Applying Theoretical Advances in Privacy to Computational Social Science Practice

1. Vision and Significance

The goal of this proposed project is to improve replicability and reproducibility in social science by developing easy-to-use tools for researchers to share confidential research data in a privacy-protective manner, supported by rigorous computational, institutional, and legal foundations. We are motivated by the exciting opportunities in social science enabled by data and technology, and by the threat that privacy poses to realizing their full potential. We expect not only to advance the state of the research in this area, but to enable practical improvements in the privacy, management, and sharing of big data produced by academic, commercial, and government entities. The proposed project builds on a successful, ongoing multidisciplinary collaboration support by an NSF Frontier grant *Privacy Tools for Sharing Research Data*.

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 will uncover many more phenomena, with greater detail and confidence, than allowed by traditional means such as surveys and interviews [LPA+09, Kin09]. 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 include:

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 [LKG+08]. (The privacy issues raised by this dataset are what initially sparked our collaboration.)

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 or even millions of subjects from around the world, affordably. No longer does experimental social science need to sample primarily from college students, 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 [HRZ11].

Sharing Data. The Internet makes it easy to share 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, described below), by companies (such as when Netflix shared movie-preference data to challenge researchers to improve its recommendation engine [BL07]), or by 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 develops

and hosts the Dataverse [IQSS06]—the world's largest repository for social science research datasets, including over 54,000 data sets.

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 we 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, the best practice for "anonymizing" medical records on patients as late as 1997 was to remove explicit identifiers such as names, addresses, and Social Security numbers. Our collaborator Latanya Sweeney showed that other information, such as date of birth, gender, and ZIP code, 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 [Swe97b].

[1] Although some regulations differentiate between the terms "privacy" and "confidentiality," we use the term "privacy" more inclusively to encompass all issues of collecting, managing, and disseminating sensitive information about individuals across the research lifecycle.

Her group went on to expose re-identification vulnerabilities in health data, including clinical trial data [Swe09], DNA data [MS02, MS00, MS01], pharmacy records [Swe03], text such as clinical letters and notes [Swe96], and registry information [Swe97a]).

In another instance, Netflix released an anonymized version of its movie preference database for the contest mentioned above (challenging researchers to 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 [NS08] were able to re-identify individuals in the Netflix dataset, and thereby learn about their entire rental history (revealing sensitive information about individuals, such as their sexual or political preferences, religious beliefs, substance abuse, etc). As a result, a class-action lawsuit was filed against Netflix, and, as part of the settlement, Netflix cancelled a second planned contest [Sin10].

There are numerous other examples of reidentification vulnerabilities discovered in "anonymized" data, including geographical information system data [GS07], genetic databases [Fel08, MS00, MS01], and Internet search engine logs [BZ06]. More generally, the size, structure, and statistical properties of data used in computational social science cause "traditional" approaches to anonymization to fail, yielding data that violate confidentiality, are 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 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.

A sobering example is the recent failure of *InBloom*. With extensive seed funding from multiple foundations, *InBloom* was established to help school districts to manage and share student data. Although the protections applied and the transparency provided by *InBloom* likely exceeded that of many participating schools [RRK+13], *InBloom* was forced to shut down because of protests, lawsuits, and changes in law motivated by concerns for student privacy.[2]

1.3 A Multidisciplinary Approach. It is possible to move beyond the seemingly bleak privacy situation described above, and to achieve wide sharing of social science data while ensuring privacy for individuals. Doing so will require the combined efforts of many disciplines, as summarized below. (More information about the investigators and the institutions that support their multidisciplinary collaboration can be found in **Appendix 7**.)

Information Science & Social Science (co-PIs Altman, Crosas, and King). Effective and practical privacy protection for research will need to integrate technically and institutionally with the tools, practices, and infrastructure of research organizations, and be designed and implemented with an awareness of how information is used and managed across the lifecycle of research and scholarly communication. The information science challenge is to systematically analyze the ways in which new approaches to privacy protection affect the analysis, collection, manipulation, movement, and dissemination of information throughout its lifecycle; and to collaborate in developing classifications of privacy restrictions and protections that enable scalable and coherent treatment of confidential information across the scholarly communications ecosystem (with a focus on the workflow of social science).

Computer Science (co-PI Vadhan, visiting scholar Nissim). Over the past decade, there has been a rich body of work developed around "differential privacy," which provides a strong

[2]http://www.slate.com/blogs/future_tense/2014/04/24/what_the_failure_of_inbloom_means_for_the_student_data_industry.html

model of privacy protection for the analysis of privacy-sensitive datasets (ensuring that individual-specific information does not leak, regardless of what auxiliary information a potential adversary has) along with algorithms that minimally distort data while achieving this protection. The computer science challenge is to translate the theoretical promise of differential privacy into efficient and usable tools that enable the wider sharing of privacy-sensitive data.

Law and Policy (co-PIs Gasser and Altman). Researchers need tools to help them navigate the complexities of the numerous privacy laws, regulations, and contracts that apply to the handling of confidential research data. Further, 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 protecting against privacy risks of varying natures and degrees.

1.4 Privacy Tools for Sharing Research Data. Our multidisciplinary collaboration has already been making successful advances under the support of an NSF Frontier grant *Privacy Tools for Sharing Research Data* (http://privacytools.seas.harvard.edu). In particular, we have begun the design and implementation of an integrated collection of easy-to-use technological, legal, and policy tools to assist social science researchers in the collection, analysis, and sharing of confidential research data. Our target use case is that of a data repository, where researchers deposit datasets in order to make them available to others for purposes of replicating, reproducing, extending, and citing their work. As such, our tools are designed to integrate into the open-source Dataverse software, so that they can be easily deployed at Dataverse repositories (such as the one hosted at Harvard). At the same time, many of the tools will be useful as separate artifacts that can be used by other repositories or even in other contexts.

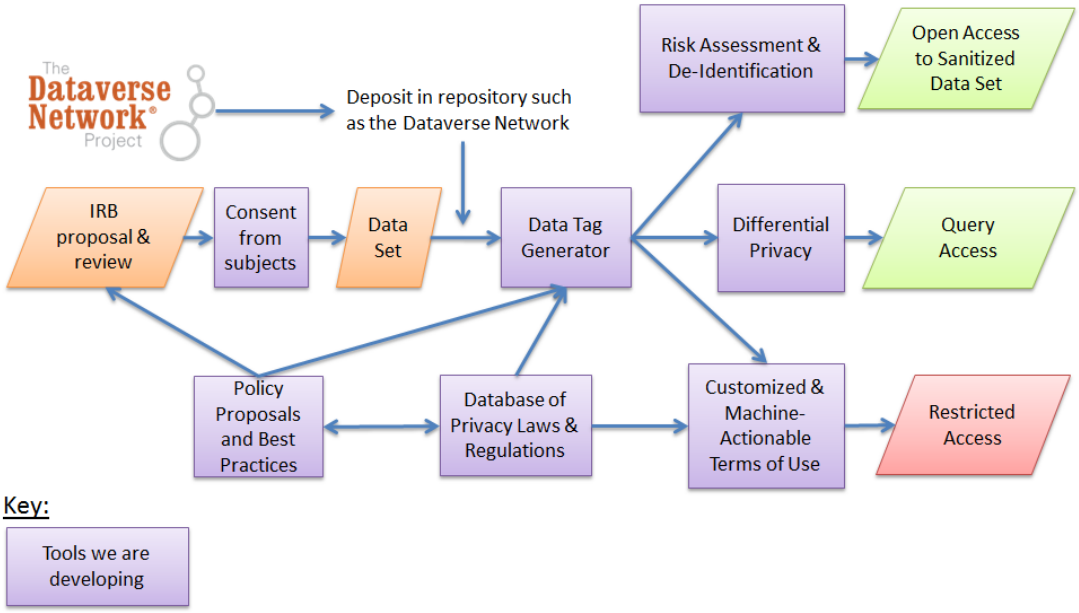

Figure A: Privacy-preserving tools for sharing sensitive data integrated through the use of the Dataverse repository and DataTags systems as the infrastructure.

Our vision for how these tools will work is illustrated in Figure A above. When a researcher comes to deposit a dataset in a Dataverse repository, she will first interact with our DataTags tool, which will lead her through a series of simple questions to determine the legal and technical constraints on the handling of the dataset (e.g., given by the many privacy laws and regulations in the US), and offer her options for the subsequent handling of the dataset by the repository and future users (which will be represented by simple, Creative-Commons-like tags). At the conclusion of the interview, the DataTags tool determines machine-readable tags to apply to the dataset based on the constraints. Depending upon the tags selected, the repository will make the data accessible to future users in any of a variety of ways. In cases of low risk, a deidentified dataset may be made widely available (to the public, or to authenticated users under a click-through agreement). In cases of higher risk, we may limit the public to differentially private statistical queries, and provide access to raw data only by application (e.g., requiring IRB approval) and with automatically generated terms of use.

1.5 Our Goals

By leveraging our ongoing multidisciplinary collaborations and theoretical advances in computation, statistics, law, and social science, the proposed project aims to extend the Privacy Tools for Sharing Research Data project. Specifically, the goals of the proposed project include (1) an incentives analysis, (2) a blueprint for massive data, (3) a study of selected data privacy use cases, and (4) an expansion of our research collaborations. Each of these goals is described in turn below, and further detail is provided in **Section 3**.

(1) An incentives analysis. The first goal of the proposed project is to engage in a systematic analysis of institutional and stakeholder incentives for managing research data privacy and the policy consequences of implementing new computational and legal privacy tools and concepts. Building on our recent work, we will study game-theoretic models of differential privacy, in order to better understand how one should incentivize and compensate subjects for allowing their data to be used for research, how privacy trades off with other objectives, and the "meaning" of differential privacy as a measure of protection. Extending the policy frameworks developed in the project, we will apply insights from these formal models, and, from a systematic analysis of the use cases discussed below, to develop models of the curation of private information across the entire research and scholarly communications lifecycle. We will use this to identify gaps in tools and workflows and to delimit the range of computational, policy, and legal interventions that are feasible at each lifecycle stage.

(2) A blueprint for massive data. The second goal of the proposed project is to design a blueprint for securing massive confidential data in the Dataverse repository. This involves (a) ensuring secure storage of protected files in the Dataverse repository and evaluating how to extend Dataverse to store datasets in the hundreds of gigabytes to terabyte range, (b) designing a

distributed computational architecture for exploring and interpreting big data in Scala and Java that mirrors the statistical architecture we have built in R, and (c) examining the adaptation of machine learning algorithms commonly used for exploring big data through provably differentially private mechanisms.

(3) A study of selected data privacy use cases. The third goal of the proposed project is to study how the new computational and legal privacy tools we are developing can be applied or extended to handle massive data and selected data privacy use cases. Big data are collected, stored, analyzed, and shared in a wide range of contexts, each of which pose different practical challenges for managing confidentiality. For this reason, we have selected three data privacy use cases for deeper analysis: (a) data collected through online education programs, (b) human subjects research data subject to institutional review board policies and consent agreements, and (c) economic data protected by nondisclosure agreements. This study will enhance our understanding of a range of categories of data, stakeholders, and legal constraints relevant to computational social science research, including the privacy risks and potential harms of disclosure, the legal and technical approaches to confidentiality currently in use, and the barriers and incentives associated with the adoption of more advanced privacy-preserving tools.

(4) An expansion of our research collaborations. The fourth goal of the proposed project is to expand our collaborations with other differential privacy and privacy law experts, ongoing data privacy and dissemination efforts at MIT and Harvard, and several related Sloan projects. This project will enable us to collaborate more closely with a number of related efforts, in order to share expertise and increase the reach of our project, contributing to the success of both our work and that of our collaborators. Examples include (a) continuing to host visit scholar Kobbi Nissim (Ben-Gurion University), who is one of the founders of differential privacy and has

contributed invaluable leadership to our project; (b) continuing our ongoing collaboration with Cynthia Dwork (Microsoft Research SVC), who is also one of the founders of differential privacy and is a co-PI on the Sloan project "Towards Practicing Privacy;" (c) continuing our ongoing collaborations with the Computer Science and Artificial Intelligence Laboratory (CSAIL) and Media Lab at MIT; and (d) pursuing new and ongoing joint research opportunities with HarvardX, the Berkman Center Student Privacy Initiative, and the Berkman Center Cyber Learning Project. These collaborations are described in more detail in **Appendices 7 and 8.**

2. Our Approach and Prior Work

2.1 Differential privacy: mathematical theory and practical tools. Differential privacy is a promising mathematical framework for privacy-preserving data analysis introduced in a line of work by collaborator Cynthia Dwork, visiting scholar Kobbi Nissim, and others about a decade ago [DN03, DN04, BDM+05, DMN+06]. Instead of trying to publish a de-identified dataset, differential privacy proposes to mediate access to the data through an interface that takes queries from data analysts and provides results in a way that provably protects privacy (by carefully injecting random "noise" into the computation). We have contributed significantly to delineating the theoretical limits of what is possible with differential privacy [BUV14, BNS14, BNS13a, BNS13b, KNR+13, DLM+12, DNV12, TUV12, UV11, MMP+10, DRV10, DNR+09, MPR+09].

However, there remain significant challenges in bringing the theoretical promise of differential privacy to practice in enabling the wider sharing of privacy-sensitive research data, in terms of both usability and performance. We have been addressing these challenges as part of our Privacy Tools project, as we design and implement easy-to-use tools to provide differentially private statistics about privacy-sensitive datasets in a repository like those in the Dataverse.

When a user deposits a privacy-sensitive dataset in the repository, our tools will walk that user through the process of setting parameters such as the level of privacy protection and deciding on the types of statistical queries that would be most useful to future analysts. Subsequently, users accessing the dataset's page in the repository will be able to obtain such statistics through adaptation of an easy-to-use data analysis interface, called TwoRavens, that is already part of the Dataverse infrastructure and was constructed by our team members Honaker and D'Orazio [HD14].

Currently (2014), we are implementing prototype tools for basic descriptive statistics (means, modes, medians, variances, quantiles, and histograms), as well as computing covariance matrices (which in turn can support arbitrary least-squares regressions). Next year (2015), we plan to start implementing prototypes to support more sophisticated statistics (e.g. higher-order marginal statistics, to allow querying about the frequency of co-occurrence of any desired set of values to a set of variables), synthetic data generation, and richer forms of regression.

What distinguishes our work from other efforts to bring differential privacy to practice is the focus on incorporating differential privacy into the infrastructure of a data repository (Dataverse) in way that fits directly into the workflow of researchers sharing data, and that is optimized for the wide variety of datasets, users, and analyses that a general data repository serves. Our goal is to enable preliminary analysis, whereby a user can determine whether it is worth applying for richer access (a more involved process that will involve a data use agreement and possibly IRB approval). More detail on the mathematical theory and practical tools is provided in **Appendix 9.1**.

2.2 Legal research, analysis, and tools. Social science research with privacy-sensitive data is greatly hindered by the lack of an effective legal and regulatory framework. A complex and

uncoordinated collection of data privacy laws and regulations, data use agreements, and legal precedents that vary across jurisdictions and according to the nature and provenance of the data often creates uncertainty and discourages data sharing. To analyze and begin to address these issues, we are examining the federal statutes, regulations, and policies that protect human subjects research, medical, education, and government records; the current contractual approaches to data sharing across academic, government, non-profit, and commercial settings; and the implications of such restrictions for researchers and data repositories. In addition, we are monitoring legislative and regulatory developments in data privacy; cataloguing privacy-related treatises, articles, and US laws and regulations to be made available as a public resource; and analyzing definitions of privacy from statutes, regulations, and legal scholarship.

Our cross-disciplinary team has begun translating this legal research and analysis into practical tools for tagging datasets and generating modular license agreements for data sharing. These will be integrated into the Dataverse infrastructure (to support our vision for Integrated Privacy Tools, as described in **Section 1.4**) and also be made available as standalone tools that can be integrated into other data storage, use, or dissemination systems. Over the past year, we have developed a prototype of a *DataTags* tool, which handles the deposit of datasets containing medical, education, and government records protected under many of the key federal data privacy laws and regulations in these areas. When a user deposits such a dataset, the tool asks the user a series of questions to determine the constraints on the handling of the data, and generates a machine-readable "tag" to capture these constraints. More detail is provided in **Appendix 9.2**.

2.3 Development of new frameworks for policy analysis. We have submitted joint comments to federal agencies in proceedings related to big data and privacy, including the White

House Office of Science and Technology Policy review of big data and privacy, the Federal Trade Commission review of mobile device tracking, and the Occupational Safety and Health Administration's proposal for the public release of workplace injury and illness reports. With these comments, we provide an overview of privacy-aware models that are used to share sensitive personal data and recommend that future data releases be informed by an analysis of re-identification risks, information sensitivity, potential harms, mitigation techniques, and legal remedies. This work brings together insights from computer science, statistics, law, policy, and social science research to describe a modern framework for privacy analysis and privacy-preserving data releases. This approach can also be used to systematically analyze the institutional implications of new computational approaches to privacy.

3. Our Approach

3.1 Incentives and differential privacy. Much of the computer science work on data privacy (such as *k-anonymity* [Swe02b, Swe02a] and *differential privacy* [DN03, DN04, BDM+05, BLR08]) is aimed at providing a single quantitative bound on how much privacy loss occurs when using a given mechanism, similar to the way security notions in cryptography bound the amount of computing power is needed. However, in privacy, unlike cryptography, the amount of privacy loss generally cannot be treated as "negligible" and instead needs to be traded off with other objectives (such as the accuracy of statistical computations that are enabled). Consequently, it is important to gain a clear understanding of the meaning of a particular quantification of privacy (such as differential privacy) and how it can be compared to the values of other objectives. Economics and game theory provide an informative lens for this task, as they provide a framework to model and compare the data subjects' loss due to privacy, the data

analyst's gain from an accurate computation, and the data subjects' gain from any compensation provided.

A line of work in this direction, started by Ghosh and Roth [GR11], and substantially advanced by co-PI Vadhan and visiting scholar Nissim [NST12,NOS12,CCK+13,NVX14], provides more realistic models of privacy loss and numerous theorems about what can and cannot be achieved. We plan to continue investigating game-theoretic models for understanding privacy loss and trading it off with other objectives. In particular, Nissim and Vadhan are using insights gained from the game-theoretic study to develop a theory of *individualized differential privacy*, where each data subject can specify her own level of privacy. With postdoc Or Sheffet, we are exploring whether differential privacy arises "naturally" due to privacy concerns in certain repeated games. We expect that this line of work will continue to provide insight into the meaning and value of (differential) privacy, and guide in setting the level of privacy protection. More details are in **Appendix 10.1.**

3.2 Incentives and a lifecycle analysis. Most traditional approaches to statistical disclosure control [cf. HDF10; WW01; Uni04], as well as newer approaches such as k-anonymity, frame privacy protection as a single-stage data transformation problem in which sensitive input information is transformed so that the resulting output satisfies a pre-designated privacy concept. This contrasts with the general approach used in information science to analyze the treatment of data: the lifecycle model. Information lifecycles trace actions on information from the stage of creation through to long-term access (see Figure B). Lifecycle approaches explicitly model the information objects, actors, actions on the data, stakeholders, and interactions at and between each stage. [e.g. Hig08, Alt12]

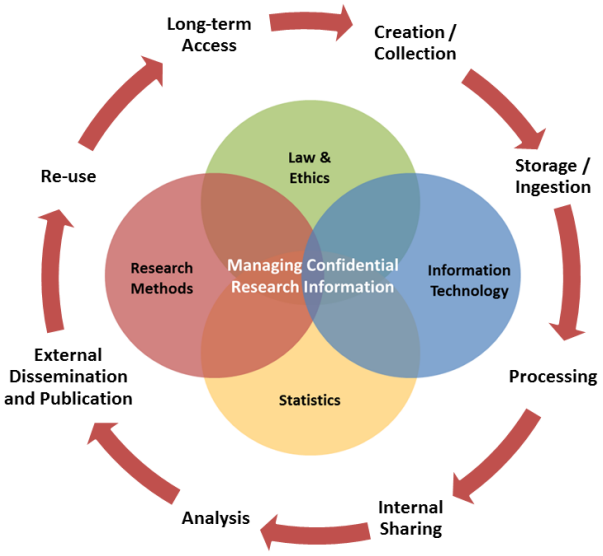

Figure B: Example Research Information Lifecycle (The center of the lifecycle shows categories of approaches to managing confidential information.)

An objective of the proposed information science work is to analyze each privacy use case in terms of the scope of information covered, the sensitivity of that information (potential for harm), the risk that sensitive information will be disclosed (by re-identification or other means), the availability of control and accountability mechanisms (e.g., review and audit), and the suitability of existing data sharing models as applied across the *entire lifecycle of information use*. This lifecycle approach suggests a range of research questions to be explored:

- *Information infrastructure questions:* How do new forms of privacy controls affect the vulnerability surfaces of infrastructure -- in decreasing statistical threats, how much are information security threats increased? In what ways could new forms of privacy control affect system development and maintenance costs?
- *Behavioral questions:* At which points in the lifecycle do stakeholders gain or transmit knowledge of the privacy risks or constraints downstream? At which points do network effects (or other increasing returns to scale) impact stakeholders and infrastructure?
- *Policy/institutional questions:* What is the range of regulatory frameworks (and regulatory

language) that are capable of implementing emerging privacy tools? What are possible systemic effects of adopting regulation mandating a particular privacy concept? What are potential sustainability models for new resource-intensive methods?

We have developed a use case methodology (see **Appendix 10.2** for details) with an extensive set of measures to characterize the conditions affecting data privacy at each lifecycle stage.

3.3 Massive data. Large datasets present both an opportunity and a challenge for our privacy tools. On one hand, big data is good for privacy, intuitively, as each person can "hide" within a larger crowd. Formally, as the number *n* of individuals in a dataset grows, the accuracy of statistical computations due to differential privacy increases, often with similar asymptotic behavior as the best non-private algorithms (e.g., [Smi09]). On the other hand, big data raises scalability issues, such as computation time and secure storage, that can prevent computational tools designed for smaller datasets from being usable. In particular, the tools we are developing rely on secure data storage in Dataverse, and secure remote computation in R on servers using the Zelig libraries for statistical modeling, which neither scale to massive data nor provide tools for big data style analytics. We plan to take the architecture we have developed for Zelig in R, and blueprint how that same architecture could be mirrored in Scala and Java for increased computational ability in distributed settings. We also plan to continue investigation of the extent to which big data analytics tools, such as clustering and supervised/semi-supervised learning, can be carried out with differential privacy, building on extensive prior work by visiting scholar Nissim [BDM+05, NRS07, KLN+11, FFK+09, BNS13a, BNS13b, BNS14]. See **Appendix 10.3** for more detail on our proposed approach to massive data.

3.4 Data privacy use cases. As discussed above, our work on implementations of differential privacy tools is not optimized for data from specific domains but rather is aimed at

providing general-purpose tools that can handle the wide variety of datasets that are deposited in repositories such as Dataverse. However, it can be very informative to explore how much utility these general tools offer in the context of specific real-world use cases that are highly-relevant to computational social science practice. We expect such an investigation to demonstrate concrete applications for the use of integrated privacy-preserving tools; provide a baseline for developing systems optimized for the online education, human subjects, and economic research domains; and point the way to future extensions of our tools.

(1) Online education data. MOOCs and other new educational technologies offer great possibilities for education research. For instance, relationships between student patterns of engagement with course materials, performance on assessments, and pedagogical interventions can be tracked and studied over time, often with a very large sample size. However, the sharing of educational data for research raises significant privacy concerns and is highly constrained by statutes and regulations. In our proposed project, we will study the laws, policies, and agreements relevant to the collection, use, and disclosure of datasets from online education programs. We will also conduct outreach, such as hosting workshops, working meetings, focus groups, or usability testing sessions, with stakeholders in the educational technologies space, to gain a better understanding of the types of data they seek to collect, analyze, and disclose, and the incentives and barriers to maximizing the research potential of these categories of data. (See **Appendix 10.4** for more detail). Potential participants and collaborators include researchers collecting and analyzing online education datasets as well as those affiliated with HarvardX, MITx, and related projects mentioned above in **Section 1.3** and in **Appendix 7**.

(2) Common Rule data. Most social science research is governed by the Common Rule,[3] which requires researchers to obtain the informed consent of human subjects and establishes guidelines for institutional review boards (IRBs) which determine whether research studies at a given institution are ethical and assess the level of risk posed to the subjects. With the shift towards data-driven, computational social science, IRBs are increasingly being forced to evaluate informational risks, or those that arise from inappropriate use or disclosure of information. In our Privacy Tools project, we are studying how our legal and technical tools could be used to enhance the collection, analysis, and dissemination of data subject to the Common Rule and IRB review processes. We propose to expand our research by conducting a systematic analysis of institutional policies and consent agreements from a wide range of institutions and studies. To do this, we will obtain collections of IRB policies and research consent forms from university sources, such as the data repositories at Harvard and MIT, and conduct a type analysis to cluster provisions from these policies and agreements in order to identify common typologies in their approaches. We expect this research, in combination with outreach with appropriate stakeholders, to inform the development of legal and technical tools that will improve the practice of human subjects research. More detail on the Common Rule data use case appears in **Appendix 10.5**.

(3) Economic data. The proprietary nature of many new sources of economic data poses a significant barrier to advances in computational economic research. For this reason, we propose to explore a third data privacy use case involving economic data disclosed by businesses and protected by nondisclosure agreements (NDAs). This use case contemplates datasets such as online auction data from eBay and pricing data from Amazon that provide a rich source of

[3] The Federal Policy for the Protection of Human Subjects, also known as the Common Rule, 45 C.F.R. part 46, is aimed at protecting human subjects in federally funded research.

information for research in the fields of microeconomics, game theory, and auction theory. We will begin a review of the common approaches to obtaining and analyzing commercial economic data, and the incentives and barriers to the disclosure of such data for research purposes. As part of this review, we will conduct outreach, such as by holding workshops, working meetings, or focus groups, with industry stakeholders and, where possible, obtain and analyze collections of NDAs for economic data from university sources, such as the Offices of Sponsored Programs at Harvard and MIT. These efforts will enhance the development of appropriate technical and legal approaches for sharing data from a range of commercial sources.

4. Outcomes

To summarize, the outcomes of this project will include: (1) practical tools for improving privacy and utility of data dissemination, enhancing those developed in our NSF Privacy Tools project and informed by the use cases studied here; (2) reports articulating blueprints for managing confidentiality in massive data; (3) revised metadata schemas, legal agreements, "good practices," or draft regulatory language developed based on our research collaborations and study of selected data privacy use cases (drawn from online education data, Common Rule data, and economic data); (4) integration of the legal and computational privacy tools we develop with the Dataverse; (5) theoretical advances in our understanding of privacy, informed by the use cases above; (6) publications in peer-reviewed outlets in law, computer science, policy, and other areas; and (7) training of postdocs, graduate students, and undergraduates in a multidisciplinary setting.

As an initial estimate of the impact of these outputs, we will track download counts on written and software products, citation rates, media mentions, and usage of privacy features deployed in the Dataverse system. Overall, we expect this project to enable practical

improvements in replicability of, access to, and confidentiality of big data produced by university, commercial, and government entities, and to advance the state of the art of data privacy research.

5. Budget Justification Overview

The budget is allocated almost entirely to personnel, in the form of existing staff to devote additional effort (0.25 FTEs/year), support for key collaborators in residence (1.25 FTEs/year), additional postdoctoral fellows (1 FTE/year), project management and coordination across project units (1 FTE/year), and research assistance hours (0.5 FTEs/year). Detailed budget justifications for each institutions are supplied in **Appendix 11**. Line-item budgets are attached separately.

6. Overview of Other Sources of Support

As described in Section 1.4, this proposal builds upon a successful multidisciplinary collaboration, *Privacy Tools for Sharing Research Data,* that we have begun in this area under the support of an NSF Frontier grant (10/12-9/16, $4.8m total). The proposed work would not be possible without the foundation provided by the Privacy Tools project, and conversely the proposed work will enhance the results of the Privacy Tools project.

In addition, we and our institutions are engaged in a number of synergistic projects that will increase the impact of the project, as summarized in **Appendices 7 and 8.** These synergistic efforts also reduce the costs necessary for convening of meetings and other collaborative activities, and provide established channels which can be leveraged for effective outreach.