Pr[A(D_2) \in S] .$$
>
> Where the probability is taken over the randomness of algorithm.

Differential privacy implies that what can be learned about an individual could also be learned *without the algorithm having access to that individual's data*. This creates a veil of deniability: since the changes to one's data are almost not noticeable by the data users, the individual could claim that his or her data is different from what it really is, or that he or she did or did not participate, at will. Differential privacy can also be formalized (equivalently) as a utility-theoretic guarantee: the decision to participate incurs a bounded utility loss.

Since the introduction of the concept in 2006, research of differential privacy has yielded many algorithms and tied it closely with many areas of research from algorithm design and machine learning to statistics, complexity theory, game theory, and geometry. A partial list of what may be computed with differential privacy includes:

- Statistical estimations
- Learning discrete classifiers
- Clustering
- Spectral analysis
- Synthetic data generation
- Convex Optimization
- Frequent Pattern Mining
- Genome-wide association studies

A couple of points to note are:

1. Differential privacy protects not only individuals but also small groups of individuals, but the quality of protection deteriorates with the group size, namely the privacy parameter deteriorates by a factor of ? for groups of size ?. Hence, differential privacy does not protect sensitive *global* information. For example, in a clinical survey one may learn that cancer and smoking are related. If you then see me smoking in public, you know that I am more prone to having cancer and this may have an effect on my health insurance premium. One may argue whether this is a bug or a feature: We do wish to recover information about the data (as a result - I may stop smoking).

   Other examples include: financial transactions – e.g. firm-level trading strategies; social data, where a single presence affects everyone else; and genomic data, where information about an individual may be revealed if enough family members participate.

2. One of the strong points of differential privacy is that it composes. With earlier attempts at defining privacy the independent execution of two "private" algorithms could potentially result in total loss of privacy. With differential privacy we can prove that privacy is still preserved, but to a lesser extent.

This deterioration in privacy is a significant shortcoming – the leakage of individual information may add up over the course of many releases. The privacy parameter resulting from executions of an ε-differentially private computations would be (roughly) . Hence, to preserve differential privacy we may need to eventually stop answering queries about a dataset! We can show that this is inevitable in some form under *any definition of privacy*. The question is, hence, how to set the privacy parameter to resolve the privacy – usability tradeoff. We note that differentially private sanitization (Blum, Ligett, Roth 2008) is one way of getting around this problem.

Is differential privacy practical? I believe differential privacy would prove to be practical, in particular in contexts where data is abundant. Differential privacy is concept under research and experimentation. It is racing towards mature theory, but, there is (currently!) little off-the-shelf software, and each application requires fresh thinking. Among the several systems available to make its use easier we mention PINQ (McSherry '09); a programming language with privacy enforced by a type system (Haeberlen et al '11); and systems for restricted classes of queries (Roy et al '10 and Moharan et al '12.)

Conclusion: The heuristic treatment of privacy leads to failures and this is why we need a formal treatment of privacy. Differential privacy (and variants) is currently the only suit of rigorous privacy solutions where privacy is defined in terms of the effect of an individual's data on output.

Differentially private algorithms introduce noise into analyses, and a rough rule of thumb is that these give better overall accuracy with more data. In particular, there is a high potential for a good match with Big Data.

Differential privacy @ Boston area:

- Harvard project involving research in differential privacy: *Privacy tools for sharing research data*. http://privacytools.seas.harvard.edu/

- *Privacy year* at the Hariri Institute for Computing and Computational Science & Engineering, Boston University (2013-2014).

Additional Resources:

- Erica Klarreich "Privacy by the Numbers: A New Approach to Safeguarding Data". Scientific American, December 31, 2012.

- Blum's brief tour of Differential Privacy. http://www.cs.cmu.edu/~avrim/Randalgs11/lectures/lect0420.pdf

- Cynthia Dwork's survey papers.

- Smith's Tutorial from CRYPTO 2012. http://www.cse.psu.edu/~asmith/talks/2012-08-21-crypto-tutorial.pdf

- Roth's lecture notes. http://www.cis.upenn.edu/~aaroth/courses/privacyF11.html

- Course by S. Raskhodnikova, A. Smith. http://www.cse.psu.edu/~asmith/privacy598

- DIMACS Workshop on Data Privacy (October 2012). Tutorials by M.Hardt, G. Miklau, B. Pierce, A. Roth http://dimacs.rutgers.edu/Workshops/DifferentialPrivacy/

- Intro to Differential Privacy. https://www.simonsfoundation.org/features/science-news/privacy-by-the-numbers-a-new-approach-to-safeguarding-data

No Free Lunch and the Pufferfish Approach to Privacy

Ashwin Machanavajjhala - Duke University

The issues relating to data privacy in the real world exhibit some similarities and some differences across all kinds of different domains. In the medical environment, there is medical data, genomic data or other research data based on private information about patients and subjects. Functional uses of the private information include finding correlations between a disease and a geographic region, or between a genome and disease. In an advertising context, social media firms like Google, Facebook, and Yahoo focus on the clicks and browsing habits of their users, assessing trends by region, age, gender, or other distinguishing features of the user population. Private information includes an individual's personal profile and "friends". Functional uses of the data could entail a prompt to recommend certain things to certain groups of users or to produce ads targeted to users based on their social networks. Different applications will have different requirements for the level of privacy needed. The challenge in any of these applications is: "How do you really trade off privacy for utility?"

One technique for managing privacy protection is Differential Privacy, which is based on an application-independent definition of privacy. It tolerates many challenges that other definitions are susceptible to and provides a nice tuning parameter, which we refer to here as the epsilon ("ε") term, allowing you to "tune" the trade off between privacy and utility. The ε term is essentially a measure of information disclosure; the larger the ε, the less privacy and the greater the utility present in the environment. Differential Privacy is an emerging technique and is slowly coming into prime time now. However, there are some drawbacks: differential privacy's ε is still not entirely sufficient for trading privacy off against utility in many real-world applications. For example, a small ε may not limit the ability of an attacker to learn sensitive information, especially when data are correlated, and a large ε may not yield sufficient utility, particularly in cases of sparse high-dimensional data.

Thus there is "no free lunch!" The no free lunch result says that it is not possible to guarantee any utility in addition to privacy without making assumptions about what the adversary knows about you (as the data collector, an individual whose information is in the data), the data, and how the data is generated. Unless you know something about the adversary's prior knowledge, you cannot guarantee both privacy as well as utility. So the question is "How can we construct principled privacy definitions with sufficient controls for trading off privacy for utility in various contexts?"

I will use the Pufferfish as an analogy. The pufferfish (the data) contains toxins (sensitive information). You must process the pufferfish (data) and follow restrictions to guarantee that there is something left in the output and that it is (relatively) safe. In the end, we are left with our dish, Fugu (the sanitized data), which is safe (minimal leakage of sensitive information) and tasty (high utility)!

In the Pufferfish framework, we include three controls (instead of just one) for tuning the privacy-utility trade-off, focusing on the following questions:

- What is being kept secret? what sensitive information about the individual do you want to keep secret?
- Who are the adversaries? and what information do they possess about the data set?
- How is information disclosure bounded? (akin to Differential Privacy's epsilon term)

In Pufferfish, we define the information that must be kept secret as a set of discriminative pairs. Each pair represents two mutually exclusive statements about an individual (e.g. Bob has cancer vs. Bob has diabetes) that the adversary should not be able to distinguish between. An adversary is completely characterized by his or her prior information about the data; we do not assume computational limits. This prior information is captured using data evolution scenarios data evolution scenarios – the set of all probability distributions that could have generated the data. Again, there are no limiting assumptions – all probability distributions over the data instances are considered possible. Disclosure is tunable and the ratio of the prior and posterior odds of the adversary about any discriminative pair must be bounded by e raised to the ε term. In fact, differential privacy is an instantiation of this framework that makes one specific choice for sensitive information (every statement about an individual and its negation is a discriminative pair) and adversarial knowledge (namely data devoid of any constraints or correlations).

Ongoing work in this area is focusing on utilizing the new tuning knobs in special ways. One area of interest lies in declaring a more fine-grained specification of sensitive information; another considers realistic adversaries who may know of constraints or correlations in the data. Tuning the knob of adversarial knowledge is non-trivial, because we need to specify the sets of probability distributions and we do not know what the adversary knows. Further, there are publicly known constraints in the data including count and marginal constraints, structural zeros, functional and other dependencies in relational data, and degree distribution in a graph. It is also important to develop a view on what weaker adversary models might be. Finally, work continues on controlling the amount of disclosure.

In summary, there is no absolute privacy, you must carefully trade-off privacy for utility. The No Free Lunch theorem says utility depends on assumptions about the adversary. The Pufferfish framework for privacy allows you to tune the privacy-utility tradeoff in more than one way. We hope real world applications can provide services on sensitive data with increased utility and provable privacy by leveraging these additional knobs in Pufferfish.

Additional Resources:

A. Machanavajjhala, A. Korolova, A. Das Sarma, "*Personalized Social RecommendaKons – Accurate or Private?*", PVLDB 4(7) 2011

A. D. Kifer, D. Kifer, A. Machanavajjhala, "*No Free Lunch in Data Privacy*", SIGMOD 2011

A. D. Kifer, A. Machanavajjhala, "*A Rigorous and Customizable Framework for Privacy*", PODS 2012

A. Machanavajjhala, B.Ding, X. He, "*Blowfish Privacy: A Policy Driven Approach to Rigorous and PracKcal Privacy* (*Work in progress)*", 2013

Accountable Systems

Lalana Kagal - Distributed Information Group, CSAIL, MIT

Much of the research in the area of data privacy focuses on controlling access to the data. But, as we have seen, it is possible to break these kinds of systems. You can in fact infer private information from anonymized data sets, examples include the re-identification of medical records, exposure of sexual orientation on Facebook, and breaking the anonymity of the Netflix prize dataset. What we are proposing is an accountability approach to privacy, when security approaches are insufficient. The accountability approach is a supplement to, and not a replacement for upfront prevention.

The approach that we are suggesting is similar to how social and legal rules work in society. Most of these rules are not perfectly enforced, and they are not enforced at the time when an action happens. If you are parking in a handicapped zone, there is nothing that prevents you from doing it, but it's possible that you're going to get caught and fined. This enforcement process is similar to the vision of information accountability. If you do something wrong, it's possible that the system will identify you and hold you accountable for your inappropriate actions. So instead of focusing on access control, we are looking at usage control.

A principle tenant of the approach is that when information has been used, it should be possible to determine what happened and to pinpoint use that is inappropriate. This requires the ability to express information use policies, to monitor and reason over information use, and to provide redress. There is a shift in focus from what is known about an individual or group to what is being done with that information.

In order to enable accountable systems, there are some specific areas we must develop:

1. **Data Provenance.** It must be possible to track data as it flows through the system. Information must be annotated with data that identifies the source. Data transfers and uses must be logged so that chains of transfer have audit trails. All provenance information should be machine-readable because we want tools to reason over it in order to identify misuse.

2. **Policy Compliance.** Data providers must supply machine-readable policies that govern permissible uses of the data and automated reasoning engines must use policies to determine whether a given data use is appropriate. Along with having a tool that identifies misuse, it is also useful to have a set of tools that are able to tell you how you should behave appropriately with respect to the data, so you can ask questions like: "Can this data be used for research? Can I use this kind of data for commercial purposes? What are the consequences of misuse?"

3. **Policy Awareness.** All participants in an accountable system must have access to an intuitive view of the policies that are acting on the data they are using. This introduces a human component such that users can manipulate information via policy-aware interfaces that signal compliant and non-compliant uses. A good example of this is the Creative Commons, which gives data curators a kind of license that can be associated

with their data using very easy to read icons to illustrate what purposes the data owner allows.

4. **Violation detection and identification.** When data is misused in a system, you must be able to identify misuse and determine who the violator is.

5. **Privacy Audit of the System** which studies how the system collects and uses private information. This is required to develop trust between the accountable system and the end user.

We have a number of research projects going on in each of these areas, for example the "Semantic Clipboard" [ref: http://dig.csail.mit.edu/2009/Clipboard/] deals with Policy Awareness and works with Creative Commons licenses. It extracts the licenses and allows people to search for media based on a certain purpose.

For all of our data modeling requirements we use Semantic Web Technologies. "Linked Data" principles provide information management at Web-scale by leveraging Web protocols and technologies. All of the specifications for Linked Data are open Web standards; the W3 specifications are supported by the community and are free for use. It also entails the reuse of existing well-developed and studied Web technologies including the HTTP protocol and Uniform Resources Identifiers (URIs).

We also make use of the "AIR Rule Language and Reasoner" [ref: http://dig.csail.mit.edu/2009/AIR/], a machine-readable rule and policy language based on Linked Data technologies, it allows reasoning over distributed information systems. It is focused on justification generation, ease of specification, rule reuse, and built-ins for use on distributed data. It has been used in various projects for information accountability, policy compliance, trust frameworks, and access control. The AIR Reasoner generates proofs for all of its conclusions, although they are not easy to understand. There are also constructs for manipulating justifications, including a graphical justification interface that provides an explorable structured natural language explanation for policy decisions.

There are a number of challenges in building accountability systems -- there is not a purely technical solution, our solutions must incorporate technology with social and regulatory components. Significant technology challenges include:

- Tracking data across different systems and at different levels
- Tracking data that has been modified significantly
- Handling provenance information that can be sensitive in itself and must be protected
- Generating machine-understandable policy
- Identifying the purpose of the data use and subsequent misuse (semi)-automatically
- Deploying efficient reasoning techniques that provide IA functionality

On the social front, there are challenges in terms of how to incentivize accountability and how to determine the right punishment for information misuse.

De-identification Methods – Anonymization of Patient Spatial Data

Shannon Wieland - Visiting Scholar, MIT

This discussion will focus on the anonymization of patient medical data for use in epidemiology, which is the study of diseases in populations. The first spatial epidemiological studies were conducted with simple dot maps of diseases. There is a famous map made by a British physician, John Snow, who plotted cases of cholera in 1856 during an outbreak in London. He noticed that the cases tended to cluster by one water pump on Broad Street and arranged for the removal of the pump's handle. This effectively ended the epidemic. It was quite an impressive insight because at that time the mechanism for cholera transmission was not well-understood; it was generally believed to be an airborne disease.

In addition to disease mapping methods, which have grown in sophistication, there are also tests of the tendency of cases to cluster as a global phenomenon in space. There are detection tests that look for localized clusters and statistical tests of the likelihood that cases are situated close to a putative environmental source. The hope is that with larger data sets and increased computing power we will be able to detect and analyze weaker associations in the data.

One of the challenges in using spatial information, although ultimately to be used for good in the medical context, is that it also identifies sensitive information that poses a risk to patient privacy. One study that looked at disease maps that were published in the medical literature found that patient locations could be traced to single addresses in several cases. The challenge is to create data sets that can be disseminated without violating confidentiality and yet still shine light on the patterns of disease distribution.

HIPAA (the Health Insurance Portability and Accountability Act of 1996) details specific information disclosures that violate privacy and defines a category of non-identifiable data sets which can be freely shared. For a data set to qualify as de-identified, either of two criteria must be met. The first is that any of eighteen specific identifiers, including five-digit zip codes, must be removed from the data. One can include the first three digits of a zip code provided at least 20,000 people share the same first three digits. The second criterion is that a qualified individual must determine that there is a very small risk that a recipient of the data set could use that data to identify an individual subject within the set.

There is no consensus for how to best calculate the risk of re-identifying an individual in spatial data, and several different measures of privacy have been proposed. K-anonymity was developed in the setting of tabular data, but in spatial data it has developed its own unique meaning. If a data subject is moved from his or her original location to a new location to create a de-identified set, then the de-identified set has k-anonymity if at least k people in the underlying population live closer to the original location than to the new location. To calculate the k-anonymity, a circle is drawn around the original point with radius equal to the distance moved, and the number of people living within the circle is counted. One challenge with k-anonymity is that one can devise de-identification strategies that are poor strategies, but have a high value of k-anonymity. For example, if everyone in the original data set is moved one mile to the west, then this yields an extremely high value of k-anonymity, but a data recipient could just move all the data points one mile to the east to reconstruct the original locations. It is possible that k-anonymity is a necessary condition for de-identification, but to make it sufficient, some other quality of the data needs to be captured.

A second measure, developed by Spruill, measures the fraction of records in the de-identified set that are closer to their original location than to all of the locations in the original set. However, it is possible, once again, to devise de-identification techniques that do not provide privacy, and yet have a good score by this measure. For example, if a list of exact locations is simply permuted, then the new set has a perfect measure of zero. However, the permuted set still consists of the entire set of patient locations, so that does not de-identify the data at all. A third measure, differential privacy, has been used recently in the setting of spatial data and is widely considered to be a robust guarantee of privacy. A differentially private mechanism guarantees similar output when applied to two data sets that are identical except at one position.

Finally, for a de-identification strategy that is determined by a matrix of transition probabilities, we can calculate the re-identification probability. This is the probability that an individual in the population is in the original list of locations, given the de-identified list, based on a straightforward application of Bayes Rule. [ref: Wieland et al. 2008]

Some of the various approaches used to manage privacy in patient data sets include:

1) **Aggregation**. This is the prevailing approach and is a well accepted method today. From the original set of exact patient addresses, the number of patients in each administrative region (e.g. county) is calculated. The de-identified set consists of the list of regions and patient counts. If the number of people in each underlying region is great enough then it usually preserves privacy. The main problem with the method is that there is a substantial loss of information. For example, aggregating data before applying a cluster detection algorithm degrades the sensitivity and specificity of detection.

2) **Adding Random Noise.** These methods add random noise chosen from a family of distributions to each point. For example, Gaussian noise can be added to each original location, where the variance of the Gaussian depends on the population density. Other researchers have added noise in a donut shape around the original point. This guarantees k-anonymity where k is the number of people that are living in the donut hole. One of the challenges of these methods is that the re-identification probability depends on the underlying population density, and it also depends on geographical features on the map. It can be hard to calculate that re-identification probability and, if it is hard to calculate, it is hard to guarantee.

3) **LP Method.** This is a new method we have developed based on linear programming. This method defines a matrix of transition probabilities that assigns patients from one set of positions to another set of positions. We want to assign values to the matrix entries so that the re-identification probability is low and so that patients are not moved a great distance; this preserves the utility of the data for further study. This process translates mathematically into a set of constraint equations and an objective function, and the equations are in a form that may be solved by existing linear programming techniques. Even though linear programming techniques do not scale well in general, this method scales well for large data sets, because we can take advantage of the sparse nature of these data sets and matrices to decompose the problem into smaller problems; we solve those smaller problems and then recompose the results into a solution.

Each anonymization method has strengths and weaknesses. Aggregation is the prevailing method and it does preserve privacy, but it also leads to significant information loss. Adding noise in a heuristic fashion is problematic; it supports a smaller distance move, but offers an uncertain privacy guarantee. Differential privacy offers a strong assurance of privacy, but masks local phenomena. Our new LP method based on linear programming preserves privacy and tends to

move patients only a short distance, but it does make some assumptions about the data collection process, which are similar to the assumptions made when aggregating data.

Encrypted Query Processing

Raluca Ada Popa - CSAIL, MIT

In 2012 hackers extracted 6.5 million hashed passwords from LinkedIn's database and were able to reverse most of them. This is a problem we all are familiar with: Confidentiality Leaks. There are many reasons why data leaks, and for the purpose of this talk, I'm going to group them into two threats. First, consider the layout of an application that has data stored in a database. The first threat is attacks to the database server. These are attacks in which an adversary could get full access to the database server, but does not modify the data, it just reads it. For example, system administrators today often times have root access to the database servers and they can read confidential data of the company, such as financial or medical data. The second threat is more general and includes any attacks, passive or active, to any part of the servers. For example, hackers today can infiltrate the application systems and even obtain root access.

How do we protect data confidentiality in the face of these threats? Let's consider a simple client server model. Many existing systems thought of specific attacks and developed techniques for those specific attacks, but don't protect against others. Other systems assume that some part of the server is trusted, but if that part of the server is compromised, then attackers still get access to confidential data. For example, some systems assume that the operating system is secure, but the operating system also gets hacked into today. A promising approach to this problem is to encrypt the data on the server with a key that only the client has. In this way, even if the adversary steals all the data from the server, it is all encrypted data. A major challenge with this approach is that many times the server can no longer provide service. What do you do with encrypted data? How can you manipulate it? One idea is to give the encryption key to the server so that it decrypts the data before processing it. The problem is that then the adversary can also steal the key or steal the data from main memory when the data is being processed. A better idea is to compute on the encrypted data: the client sends an encrypted request, the server computes the request on the encrypted data and returns encrypted results that the client can decrypt with the key. If the server gets compromised, the attacker only gets encrypted data.

Based on this insight, we have developed CryptDB, which is a database management system for computing SQL queries over an encrypted database. CryptDB is the first practical DBMS to process most SQL queries on encrypted data. You can use it, for example, to hide a database from system administrators while allowing them to still maintain the servers, such as to perform load balancing. You can also use it for outsourcing the database at the cloud while not allowing the cloud to see the confidential data. CryptDB has a modest overhead -- using CryptDB, we see only 26% throughput loss (TPC-C performance) than with MySQL, which is an industry standard benchmark.

CryptDB requires no changes to the DBMS (Postgres, MySQL) and no changes to applications. If we compare CryptDB to fully homomorphic encryption (FHE), we know that FHE can compute any function on encrypted data and provides very strong security; however, it is prohibitively slow, with greater than 10 ^9 slowdown currently.

Shifting to attacks on the application server, CryptDB protects the data of logged out users during the attack. Ongoing work includes CryptWeb, which will protect the confidentiality of users'

data at any time against any attacks to the application servers. The team is also working on functional encryption, with a goal of computing any general function on encrypted data.

Raluca Ada Popa, Catherine M. S. Redfield, Nickolai Zeldovich, and Hari Balakrishnan.
CryptDB: Protecting Confidentiality with Encrypted Query Processing.
In *Proceedings of the 23rd ACM Symposium on Operating Systems Principles (SOSP 2011)*, Cascais, Portugal, October 2011.

Encrypted Databases – Private Information Retrieval Using Secure Hardware

Srini Devadas - CSAIL, MIT

Private Information Retrieval (PIR) is a technique that allows a user to retrieve data from an untrusted server without the server being able to tell which data the user is interested in.

With PIR, the data could be public data, it could be de-identified data, it doesn't have to be encrypted, but the server should not be able to tell what data the user was interested in. This is a way to protect the query -- the query could be an image, a genome, or stock price fluctuations. Or, for example, if you're constantly asking for data from a particular geographical region, you are giving away your interest in that region. PIR will help conceal that interest.

Cryptography-based approaches include homomorphic encryption and fully homomorphic encryption (FHE). The limitations to the homomorphic approach are that the queries are limited in terms of complexity and the performance overhead is not acceptable today in real world applications.

With PIR, the limitations are that the matching programs need to be trusted not to leak information through the I/O channels and that a curious server could run arbitrary programs on the private data to discover the data in question. Possible applications for PIR include document matching, DNA sequence matching, and content-based image retrieval (CBIR). In document matching, the user provides a private, encrypted set of document features and a private, encrypted distance metric. For each public document, the document matching program calculates its features and then compares the features of the public document to those of the encrypted document, using the private distance metric. The documents with the highest score are returned to the user.

In DNA matching, the user sends his or her private DNA sequence to the server. The server compares the user's DNA with its public DNA sequences and returns the sequences that share the longest common substring with the user query. In CBIR, the user sends a private image to the server, which compares that image to all of the images that it has and returns the ones that are most similar to the user image.

In one security/threat model, data is processed as follows: the user's sensitive data should be protected and so input data is encrypted. The memory access pattern of the secure hardware could reveal the sensitive data, so memory accesses are made oblivious. Finally, the public records that are touched by the query may also leak information, so the server streams all the records in for any query.

In oblivious computation, one method is to use an Ascend secure processor and path-oblivious RAM – each ORAM access is translated to multiple DRAM accesses. In this case, the ORAM incurs up to 30x access latency overhead. In contrast, with PIR, each public record is touched exactly once and there is no need to store the public record stream in ORAM. The system reads the record that is being processed into the on-chip secure cache and discards it after processing. It then stores pointers to the best records in ORAM. Thus ORAM only stores the working set and is accessed only on cache spills.

PIR is possible with reasonable overhead, trusting only hardware. The queries are fairly general, but we can assume that each record is only touched once. The results were obtained using simulation models and hardware prototyping is underway.

C. Fletcher, M. van Dijk, and S. Devadas, **"A Secure Processor Architecture for Encrypted Computation on Untrusted Programs"**, *ACM Scalable Trusted Computing Workshop (STC)*, October 2012.

http://csg.csail.mit.edu/pubs/memos/Memo-509/memo509.pdf