TWC: Frontier: Privacy for Social Science Research

Information technology, advances in statistical computing, and the deluge of data available through the Internet are transforming social science. With the ability to collect and analyze massive amounts of data on human behavior and interactions, social scientists can hope to uncover many more phenomena, with greater detail and confidence, than allowed by traditional means such as surveys and interviews. In addition to advancing the state of knowledge, the rich analysis of behavioral data can enable companies to better serve their customers, and governments their citizenry.

However, a major challenge for computational social science is maintaining the *privacy of human subjects.* At present, an individual social science researcher is left to devise her own privacy shields, such as stripping the dataset of "personally identifiable information" (PII). However, such privacy shields are often ineffective and provide limited or no real-world privacy protection. Indeed, there have been a number of cases where the individuals in a supposedly anonymized dataset have been re-identified. At the same time, social scientists are increasingly analyzing complex forms of data, such as large social networks, spatial trajectories, and semi-structured text, that are even less amenable to naive attempts at anonymization.

Beyond harm that may be suffered by the subjects themselves, such privacy violations are a serious threat to the future of computational social science research. After a few serious and highly publicized incidents, it may become much harder for researchers to obtain good social science data. Subjects may be reluctant to participate in experiments, data holders may become subject to stifling regulation, and companies may refuse to share proprietary data out of fear of lawsuits or bad public relations.

This project is a broad, multidisciplinary effort to help enable the collection, analysis, and sharing of social science data while providing privacy for individual subjects. Bringing together computer science, social science, statistics, and law, the investigators seek to refine and develop definitions and measures of privacy and data utility, and design an array of technological, legal, and policy tools for social scientists to use when dealing with sensitive data.

These tools will be tested and deployed at the Harvard Institute for Quantitative Social Science's *Dataverse Network*, an open-source digital repository that offers the largest catalogue of social science datasets in the world. Our aim is to provide social scientists with a technological and legal framework that embodies the modern computational understanding of privacy, and a reliable open infrastructure that aids in the management of confidential research data from collection through dissemination.

Intellectual Merit. While data privacy has been examined extensively within the individual fields of computer science, law, statistics, and social science, this project is unique in the way it integrates all of these perspectives, both in identifying the goals for privacy and data utility and in developing tools to achieve these goals. This effort is likely to yield solutions that are more viable in practice than those coming from a single approach, in addition to raising fundamental new questions in the individual disciplines.

Broader Impacts. The tools developed and deployed at the Dataverse Network will contribute to research infrastructure for social scientists around the world. Moreover, the underlying ideas will benefit society more broadly as it grapples with data privacy issues in many other domains, including public health and electronic commerce. In addition, the project will support the development of new curricular material and train a new generation of researchers and citizens with the multidisciplinary perspectives required to address the complex issues surrounding data privacy.

Keywords: Data Privacy; Algorithms; Theory; Social Science; Law; Statistics

1 Introduction

1.1 Computational Social Science

Information technology and the Internet are transforming social science. With the ability to collect and analyze massive amounts of data on human behavior and interactions, social scientists can hope to uncover many more phenomena, with greater detail and confidence, than allowed by traditional means such as surveys and interviews [51, 49]. In addition to advancing the state of knowledge, the rich analysis of behavioral data can enable companies to better serve their customers, and governments their citizenry. The potential benefits of information technology for social science come in several forms:

Digital Traces. Nowadays, human activity leaves a continual "digital trace," in the form of emails exchanged, social network postings and interactions, web-search and browsing histories, cell phone calls, credit-card purchases, video surveillance, and much more, all of which present tremendous sources of data and opportunities for social science analysis. While much of this data is proprietary, some companies may be willing or even eager to share this data with researchers with the right controls in place. For example, Facebook allowed Jason Kaufman, a sociologist and Fellow at Harvard's Berkman Center for Internet & Society, to track social network data and cultural preferences of a full cohort of undergraduates over their 4-year time in college, and link it with student room assignments and demographics [52]. (The privacy issues raised by this dataset sparked the collaboration that eventually led to this proposal.)

Online Experiments. Through systems such as Amazon's Mechanical Turk, the Internet provides an opportunity to run experiments on human behavior and interaction with thousands and even millions of subjects from around the world, at an affordable cost. No longer does experimental social science need to focus primarily on the college-student demographic, but can reach out in novel ways to diverse groups. If these platforms are further enabled with best-practice protocols for data gathering and sharing, the number and scale of online experiments could flourish — producing robust, responsible, lower cost and higher speed results, while supporting consistency and replication [44]

Sharing Data. The Internet makes it extremely easy to share social science datasets, so that they can be analyzed by many different researchers in order to replicate or improve the original analysis, or to ask entirely new questions. This sharing can be done by researchers themselves (e.g. using the Dataverse Network, described below), by companies (such as when Netflix shared movie-preference data to challenge researchers to improve its recommendation engine [10]), or public institutions (e.g. the U.S. Census Bureau regularly publishes aggregate data from its various surveys of the U.S. population).

These technological changes have led to an emerging field of *computational social science*. The development of this field is enabled by centers such as the Harvard Institute for Quantitative Social Science (IQSS), founded and directed by co-PI Gary King. In particular, IQSS developed and hosts the *Dataverse Network* [1] — the world's largest repository for social science research datasets, containing data from over 35,000 studies.

As suggested by the examples above, the benefits of analyzing and sharing human behavioral data are not limited to social science researchers. Companies are increasingly analyzing their customers' data and sharing it with partners in order to provide enhanced services, and public institutions need to do the same for the sake of transparency and accountability. In this project, we focus on *research* data in order to provide concrete goals and a testbed for our efforts, but expect that the ideas and tools we develop will also apply to other contexts.

1.2 The Problem: Privacy

A major challenge for computational social science is maintaining the *privacy of human subjects.*[1] At present, an individual social science researcher is left to devise her own privacy shields, such as stripping the dataset of "personally identifiable information" (PII). However, such privacy shields are often ineffective and provide limited (or no) real-world privacy protection. Indeed, there have been a number of cases where the individuals in a supposedly anonymized dataset have been re-identified. For example:

[1] Although in legal terminology, "privacy" and "confidentiality" are distinct, we use the term "privacy" more inclusively to encompass all issues of collecting, managing, and disseminating sensitive information about individuals across the research lifecycle.

• The best practice for "anonymizing" medical records on patients as late as 1997 was to remove explicit identifiers, such as {name, address, Social Security number}. Co-PI Latanya Sweeney showed that other information, such as {date of birth, gender, ZIP}, which remained in the records could be linked to other publicly available data to re-identify patients. As evidence, she demonstrated how Gov. William Weld's record could be uniquely re-identified by linking his demographics to a publicly available voter list [82]. More generally, her analysis revealed that 87% of the US population is uniquely identified by those seemingly innocent demographics [84]. Her group went on to expose re-identification vulnerabilities in health data, including clinical trial data [89], DNA [63, 61, 62], pharmacy data [87], text (clinical letters and notes) [80] and registry information [81]).

• Netflix released an anonymized version of its movie preference database for the contest mentioned above (challenging researchers to try and improve its recommendation engine). By comparing rental dates and ratings in the Netflix database with reviews posted on the Internet Movie Database (IMDb), Narayanan and Shmatikov [68] were able to re-identify individuals in the Netflix dataset, and thereby learn about their entire rental history (potentially revealing sensitive information about the individuals such as their sexual or political preferences, religious beliefs, substance abuse, etc). As a result of this re-identification, a class-action lawsuit was filed against Netflix, and, as part of the settlement, Netflix cancelled a second planned contest [77].

• Berkman Fellow Jason Kaufman intended to make an anonymized version of his Facebook dataset [52] publicly available in the IQSS Dataverse Network so that other social science researchers could benefit from it. Indeed, this was also an expectation associated with his NSF funding (grant #SES-0819400). Working with the Harvard Institutional Review Board (IRB) and the IQSS, Kaufman engaged in due diligence, making reasonable but ad hoc decisions about how to anonymize the data [53]. However, Michael Zimmer [98], co-PI Latanya Sweeney, and other researchers raised the possibility of student re-identification, and the Berkman Center Cyberlaw Clinic (directed by co-PI Phil Malone) helped assess issues regarding controls on data flow and sharing. As a result, social scientists are currently unable to access Kaufman's database, pending a satisfactory long-term resolution for sharing useable data with a limited community of scholars. This threat is not hypothetical; Sweeney and undergraduate Jeff Solnet recently found a simple method for re-identifying 95% of the students in the dataset (verified by Kaufman).

There are numerous other examples of reidentification vulnerabilities discovered in "anonymized" data, including geographical information system data [34], genetic databases [27, 61, 62], and internet search engine logs [9]. What these examples indicate is that the size, structure and statistical properties of data used in computational social science cause "traditional" approaches to anonymization to fail – yielding data that violates confidentiality, is useless for research, or both.

Beyond harm that may be suffered by the subjects themselves, such privacy violations are a serious threat to the future of computational social science research. After a few serious and highly publicized incidents, it may become much harder for researchers to obtain good social science data. Subjects may be reluctant to participate in experiments, data holders may become subject to stifling regulation, and companies may refuse to share proprietary data out of fear of lawsuits or bad public relations.

1.3 Our Approach

We believe that it *is* possible to move beyond the seemingly bleak privacy situation described above, and achieve wide sharing of social science data while ensuring privacy for the individual subjects. Doing so will require the combined efforts of many disciplines:

Computer Science (PI Vadhan, co-PI Sweeney, Sr. Personnel Chong). There is a growing understanding in the computer science literature of how to distort data so that a desired level of individual privacy is guaranteed even when the adversary has access to unknown auxiliary data.

Social Science (co-PI King, Sr. Personnel Crosas). Before modifying the data for reasons of confidentiality, we need to understand how it will be used in research and the types of analytic methods that could be applied to it, so that we can maintain the utility of the data.

Statistics (co-PI Airoldi). Both "individual privacy" and "data utility" are not absolute notions, but rather need to be measured in a quantitative manner. Statistics provides formal methods for reasoning about the significance of results derived from data, or similarly the certainty with which some sensitive attributes of a user may be compromised.

Law and Policy (co-PI Malone, co-PI Sweeney). New privacy-preserving data sharing technologies will require supportive and enabling governance mechanisms, and the development of appropriate instruments and practices. The policy challenge is to strike a socially optimal balance between creating research opportunity and the nature and degree of associated privacy risks.

A concrete goal for our effort is to design and build tools that can be integrated into the open source Dataverse software and deployed at the Harvard IQSS Dataverse. These tools will provide social science researchers with a combination of technological, legal, and policy means to help ensure the privacy of their subjects as they collect, analyze, and share data. While our target application refers to social science research data, we expect that many of the methods we develop will be useful for other types of data (e.g. health data) and other contexts (e.g. corporate data on customers).

1.4 Research Topics

In the pipeline of computational social science research, data moves through several stages:

$$\text{human subject} \rightarrow \boxed{\text{data collection}} \rightarrow \boxed{\text{data sharing}} \rightarrow \boxed{\text{data analysis}} \rightarrow \text{research results.}$$

The data collection may be done by a researcher running experiments (e.g. on Amazon Mechanical Turk), or a company/institution collecting the data for other purposes. The sharing may be done with specific researchers under bilateral agreements, with the public at large by on-line posting (e.g. in a repository such as the IQSS Dataverse Network), or any of a range of intermediate options (a number of which are supported by IQSS). The data analysis is typically done by social science researchers and statisticians, using statistical software systems such as R and Zelig (co-authored by co-PI Gary King [45] and supported at IQSS). Here data analysis refers to "secondary" analysis done by researchers other than the ones that collected the data, who typically perform and publish their analysis before sharing.

In this project, we aim to augment this pipeline with tools to help satisfy three requirements, which are often in tension with each other:

- Privacy: What parties can learn about a subject's identity and other sensitive information should be bounded within acceptable limits.
- Utility: The accuracy of the research results is not substantially diminished by whatever privacy protections are put in place.
- Efficiency: The processes of data collection, data sharing, and data analysis do not require excessive resources, whether computational, human, or monetary.

The various entities in the research pipeline often have different views on these requirements. Privacy has the greatest direct impact on the subjects themselves, and utility on the data analyst. The entities collecting and sharing the data are often also concerned about privacy and utility, but for different reasons, coming from ethics, liability, and/or public relations considerations. Even if all parties are well-intentioned, the relative weight placed on privacy vs. utility is likely to differ among them. Thus deciding on the "socially optimal" balance should not be left solely to one of the entities, but instead should be done with a broader perspective, guided by appropriate Institutional Review Boards (IRBs),[2] Law, and Best Practices. With this in mind, we propose to investigate the following:

Defining and Measuring Privacy. How should we define and measure privacy in both mathematical and legal terms, and communicate it in language that can be understood by social scientists, IRBs, and policy makers? Of course there is a large computer science literature on mathematical definitions of privacy, starting with the notion of *k-anonymity* introduced by co-PI Sweeney [86, 85] and including the compelling recent notion of *differential privacy* [19, 25, 12, 23]. We wish to extend this line of work by taking a more complete view of

[2] An Institutional Review Board (IRB) is a committee that oversees research involving human subjects, as required at federally funded research institutions.

the research pipeline (including the data collection stage), bridging the gap between mathematical and legal notions of privacy, conveying these notions in lay terms, and exploring alternate definitions of privacy that may be more general (e.g. for data that cannot be decomposed into pieces corresponding to distinct individuals, such as social network data) or more practical (e.g. for contexts where differential privacy turns out to be infeasible).

Defining and Measuring Utility. When do data sharing and analysis mechanisms provide sufficient utility to the data analyst, enabling them to answer the questions in which they are interested with sufficiently accurate results? In the literature, there is a rich collection of differentially private algorithms for natural types of database queries with asymptotic guarantees on natural notions of accuracy. However, it is still not clear whether the types of queries that have been studied so far and the concrete (non-asymptotic) accuracy achieved are sufficient for a typical social scientist or statistician to perform their analyses. We aim to narrow this gap. To do so, we first need to understand the kinds of analyses that are done in practice and the type and amount of accuracy that is required. Then we can try to match existing algorithms to these needs, both theoretically and in practice, or design new algorithms as needed.

Tools for Privacy. Informed by the above, can we develop an array of technological, legal, and policy tools to enable a better balancing of privacy and utility in social science research? We will do so by designing new algorithms (in particular trying to overcome computational complexity barriers identified by PI Salil Vadhan [24, 92]); recommending best practices for researchers and IRBs; developing custom licensing, contracts, and other legal agreements tailored to the specific needs of entities in the social science research pipeline; and implementing and experimenting with a number of these tools at the Dataverse Network.

All of the above research directions require the participation of multiple disciplines from computer science, social science, law, and statistics, and indeed each of the co-PIs will be involved in more than one.

2 Institutional Context

This project is part of a collaboration between the Center for Research in Computation and Society (CRCS) in the Harvard School of Engineering and Applied Sciences (SEAS); the Institute for Quantitative Social Science (IQSS), a university-wide center that is based in the Harvard Faculty of Arts & Sciences; and the Berkman Center for Internet & Society, a university-wide center that was founded at Harvard Law School.

CRCS was founded in 2004 to drive innovative computer science research and technology towards problems of importance to society. It has done this by bringing Harvard's computer science faculty and students together with postdoctoral fellows and visiting scholars that have expertise in relevant areas, and hosting an interdisciplinary seminar series to enable interaction with social scientists, legal scholars, policy makers, and other computer scientists. In addition to numerous publications by faculty and fellows in first-tier computer science venues, past contributions of CRCS include the Helios web-based open-audit voting system by CRCS fellow Ben Adida [3], a workshop on Data Surveillance led by CRCS fellow Simson Garfinkel that led to a special issue of *IEEE Security & Privacy* [30], and the development of courses on privacy, usable security, and cryptography co-taught by CRCS fellows. The current activities of CRCS are structured around two focus areas: Privacy & Security and Economics & Computer Science. PI Salil Vadhan served as the faculty director of CRCS from 2008 to 2011, has extensive experience in the foundations of cryptography, and has recently been very active in theoretical research on differential privacy. Co-PI Latanya Sweeney has been a CRCS Visiting Scholar for three years, during which has she developed collaborations within both CRCS and Berkman, and brings extensive expertise in data privacy technology and policy. CRCS faculty member Stephen Chong (Sr. Personnel) researches language-based information security, including studying the specification and enforcement of expressive and practical confidentiality policies, and has recently collaborated with PI Vadhan on the interaction between differential privacy and economic mechanism design [17].

IQSS was founded in 2005 by co-PI Gary King as a university-wide institute with a dual scientific mission. First, IQSS catalyzes research to understand and solve major problems that affect society and the well-being of human populations, by bringing together diverse researchers and approaches from multiple disciplines. Second, IQSS develops analytical tools for social and health sciences, focusing on open collaborative tools for computational social science, statistical analysis, and data sharing and preservation. One of IQSS's initiatives is the Dataverse Network project – which is open source, web 2.0 software for data sharing, preservation, citation, and analysis now hosted at universities around the world. Each Dataverse Network distributes virtual archives (called "dataverses") to hundreds of researchers and institutions; each dataverse provides all the services of a professional archive on your web site, and with your branding, but without having to install anything locally. IQSS hosts an instance of the Dataverse Network software which now comprises the largest catalogue of social science research data in the world. Dataverse Network software also integrates with Zelig, developed by co-PI King and his collaborators [45], which offers a unified framework for applying the wide variety of statistical analyses implement in the statistical language R. Sr. Personnel Merce Crosas is Director of Product Development at IQSS. She has led the design, architecture and implementation of the Dataverse Network since the project started 6 years ago, and has given multiple talks and training sessions on data sharing, analysis, management and preservation. She has contributed to defining and implementing policies for the management and dissemination of public and confidential research data for IQSS and the Dataverse Network. In 2011, co-PI Sweeney's *Data Privacy Lab* moved to IQSS from Carnegie Mellon University, reflecting IQSS' strong interest in taking on privacy-sensitive datasets. Since its founding in 2001 at CMU, the Data Privacy Lab has had a substantial impact on both privacy technology and privacy policy.

The Berkman Center for Internet & Society was founded at the Harvard Law School in 1997 and is dedicated to pursuing the highest quality teaching and scholarship about the impact of Internet technologies and their use in society. Its mission is to recognize, study, and engage the most fundamental problems of the digital age, and to share in their resolution in ways that advance the public interest. The Berkman Center investigates the real and possible boundaries in cyberspace between open and closed systems of code, commerce, governance, and education, and the relationship of law to each. It does this through active rather than passive research, believing that the best way to understand cyberspace is to actually build out into it. Indeed, the Berkman Center has deployed several widely used systems, including HerdictWeb, a crowdsourced platform for compiling reports of internet censorship, and StopBadware, which provides data on the spread of malware and serves as a community forum for researchers and the public. The Berkman Center Cyberlaw Clinic, directed by co-PI Malone, provides innovative, hands-on training and course credit to Harvard Law students who, under careful supervision, offer legal and policy research, guidance and representation to a variety of real-world clients, including research projects and institutional entities. Many of the Clinic's clients and projects have directly involved data and information privacy and security, and co-PI Malone and the Clinic's students frequently address novel questions about the practical application of privacy and data security laws, regulations and legal instruments to social science research initiatives. It counsels research projects on system design and legal measures to assess and mitigate privacy concerns raised by the collection and sharing of data and on compliance with the new Massachusetts Data Protection regulations and other state data breach laws. The Clinic also drafts privacy policies, data sharing agreements and other legal agreements to assist data control and privacy.

Co-PI Edo Airoldi is a member of the Harvard Department of Statistics. His research agenda focuses on statistical methodology and theory with applications to problems in genomics, the computational social sciences, and statistical analysis of large biological and information networks. He has recently developed statistical approaches to disclosure risk assessment, the effects of location access behavior on re-identification risk in a distributed environment, and protocols for multi-party data analysis with malicious participants [60, 31, 58, 59, 4]. He has written a joint paper with co-PI Sweeney [31] and has a joint PhD student with co-PI King. In Summer 2011, an undergraduate in his lab (Stephen Kent) worked with a PhD student (Jon Ullman) of PI Salil Vadhan on preliminary work for this project. (See Section 5.1.)

In Fall 2009, CRCS and Berkman formally joined their fellowship programs to enable the greater interdisciplinary interaction sought by both sides. CRCS and Berkman faculty and fellows have shared a weekly "Fellows hour," a weekly discussion seminar, and a biweekly technical seminar series, and interdisciplinary research group meetings on Data Privacy, Trustworthy Crowdsourcing, and Data Marketplaces. This proposal emerged from the research group meetings on Data Privacy, along with the shared experience of the three centers (CRCS, Berkman, IQSS) on Kaufman's Facebook dataset [52] mentioned earlier. Co-PI Sweeney's Data Privacy Lab at IQSS now hosts the Data Privacy Meetings, as part of its "Topics in Privacy" series.

In addition to the funded personnel and institutes mentioned above, we also plan to collaborate extensively with Cynthia Dwork (Distinguished Scientist, Microsoft Research) and Micah Altman (Director of Research, MIT Libraries). Dwork is one of the primary founders of differential privacy, has collaborated with PI Vadhan extensively for the past several years, and has recently worked with a group of CRCS, Berkman, and IQSS researchers on bridging the computer science and legal formulations of privacy [94]. Until recently, Altman was a Senior Research Scientist at IQSS and Archival Director of the Murray Research Archive, responsible for the management and dissemination of confidential research data. He has been collaborating with the personnel on this project for the past two years, also including [94].

A small amount of seed funding ($200k) from Google has enabled the personnel to start working on preliminary aspects of this project over the past year.

3 Measuring and Defining Privacy

The first component of our research will involve defining and measuring privacy in both mathematical and legal terms, and communicating it in language that can be understood by social scientists, Institutional Review Boards (IRBs), and policy makers. While there has been much progress on definitions of privacy in recent years, there are still important gaps that need to be filled in order to have a robust approach to data privacy in contexts such as social science research.

3.1 Background

Traditional legal formulations of data privacy focus on the *identifiability* of data, meaning whether a particular individual can be linked to a record in a dataset. For example, one of the Fair Information Practices [2] is to "provide the least identifiable version of the data needed." Traditionally, de-identification has been interpreted to mean stripping a dataset of "personally identifying information (PII)." For example, the HIPAA Privacy Rule specifies 18 fields (e.g. name, address, birth date and month, biometric identifiers, etc.) whose removal render a dataset presumptively de-identified.

Most social science research is governed by the "Common Rule" (Code of Federal Regulations, Title 45, Part 46, Subpart A), which is aimed at protecting human subjects in federally funded research. The Common Rule includes requirements for obtaining the informed consent of participants, and for the operation of Institutional Review Boards (IRBs), which are tasked with determining whether research studies at a given institution are ethical, and in particular assessing the level of risk posed to the subjects. With the shift towards data-driven, computational social science research, IRBs are increasingly being forced to evaluate "informational risks" — those that arise from inappropriate use or disclosure of information (i.e. failure of data privacy or security). Positing that most current IRBs are ill-equipped to evaluate informational risks, the Department of Health and Human Services (HHS) has recently proposed to adopt the HIPAA de-identification standard in the Common Rule (i.e. removal certain fields would render a dataset presumptively de-identified and thus free of informational risks) [93]. In addition to providing a simple rule that is easy for researchers and IRBs to apply, this proposal has the potential advantage of harmonizing the Common Rule and HIPAA (both of which cover research done with data that is gathered in a clinical setting).

Unfortunately, as the examples discussed in the introduction illustrate, removal of PII does not suffice to render a dataset anonymous. The numerous re-identifications by co-PI Sweeney and others show that data that is seemingly anonymous on its own can often be linked with externally available datasets to reidentify individuals. It is extremely difficult for even a privacy expert, much less a social science researcher or company wishing to share data, to anticipate the kinds of datasets that will become available for linkage or the clever re-identification methods that an adversary may employ in the future. And it is difficult to estimate the degree of harm posed by potential linkages — which often leads institutions to restrict data based on the worst case harm. Thus it is important to have *formal protection models* that, if satisfied by an algorithm used to anonymize data or answer queries about data, provide *general* privacy assurances regardless of what methods or data a future adversary may employ.

Co-PI Sweeney formulated one of the first formal mathematical models of privacy, known as *k-anonymity* [86, 85], designed to protect against the kind of linkage attacks that have been so successful in the past. This model was strengthened in work by Machanavajjhala et al. [55] to yield the model of *ℓ-diversity*. Many other privacy models have been proposed in the literature; for a review see [29].

In recent years, a compelling new model for data privacy has emerged, known as *differential privacy* [19, 25, 12, 23]. (See also the surveys [20, 21].) Let M be a randomized algorithm that takes a dataset D and produces some information about it, which could be an answer to a given query, an "anonymization" of D, or some other statistical summary of D (e.g. a "contingency table"). We say that M is *differentially private* if for every two datasets D and D' that differ on a single individual's record, the output distribution of $M(D)$ is "similar" to the output distribution of $M(D')$. Specifically, we say that M is *ε-differentially private* if for every set T, we have $e^{-\varepsilon} \cdot \Pr[M(D') \in T] \leq \Pr[M(D) \in T] \leq e^{\varepsilon} \cdot \Pr[M(D') \in T]$. Here ε should be thought of as a small constant (e.g. 1/10), in which case $e^{\varepsilon} \approx 1 + \varepsilon$. Intuitively, this protects an individual's privacy because the output of M looks essentially the same whether or not that individual's data is included in the database. Note that this interpretation holds regardless of how much externally available information an adversary has. Thus, differential privacy is considered a very strong privacy protection model.

Researchers have discovered differentially private algorithms for a wide variety of computations that one might want to use to analyze data, such as histograms [23], contingency tables [8, 32], machine learning [12, 46], logistic regression [16], and clustering [12]. Typically, these algorithms produce approximate answers, but the error introduced is often asymptotically smaller than the statistical error that would be present in any database obtained by sampling a population. The rich collection of differentially private algorithms illustrates another benefit of having a formal protection model — it encourages innovation, because it provides a clear criterion within which researchers can freely explore the design space without having to guess whether their algorithm will eventually be deemed "private".

3.2 Bridging Mathematical and Legal Definitions of Privacy

Given the current interest in revising privacy regulation such as the Common Rule and the substantial computer science advances in understanding data privacy, it is an opportune time to try and bridge the mathematical and legal conceptions of data privacy. For this reason, our group recently submitted extensive comments on the proposed revisions to the Common Rule [94, 90]. Our main thesis is that the vast majority of privacy regulations, including HIPAA and the proposed changes to the Common Rule, are flawed due to an implicit assumption that all data is *microdata*, meaning a collection of records, each of which pertains to a single individual (or household or business). Even if one's original dataset is in microdata form (which does not hold for social network data), the data can be processed into many other forms before sharing — such as contingency tables, synthetic datasets, data visualizations, interactive mechanisms, multiparty computations — rendering rules designed for microdata inapplicable.

Sharing data in forms such as those described above (rather than restricting to microdata format) is attractive from the perspectives of both data utility and data privacy. For data utility, the above formats often allow for researchers to obtain meaningful answers to questions that they cannot obtain using microdata that has been de-identified in the manner HIPAA requires. For example: As HIPAA requires generalizing birthdates to the year, and geography to the state level, appropriately de-identified data could not be used to answer questions like, "How many babies were born with birth defects in Dauphin County, Pennsylvania during the three months after the Three Mile Island nuclear meltdown?" Nonetheless, an (approximate) answer to this question is unlikely to violate privacy – and could be of significant utility in public health research.

In addition, many of the above forms of data sharing can often provide much stronger privacy protections for subjects than de-identified microdata. As discussed earlier, robust de-identification of microdata by removing and generalizing fields is quite difficult, and this has led to a heated debate among privacy law scholars about how to balance the risks and value of data sharing in a de-identification regime [69, 97, 15]. In contrast, all of the non-microdata formats mentioned above have been successfully used in practice to share data while protecting privacy. For example, synthetic data has been used by both the U.S. Census Bureau [56, 50] and the German IAB [71], and multiparty computations have been used to aggregate data across homeless programs [88] as well as in industry [48]. Moreover, many of these forms of data sharing have been shown to be compatible with differential privacy [23, 22, 8, 13]. Although no form of sharing is completely free of risk, it seems clear that we would want to make non-microdata forms of sharing an option for researchers in cases where they offer both better privacy and better utility than HIPAA-style de-identification.

Unfortunately, the HIPAA Privacy Rule (indeed, most privacy regulation) provides no guidance on how to evaluate privacy protections when data is shared in non-microdata formats, and it is difficult to imagine a simple rule that would be appropriate for all contexts. Methods for sharing privacy-sensitive data should be tailored to the structure of the data (e.g. standard relational microdata vs. social network data vs. text), the sensitivity of the information and potential harms of disclosure, the level of consent obtained from subjects, and the intended recipients of shared data. Indeed, sharing with researchers governed by IRBs, sharing with the public, and sharing under limited data-use agreements should all be treated differently. On the other hand, a case-by-case approval process by IRBs or by expert statisticians is likely to be inefficient, time-consuming, and inconsistent. Thus, in [94] we proposed that a body of experts be tasked with maintaining an evolving *safe-harbor catalogue,* listing a variety of methods that are deemed to provide sufficient privacy protection in a variety of corresponding contexts.

We view these comments as simply a first step in bridging the computer science and legal understandings of data privacy. Our comments focus mainly on reconciling differences in *scope* (microdata vs. a broader conception of data). Beyond that, we will examine the deeper question of what the requirements should be and the normative basis for such requirements (e.g. rights-based vs. harms-based). One issue we will explore is how to describe privacy protections in a way that facilitates weighing them against the social value of the research being done (since no mechanism that provides utility can provide perfect privacy protection, or prevent individuals from being harmed by use of the global knowledge gained). Another issue is how to determine, for a given data collection process, *what* information in the resulting dataset should be protected. The standard notion of differential privacy protects any information that is “localized” to a small number of records in a database, since traditional datasets associate one record per individual. But in other contexts this is insufficient, and we need variants of differential privacy. For example, in social networks, a record may correspond to a single edge or relationship in the network, but we really want to provide privacy at the level of individuals, which correspond to nodes (each of which can involve many edges). This leads to the notions of “edge-level privacy” (weak) and “node-level privacy” (strong) considered by Hay et al. [43]. Similar issues can arise also in datasets which at first seem to be of the traditional “one individual per record form” — for example, results from a small-town survey that asks everyone about a particular individual X’s alcoholism status (so information about X is spread throughout the dataset). We expect that exploring these kinds of issues will be informative for both the computer science and legal understandings of privacy.

In addition, we will continue to engage on the policy side, as the process for revising the Common Rule and other privacy regulation moves forward over the next few years. Part of this effort will involve refining our proposal of a safe-harbor catalogue, and exploring which existing methods and contexts are sufficiently well-understood to be included in such a catalogue at present.

3.3 Estimating Privacy Risk

While it is preferable to share data using algorithms that are mathematically proven to satisfy a strong privacy protection model, sometimes such guarantees come at too great a cost in data utility. Thus, it is also of interest to be able to heuristically estimate the privacy risk in a (de-identified) dataset given the information that is likely to be available to an adversary at present or in the near future. This is analogous to how work in cryptanalysis (breaking cryptographic algorithms) is a useful guide for the design of heuristic cryptographic algorithms when existing ones with provable guarantees are insufficiently efficient for some practical applications. Also heuristic estimates of privacy risk can also provide a guide for setting the privacy parameter (e.g., ε in differential privacy or k in k-anonymity) when using a formal protection model, similarly to how cryptanalysis guides the choice of key length even for provably secure systems.

Co-PIs Airoldi and Sweeney have extensive experience in estimating privacy risk. Given a description of the fields of a dataset and the population from which subjects are drawn, co-PI Sweeney introduced a computational model for predicting the number of individuals in the dataset that could be uniquely identified (or identified among some small number of others). This model (“the Risk Assessment Server”) uses population demographics and meta-data about other available datasets to infer identifiability risk by computing inferential linkages across the datasets, and is the primary method used in practice for determining whether data are sufficiently de-identified under the HIPAA Privacy Rule. Co-PI Sweeney also introduced the Iterative Profiler [81], which sifts through publicly available data, using inferential linkages of data fragments across data sources to construct profiles of people whose information appears in the data. After it re-identified the names of children who appeared in a cancer registry, an Illinois court ordered the approach sealed.[3]

Starting with Sweeney's work, co-PI Airoldi has investigated the statistical basis of various re-identification algorithms. He has shown how the strength of these algorithms depends on statistical properties of how attributes occur across the population and within a database; intuitively, heavy-tailed attributes lead to greater re-identification risk than uniform ones [59]. He has also quantitatively characterized how these algorithms leverage statistical properties of the dissemination of individual information in distributed databases [58].

We will combine this body of results with previous work on individual measures of utility, which have been explored in an image de-identification context [31] and need to be generalized, to produce a new framework for exploring the privacy-utility trade-offs at the individual level. We will develop methodology to assess risk exposure of individuals with respect to individual databases in which their data is recorded, in the context of the identified and de-identified data available to the public and to private corporations for a price. We will develop publicly available tools to explore possible futures and what-if scenarios in terms of the increasing amount of data made available and the implied individual privacy risks. Importantly, the proposed methodology does not depend on knowing which identified data sets will be available to attackers in the future. Rather, re-identification risk projections only depend on the number of *quasi-identifiers* that will be present, or on their projected growth as a percentage of the fields in de-identified records, and on the distribution of the population over possible combinations of quasi-identifier values. This effort will build on a recently proposed theory of re-identification in a distributed environment, which leverages co-PI Airoldi's entropy measures of risk [4] within the framework of trail disclosure [57].

4 Measuring and Defining Utility

As hinted in the previous section, computational privacy protection mechanisms inherently involve a tradeoff between privacy and utility. That is, answers to queries on the data or the data elements themselves are distorted in some way to provide privacy. For example, k-anonymity is often achieved by generalizing and suppressing cells in a database [75, 85], and differential privacy is often achieved by adding noise to statistics computed on the database [12, 23].

Thus, a goal is to achieve an optimal privacy-utility tradeoff. For any desired level of privacy, we want to have data that is accurate as possible. However, the appropriate metric for "accuracy" will generally depend on the computation being done and the particular application. Existing work on differential privacy and k-anonymity has sought to optimize privacy-utility tradeoffs for natural measures of accuracy (e.g. expected squared error [78] or the number of cells suppressed [66]). However, there remains an empirical question of how well these utility measures correspond to what quantitative social scientists and statisticians need to do their work, and whether the concrete, non-asymptotic error introduced by the algorithms is affordable.

We plan to address this question in a series of steps. First, for each dataset it contains, the IQSS Dataverse Network maintains a list of papers published using that dataset. Thus, we plan to examine a sample of these papers, and for each, understand the kinds of computations that needed to be done on the dataset to produce the published results. If there are privacy mechanisms in the literature (e.g. satisfying differential privacy or k-anonymity) that are appropriate for the analysis, then we plan to reproduce a sample of the results using those mechanisms. If the results are not substantially affected, then the privacy mechanisms are providing high utility (which is good news). However, we expect that in a number of cases, the results will deteriorate significantly under privacy mechanisms, or the literature will not even have a private algorithm for the kind of analysis that was done. Both of these situations lead to interesting research questions that we would pursue. In the former situation, we would ask whether the measure of utility in the literature is the right one for the particular analysis at hand and whether the known algorithms are optimal with respect to the right measure. In the latter situation,

[3] http://www.state.il.us/court/Opinions/AppellateCourt/2004/5thDistrict/June/Html/5020836.htm

we would try to design a new algorithm for the given data analysis task. We expect this to occur quite often when we don't have a traditional row-structured dataset with numerical or categorical fields, but have other types of data, such as social network data, text, or media.

Second, we note that being able to reproduce even all of the results in the quantitative social science literature with private computations does not imply that social scientists can do their work given only differentially private or k-anonymous access to the data, as the final published results do not reflect all of examination of the data that led the researchers to the final analysis that they published.[4] Thus, in addition to the above, we will also carry out usability studies, with real students and researchers in quantitative social science and statistics using real datasets in the Dataverse Network, to see how well they can work with the privacy tools.

Finally, we aim to eventually classify a larger collection of datasets on the Dataverse Network into categories based on types of data they contain and the statistical analyses they were intended to support, and develop a set of benchmark analyses for testing the utility of private data analysis mechanisms in the future.

5 Tools

Once we have determined our privacy and utility goals, we aim to develop an array of technological, legal, and policy tools to enable a more informed and better balancing of privacy and utility in social science research.

5.1 Algorithm Design and Testing

We will invest a substantial effort in developing, implementing, testing, optimizing and refining a variety of algorithms for privacy-preserving analysis and sharing of data, as well as understanding the limitations on such algorithms. We will focus on algorithms that we expect to be particularly useful for social science research, including the following:

Statistical Estimation. *Zelig*, developed by co-PI King and his collaborators [45], is a tool for statistical analysis that offers a simple and unified framework for applying a wide variety of statistical estimation algorithms in the statistical language R, and is commonly used by social scientists. Zelig is already supported on the IQSS Dataverse Network, meaning researchers can run Zelig methods on datasets without downloading them locally.

Most of Zelig's functionality matches remarkably well with a recent paper of Smith [78], which presents a general technique for converting (non-private) statistical estimation algorithms into differentially private ones. The advantage of Smith's technique is that it does not require a case-by-case investigation of each statistical estimation strategy. Rather, it applies if the original, non-private strategy has a certain kind of syntax, which happens to be provided by the Zelig framework. Privacy is always guaranteed, and utility (namely that the private statistical estimator converges to the same distribution as the non-private statistical estimator) is guaranteed provided that the original algorithm satisfies certain natural properties, e.g. "asymptotic normality." Thus, in principle, we can hope to construct a single "wrapper" around Zelig, and at once make many differentially private statistical analyses available to social scientists on the IQSS Dataverse Network.

The catch, of course, is that Smith's analysis refers to asymptotic behavior as n, the number of data samples, tends to infinity. Thus, we will attempt to optimize Smith's technique to provide useful results for datasets of practical size. Over the past year, Jon Ullman (Ph.D. student of PI Vadhan) has worked with undergraduates Stephen Kent (from co-PI Airoldi's lab) and Rebecca Goldstein (undergraduate RA at IQSS) to implement a prototype of Smith's technique in R, and have run initial experiments to test it with Zelig's implementation of logistic regression. The results of initial experiments are promising but indicate that substantially more work needs to be done. For example, when performing logistic regression on $d = 3$ covariates and $n = 10,000$ data points (synthetically generated), use of the differential privacy wrapper with privacy parameter $\varepsilon = 3$ only makes the per-sample log-likelihood about 10% worse than non-private logistic regression. This result was achieved with very little optimization of the algorithm. After studying the sources of inaccuracy in Smith's technique and making the appropriate optimizations, we aim to improve all the parameters (increase d, reduce

[4] We note that Cynthia Dwork, Moni Naor, Guy Rothblum, and Jon Ullman (Ph.D. student of PI Vadhan) have some promising initial results for how to provide differential privacy even when the data analyst is allowed to look at the raw data, but is trusted to not to reveal any information beyond the final published results.

n, and reduce ε) by an order of magnitude each. By reducing n to a few thousand, we will able to handle the large datasets in the IQSS DVN, such as the General Social Survey [79]. We note that Chaudhuri, Monteleoni, and Sarwate [16] have previously worked on differentially private logistic regression, and achieved promising results for data of much higher-dimensionality ($d \approx 100$) and for much better privacy parameters (ε ranging from .01 to .5), on somewhat larger datasets (n ranging from 45,000 to 600,000). A key difference is that their methods are tailored for statistical estimation strategies that utilize "Empirical Risk Minimization," whereas we are examining methods that apply to a much wider family of strategies. Of course, if we discover that the general methods cannot be sufficiently optimized to provide practical results, we may need to turn to more specific approaches.

Managing the Privacy Budget. One of the common concerns with differential privacy is that the level of privacy protection degrades as more queries are asked about a particular dataset. Eventually, the mechanism must stop answering questions (or distort the answers so much that they are useless). Thus, an important problem is to ensure that the "privacy budget" is not depleted too quickly. The original paper that defined differential privacy [23] showed that the privacy deteriorates at most linearly with the number of questions answered. With Dwork and Rothblum, PI Vadhan [26] has shown that the privacy actually deteriorates only with the *square-root* of the number of questions, generalizing previous results known for specific differentially private mechanisms [19, 25]. And a series of beautiful results [13, 24, 72, 26, 41] have shown how to construct particular differentially private mechanisms where the privacy deteriorates only *logarithmically* with the number of questions asked (i.e. the mechanism can answer a number of queries that is *exponential* in the size of the database).

We will attempt to combine these techniques with differentially private statistical estimation, as discussed above. Another approach to the privacy budget that we will explore is via modelling of the adversary. If we have each data analyst register with the system, limit the number of questions asked by each analyst, and can be reasonably confident that analysts do not collude in large groups to compromise privacy, then we can give each analyst their own privacy budget.

Synthetic Data Generation. A *synthetic or modified dataset* that contains incorrect, suppressed, or generalized information, but reflects a wide variety of statistical properties of the original dataset, e.g. as in [75, 85, 8, 13, 56]. Synthetic data is a very attractive option. The original, non-private data can be discarded or kept in a very secure location, and the data analysts can work with the synthetic data as if it were the original (eliminating all privacy budget issues!). Thus synthetic data has long been considered as a viable approach to privacy in the statistics community [74, 28], albeit generally without formal protection models. A remarkable result of Blum, Ligett, and Roth (BLR) [13] shows that it is possible in possible to produce *differentially private* synthetic data that approximately preserves a huge collection of statistics about the original dataset (even the full contingency table, with all multi-way correlations between attributes). Unfortunately, PI Vadhan has shown producing such synthetic data inherently requires computation time exponential in the dimensionality of the data (under standard cryptographic assumptions) [24, 92]. (There are also computational intractability results for producing accurate k-anonymous datasets using suppression and generalization [83, 85, 66, 11].)

Nevertheless, producing differentially private synthetic data has been shown to be practical for low-dimensional data sets of interest, even with under 700 records [40]. We will explore whether these algorithms can produce useful results on social science datasets in the IQSS Dataverse Network, and whether further theoretical improvements can be obtained under reasonable assumptions about the statistical properties of the dataset (the aforementioned complexity lower bounds are for worst-case utility guarantees), as has been partly explored in [41]. In addition, we will examine algorithms for producing differentially private synthetic graphs, such as the work of PI Vadhan's PhD student Jon Ullman [33] which approximately preserves all "graph cuts," as such algorithms may be useful when applied to social networks.

We will also explore algorithms that modify datasets so as to minimize the statistical measures of risk discussed in Section 3.3. While these may not provide the attractive worst-case privacy guarantees of differential privacy, they may be best choice when differential privacy is found to be impractical and the social value of the data sharing is deemed to be sufficiently high.

Data Summaries. The aforementioned results showing that producing synthetic data requires exponential time (in the worst case) do not rule out the possibility of producing other types of differentially private *data summaries* that allow for answering a huge (even exponential) number of questions about the original dataset. Indeed, building on [42], PI Vadhan and PhD students [91] have recently shown how to take a dataset with d boolean attributes per row and an integer $k \leq d$ and produce in time roughly $d^{\sqrt{k}}$ a summary that allows one to approximately compute any k-way marginal (the fraction of rows with given values for a given set of k attributes). This substantially improves over the time roughly 2^d needed to produce synthetic data even for $k = 2$ [8, 92] and the time roughly d^k needed to explicitly compute and add noise to all k-way marginals. We will test the practical efficiency of this algorithm, and try to find better algorithms (e.g. one running in time polynomial in d and k).

5.2 Algorithm Implementation

As discussed throughout the proposal, a concrete goal for our work is to implement algorithmic and legal tools for privacy protection at the IQSS Dataverse Network (DVN) and make them available for use by the social science research community. For the algorithms, the first step will be to write prototypes of the algorithms in the statistical programming language R. These prototypes will enable us to test the algorithms' performance on datasets in the DVN, map out the privacy-utility-efficiency tradeoffs, run usability studies with social scientists, and iteratively refine and optimize the algorithms.

For the algorithms that emerge as useful in practice, the next step is to design and produce a secure implementation. For this, we will draw upon the variety of recent efforts to bring differential privacy closer to practice, such as the PinQ programming language for private-integrated database queries [65], the Airavat system for MapReduce computations [73], the generation of synthetic commuter data from the Census [56], and the Fuzz language for differentially private computations [35]. As demonstrated by [35], even closing known side channels when providing differential privacy is a highly nontrivial task. Sr. Personnel Chong's expertise in language-based security will be crucial in this process. His group will work with Sr. Personnel Crosas and an IQSS software developer to to design a secure implementation that will interface with the DVN software.

The final step is to make the implemented algorithms available for public use. Our algorithms for automatic risk assessment (as discussed in Section 3.3) may be integrated into the ingest tools provided by the DVN in order to alert researchers to reidentification risks in data that they deposit. Our algorithms for private data analysis and sharing may be integrated into the Zelig package, in order to support dynamic analysis of sensitive data in the system. Integrating these with Zelig will make them available both for use in the DVN system and will enable use with other systems using the R programming language for statistics. Our goal is to introduce our private data analysis and sharing tools as an *additional* mechanism for accessing privacy-sensitive datasets in the DVN. Prior to this project, the IQSS Dataverse Network planned to support access to privacy-sensitive data by offering access control mechanisms that would be controlled by the researchers who deposited the data set (and might be very stringent, e.g. requiring IRB approval). We aim to enable more people to study some of these datasets using our tools. (For example, a graduate student might do some preliminary analysis using our tools, and if it turns out to be promising, the advisor would go through the process of obtaining full access.)

5.3 Legal Instruments

Social science research with privacy-sensitive data is greatly hindered by the lack of an effective legal and regulatory framework. As discussed in Section 3.1, researchers and IRBs operating under the Common Rule are ill-equipped to evaluate privacy risks. Companies that wish to share data with researchers are faced with a complex and uncoordinated collection of laws, regulations, and precedents that vary across jurisdictions and according to the nature of data (e.g. health information vs. video rentals vs. census data). Most American laws and regulations addressing data gathering and sharing are premised on some form of notice and consent, a flawed approach even without the additional challenges presented by re-identification.

With our interdisciplinary mix of computer scientists, legal scholars, and social scientists, we expect to address these legal shortcomings in several ways. First, as discussed in Section 3.2 in the case of the Common Rule, we will explore and propose new legal and regulatory frameworks for balancing research utility against the potential harms of particular practices. This includes a more specific set of best practices for researchers and IRBs deciding how to handle privacy-sensitive data that will, in part, articulate in non-technical terms what the guarantees of various formal protection models (such as differential privacy) mean, and what an IRB should consider (e.g. about the population and data being collected) to determine whether the model would provide sufficient privacy, balanced against the social benefits of the study.

Next, we will develop a suite of legal tools to facilitate and complement the computational privacy tools developed through this project. These legal instruments will include custom policies, licenses, contracts, and other legal agreements carefully tailored to the specific needs of researchers (and their subjects) working with specific types of data under different technical approaches. For example, the computational tools we will build and deploy at the Dataverse Network must be accompanied by a facilitating and supporting set of legal instruments designed to ensure their consistent application in appropriate cases. These legal instruments also should provide additional protection, such as transparent communication of subjects' expectations and consent, proper limitations on use and sharing of the data, protections for data integrity and security, and/or audit mechanisms to regulate compliance, in instances where assurances of anonymity are more difficult. Designing these legal instruments will build on previous work by the Berkman Center and Cyberlaw Clinic in evaluating and drafting licenses and policies to secure protection of data where anonymization was uncertain, and on current approaches by Harvard and other institutions to enhance privacy protection.

For implementing the legal instruments, the Dataverse Network already currently provides the capability of associating data and collections with specific, machine-actionable, terms of use. We will add support for associating data with sensitivity-based licenses and terms of use. Creative Commons (which was founded in part through an early gathering at the Berkman Center and included Berkman Center faculty) is an exemplar of modular, standardized licensing terms with machine-parsable metadata, and we will draw upon their example. We will also explore the design of a richer language for specifying privacy policies, drawing on Sr. Personnel Chong's extensive experience with specifying and enforcing information flow policies (e.g. [18]).

6 Curriculum Development, Education, and Outreach

The research on Privacy for Social Science Research funded by this proposal will drive the creation of new educational materials and modules in Harvard's ongoing effort to enhance awareness of the opportunities and risks associated with the development of digital culture. Our existing and planned undergraduate, graduate and open-access curricula will immediately incorporate new ideas developed through this grant on how and why the information all around us can be made trustworthy and used for personal and public benefit. The materials will likewise be inserted in the stream of other events, publications, communications and other interactions that have become core to the University's efforts to disseminate.

The investigators on this proposal teach a number of courses related to data privacy; their teaching and research efforts will naturally inform each other.

Co-PI Sweeney co-teaches the course Computer Science 105 "Privacy and Technology," which was developed by CRCS faculty and fellows. This course examines several areas in which privacy and technology are thought to be in conflict. In the course, students from disparate disciplines jointly explore whether these conflicts are real and, if so, discuss what could reasonably be done about them in the areas of both technology and policy. The reidentification of Kaufman's Facebook dataset discussed in Section 1.2 began with a CS 105 course project by undergraduate Jeff Solnet. The research in this proposal will likely suggest a variety other student projects for the course. In Spring 2013, PI Vadhan will teach a new, graduate-level course on Mathematical Approaches to Privacy (focusing on Differential Privacy). This course will provide training to students and postdocs interested in working on this project, and inform other students about this rapidly evolving line of work. Sr. Personnel Chong teaches a course on the use of programming language techniques to achieve information security guarantees.

Numerous recent Berkman Center-taught courses addressing in part various privacy and technology/internet issues have been offered at Harvard College, Harvard Law School, Harvard Extension School and jointly with MIT and Northwestern, often streamed openly and shared online. Co-PI Phil Malone teaches an annual Freshman Seminar at Harvard College that includes consideration of privacy policy and regulation, as well as at

Harvard Law School a Reading Group on Technology and Privacy and other courses in which legal, regulatory and technological approaches to information privacy and security are explored. His law students in the *Cyberlaw Clinic* learn the intersection of law and privacy first-hand as they solve real-world privacy problems for real clients, counseling researchers and others on data security, compliance with security and privacy laws, drafting privacy policies, developing licenses and other legal instruments to protect sensitive data, among other projects. The students of the Cyberlaw Clinic will be an integral part of the current project, as the Clinic will design and implement the practical legal and policy instruments to be deployed at the Dataverse Network.

As the research in this proposal informs the development of educational materials in our existing and new courses, we will also continue efforts to reach a wide range of audiences beyond the Harvard campus. In particular, we propose to record, tag, and edit videoed lectures on Data Privacy from a variety of courses and make them public through Harvard's Open Learning Initiative.[5] The materials will be made available as modules that can be reassembled for use in targeted contexts for specific audiences.

In addition, we will also invest substantial effort in training the next generation of researchers to take a broad, multidisciplinary perspective on data privacy issues. Indeed, undergraduates, graduate students, and postdoctoral fellows will be fully integrated into our research efforts. The co-PIs have extensive experience in supervising junior researchers, including research with undergraduates that has led to publications in first-tier venues. The involvement of researchers at different stages of training will be beneficial to all. For example, Ph.D. students will have the opportunity to supervise two undergraduate projects in the Harvard SEAS REU Site every summer, while having postdoctoral fellows as mentors in addition to the faculty. They will also benefit from many of the same mentoring mechanisms described in the separate postdoc mentoring plan. We will make an effort to recruit women and under-represented minorities at all levels, and co-PI Sweeney (a female African-American) can serve as a role model for these students and postdocs.

7 Results from Prior NSF Support

PI Vadhan's grants most closely related to this project are CNS-0831289 "CT-ISG: The Assumptions for Cryptography" $399,877, 9/1/08–8/31/11 and CNS-0430336 "New Complexity-Theoretic Techniques in Cryptography", $399,999, 7/1/04–8/31/08. These projects substantially improved our understanding of a number of fundamental cryptographic primitives. Through a series of papers [95, 37, 70, 39], tight characterizations were given for the power of zero-knowledge proofs and commitment schemes and the assumptions needed to construct them, resolving a 15-year old open problem and receiving the Best Paper Award at EUROCRYPT 2007 and the SIAM Outstanding Paper Prize 2011. The works [39, 38, 36] substantially simplified, unified, and improved the classic constructions of commitment schemes, pseudorandom generators, and universal one-way hash functions from arbitrary one-way functions, using new notions of "computational entropy." In these projects, Vadhan also began to bring a complexity-theoretic lens on differential privacy, identifying settings where computational complexity is a barrier for private data analysis [24, 92], developing new efficient algorithms for data sanitization [24, 26], and elucidating the benefits of analogues of differential privacy against computationally bounded adversaries [67, 64].

Among NSF grants received by co-PI King during the past five years on various topics includes two awards directly related to this project: DMS-0835500, "CDI Type II: Collaborative Research: Bibliographic Knowledge Network," $217,095, 10/1/08–9/30/11; SES-0112972, "A Feasible Uniform Standard for Deep Citation of Social Science Data," $805,102, 7/17/01–8/12/2005. The second, most closely related award, led to the development of a citation standard for quantitative data, and contributed to the development, distribution and deployment of the Dataverse Network system for data sharing, archiving, and analysis. This system produced a permanent open source infrastructure for data sharing that is now used by 249 universities around the world.

Co-PI Airoldi has been supported by two NSF grants during the past two years: DMS-0907009 "Models of network growth: Wikipedia" $59,485, 07/01/09–09/30/11; IIS-1017967 "Representation, modeling and inference for large biological and information networks", $497,780, 08/01/10–07/31/2013. These projects have generated a number of results published in high-profile venues, including Nature, Nature Methods, Proceedings

[5] `http://www.extension.harvard.edu/openlearning/`

of the National Academy of Sciences, Journal of Machine Learning Research, Annals of Applied Statistics, Public Library of Science journals, and other primary machine learning and computational biology journals [14, 7, 5, 76, 54, 6, 47]. The proposed work capitalizes on existing results that the co-PI has accumulated over the years, independently of these NSF-funded projects. The grant most closely related to this proposal is IIS-1017967, in which the co-PI explores representation and compression techniques for large information networks, partly in collaboration with Facebook and AT&T. As part of the proposed work, the co-PI will be able to leverage data from these companies to explore individual privacy issues.

Sr. Personnel Chong is supported in part by an NSF CAREER Award: CCF-1054172 "CAREER: Practical, expressive, language-based information security," $466,074, 02-01-2011 to 01-31-2016) that explores the specification and enforcement of expressive and practical information security policies. This project has produced a method for the automatic inference of expressive information security policies with the aim of understanding what information security guarantees a Java program offers [96].

Co-PIs Malone and Sweeney and Sr. Personnel Crosas have not received NSF funding in the past five years.

8 Summary: Intellectual Merit and Broader Impact

To summarize, the contributions of this project will include:

- An understanding of the relation between mathematical and legal notions of privacy, and proposals to bridge the gap between them,
- An engagement with and impact on the policy processes for revising privacy law and regulation,
- Variants of differential privacy for use in contexts where it is inappropriate or impractical,
- New statistical measures of risk and utility, and an understanding the tradeoffs between these,
- New theoretical results on differential privacy (both upper and lower bounds), aimed at issues relevant to social science research data,
- An understanding of the practical performance and usability of a variety of algorithms for analyzing and sharing privacy-sensitive data,
- New legal instruments, such licenses and terms of use, to accompany privacy-sensitive datasets and the algorithms used to share and analyze them,
- Secure implementations of some of these algorithms and legal instruments, made publicly available and used to enable wider access to privacy-sensitive data sets at the IQSS Dataverse Network,
- Open-access course materials and videos on data privacy from a variety of disciplinary perspectives, and
- Training of postdocs, graduate students, and undergraduates in a multidisciplinary approach to data privacy.

While we are optimistic that differential privacy will prove to be practical on the larger datasets in the Dataverse Network, we note that most of the outcomes above are *not* dependent on this, and the lessons learned in trying to bring differential privacy to practice will inform the next generation of privacy definitions and tools.

Intellectual Merit. While data privacy has been examined extensively within the individual fields of computer science, law, statistics, and social science, this project is unique in the way it integrates all of these perspectives, both in identifying the goals for privacy and data utility and in developing tools to achieve these goals. This effort is likely to yield solutions that are more viable in practice than those coming from a single approach, in addition to raising fundamental new questions in the individual disciplines.

Broader Impacts. The tools developed and deployed at the IQSS Dataverse Network will contribute to research infrastructure for social scientists around the world. Moreover, the underlying ideas will benefit society more broadly as it grapples with data privacy issues in many other domains, including public health and electronic commerce. In addition, the project will support the development of new curricular material and train a new generation of researchers and citizens with the multidisciplinary perspectives required to address the complex issues surrounding data privacy.

9 Collaboration Plan

9.1 Team Members and their Roles

As discussed throughout the proposal, all of the co-PIs will be involved in multiple aspects of the project, most of which require multidisciplinary collaboration (which the co-PIs have already begun). Nevertheless, each participant will take a leadership role on specific aspects of the project, to make sure steady progress is made on all fronts. These roles are as follows:

- Salil Vadhan (CRCS, PI): Responsible for cohesion of entire project. Lead for definitional and algorithmic work on formal privacy protection models.
- Latanya Sweeney (IQSS & CRCS, co-PI): Shares responsibility for cohesion of project. Lead for development of technological tools and the experimental validation of their privacy and utility.
- Gary King (IQSS, co-PI): Lead for defining and measuring data utility for social scientists, and the integration of private data analysis algorithms into Zelig.
- Phil Malone (Berkman, co-PI): Lead for revisiting legal definitions of privacy, and for the development of legal instruments for handling sensitive data.
- Edo Airoldi (Statistics, co-PI): Lead for developing statistical measures of privacy risk and the tools to estimate such risk for given datasets.
- Stephen Chong (CRCS, Sr. Personnel): Lead for language techniques to specify privacy policies and enforce them in system implementation.
- Merce Crosas (IQSS, Sr. Personnel): Lead for the integration of the technological and legal tools for handling privacy-sensitive data into the Dataverse Network.
- 3 postdoctoral fellows (1.5 at CRCS, 1.5 at Berkman), will also be fellows in IQSS Data Privacy Lab): Will be recruited to bring a mix of computer science, social science, legal, and statistical expertise to the project. Responsible for various aspects of the project according to individual expertise. Will ensure the steady progress of the project at all levels, from long-term academic research to the concrete goals of building tools for the Dataverse Network.
- 4 Ph.D. students (2 at CRCS, 1 at Statistics, 1 at IQSS): Will work closely with the co-PIs and postdoctoral fellows to advance specific research directions in the project, and carry out the needed implementation and experiments.
- 2 Harvard Law students (Berkman): Will work on the development of legal instruments for use in the Dataverse Network, through the Berkman Cyberlaw Clinic.
- 4+ student interns (2 at SEAS REU site, 2 at Berkman, plus Harvard College undergraduates working for course/thesis credit): Will work on smaller, self-contained projects drawn from the range of the entire research agenda, often in close conjunction with one of the Ph.D. students or postdoctoral fellows.
- 1 software developer (IQSS): Responsible for expanding the Dataverse Network software to support privacy-sensitive social science data. Will work very closely with the researchers above to integrate the software tools needed to manage privacy-sensitive data with the Dataverse Network software. This will include modifying existing Dataverse APIs and implementing additional interfaces to connect with the mathematical code provided by the researchers.
- .5 project manager (SEAS): Will coordinate all logistical aspects of the project, including the meetings between the researchers above, the organization of workshops, arrangements for visitors and external

speakers relevant to the project, collection and organization of material for project reports and project-specific financial summaries. Will record notes at project meetings and share them in a way that enables all participants to stay up to date on the group's progress. Will also collect educational materials and videos from privacy-related courses for tagging and editing by the students, and for online dissemination.

9.2 Team Coordination

The entire research team above will have a monthly meeting to discuss concrete progress on the project. This will be in addition to smaller, more focused meetings between the senior and junior personnel about progress on specific aspects of the project and substantial on-line interaction.

As mentioned in Section 2, the collaboration between CRCS, IQSS, and the Berkman Center has already begun with the integration of the CRCS & Berkman fellowship programs in Fall 2009, weekly research meetings on data privacy, and the moving of co-PI Sweeney's Data Privacy Lab from CMU to IQSS in Fall 2011. Thus, the communities are already interacting productively through a number of regular mechanisms, and these mechanisms will be leveraged to further advance the project:

- A weekly "fellows' hour" (hosted at the Berkman Center) in which fellows and faculty from both centers get together to discuss and develop speculative research ideas.
- A weekly lunch discussion series (hosted at the Berkman Center), in which internal and external speakers give short, generally non-technical presentations followed by open discussion. These are open to the public and webcast live.
- A twice-monthly lunch seminar series (hosted at CRCS) in which internal and external speakers present work focusing on computer science research directions related to computation and society. These are open to the public and the videos are posted on-line.
- A number of Berkman community events that provide opportunities to learn about other projects and people and develop a research network. These include an annual open house, an annual retreat, and several social events.
- A weekly, multidisciplinary "Topics in Privacy" research seminar hosted at the Data Privacy Lab in IQSS.

The close proximity of all the institutes (just 1 block from CRCS to Berkman, and 4 blocks to IQSS) has made these interactions very easy, and participation from both sides has been high.

The several existing seminar series will be used to bring in outside speakers relevant to the project, and progress on the projects will be periodically discussed with the wider CRCS, Berkman, and IQSS communities. Researchers will also have frequent interaction with the Berkman Center's research director (Robert Faris), program coordinator (Amar Ashar), and communications lead (Seth Young) to form another layer of support and networking. Berkman's technical support team and Chief Technologist Sebastian Diaz will implement a supportive technology platform that allows for various forms of activity management, planning, and communication.

9.3 Timeline

The following (ambitious) timeline describes our goals for the project. Naturally, most aspects of the project will not be confined to a particular year, but the following reflects our expected emphases during different stages of the project.

Year 1:

- Examine privacy-sensitive IQSS datasets in order to understand the common data types and legal constraints associated with them.
- Work on statistical measures of privacy risk.

- Survey past research studies done on IQSS datasets to determine the analytic methods used, and develop metrics for data utility.

- Continue implementing prototypes of private data analysis algorithms that can be used to reproduce past studies.

- Host workshop on “Measures of Data Utility for Social Science Research.” (Majority of participants are social scientists and statisticians.)

- Engage with policy process for the revision of the Common Rule.

Year 2:

- Map out limits of what can be achieved in terms of privacy-utility-efficiency tradeoffs, including design of new algorithms and theoretical bounds.

- Begin systematic utility and usability experiments, comparing the quality of results and ease of obtaining with and without the privacy tools.

- Test statistical measures of risk on IQSS datasets.

- Work on bridging mathematical and legal definitions of privacy.

- Host workshop on “Balancing Privacy and Utility in Social Science Research.” (Majority of participants are privacy law scholars, computer scientists, and social scientists.)

- Begin to design legal instruments that are tailored to the nature of data in the Dataverse Network, and to the private data analysis algorithms that are being developed.

Year 3:

- Iteratively refine the utility metrics, the algorithmic tools, and legal instruments according to the results of the experiments.

- Develop and carry out “cryptanalytic” attacks to compromise privacy and gauge the practical consequences of various settings of the mathematical privacy parameters (e.g. in differential privacy or statistical measures of risk).

- Begin secure implementations of the algorithms that perform well in the utility and usability experiments, as well as the legal instruments that can be automated.

- Host workshop on “Secure Implementation of Privacy Tools.” (Majority of participants are computer scientists.)

Year 4:

- Complete secure implementations.

- Package the algorithms and legal instruments into a collection of integrated and usable solutions for sharing and analyzing privacy-sensitive data, accompanied by best practices and policy recommendations. Deploy for use by other researchers.

- Write and publish research papers on the results of the research, and disseminate widely among the relevant communities.

References

[1] IQSS Dataverse Network. The Institute for Quantiative Social Science at Harvard University.

[2] Fair information practice principles. U.S. 1974 Privacy Act, 1974. http://www.ftc.gov/reports/privacy3/fairinfo.shtm.

[3] B. Adida. Helios: Web-based open-audit voting. In P. C. van Oorschot, editor, *USENIX Security Symposium*, pages 335–348. USENIX Association, 2008.

[4] E. M. Airoldi, X. Bai, and B. Malin. An entropy approach to disclosure risk assessment: Lessons from real applications and simulated domains. *Decision Support Systems*, 51:10–20, 2010.

[5] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

[6] E. M. Airoldi, E. A. Erosheva, S. E. Fienberg, C. Joutard, T. M. Love, and S. Srhingarpure. Reconceptualizing the classification of PNAS articles. *Proceedings of the National Academy of Sciences*, 107:20899–20904, 2010.

[7] E. M. Airoldi, C. Huttenhower, D. Gresham, C. Lu, A. Caudy, M. Dunham, J. R. Broach, D. Botstein, and O. G. Troyanskaya. Predicting cellular growth from gene expression signatures. *PLoS Computational Biology*, 5(1):e1000257, 2009.

[8] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In L. Libkin, editor, *PODS*, pages 273–282. ACM, 2007.

[9] M. Barbarao and T. Zeller. A face is exposed for aol searcher 4417749. *New York Times*, page A1, 9 August 2006.

[10] J. Bennett and S. Lanning. The Netflix prize. In *Proceedings of KDD Cup and Workshop*, 2007.

[11] J. Blocki and R. Williams. Resolving the complexity of some data privacy problems. In S. Abramsky, C. Gavoille, C. Kirchner, F. M. auf der Heide, and P. G. Spirakis, editors, *ICALP (2)*, volume 6199 of *Lecture Notes in Computer Science*, pages 393–404. Springer, 2010.

[12] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In C. Li, editor, *PODS*, pages 128–138. ACM, 2005.

[13] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In R. E. Ladner and C. Dwork, editors, *STOC*, pages 609–618. ACM, 2008.

[14] M. J. Brauer, C. Huttenhower, E. M. Airoldi, R. R. J. C. Matese, D. Gresham, V. M. Boer, O. G. Troyanskaya, and D. Botstein. Coordination of growth rate, cell cycle, stress response and metabolic activity in yeast. *Molecular Biology of the Cell*, 19:352–367, 2008.

[15] A. Cavoukian and K. E. Emam. Dispelling the myths surrounding de-identification: Anonymization remains a strong tool for protecting privacy. Discussion Paper, Information & Privacy Commissioner, Ontario, Canada, June 2011. . http://www.ipc.on.ca/English/Resources/Discussion-Papers/Discussion-Papers-Summary/?id=1084.

[16] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

[17] Y. Chen, S. Chong, I. A. Kash, T. Moran, and S. P. Vadhan. Truthful mechanisms for agents that value privacy. *CoRR*, abs/1111.5472, November 2011.

[18] S. Chong and A. C. Myers. Language-based information erasure. In *Proceedings of the 18th IEEE Computer Security Foundations Workshop*, pages 241–254. IEEE Computer Society, June 2005.

[19] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 202–210, 2003.

[20] C. Dwork. Differential privacy. Invited talk. In *ICALP (2)*, 2006.

[21] C. Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation, TAMC 2008*, volume 4978, pages 1–19. Springer, 2008.

[22] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: privacy via distributed noise generation. In *Advances in cryptology—EUROCRYPT 2006*, volume 4004 of *Lecture Notes in Comput. Sci.*, pages 486–503. Springer, Berlin, 2006.

[23] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors, *TCC*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

[24] C. Dwork, M. Naor, O. Reingold, G. Rothblum, and S. Vadhan. On the complexity of differentially private data release: Efficient algorithms and hardness results. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC '09)*, pages 381–390, 31 May–2 June 2009.

[25] C. Dwork and K. Nissim. Privacy-preserving datamining on vertically partitioned databases. In *Proceedings of CRYPTO 2004*, volume 3152, pages 528–544, 2004.

[26] C. Dwork, G. Rothblum, and S. Vadhan. Boosting and differential privacy. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS '08)*, pages 51–60. IEEE, 23–26 October 2010.

[27] J. Felch. DNA databases blocked from the public. *Los Angeles Times*, page A31, 29 August 2008.

[28] S. E. Fienberg. Conflicts between the needs for access to statistical information and demands for confidentiality. *Journal of Official Statistics*, 10:115–132, 1994.

[29] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42:14:1–14:53, June 2010.

[30] S. L. Garfinkel and M. D. Smith, editors. *Data Surveillance, IEEE Security & Privacy*, volume 4. IEEE, 2006. Special Issue.

[31] R. Gross, E. M. Airoldi, B. Malin, and L. A. Sweeney. Integrating utility into face de-identification. In *Privacy Enhancing Technologies (Revised Selected Papers)*, volume 3856, pages 227–242, 2005.

[32] A. Gupta, M. Hardt, A. Roth, and J. Ullman. Privately releasing conjunctions and the statistical query barrier. In *Proceedings of the 43rd ACM Symposium on Theory of Computing (STOC '11)*, pages 803–812. ACM, 6–8 June 2011. Full version posted as CoRR abs/1011.1296.

[33] A. Gupta, A. Roth, and J. Ullman. Iterative constructions and private data release. *CoRR*, abs/1107.3731, 2011. To appear in *TCC 2012*.

[34] M. Gutmann and P. Stern. *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data*. National Academy Press, Washington, DC, 2007.

[35] A. Haeberlen, B. C. Pierce, and A. Narayan. Differential privacy under fire. In *Proceedings of the 20th USENIX Security Symposium*, Aug. 2011.

[36] I. Haitner, T. Holenstein, O. Reingold, S. Vadhan, and H. Wee. Universal one-way hash functions via inaccessible entropy. In H. Gilbert, editor, *Advances in Cryptology—EUROCRYPT '10*, volume 6110 of *Lecture Notes on Computer Science*, pages 616–637. Springer-Verlag, 30 May–3 June 2010.

[37] I. Haitner, M. Nguyen, S. J. Ong, O. Reingold, and S. Vadhan. Statistically hiding commitments and statistical zero-knowledge arguments from any one-way function. *SIAM Journal on Computing*, 39(3):1153–1218, 2009. Special Issue on *STOC '07*. Merge of papers from FOCS '06 and STOC '07. Received *SIAM Outstanding Paper Prize 2011*.

[38] I. Haitner, O. Reingold, and S. Vadhan. Efficiency improvements in constructing pseudorandom generators from one-way functions. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing (STOC '10)*, pages 437–446, 6–8 June 2010.

[39] I. Haitner, O. Reingold, S. Vadhan, and H. Wee. Inaccessible entropy. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC '09)*, pages 611–620, 31 May–2 June 2009.

[40] M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. *CoRR*, abs/1012.4763, 2010.

[41] M. Hardt and G. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS '08)*, pages 61–70. IEEE, 23–26 October 2010.

[42] M. Hardt, G. N. Rothblum, and R. A. Servedio. Private data release via learning thresholds. In D. Randall, editor, *SODA*, pages 168–187. SIAM, 2012.

[43] M. Hay, C. Li, G. Miklau, and D. Jensen. Accurate estimation of the degree distribution of private networks. In W. Wang, H. Kargupta, S. Ranka, P. S. Yu, and X. Wu, editors, *ICDM*, pages 169–178. IEEE Computer Society, 2009.

[44] J. J. Horton, D. G. Rand, and R. J. Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14:399–425, 2011. Available at SSRN: http://ssrn.com/abstract=1591202.

[45] K. Imai, G. King, and O. Lau. Toward a common framework for statistical analysis and development. *Journal of Computational and Graphical Statistics*, 17(4):892–913, 2008.

[46] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? In *FOCS*, pages 531–540. IEEE Computer Society, 2008.

[47] Y. Katz, E. Wang, E. M. Airoldi, and C. B. Burge. Analysis and design of rna sequencing experiments for identifying mrna isoform regulation. *Nature Methods*, 2010. In press.

[48] F. Kerschbaum, A. Schröpfer, A. Zilli, R. Pibernik, O. Catrina, S. de Hoogh, B. Schoenmakers, S. Cimato, and E. Damiani. Secure collaborative supply-chain management. *IEEE Computer*, 44(9):38–43, 2011.

[49] G. King. The changing evidence base of social science research. In K. Schlozman and N. Nie, editors, *The Future of Political Science: 100 Perspectives*. Routledge Press, 2009.

[50] S. K. Kinney, J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd. Towards unrestricted public use business microdata: The synthetic longitudinal business database. Technical Report Discussion Paper CES-WP-11-04, Center for Economic Studies, 2011. Now in use by the Census Bureau for distribution of business establishment data through the Synthetic Longitudinal Business Database. http://www.census.gov/ces/dataproducts/synlbd/.

[51] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, J. F. N. Contractor, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, , and M. V. Alstyne. Computational social science. *Science*, 2009.

[52] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. Tastes, ties and time: a new social network dataset using facebook. In *Social Networks*, volume 30, 2008.

[53] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. Tastes, ties and time dataverse. Technical report, Harvard IQSS, Cambridge, MA, 2009.

[54] R. Lu, F. Markowetz, R. D. Unwin, J. T. Leek, E. M. Airoldi, B. D. MacArthur, A. Lachmann, R. Rozov, A. Ma'ayan, L. A. Boyer, O. G. Troyanskaya, A. D. Whetton, and I. R. Lemischka. Systems-level dynamic analyses of fate change in murine embryonic stem cells. *Nature*, 462:358–362, 2009.

[55] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In L. Liu, A. Reuter, K.-Y. Whang, and J. Zhang, editors, *ICDE*, page 24. IEEE Computer Society, 2006.

[56] A. Machanavajjhala, D. Kifer, J. M. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. In *Proceedings of the 24th International Conference on Data Engineering, ICDE 2008, April 7-12, 2008, Cancún, México*, pages 277–286, 2008.

[57] B. Malin. *Trail Re-identification and Unlinkability in Distributed Databases*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2006.

[58] B. Malin and E. M. Airoldi. The effects of location access behavior on re-identification risk in a distributed environment. In *Privacy Enhancing Technologies (Revised Selected Papers)*, volume 4258 of *Lecture Notes in Computer Science*, pages 413–429, 2006.

[59] B. Malin and E. M. Airoldi. Confidentiality preserving audits of electronic medical record access. In *World Congress on Health Medical Informatics (MEDINFO)*, Brisbane, Australia, 2007.

[60] B. Malin, E. M. Airoldi, S. Edoho-Eket, and Y. Li. Configurable security protocols for multi-party data analysis with malicious participants. In *IEEE International Conference on Data Engineering (ICDE),*, Tokyo, Japan, 2005.

[61] B. Malin and L. Sweeney. Determining the identifiability of DNA database entries. In *Proceedings, Journal of the Medical Informatics Association*, Washington, DC, 2000. Hanley Belfus.

[62] B. Malin and L. Sweeney. Reidentification of DNA through an automated linkage process. In *Proceedings, Journal of the Medical Informatics Association*, Washington, DC, 2001. Hanley Belfus.

[63] B. Malin and L. Sweeney. Pacific symposium on biocomputing 2002. In *Inferring Genotype from Clinical Phenotype through a Knowledge Based Algorithm*, Singapore, 2002. World Scientific.

[64] A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. Vadhan. The limits of two-party differential privacy. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS '08)*, pages 81–90. IEEE, 23–26 October 2010.

[65] F. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. *Commun. ACM*, 53(9):89–97, 2010.

[66] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In A. Deutsch, editor, *PODS*, pages 223–228. ACM, 2004.

[67] I. Mironov, O. Pandey, O. Reingold, and S. Vadhan. Computational differential privacy. In S. Halevi, editor, *Advances in Cryptology—CRYPTO '09*, volume 5677 of *Lecture Notes in Computer Science*, pages 126–142. Springer-Verlag, 16–20 August 2009.

[68] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Research in Security and Privacy*, Oakland, CA, 2008. IEEE.

[69] P. Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57:1701–1777, 2010.

[70] S. J. Ong and S. Vadhan. Zero knowledge and soundness are symmetric. In M. Naor, editor, *Advances in Cryptology—EUROCRYPT '07*, volume 4515 of *Lecture Notes in Computer Science*, pages 187–209. Springer-Verlag, 20–24 May 2007. Recipient of Best Paper Award.

[71] J. Reiter and J. Drechsler. Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. Technical report, Intitut fur Arbeitsmarkt und Berufsforschung (IAB), Nurnberg (Institute for Employment Research, Nuremberg, Germany), 2007. http://ideas.repec.org/p/iab/iabdpa/200720.html.

[72] A. Roth and T. Roughgarden. Interactive privacy via the median mechanism. In L. J. Schulman, editor, *STOC*, pages 765–774. ACM, 2010.

[73] I. Roy, S. T. V. Setty, A. Kilzer, V. Shmatikov, and E. Witchel. Airavat: Security and privacy for mapreduce. In *NSDI*, pages 297–312. USENIX Association, 2010.

[74] D. Rubin. Discussion: statistical disclosure limitation. *Journal of Official Statistics*, 9(2):461–468, 1993.

[75] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In *PODS*, page 188. ACM Press, 1998.

[76] R. Silva, K. A. Heller, Z. Ghahramani, and E. M. Airoldi. Ranking relations using analogies in biological and information networks. *Annals of Applied Statistics*, 4:615–644, 2010.

[77] R. Singel. NetFlix cancels recommendation contest after privacy lawsuit. Wired.com ThreatLevel Blog, 12 March 2010. http://www.wired.com/threatlevel/2010/03/netflix-cancels-contest/.

[78] A. Smith. Privacy-preserving statistical estimation with optimal convergence rates. In L. Fortnow and S. P. Vadhan, editors, *STOC*, pages 813–822. ACM, 2011.

[79] T. W. Smith, P. V. Marsden, and M. Hout. General social survey, 1972-2010 [cumulative file]. Inter-university Consortium for Political and Social Research, August 2011. http://hdl.handle.net/1902.2/31521.

[80] L. Sweeney. Replacing personally-identifying information in medical records: the scrub system. In *Proceedings, Journal of the Medical Informatics Association*, Washington, DC, 1996. Hanley Belfus.

[81] L. Sweeney. Iterative profiler. Technical report, MIT, Cambridge, MA, 1997. Sealed by order of court in Southern Illinoisan v Department of Health.

[82] L. Sweeney. Weaving technology and policy together to maintain confidentiality. In *Journal of Law. Medicine and Ethics*, volume 25, 1997.

[83] L. Sweeney. Towards the optimal suppression of details when disclosing medical data: the use of subcombination analysis. In *MEDINFO*, 1998.

[84] L. Sweeney. Uniqueness of Simple Demographics in the US Population. Technical report, Carnegie Mellon University, Data Privacy Lab, Pittsburgh, PA, 2000.

[85] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):571–588, 2002.

[86] L. Sweeney. k-anonymity: a model for protecting privacy. In *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, volume 10, 2002.

[87] L. Sweeney. Identifiability of de-identified pharmacy data. Technical report, Carnegie Mellon University, Data Privacy Lab, Pittsburgh, PA, 2003.

[88] L. Sweeney. Demonstration of a privacy-preserving system that performs an unduplicated accounting of services across homeless programs. Technical Report Working Paper 902, Data Privacy Lab, 2008. http://dataprivacylab.org/projects/homeless/index2.html.

[89] L. Sweeney. Identifiability of de-identified clinical trial data. Technical report, Carnegie Mellon University, Data Privacy Lab, Pittsburgh, PA, 2009.

[90] L. Sweeney and Data Privacy Researchers. Comments on advance notice of proposed rulemaking: Human subjects research protections: Enhancing protections for research subjects and reducing burden, delay, and ambiguity for investigators, Docket ID number HHS-OPHS20110005. http://www.regulations.gov/#!documentDetail;D=HHS-OPHS-2011-0005-1108, October 2011.

[91] J. Thaler, J. Ullman, and S. Vadhan. Faster algorithms for privately releasing marginals. In preparation, February 2012.

[92] J. Ullman and S. Vadhan. PCPs and the hardness of generating synthetic data. In Y. Ishai, editor, *Proceedings of the 8th IACR Theory of Cryptography Conference (TCC ‘11)*, volume 5978 of *Lecture Notes on Computer Science*, pages 572–587. Springer-Verlag, 28–30 March 2011. Full version posted as *ECCC* TR10-017.

[93] U.S. Department of Health and Human Services (HHS). Human subjects research protections: Enhancing protections for research subjects and reducing burden, delay, and ambiguity for investigator (advance notice of proposed rulemaking). *Federal Register*, 76(143):44512–44531, July 2011. http://www.hhs.gov/ohrp/humansubjects/anprm2011page.html.

[94] S. Vadhan, D. Abrams, M. Altman, C. Dwork, P. Kominers, S. D. Kominers, H. R. Lewis, T. Moran, G. Rothblum, and S. Vadhan. Comments on advance notice of proposed rulemaking: Human subjects research protections: Enhancing protections for research subjects and reducing burden, delay, and ambiguity for investigators, Docket ID number HHS-OPHS20110005. http://www.regulations.gov/#!documentDetail;D=HHS-OPHS-2011-0005-1101, October 2011.

[95] S. P. Vadhan. An unconditional study of computational zero knowledge. *SIAM Journal on Computing*, 36(4):1160–1214, 2006. Special Issue on Randomness and Complexity. Extended abstract in *FOCS ‘04*.

[96] J. A. Vaughan and S. Chong. Inference of expressive declassification policies. In *Proceedings of the 2011 IEEE Symposium on Security and Privacy*, pages 180–195, May 2011.

[97] J. Yakowitz. Tragedy of the data commons. *Harvard Journal of Law and Technology*, 25, 2011. Available at http://dx.doi.org/10.2139/ssrn.1789749.

[98] M. Zimmer. but the data is already public: on the ethics of research in facebook. *Ethics and Information Technology*, 12:313–325, 2010. 10.1007/s10676-010-9227-5.