HHS Public Access

Author manuscript

Socius. Author manuscript; available in PMC 2023 June 21.

Published in final edited form as:

Socius. 2019 ; 5: . doi:10.1177/2378023118813023.

Privacy, Ethics, and Data Access: A Case Study of the Fragile Families Challenge

Ian Lundberg[1], **Arvind Narayanan**[1], **Karen Levy**[2], **Matthew J. Salganik**[1]

[1]Princeton University, Princeton, NJ, USA

[2]Cornell University, Ithaca, NY, USA

Abstract

Stewards of social data face a fundamental tension. On one hand, they want to make their data accessible to as many researchers as possible to facilitate new discoveries. At the same time, they want to restrict access to their data as much as possible to protect the people represented in the data. In this article, we provide a case study addressing this common tension in an uncommon setting: the Fragile Families Challenge, a scientific mass collaboration designed to yield insights that could improve the lives of disadvantaged children in the United States. We describe our process of threat modeling, threat mitigation, and third-party guidance. We also describe the ethical principles that formed the basis of our process. We are open about our process and the trade-offs we made in the hope that others can improve on what we have done.

Keywords

privacy; ethics; mass collaboration; social data

Social data—data about people—can be both valuable and dangerous. On one hand, they can be used to advance scientific understanding and yield insights that can benefit society. On the other hand, they can be used in ways that violate privacy and lead to other harms. Stewards of social data, therefore, face a fundamental tension. At one extreme, a data steward could share a complete data set publicly with everyone. This *full-release* approach maximizes the potential for scientific discovery, but it also maximizes risk to the people whose information is in the data set. At the other extreme, a data steward could share the data with no one. This *no-release* approach minimizes risk to participants, but it also eliminates benefits that could come from the responsible use of the data. In between these two extremes, no release and full release, there are a variety of intermediate solutions, which involve balancing risk to participants and benefits to science (Figure 1). In this article, we

Corresponding Author: Ian Lundberg, Princeton University, Department of Sociology, 227 Wallace Hall, Princeton, NJ 08544, USA, ilundberg@princeton.edu.

Minor updates have been made since first publication: in the abstract, the last 3 sentences have been changed from third person to first person.

present a case study describing how we balanced this trade-off between risks and benefits when we served as data stewards during the Fragile Families Challenge. We hope this case study will benefit a variety of data stewards, including those within universities, companies, and governments. We also hope that this case study will benefit policy makers who seek to enable responsible data access and researchers who seek responsible access to detailed and potentially sensitive social data.

The Fragile Families Challenge is a scientific mass collaboration involving hundreds of researchers. During this mass collaboration, a diverse group of social scientists and data scientists worked with a common data set that contained detailed information about the lives of 4,242 families in the United States, many of whom were disadvantaged. The detailed nature of the data made the project particularly valuable for developing knowledge about the lives of disadvantaged families, yet these very features also heightened concerns about privacy and ethics. In other words, the Challenge brought the tension between risks and benefits into sharp focus. In this article, we provide no single solution to the fundamental tension between access and privacy; instead, we describe our process of addressing it. More specifically, we describe the privacy and ethics audit we conducted from December 2016 through March 2017, as well as steps we carried out during the Challenge from March through August 2017 (see timeline in Figure 2).

The article is divided into eight sections. The current section introduces the paper. The next section provides more background about the data and the Challenge. We then describe our threat modeling and threat mitigation strategies, followed by our response plan in case our mitigations were ineffective. The next section summarizes our mechanisms for third-party guidance. We then discusses the ethical principles that guided our thinking. The next section describes our ultimate decision to conduct the Challenge. The final section concludes. Although the article is written linearly, the real process, summarized in Figure 3, cycled through all of these steps many times. We are open about our process and the trade-offs we made in the hope that others can improve on what we have done.

Background

Fragile Families and Child Wellbeing Study

The Fragile Families Challenge builds on the Fragile Families and Child Wellbeing Study (hereafter Fragile Families Study), a longitudinal study of 4,898 families. The study began with a probability sample of newborns in 20 large U.S. cities, of which 16 cities form a probability sample of all U.S. cities with populations greater than 200,000 (Reichman et al. 2001). For more than 15 years, researchers have followed these families to collect information related to child and family development as reported by the children as well as the children's mothers, fathers, primary caregivers, and teachers. These rich longitudinal data have already been used in hundreds of published articles and dozens of dissertations on aspects of urban poverty, including multiple-partner fertility (Carlson and Furstenberg 2006), multigenerational households (Pilkauskas 2012), paternal incarceration (Wildeman 2009), housing instability (Desmond and Kimbro 2015), and neighborhood disadvantage (Donnelly et al. 2017).[1]

Four features of the Fragile Families Study were particularly relevant to the design and conduct of the Fragile Families Challenge. Many of these features are common in large-scale social data sets but may not be common in data sets held by companies and governments. These features also make these data different from some types of data commonly considered in privacy research.

First, these data were collected with informed consent. Parents explicitly agreed to join the study and made this agreement on behalf of their children. Furthermore, the children themselves provided their assent to participate once they were old enough. These procedures were overseen by the Institutional Review Board of Princeton University. Informed consent makes the Fragile Families Study different from many other cases privacy scholars have considered. For instance, a main critique of the use of Facebook data for research purposes has been the lack of informed consent; participants may not expect their activity on Facebook to be used in research (Zimmer 2010). In this case, however, respondents learned about the goals of the study and gave explicit permission for the information they provided to be used by researchers.

Second, these data are already available to researchers through an established system. This data access system, which is overseen by the Institutional Review Board of Princeton University, has already been used by thousands of researchers for more than 15 years. The current data access system follows a tiered model in which there are basic files and restricted files (Figure 4). The two main differences between the basic and restricted files are the application process and the types of data that are provided; in all cases, the data are stripped of obviously personally identifying information.[2] This system of tiered access served as our baseline as we designed and implemented the Fragile Families Challenge. We think it is reasonable to accept the current system as our baseline because this system developed over many years in the full view of the scientific community.

The third feature of the Fragile Families Study that shaped our design of the Challenge is that these data contain information about many people around the focal child—such as the mother, father, primary caregiver, and teacher—and contain information about many domains of the respondents' lives. For example, the study collects information about the home environment, the school environment, teacher characteristics, parental criminal history, and child health, to name just a few domains (Figure 5). The multidomain, linked

[1]For a complete list of research using the Fragile Families Study, see https://ffpubs.princeton.edu.

[2]The restricted files may contain detailed geographic, genetic, or other data deemed especially sensitive or identifiable. Access to these files requires a detailed restricted data contract and a carefully vetted research proposal. The Fragile Families staff grants approval only to projects with research merit that can only be achieved with the restricted files. Data are shared only after researchers sign a data protection agreement, show that they have completed National Institutes of Health-approved training in protecting human research participants, provide evidence of approval from the institutional review boards of their institutions, and detail a data protection plan summarizing how data will be used. Student applicants must apply with faculty mentors, who bear responsibility for violations of the agreement. Researchers approved in this process are given access to a tailored version of the data with detailed information only on the domains relevant to their research. For instance, a researcher studying neighborhood effects would be given neighborhood information but not genetic information, while a researcher studying epigenetics would be given genetic information but not neighborhood information; only researchers with projects involving both fields would be given data covering both domains. Fragile Families staff members work with researchers to determine the particular variables needed for any given study. The basic files exclude obvious personally identifying information and obvious geographic information, such as city of birth and residential location. To obtain these files, researchers must agree to a set of terms and conditions and propose a viable project, which Fragile Families staff members approve as potentially important social science research. Several thousand researchers have completed these procedure and were already using the data before the Challenge began.

nature of the data increases its scientific value, but it also increases risks for two main reasons. First, it creates many possible entry points for a reidentification attack (this risk is described in more detail later). Second, the multidomain nature of the data increases the harm that could occur if a reidentification attack were successful, because many potentially sensitive pieces of information would be revealed. The linked, multidomain nature of the data differs from the cases normally considered by privacy researchers, which usually involve a collection of individuals with information from a single domain (e.g., medical records).

The fourth and final feature that is relevant to the Fragile Families Challenge is that these data are frequently used in scientific and policy debates. For example, a recent National Academies of Sciences report on the effects of parental incarceration on children drew heavily on the Fragile Families Study (National Research Council 2014). Although these data have already been used extensively by social scientists, we thought it would be possible to learn even more if a larger, more diverse group of researchers approached the data in a very different way. This goal of increased scientific and policy impact was one of the main motivations for the Fragile Families Challenge.

Fragile Families Challenge

The Fragile Families Challenge is a mass collaboration that combines predictive modeling, causal inference, and in-depth interviews to yield insights that can improve the lives of disadvantaged children in the United States. This article, and this special collection, describes the first stage of the Fragile Families Challenge, which focuses on predictive modeling. This predictive modeling stage follows an approach called the common task framework, which is used frequently in computer science (Donoho 2017) and biomedical research (Saez-Rodriguez et al. 2016). The common task framework is a process that invites many researchers to participate in a unified task characterized by three key aspects: (1) a common predictive modeling goal using (2) a single data set made available to all, with (3) a well-defined scoring metric to evaluate contributions. This process often yields better predictive performance than any individual researcher can realize working alone (e.g., Bennett and Lanning 2007) and often leads to new scientific and methodological insights (e.g., Feuerverger, He, and Khatri 2012). Donoho (2017) went so far as to describe the common task framework as part of “the ‘secret sauce’ of machine learning.”

As described in more detail in the introduction to this special collection (Salganik et al. 2019), we set out the goal of using the data collected from a family at the birth of the child up to when the child was 9 years old to predict data from the family when the child was 15 years old. These age 15 data had been collected but were not yet available to participants (Figure 6). The existence of these collected but otherwise unavailable data is critical for the common task framework, and fortunately all longitudinal social surveys present this possibility every time a new wave of data has been collected. Among the many possible outcomes in the year 15 data, we asked participants in the Challenge to predict six key outcomes: grade point average (GPA) of the child, grit of the child, material hardship of the family, whether the family was evicted from their home, whether the primary caregiver participated in job training, and whether the primary caregiver lost his or her job. The choice

of these outcomes was driven by ethical considerations and our scientific goals, and each outcome is described in more detail in the introduction to this special collection.

As is typical in projects using the common task framework, we split the year 15 data for these six outcomes into three groups: (1) a training set that we provided to participants, (2) a leaderboard set that participants could access during the Challenge, and (3) a holdout set that participants could not access until the first stage of the Challenge was complete (Hardt and Blum 2015) (Figure 6). Participants in the Challenge received the training set and a specially constructed background data file that contained information about the family from birth to age 9 (the construction of this file is described later). This background file included 4,242 families and 12,942 variables about each family. The high-dimensional nature of the data—more predictors than observations—is a result of the linked, multiple-domain nature of the data (as discussed earlier) and has important implications for privacy (as discussed later).

Challenge participants used the background data file and training data to build statistical or machine learning models. They used these models to predict the holdout data (e.g., GPA at age 15). We measured the quality of these predictions using mean squared error.[3]

The immediate goal of this stage of the Fragile Families Challenge was to find the most accurate predictive model for the six outcomes. Given the nature and size of the data—thousands rather than millions of observations—we wanted to learn the extent to which machine learning methods would improve predictive performance beyond the generalized linear models typically applied by social scientists. This explicit focus on prediction is atypical for social science, but substantial current scholarship argues that prediction is important for both scientific and policy reasons (Breiman 2001; Hofman, Sharma, and Watts 2017; Kleinberg et al. 2015, 2017; Mullainathan and Spiess 2017; Shmueli 2010; Watts 2014).

In our case, the immediate goal of prediction was important to prepare for a long-term goal of explanation and hypothesis generation. As will be described in future articles, these predictions will be used to target qualitative, in-depth interviews with families that reported unexpected outcomes. We hope these interviews will help us discover important and currently unmeasured factors and generate hypotheses about how these may affect the lives of disadvantaged families. We also hope the interviews will inform the credibility of the assumptions required to draw causal inferences from survey data in a setting in which thousands of pretreatment variables are potentially involved in confounding.

Why Worry?

The data from the Fragile Families Study were collected with informed consent and are already provided to researchers under a well-established system, all of which has been overseen by the Institutional Review Board of Princeton University. More generally, survey

[3]The mean squared error is a common scoring metric, and it can be written as $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is the predicted value for person i, y_i is the true value for person i, and n is the number of people in the holdout set. For binary outcomes, mean squared error is sometimes called Brier score (Brier 1950).

data of this type have been provided to researchers in a similar fashion for more than 50 years. Why should we worry about the risk for reidentification in the Fragile Families Challenge?

Quite simply, all data are potentially identifiable. This possibility was made clear to us in one of our first meetings, when a member of our team (Narayanan) proposed the following hypothetical scenario. Imagine an all-powerful and evil business magnate such as Lex Luthor, Superman's archenemy. Further imagine that Lex Luthor heard about the Fragile Families Challenge and wished to reidentify the data. Lex could invest billions of dollars to conduct a census of every child born from 1998 through 2000 in the cities covered by the Fragile Families Study. Then Lex could merge his census with the Fragile Families Challenge data, identify everyone, and learn everything about them in the Fragile Families Study.

This hypothetical attack from Lex Luthor illustrates two important points. First, it illustrates a common pattern in reidentification attacks. Data that have been *deidentified*, meaning stripped of obviously identifying information in an effort to protect privacy, can often be *reidentified* by linkage to an auxiliary data set that contains identifying information. Through this process of merging, the apparently anonymous data are reidentified (Ohm 2010). This hypothetical attack also illustrates that the safety of a given data set depends not just on that data set but on all the auxiliary data that exist today and may exist in the future (Narayanan and Shmatikov 2010). Deidentification of a data set does not guarantee anonymity.

Reidentification attacks such as the one performed by Lex Luthor are not merely hypothetical risks, unfortunately. Although we do not know the frequency with which these attacks occur "in the wild," we do know that academic privacy researchers have conducted and published similar attacks (we will discuss their motivations for these attacks later).[4] Two prominent examples come from the research of Latanya Sweeney. First, while a graduate student at MIT, Sweeney (2002b) was able to reidentify apparently anonymous medical records that were provided to researchers by the Massachusetts Group Insurance Commission. She did this by combining the apparently anonymous medical records, which contained date of birth, ZIP code, and sex, with nonanonymous voter registration data, which also contained date of birth, ZIP code, and sex. Because these three variables were available in both the deidentified data and the identified auxiliary data, Sweeney was able to merge them together (Figure 7). Fortunately, rather than posting all of the records online, Sweeney mailed Massachusetts governor William Weld a copy of his own records. More important, Sweeney published an article describing her attack and a proposing a defense against that kind of attack (Sweeney 2002b, 2005; Ohm, 2010).

Beyond simple merges between two files, reidentification attacks can also involve multiple sources of auxiliary information. For example, Malin and Sweeney (2004) combined databases of DNA records, which contained time and place of collection, with publicly

[4]For an attempt to estimate the rate of reidentification attacks that have been published in the academic literature, see El Emam et al. (2011).

available hospital discharge data. These hospital discharge data in turn contained basic demographic information that could be used to combine them with identified voter registration records (Malin and Sweeney 2004) (Figure 7). In other words, the hospital discharge data served as a critical middle step as Malin and Sweeney linked the deidentified DNA records with identified voting records.

The hypothetical example of Lex Luthor clarifies that all data are potentially identifiable when facing a powerful adversary. Furthermore, Sweeney's reidentification attacks, as well as attacks by other privacy and security researchers (Narayanan et al. 2016), show that reidentification attacks are possible in the real world and that these attacks benefit from the presence of rich auxiliary data sets. These facts alone suggest that social scientists should be concerned about the possibility of reidentification attacks. A further cause for concern is that reidentification attacks are probably easier today than at any time in the past. Just as scientists are excited about big data sources for research, adversaries can use these same big data sources for reidentification attacks. Because all data are potentially identifiable and because reidentification attacks are easier now than at any time in the past, we were concerned about the possibility of a reidentification attack during the Fragile Families Challenge.

Potential Technical Solutions

Because the risk for reidentification attacks occurs in many situations, researchers have developed techniques that facilitate the analysis of data while protecting the privacy of the people described in the data (Duncan, Elliot, and Salazar-González 2011; Dwork and Roth 2014; Willenborg and de Waal 2001). We believe that many social scientists, following the lead of statisticians, would organize these privacy-preserving techniques into two main groups: those that focus on modifications to the data and those that focus on access to the data (Duncan et al. 2011; Reiter and Kinney 2011). However, our own deliberations were more influenced by the literature in computer science, and so we organized these privacy-preserving techniques into two different groups: those that offer provable guarantees and those that do not. Techniques that offer provable guarantees would enable us to provide specific, mathematical bounds on what an adversary might be able to learn about individuals in the data set, while making minimal assumptions about the adversary's knowledge, capability, or behavior. During our deliberations, we considered two approaches that offer provable guarantees: differential privacy and cryptography. Unfortunately, as we will now describe, we did not believe either approach was feasible in our setting. For the Fragile Families Challenge, we relied on many techniques that did not offer provable guarantees. Before describing the techniques we used, we summarize the approaches we considered that offer provable guarantees, and we describe why we did not think they were appropriate in our setting.

The first main approach we considered that offers provable guarantees is differential privacy (Dwork 2008; Dwork and Roth 2014). Differential privacy is a set of techniques for developing data release algorithms that achieve the following privacy guarantee: any output of the data release algorithm would have been roughly as likely even if any particular record had been removed from the data. Intuitively, this means that the adversary cannot

tell from the data release whether any individual's record was included in the sample, and this property is definitionally treated as a kind of individual privacy. Within the broad area of differential privacy, we considered two families of approaches. Under one family of approaches, called *noninteractive approaches*, we would release a modified form of the data to Challenge participants.[5] Under the other family of approaches, called *interactive approaches*, we would not release any data; rather, we would keep the data on a secure server and allow researchers to send queries to the server.[6]

Within the family of noninteractive approaches, we considered two subfamilies of approaches, transformed data[7] and synthetic data,[8] but we concluded that neither approach was feasible in our setting.

In addition, we also considered interactive approaches whereby we would have hosted the Challenge data on a server and allowed Challenge participants to query the data (e.g., request a specific regression model). We would then return results with carefully generated noise that would satisfy differential privacy. We found that interactive approaches to differential privacy required major changes to the work flow of analysts and restricted the types of analysis that were possible.[9] In conclusion, although we think differential privacy is an elegant and promising approach to providing provable guarantees, we did not think these approaches, either noninteractive or interactive, were feasible for the Fragile Families Challenge.[10]

[5]The idea of releasing a modified form of the data to increase privacy protections is common in social science. However, many of the approaches that are typically used with social data, such as top-coding and coarsening, do not generally offer provable guarantees within the framework of differential privacy. This does not mean that these approaches should not be used. In practice, we think they often make reidentification attacks more difficult, and we used some of them as described later.

[6]Some researchers refer to these two families as offline approaches and online approaches (Dwork and Roth 2014).

[7]Under the transformed data approach, we would not release individual-level data but rather some aggregated form of data that could then be used for analysis. For example, for a movie recommendation task, McSherry and Mironov (2009) argued that many predictive algorithms operate on the movie-movie covariance matrix, and they demonstrated how to achieve differential privacy for this class of algorithms by releasing a perturbed version of the covariance matrix. We did not believe that such an approach was possible in the Fragile Families Challenge, because we are not aware of an aggregated data structure that would not substantially limit the modeling techniques that would be available to Challenge participants.

[8]Under the synthetic data approach, we would have to build a generative model that, when sampled, produces data from the same joint distribution as the Fragile Families Study data. Furthermore, we would have to build this generative model in a way that is differentially private. There are two main approaches to building such models (differentially private or not): using domain expertise (Drechsler 2011) and algorithmic learning (Hardt, Ligett, and McSherry 2012). We did not believe we had sufficient social science domain expertise to create a realistic data-generating process for the joint distribution of all 12,942 variables in the Challenge data set. To the best of our knowledge, existing applications of synthetic data created on the basis of domain expertise generally involve a much smaller number of variables. For example, the U.S. Census Bureau released a longitudinal data set of businesses that contained five synthetic variables (Kinney et al. 2011). Although algorithmic learning approaches do not require domain expertise, they too are generally limited to data sets with small numbers of variables. For example, a recent technique based on generative adversarial neural networks was applied to a clinical trial with 36 variables and 6,502 observations (Beaulieu-Jones et al. 2017). To the best of our understanding, this and other related algorithmic approaches will not be effective in generating synthetic data with 12,942 variables.

[9]An early tool for interactive differentially private data analysis is PINQ (Privacy Integrated Queries) (McSherry 2009). PINQ is geared toward data analysis and summary statistics such as Structured Query Language (SQL). It also supports basic machine learning algorithms, but it is not clear if it allows building complex machine learning models with many predictors. PINQ requires data analysts to learn a new programing language—a combination of C# and LINQ, a SQL-like language—to express their queries. Furthermore, like all differentially private systems, PINQ imposes a privacy budget whereby each query has a query-specific privacy cost, and analysts have a fixed budget. This budget ensures that the claimed differential privacy guarantees can be met, but it introduces substantial complexity for analysts who are not already familiar with differential privacy. More recently, the Harvard University Privacy Tools Project is attempting to build techniques for differentially private access to social data sets. They have developed a tool named Psi, but it did not appear to be publicly available at the time of the Fragile Families Challenge. Furthermore, on the basis of a recent paper (Gaboardi et al. 2016), it appears that Psi does not (yet) support high-dimensional statistical and machine learning models; it is not clear if these limitations are fundamental. We believe that tools like PINQ and Psi are promising approaches to interactive differential privacy, and organizers of future projects similar to the Fragile Families Challenge should consider these tools and related tools that may be developed in the future.

The second major class of provable privacy techniques we considered is based on cryptography. Specifically, we considered ideas related to homomorphic encryption, which is a technique to encrypt data in such a way that computing a function *f*() on the encrypted data and decrypting the result yields the same output as computing *f*() on the original data.[11] For the Fragile Families Challenge, we imagined using homomorphic encryption as follows: (1) we could homomorphically encrypt the Challenge data and release them to everyone because the encryption would make reidentification attacks unfeasible, (2) Challenge participants could build specially constructed models on the encrypted data (models designed to work on the unencrypted data cannot be run on the encrypted data), (3) participants could upload their encrypted predictions to the Challenge server, and (4) the Challenge server would decrypt the predictions and calculate the mean square error. Although this approach sounds very promising, there a number of concerns, both conceptual[12] and practical,[13] that led us to conclude that approaches using homomorphic encryption were not appropriate for the Challenge.

To summarize, we did not think that approaches that offered provable guarantees, differential privacy or cryptography, were applicable in our setting. We hope these techniques will continue to improve and will ultimately become more useful in this type of setting in the future. Because we could not deploy a technique with provable guarantees, we undertook a process of threat modeling and threat mitigation, which we now describe.

Threats and Mitigations

Given that reidentification attacks were possible and technical solutions with provable guarantees were not available, we sought to better understand the possible risks and then reduce them as much as possible. Therefore, we undertook a process of threat modeling and threat mitigation (Shostack 2014). During the threat modeling, we tried to imagine very specific, concrete risks. When considering these risks, we found that it was helpful to separately consider the probability of harm and the magnitude of harm. Furthermore, we tried to avoid spending an inappropriate amount of time considering high harm events with a low probability (Sunstein 2002). We also found it helpful to separate risk to us as organizers

[10]There are also other parts of the differential privacy literature that appear related to our problem but turn out not to be relevant. For example, there is a large literature on techniques for differentially private machine learning. In this body of work, the privacy problem to be solved is that the model trained on private data (e.g., a set of weights for logistic regression) might itself leak information about the training data. This research assumes that the learning algorithm has direct access to the raw data; the privacy question pertains to the algorithm's outputs. Thus, it is not applicable to our setting, as it assumes that researchers have access to the raw data.

[11]More precisely, the code for *f*() is transformed into a function that operates on encrypted data, and these operations are potentially computationally expensive, requiring cryptographic operations for every bit manipulation in *f*(). Early homomorphic encryption techniques were limited in the variety of functions that could be computed under encryption, and so they were called "somewhat homomorphic encryption." A breakthrough by Gentry (2009) removed these limitations, enabling "fully homomorphic encryption." The downside is that fully homomorphic encryption introduces computational overheads that make it currently impractical in many contexts.

[12]Conceptually, there is a much simpler way to offer the same level of privacy guarantees: releasing no raw data, requiring contestants to upload code to the Challenge server, executing that code on the server, and revealing only the mean squared error. We think homomorphic encryption is useful only when the data owner is computationally limited or both parties have private inputs. In the former case, homomorphic encryption allows outsourcing of expensive computations (such as machine learning), which might be desirable even after we account for the slowdown introduced by computation under encryption. In the latter case, homomorphic encryption allows the data recipient to keep a proprietary algorithm secret. Neither of these conditions was met in our situation.

[13]Practically, at the time of the Challenge it took substantial expertise to create even extremely simple statistical models that could run on encrypted data. Experts on homomorphic encryption have likened the process of creating these models to programming in assembly language (Crawford et al. 2018).

of the Fragile Families Challenge from risk to the survey respondents. We were much more accepting of risk to ourselves (e.g., reputational risk) than risk for study participants. Once these risks were identified, we tried to design steps that would mitigate these risks.

Our threat modeling was primarily focused on reidentification attacks and revolved around two main questions: (1) Who might have the skills and rich auxiliary information that would be needed for a reidentification attack? and (2) Who might have the incentive to carry out such an attack? To help answer these questions, we conducted an in-house attack of our own data. We imagined possible data sources that could (1) be identified and (2) contain variables that also exist in the Fragile Families Study. Just as Sweeney (2002b) merged identified voting records with deidentified medical records to create identified medical records (Figure 7), we tried to find information an attacker might merge with the Fragile Families Challenge data.

After envisioning and investigating many types of auxiliary data, we used one such source to attack our own data. This in-house attack demonstrated that many respondents were unique in both the Fragile Families Study and the auxiliary data source, and it led to modifications of our data that we will describe later. We will illustrate the general structure of our attack with a hypothetical example. Suppose that the Fragile Families Study collected data on voting behavior, recording in each wave of data collection variables such as sex, age, party identification, and whether the respondent voted in the most recent primary and general elections. These variables are also available in identified voter registration databases, which would allow an adversary to merge them (see Figure 8).[14] A key feature of this hypothetical attack, as well as our real in-house attack, was that it did not involve the use of ZIP code or any other geographic information. Thus, even though the two real attacks we described previously used ZIP code (Figure 7), we wish to highlight the fact that it is still possible to reidentify data even if it does not contain obvious geographic information.

The process of attacking our data was useful for three reasons. First, it helped us realize how a small number of variables could substantially aid reidentification. In particular, continuous variables such as age made reidentification especially easy because they differentiated people into many groups. We decided to redact or modify variables that we thought were most likely to be the target of an attack, as we describe more later. Second, our attempts to attack our data made us realize the difficulty of obtaining auxiliary data at the national level. An attacker of the Fragile Families Challenge file would likely need a data set that is national in scope, but state differences in data sources make assembling such a data source difficult (but certainly not impossible). Finally, attempting to attack our own data made our fears about reidentification concrete and produced a clear way to explain the risks to our Board of Advisors and to other stakeholders.

[14]The ability of this hypothetical merge to succeed depends on how unique people are in the data set and in the population. As an extreme example, imagine a child's mother who was 24 years old in 2000. Suppose this mother registered with the Green Party in 2000, the Republican Party in 2002, and the Democratic Party in 2004. In each year, she voted in the primary election but not the general election. This respondent's particular combination of variables may be unique in the entire U.S. population, and this person would be at risk in this kind of merge. More generally, the more unique each person is, the more vulnerable he or she would be to this kind of attack.

This in-house reidentification attack was part of a larger process of moving away from a general fear of reidentification toward specific, actionable worries about particular people with a dangerous combination of capability and incentives. Next, we summarize the five biggest threats on which we focused, roughly ordered by the amount of risk we think they posed. Then, we briefly discuss other threats on which we did not focus during the Challenge but that might be important in other settings. After describing these threats, we describe the six main steps we took to mitigate those threats, roughly ordered by our perception of their importance (Table 1). When describing our threat modeling and threat mitigation process, we will be intentionally vague at certain points because versions of the Fragile Families Study data exist in the research community outside of our control, and we do not wish to increase the risk for a reidentification attack in the future.

Threats

Threat 1: A Privacy Researcher.—Privacy researchers represent an important "threat" to any social data set. They have the skills to reidentify the data and the incentives to conduct and publicize an attack. Although we describe privacy researchers as a "threat," we wish to emphasize that these researchers have good intentions. Some privacy researchers undertake attacks with the goal of developing defensive techniques that can prevent future attacks. Other privacy researchers might seek to illustrate privacy problems in a particular data set (e.g., Zimmer 2010), with the goal of encouraging other data stewards to be more careful before a true adversary finds the problems.[15]

Because privacy researchers represent an important and sometimes misunderstood threat, it is helpful to briefly describe the history and norms of this community so that other researchers can better understand their motivations. Today's information security and privacy research community traces its intellectual lineage in part to the field of cryptography and its military applications, in which secrets must be defended against powerful adversaries, with human lives at stake (Kahn 1996). The research culture of cryptography is shaped by its painful history of naively optimistic claims of "unbreakable" ciphers. Centuries of failures of such claims gradually established the importance of adversarial analysis as a scientific technique. Computer scientists today believe that scientific rigor in data privacy protection technology can be obtained only if claims are subject to adversarial scrutiny (Menezes, Van Oorschot, and Vanstone 1996). Furthermore, in the absence of a specific known adversary, privacy researchers believe that the prudent course of action is to assume the union of all such adversaries (i.e., a very powerful one). Assuming a capable adversary with complete knowledge of the system is a central principle in cryptography known as Shannon's maxim (Shannon 1949).

Privacy researchers consider real data releases to be valuable targets for demonstrations (i.e., Sweeney 2002b; Narayanan and Shmatikov 2008; Zimmer 2010), because they believe that work with toy data sets has limited ecological validity. Apart from the scientific benefit, reidentification demonstrations on real data sets are seen as ways to warn consumers of risks and disincentivize bogus claims of security. However, not all real data releases are

[15]On the basis of our experience, we believe that this strategy is effective. The threat of privacy researchers caused us to be more careful.

equally attractive targets, at least for academic privacy researchers. It is difficult to publish reidentification research unless there is novelty in the method, and this desire for novelty has served as a check on the number of such studies in practice. In other words, ironically, data sets that are too easy to reidentify are likely to escape the attention of privacy researchers.

What about the risks of reidentification research? Debate around this question is informed by the debates on the ethics of offensive computer security research more generally. In short, this community has concluded that it must engage in privacy attacks in order to anticipate and mitigate weaknesses in data protection before they are discovered by more nefarious adversaries.

Threat 2: A Nosy Neighbor.—In addition to privacy researchers, a very different kind of threat comes from a group of people with very different motivations and knowledge, a group often called *nosy neighbors*. These people often have a specific interest in a single person in the data set and already have substantial auxiliary information about that person. In the Fragile Families Challenge, we imagined that a mother might want to reidentify the data to learn about the survey answers provided by the father.[16] The linked nature of the Fragile Families data makes a nosy neighbor attack easier because the attacker could herself be in the data. They would only need to find themselves in order to learn more about the responses of the people in their family.

Threat 3: A Troll.—Third, we considered the possibility of a *troll* who might attempt to reidentify the data because he or she enjoys causing trouble or seeking attention (see Phillips 2015). Although some adversaries (i.e., a privacy researcher) might be able to harm our academic careers, a troll who posted respondents' information online might actually harm survey respondents.

We also considered the possibility of a "hacktivist" who might attempt to reidentify the data to make a larger political point. For instance, the hacker group Anonymous publicly posted the names and social media profiles of members of the Ku Klux Klan in 2015. When doing so, they wrote, "We hope Operation KKK will, in part, spark a bit of constructive dialogue about race, racism, racial terror and freedom of expression" (Franceschi-Bicchierai 2015). Might there exist adversaries who have similarly negative feelings toward social scientists doing research on a disadvantaged population? We would hope that potential hacktivists would recognize our good intentions, but we could not rule out the possibility that someone would attack the study to make a statement against our project or against social science research in general.

[16]The threat of a nosy neighbor attack can be illustrated through the example of the Netflix Prize, a mass collaboration that partially inspired the Fragile Families Challenge and that was subject to a reidentification attack (Narayanan and Shmatikov 2008). In 2006, Netflix offered a $1 million prize to the team that could most improve its movie recommendation algorithm. Netflix released a data set of ratings made on specific movies by specific users. Researchers were challenged to use the ratings in the public data to predict held-out movie ratings. Before Netflix released the data, they took some steps to prevent reidentification, and the data appeared to be anonymized; they consisted solely of movie ratings without any explicit individual identifiers. However, Narayanan and Shmatikov (2008) found that the vast majority of users had histories of movie ratings that were unique in the sample and were statistically likely to be unique in the population. If one knew some of the movies one of these individuals had watched, one could reidentify the user's row in the data and see all of the movies the user had rated. As the authors described, "a water-cooler conversation with an office colleague about her cinematographic likes and dislikes may yield enough information" (Narayanan and Shmatikov 2008). Given this water cooler information, a nosy neighbor would need only minimal technical abilities to reidentify the colleague by retrieving the user record that most closely matched the known likes and dislikes.

Threat 4: A Journalist.—A fourth adversary we considered was a journalist. For example, in 2006, America Online made the searches of thousands of users available to the public, assuming that one's search terms would not be easily traceable back to an individual's identity. Contrary to their expectations, Barbaro and Zeller (2006) wrote a widely read *New York Times* article revealing the identity of one woman whose search information had been included in the release. We worried that a journalist could do the same thing with the data from the Fragile Families Challenge. Like a privacy researcher, a journalist might be motivated to attack the data to make a larger point. However, we believe that a journalist bound by the norms of his or her profession would be unlikely to intentionally harm study participants. Therefore, we reasoned that journalists posed a greater risk to us, as organizers, than to the survey respondents.

Threat 5: A Cheater.—Finally, prior challenges have been won through strategies involving reidentification (Narayanan, Shi, and Rubinstein 2011). We worried that, if we set up the Fragile Families Challenge with a big prize and no clear prohibitions on linkage to auxiliary data, someone might try to win by reidentification. An adversary who reidentified the respondents could contact them and discover their outcomes, thereby achieving remarkably successful predictive performance.

Other Threats.—These five adversaries—privacy researcher, nosy neighbor, troll, journalist, and cheater—were the ones we considered most closely in our threat modeling. However, this list is not exhaustive of the threats we considered or the threats that might arise in other situations. For instance, three other adversaries we considered, and which might be more important in other settings, are governments, criminals, and companies.

Certain parts of the U.S. government most likely have the skills and rich auxiliary data that would be needed for a reidentification attack of our data. They might also have the incentive if our data contained information they could not find elsewhere, such as if respondents in our survey had reported on their experience as spies for a foreign government. Because we deemed the information in our data to be of little value to the U.S. government, we did not believe the it had an incentive to conduct such an attack in our setting.

A different set of attackers might be motivated by financial gain. For example, companies seeking to build databases for targeted marketing might try to acquire large social data sets. Given the size, structure, and deidentification of our data, however, we believed that it would be unattractive to companies. Furthermore, sophisticated criminals might wish to attack a data set containing credit card numbers or containing compromising information on wealthy individuals that could be used as the basis for blackmail. Our data do not contain information such as credit card numbers, and we reasoned that an adversary with the goal of finding compromising information on elites would more likely target other data sets. Therefore, we believed that an attack motivated by financial gain, either by a company or criminal, was unlikely in our setting.

Although our threat modeling was focused mainly on reidentification attacks that would occur through a merge with auxiliary information, other attacks were also possible. For example, we considered that someone might attempt to learn the identity of the survey

respondents by breaking into the Fragile Families Study offices and physically stealing computers. We deemed this possibility extremely unlikely, mostly because we did not see a clear incentive to carry out this attack. Furthermore, there are defenses in place that would make this attack more difficult than it appears. These defenses also make the possibility of accidental leak of information extremely unlikely. Overall, we would recommend that other researchers conduct a similar threat modeling exercise, keeping in mind that the threats in each situation might be different.

Threat Mitigation

There is no way with present technology to completely eliminate the risks these threats pose while achieving the scientific objectives of the mass collaboration. Nevertheless, we took six main steps to make an attack more difficult and less attractive. The columns of Table 1 represent these steps and their expected efficacy against various adversaries. We have ordered our actions in terms of our perception of their importance, from most to least important.

Low Profile.—When organizing a mass collaboration, one might seek press coverage in major national and international venues. Such publicity would help attract the widest and most diverse pool of participants possible, but it could also increase the risk for attack. High-profile studies are more likely to attract the attention of a nosy neighbor, a troll seeking attention, or a journalist or privacy researcher looking to reidentify a project that will draw interest from a wide readership. A low-profile study is less likely to be noticed by these adversaries and might be a less attractive target.

During the Fragile Families Challenge, we decided to keep a relatively low profile, and we focused our outreach on settings with a high probability of yielding participants who could contribute and a low probability of yielding participants who might attack the data.

For example, to raise awareness about the Challenge, we e-mailed the directors of population centers funded by the National Institutes of Health, and we hosted getting-started workshops at universities, in courses, and at scientific conferences. Our strategy of keeping a low profile can be easily adopted by other mass collaborations.

Careful Language.—Many data stewards may not realize it, but using careful, precise, and humble language may help prevent an attack from a privacy researcher or journalist. For instance, Zimmer (2010) made an ethical example out of the Tastes, Ties, and Time study in part because the original authors made strong statements such as “all identifying information was deleted or encoded” (quoted by Zimmer 2010 from the original study codebook). Privacy researchers may wish to correct data stewards who make overly confident statements about the deidentified nature of their data. By choosing language carefully, data stewards can be more honest about the safety of their data, thereby removing the need for privacy scholars to correct them. Instead of saying that “all identifying information was deleted,” one might say, “we removed information that was obviously identifiable.” Instead of writing that data are “anonymized,” data stewards should write that steps were taken to make the data less identifiable. Small changes in language can help convey that data stewards understand the

privacy risks and have made a reasoned judgment to proceed anyway. For example, we wrote the following on the Fragile Families Challenge Web site:

> All participants in the Fragile Families and Child Wellbeing Study have consented to have their data used for social research. These procedures, as well as procedures to make de-identified data available to researchers, have been reviewed and approved by the Institutional Review Board of Princeton University (#5767). The procedures for the Fragile Families Challenge have been reviewed and approved by the Institutional Review Board of Princeton University (#8061). In addition, we have also taken further steps to protect the participants in the Fragile Families and Child Wellbeing Study. If you would like to know more, please send us an email.

We believe that the relatively easy step of using careful language and inviting contact from potential attackers decreased the risk for attack.

Structure of the Challenge.—The structure of the Fragile Families Challenge was also designed in part to decrease the incentives to attack the data. By avoiding asking Challenge participants to predict sensitive outcomes, such as involvement in the criminal justice system or sexual behavior, we think we reduced the risk for attack from a privacy researcher, journalist, or troll. Avoiding potentially sensitive outcomes was part of keeping a low profile, and we suspect that it was important.

We also built certain things into the Challenge that would decrease the risk that a cheater would attack the data. First, in contrast to other high-profile challenges that offered large monetary prizes (Bennett and Lanning 2007), we chose to reward the best submissions with a trip to Princeton University to discuss their work. This prize was designed to emphasize an intrinsic goal of knowledge creation rather than an extrinsic financial incentive, thereby reducing incentives to cheat. Second, we emphasized in all promotional materials that the Challenge was a mass collaboration and a new way of working together, not a competition. For example, the banner at the top of our Web site read, “What would happen if hundreds of social scientists and data scientists worked together on a scientific challenge to improve the lives of disadvantaged children in the United States?” In the approval process, we also asked people about their motivations to participate, thereby encouraging them to reflect on their reasons. Nearly all participants responded with motivations that involved participation in scientific research or helping disadvantaged children. Third, we made the rules of the Challenge very clear: linking to auxiliary data sources was not allowed, and anyone who did so would be disqualified. Fourth, we required all participants to upload their code and narrative explanations along with their predictions. If a cheater successfully attacked the data and merged in outside information, this process would have to be obscured in whatever code was uploaded. Altogether, we think these aspects of the structure of the Challenge decreased incentives to attack the data.

Application Process.—In the interest of open and reproducible science, many have argued that research data should be made public. Although we agree with the spirit of this call, we join others making the more modest call for open sharing of data with allowable constraints when privacy or other concerns must be balanced against the goal of open science (Freese and Peterson 2017; Salganik 2018). In particular, an application process can

help ensure that only those who can plausibly yield scientific benefit be given access to the data and that participants who pose increased risk can be monitored more closely.

For the Fragile Families Challenge, we developed a process to screen applicants (Figure 9). Initially, people interested in participating completed an application that asked for information about their educational background, research experience, and motivations to participate in the study (see Appendix for the exact application). Responses to the application were sent to an e-mail account checked by two of the Challenge's central organizers (Lundberg and Salganik). One of the Challenge organizers would provide an initial review of the application, in many cases searching the Internet to corroborate claims made in the application or to look for important information that was excluded.[17] The reviewer would make an initial recommendation as to whether a participant should be approved and would send a summary of the application to a team of eight reviewers, including all the authors of this article, survey administrators, and the principal investigators of the Fragile Families Study. Members of this broader review team were given 24 hours to raise any opposition to the application.[18]

After this review process, we required approved applicants to type a set of statements acknowledging their agreement with our terms and conditions (the full set of statements is provided in the Appendix). The purpose of the terms and conditions was not to screen participants but to make sure they understood their ethical responsibilities.

After an approved applicant had agreed to the terms and conditions, we e-mailed him or her a link to an encrypted data file (the link automatically expired within seven days). Finally, the approved participant had to call us by phone to receive the password to decrypt the file. This last step reduced the risk that an e-mail could be intercepted and also provided an opportunity to speak in person with participants and thereby reinforce that the project was a mass collaboration involving real people, not a competition to be won at all costs.

During the Fragile Families Challenge, we received 457 applications,[19] and most were uneventful. However, about 10 applications raised yellow flags that we addressed either by sending a follow-up e-mail or by having a discussion with the applicant (in person or by video chat). These conversations helped ensure mutual understanding of the importance of respecting the privacy of respondents. For instance, one applicant told us the applicant was working on the "record linkage problem with massive data." Another applicant told us of a research interest in reidentification. In both cases, we were glad to have spoken with the applicants before providing the data. In one case, a researcher who had previous experience doing reidentification attacks did not respond to our follow-up e-mail and was therefore not granted access to the data.

[17]Applications from students were sometimes difficult to evaluate, given that students often have little research experience to report. We occasionally spoke with a student's academic adviser to ensure that participation was being overseen by a responsible individual who understood the importance of respecting the data. In one case, we pointed a student toward alternative data sets that were less sensitive and would serve as equally useful for that student's project.

[18]We occasionally omitted this waiting period when individuals were participating in a known setting, such as a class assignment or a workshop in which we spoke directly with potential participants.

[19]The first 100 of these applications came in the pilot test we ran in an undergraduate machine learning class at Princeton and followed a slightly different format from the rest because we modified the application procedure on the basis of feedback from the pilot test.

One might worry that this screening process—which involved many hours of work from the Challenge organizers, graduate students, study staff members, and principal investigators—was completely futile because people could just lie in their applications. However, from our threat modeling, we suspected that well-intentioned privacy researchers might represent our greatest threat. These individuals believe strongly in the ethical use of data, and we suspected that they would be honest in their applications. Because those who posed the greatest threat also seemed unlikely to lie in an application, we believe that our screening process was a worthwhile endeavor.

One final aspect of the application process is worth describing. Initially, we were worried about what might happen if we actually rejected an applicant. Might this rejection anger someone and turn him or her into an even more motivated adversary who would attempt find the data elsewhere (i.e., from another participant) and then attack them? Because of this concern, we made preparations to accept high-risk applicants as *local participants*. For these local participants, we prepared a special, secure computer in the office of the Fragile Families project director. This computer was configured such that it was difficult to bring data onto the computer or take data off the computer (e.g., drives were deactivated, and the computer was not connected to the Internet and was locked to the wall). However, we maintained the ability to reconnect this computer to the Internet ourselves to upload a Challenge participant's predictions if needed. Fortunately, we never had to use this secure computer, but it provided an opportunity for us to remain open to the potential participation of even the riskiest applicants.

Ethical Appeal.—It is not possible to force people to act ethically, nor is it possible in a screening process to determine with full confidence whether an applicant plans to act ethically. However, one can clarify for participants the ethical implications of misusing the data, thereby creating the possibility that those who might plan to reidentify the data will consider this plan from an ethical perspective before proceeding. After participants' applications were approved, they completed a set of terms and conditions designed to achieve this goal. Every participant read the following statement, written in boldface type, before participating.

> The Fragile Families and Child Wellbeing Study is a dataset of real people who have selflessly opened up their lives to us for the last 15 years so that their experiences can contribute to scientific research. By participating in the Fragile Families Challenge, you become a collaborator in this project. It is of the utmost importance that you respect the families in the data by using what they have told us responsibly.

After reading the statement, each participant typed a set of statements defining how he or she would use the data responsibly (see Appendix for specific statements). Participants could easily lie in this section, so it would not have stopped someone determined to act unethically. However, by outlining our expectations, we clarified a positive vision of what ethical behavior would entail.

Modifications to Data.—The final step we took to mitigate threats was to modify the data in the Fragile Families Challenge file. Overall, these modifications were relatively minor

because the preexisting Fragile Families Study files had already undergone an extensive process to promote privacy (Figure 4).

First, as described earlier, we attempted to attack the data ourselves, and this led us to make certain changes in the structure of the background data file for the Challenge. We choose not to describe these changes fully, but we note that some were inspired by prior reidentification demonstrations. We did not implement a *k*-anonymity approach (Sweeney 2002b),[20] but we did make more minor changes such as adding noise to key variables to make a simple one-to-one merge more difficult.[21] Although adding noise has a long tradition in the statistical literature (Kim 1986; Reiter 2012), it is not foolproof in the high-dimensional setting.[22] Despite its limitations, adding noise to key variables at least makes a simple one-to-one merge with auxiliary data more difficult. For instance, we worried that height, weight, and body mass index could be key linking variables if there were a breach of identified medical records data in the future, so we added a small amount of noise to these variables. We also added noise to other variables, which we choose not to identify. The noise we added came from an independent draw for each variable, with the exception of variables that measured the same construct (i.e., height in inches and height in centimeters). In these cases, we drew one noise term per child per construct, so that an adversary could not reduce the size of the noise by averaging over multiple responses. Overall, we added noise to a few hundred variables. To minimize the risk for unnecessary harm to the scientific promise of the project, the noise we added was always relatively small,[23] and Fragile Families Challenge participants were generally unaware of the added noise.[24] We suspect that adding noise made reidentification slightly harder with minimal harm to the scientific utility of the data.

In addition to seeking to make a reidentification attack more difficult, we also sought to reduce the magnitude of harm should reidentification occur by redacting some information we thought might pose a serious risk to survey respondents. When considering whether to redact information for harm reduction, we tried to weigh how much harm might come to how many people against how much the redaction might jeopardize the scientific goals of the project. While making these decisions, we were aware that in a large data set, it can be difficult to know which information will be sensitive (Salganik 2018, chap. 6).

---

[20] *k*-Anonymity states that data should be released only if each row is exactly the same as at least $k-1$ other rows, so that an adversary with perfect auxiliary data would be able to link a given individual with no fewer than *k* records (Sweeney 2002b). The higher the value of *k* chosen, the more difficult it would be for an adversary to find all candidates and discover the true match. *k*-Anonymity can be achieved by suppressing information on key features of certain individuals that would otherwise make them unique (Sweeney 2002a). Unfortunately, *k*-anonymity quickly becomes infeasible as the number of features grows, because rows are rarely exactly the same on the joint set of all features (Aggarwal, 2005), which means that a pure *k*-anonymity approach would substantially harm the utility of the data (Brickell and Shmatikov 2008). Despite the impossibility of applying *k*-anonymity to all 12,942 variables, we did apply the spirit of this approach to try to reduce the number of participants who were unique or nearly unique in the sample on a subset of variables.

[21] The noise we discuss here is distinct in two ways from the noise added in differential privacy. First, we added noise to the inputs to participants' models, whereas differential privacy adds noise to the outputs. Second, we added normally distributed noise, a choice distinct from the differential privacy approach of heavy-tailed Laplacian noise. Our approach to adding noise does not yield provable privacy guarantees.

[22] For example, Netflix added noise to some of the data used in the Netflix Prize, but this did not prevent a reidentification attack (Narayanan and Shmatikov 2008).

[23] For more on the risks of adding too much noise, see Brickell and Shmatikov (2008).

[24] In one case a participant noticed that the distribution of some variables did not match that listed in the official study documentation, thereby correctly realizing that we had changed something in those variables. This participant e-mailed us, and we were happy to explain what we had done.

Finally, we decided that we did not want the Fragile Families Challenge data files to include anything that might obviously provide information about the survey respondent's location. Unlike adding noise, we think this decision to strip geographic information may have decreased Challenge participants' ability to predict the six key outcomes. However, we think that removing obvious geographic information made a reidentification attack much more difficult. If an attacker were able to place a group of survey respondents in a particular city, the attacker would only need an identified auxiliary data source that includes that city (e.g., state voting records). However, if the respondents could be anywhere in the country, the attacker would need national data, which might be harder to acquire.

Although many privacy audits focus mostly on modifications to the data, we saw this as just one of the six steps we took to mitigate threats. In fact, we think that many of other steps, such as keeping a low profile, were more important in this setting.

Comprehensive Assessment

After implementing the threat modeling and threat mitigation described above, which evolved over a period of months, we stepped back and comprehensively assessed who might have incentives to attack the data and what barriers might stand in their way (see Table 1).

A privacy researcher would have to learn and care about the project despite our low profile. Even then, privacy researchers would have an incentive to attack the data only if they wanted to show publicly that we had done something wrong; our careful language incorporating the findings of privacy researchers mitigated against this danger. The privacy researcher would then have to lie in the application process or otherwise convince us that he or she did not intend to reidentify the data, and he or she would have to proceed in the face of our ethical appeal. To the extent that privacy researchers are motivated by an ethical call for researchers to recognize the limits of privacy and handle data with care, we expected that these steps substantially reduced the risk that a privacy researcher would reidentify the data.

Nosy neighbors are almost impossible to stop with technical barriers; enormous changes to the data would be needed to render a respondent unrecognizable to a nosy neighbor. By keeping a low profile, we reduced the likelihood that a nosy neighbor would learn about the Fragile Families Challenge. By screening participants, we increased the chance that we would notice a nosy neighbor before sharing the data. With these changes in place, we determined that the risk of nosy neighbors increased only negligibly from the Challenge, compared with the risk that already existed from the availability of the Fragile Families Study data to researchers.

A troll might lie through the application process, be undeterred by an appeal to ethics, and have the technical skills to overcome our modifications to the data. However, a troll is likely to pursue the highest profile targets for attention, so a low profile reduces the chance of a troll attacking the data.

A journalist would primarily be stopped by the application process: we expect that journalists, because of the norms of their field, would not lie when applying to use the data. Even a journalist who made it through the screening process would have to find the

project interesting despite its low profile, ignore our appeal to ethics, and succeed in the technical difficulties of reidentification despite our modifications to the data. This seemed unlikely.

The threat from a cheater was primarily mitigated by the Challenge structure, which explicitly stated that anyone who cheated would be ineligible for a prize, which was nonmonetary to begin with. With this structure, we doubted anyone would reidentify the data for the sole purpose of winning the Challenge.

To summarize, the design of the Fragile Families Challenge did not completely eliminate any of the threats we imagined, but it mitigated them to a sufficient degree that we concluded that the Challenge only slightly increased the risk to participants above that which already existed from the use of the Fragile Families Study data by independent researchers.

Response Plan

Having mitigated, but not eliminated, the risk for reidentification, we created a response plan in case something went wrong. More specifically, we took steps to ensure that we had the appropriate people involved in the project so that we could begin responding to a crisis quickly and forcefully. Our Board of Advisors included a computer scientist who had previously reidentified several data sets (Arvind Narayanan); a sociologist and lawyer with expertise in data privacy, surveillance, and inequality (Karen Levy); and a journalist (Nicholas Lemann). In the event of an attack from a privacy researcher, the organizers of the Challenge would consult with Narayanan when responding. If a social scientist or policy maker criticized the project for mishandling the private information of disadvantaged children, the organizers of the Challenge would consult with Levy when responding. If the data were reidentified by a journalist who planned to publicize the story, the organizers of the Challenge would consult with Lemann. Each of these individuals is respected in various communities that might attack the data, and we hoped that they could mediate and guide the project through any problems which might arise. We recommend that other data stewards take similar steps to prepare for the unexpected, both in the mass collaborative setting and in the more common setting of providing data for use by individual researchers.

Third-Party Guidance

It is easy for those coordinating a project to fall into group-think and ignore potential problems out of a common interest in the success of the project. To avoid this pitfall, we worked under the guidance of third parties. First, the Fragile Families Study and the Fragile Families Challenge were undertaken with the oversight of the Institutional Review Board at Princeton University.[25]

We created an additional community for third-party guidance by assembling a Board of Advisors. The board included professors of sociology, education, and social work who had each devoted much of their careers to studying disadvantaged families, a journalism

[25]Components of the Fragile Families Study were also reviewed by the institutional review boards of partner institutions (i.e., Columbia University and Westat, the data collection contractor).

professor, a machine learning researcher, and a computational social scientist, some of whom are authors of this article. During our privacy and ethics audit, we sent the board weekly updates about our progress. These weekly updates were also shared with staff members working on the Fragile Families Study, as well as other stakeholders. We found this process of weekly updates incredibly valuable both to sharpen our thinking and to ensure that all stakeholders were involved.

Finally, we also sought informal advice from a wide variety of people not involved in the Fragile Families Challenge in any way. These outsiders included a philosophy professor, a member of the military with experience planning high-risk operations, a lawyer with experience dealing with health records, and a public-interest lawyer who provides direct services to children in foster care. We found that these uninvolved third parties often provided an interesting perspective, and we would encourage other data stewards to have broad discussions if possible. Overall, we believe that third-party guidance helped improve our process.

Ethical Framework

During our process we had to make many different decisions, and these decisions were guided by the principles described in the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1978), a foundational document in social science research ethics. More specifically, the Belmont Report emphasizes three principles that should guide research ethics: respect for persons, beneficence, and justice.

Respect for Persons

The principle of respect for persons means that researchers should respect each participant's autonomy to choose whether to participate after being informed about the nature of the research. This principle is often operationalized as the process of informed consent. In this case, when parents and children agreed to participate in the Fragile Families Study, they understood that the data they provided would be used for scientific research. Therefore, when designing and conducting the Challenge, we remained within the context of scientific research as much as possible in order to honor the agreement between survey respondents and researchers. For example, when structuring the Challenge we decided not to offer prize money, because that is not generally consistent with the context of scientific research. Our thinking about preserving context was heavily influenced by Nissenbaum's (2009) work on the importance of "contextual integrity."

Beneficence

The second key principle discussed in the Belmont Report is beneficence: the potential harm of the study must be weighed against the potential benefits. Furthermore, reasonable steps should be taken to maximize benefits and minimize harms, both probability and magnitude. These ideas were central to our process of threat modeling and threat mitigation. For example, these ideas affected our changes to the data. We hoped to reduce the probability of harm by redacting or adding noise to variables that we expected might aid reidentification,

and we hoped to reduce the magnitude of harm should reidentification occur by redacting some information (e.g., illegal behaviors). However, we made these decisions while also weighing the potential benefits of including variables that might help contribute to scientific knowledge. For instance, we did not redact information about child behavior problems, because this information might be an especially important predictor of adolescent outcomes. The principle of beneficence thus affected not only our decision to release the data but also our decision about which variables ought to be released.

In addition to risks and benefits for participants in the Fragile Families Study, we also considered possible broader social impacts of our work (Zook et al. 2017). Predictive models similar to those built during the Challenge are increasingly being used for high-stakes decisions, such as in criminal justice (Berk et al. 2017) and child protective services (Chouldechova et al. 2018). Although these predictive models can improve social welfare (Kleinberg et al. 2015, 2017), they can also discriminate against protected groups (Barocas and Selbst 2016) and magnify social inequality (Eubanks 2018). Therefore, we weighed the possibility that the knowledge created in the Fragile Families Challenge could be used inappropriately against the possibility that it could be used to help policy makers who are seeking to understand the strengths and weaknesses of predictive modeling. Because of the type of data we used and the outcomes we selected, we believe that the risk for unintended secondary use is low and that it is outweighed by the possible social benefits of the research. Because the risks for the Challenge, both to survey participants and to society, were very difficult to quantify, we found that existing ethical, philosophical, and legal debates about the precautionary principle (Narayanan et al. 2016; O'Riordan and Cameron 1994; Sunstein 2003, 2005) and dual use research (National Research Council 2004; Selgelid 2009) helped guide our thinking.

Justice

The final principle in the Belmont Report is justice: risks and benefits should flow to similar populations. Unfortunately, many failures of scientific research ethics have involved disadvantaged or vulnerable populations (for examples, see Emanuel et al. 2008). Informed by this history, social scientists today recognize a special obligation to make ethical decisions about research involving vulnerable populations. The fact that some participants in the Fragile Families Study are disadvantaged or children (or both) caused us some concern. However, the nature of the Challenge meant that the population most likely to benefit from the scientific knowledge produced by the Challenge was the very population from which participants were drawn. In other words, by conducting research on a sample of disadvantaged urban families, we can generate knowledge that might help improve the life chances of future families in similar positions. Thus, our approach to the principle of justice is heavily influenced by the argument that no group should be excluded from the potential to benefit from research (Mastroianni and Kahn 2001).

Decision

Ultimately, after all the discussing, designing, and debating, we faced a decision: whether to go forward with the Fragile Families Challenge or not. When forced to make a go/no-

go decision, we conducted a comprehensive reassessment. We believed that the project was consistent with existing ethical rules and ethical principles governing social science research. This rules-based decision was made by our Institutional Review Board, and the principles-based decision was made by us and by members of our Board of Advisors. We also believed that after our threat mitigation, the risks were low in an absolute sense and in appropriate balance with the potential for scientific and societal benefit.

Before a full-scale launch, however, we conducted a pilot test in an attempt to discover any errors in our threat modeling, threat mitigation, or ethical thinking. We conducted the pilot test by deploying the Challenge as a project in an undergraduate machine learning class at Princeton University. In doing so, we faced a difficulty inherent in pilot testing: being realistic while also being safe. By conducting the test in a class taught by a trusted professor, we erred on the side of safety. In our case, the pilot test turned out to be useful; it helped us discover one variable in the Challenge data file that was both confusing to participants and increased the risk for reidentification. The pilot also gave us a chance to test our screening process, which we modified for the full launch to gather more information and facilitate the process for both participants and organizers. The full launch of the Challenge was much smoother because we started in a controlled setting, and we highly recommend that others conduct a similar pilot test.

Conclusions

This case study describes the privacy and ethics audit that we conducted as part of the Fragile Families Challenge. Our process was certainly not perfect, and we hope that by being open about it, others can improve on what we did.[26] We want to emphasize that other data stewards may reasonably come to different decisions about how to strike an appropriate balance in their own situation. Despite differences between situations, however, we believe that the key elements of our approach—threat modeling, threat mitigation, and third-party guidance, all within a specific ethical framework—may be useful to other data stewards, whether they reside in universities, companies, or governments.

Those who might wish to follow or build on our example will undoubtedly wonder about the costs of doing so. The approach we followed was time-consuming. We spent about three months preparing to launch the Challenge, and the privacy and ethics audit was the most time-consuming part (see Figure 2). The parts that took the most time were assembling a team with appropriate expertise (including for the response plan), attacking our data, debating how to balance the various trade-offs, and keeping in contact with stakeholders. Attacking the data, by itself, involved about 1.5 months of consistent work for a skilled graduate student. Beyond the amount of time involved, this process was also stressful and emotionally taxing; some of us found it difficult to spend so much time imagining all the worst-case scenarios of reidentification.

[26] For more on a process-based approach to data access, see Rubinstein and Hartzog (2016).

Although our privacy and ethics audit was time-consuming, we feel that the effort was worthwhile. In particular, this process enabled us to run the Challenge, which has already started to achieve some of its scientific objectives, as illustrated by this special collection.

Finally, we hope that this case study illustrates that there is an important middle ground between calls for no data sharing and complete data sharing. Everyone will benefit if scientists, companies, and governments can continue to develop the technical, legal, ethical, and social infrastructure to enable safe and responsible data access.

Acknowledgments

We thank the Board of Advisors of the Fragile Families Challenge for supporting the project and offering feedback throughout. We also thank Brandon Stewart, Alex Guerrero, Andrew Ledford, Julien Teitler, Sam Salganik, Amanda Slater, Prateek Mittal, and Vitaly Shmatikov for helpful conversations, and we thank Kristin Catena, Kate Jaeger, Ryan Vinh, Nathan Matias, and Jessica West for helpful feedback on drafts of this manuscript. Participants in the Princeton Sociology Proseminar, the Northeast Privacy Scholars Workshop, and the Center for Information Technology Policy reading group provided valuable feedback as well. This research was approved by the Institutional Review Board of Princeton University (Protocol 8061). The content of this paper is solely the responsibility of the authors and does not necessarily represent the views of anyone else.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Research reported in this publication was supported by the Russell Sage Foundation and the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under Award Number P2C-HD047879, and the National Science Foundation (NSF IIS Award 1704444 to AN). Funding for the Fragile Families and Child Wellbeing Study was provided by the Eunice Kennedy Shriver National Institute of Child Health and Human Development through grants R01-HD36916, R01-HD39135, and R01-HD40421 and by a consortium of private foundations, including the Robert Wood Johnson Foundation.

Biographies

Ian Lundberg is a PhD candidate in sociology and social policy at Princeton University. His research focuses on the use of statistical and machine learning methods in the study of stratification and inequality. Beyond concerns about ethical data sharing, those seeking to apply new statistical developments face several hurdles: formalizing the estimand precisely in relation to a theoretical claim, stating and defending identification assumptions, and dealing with practical problems such as missing data. Lundberg's research addresses these hurdles in substantive applications including the predictability of adolescent well-being, patterns of social mobility over multiple generations, the prevalence of housing eviction among U.S. children, and the effect of marriage on men's wages. The common goal across these applications is to improve statistical practice in stratification and inequality research.

Arvind Narayanan is an associate professor of computer science at Princeton University. He leads the Princeton Web Transparency and Accountability Project to uncover how companies collect and use our personal information. Narayanan also leads a research team investigating the security, anonymity, and stability of cryptocurrencies as well as novel applications of blockchains. He co-created a massive open online course as well as a textbook on Bitcoin and cryptocurrency technologies. His doctoral research showed the fundamental limits of deidentification, for which he received the Privacy Enhancing Technologies Award. Narayanan is an affiliated faculty member at the Center

for Information Technology Policy at Princeton and an affiliate scholar at Stanford Law School's Center for Internet and Society.

Karen Levy is an assistant professor of information science at Cornell University and associated faculty member at Cornell Law School. Her research investigates social, legal, and ethical implications of emerging technologies, with a focus on surveillance and privacy. Her work considers these topics across several domains, including the workplace, intimate relationships, and online exchange platforms. She is currently a New America Fellow, a Faculty Fellow at Cornell's Institute for the Social Sciences, and a faculty co-lead of Cornell's AI, Policy, and Practice Initiative.

Matthew J. Salganik is a professor of sociology at Princeton University, and he is affiliated with several of Princeton's interdisciplinary research centers: the Office for Population Research, the Center for Information Technology Policy, the Center for Health and Wellbeing, and the Center for Statistics and Machine Learning. His research interests include computational social science, social networks, and methodology. He is the author of *Bit by Bit: Social Research in the Digital Age* (Princeton University Press, 2018).

Appendix

This appendix includes reproductions of the web forms participants completed to (1) apply to participate and (2) agree to a set of terms and conditions of participation.

1. We used the application to assess each applicant's ability to contribute and motivations, and to gain advance notice of potential threats with whom to follow up.
2. We used the terms and conditions to reinforce norms of scientific behavior and emphasize the importance of respecting the data. The terms and conditions also include a set of items about how we would use the submissions of Challenge participants.

Application to Participate

For participants working in teams, each individual will need to complete this application.

Describe your previous research, coursework, or work experience that is related to the Fragile Families Challenge. We expect that the most relevant experience will come from social science and data science, but we are open minded. Please include links to your CV, your published papers, and your open source software (as appropriate).

Describe your previous experience working with restricted-access data. This could be in an academic, corporate, or governmental setting. Please be specific. If you have no experience with restricted-access data, please let us know, and it will not automatically disqualify you from participating.

Describe the context in which you plan to participate. For example, are you planning to participate at a university, company, or government agency? If you are doing this in a course, please tell us the name of the professor.

Please let us know your current employer or if you are a student, your current university. If you are self-employed, not employed, or retired, please say so.

Do you plan to participate in the Challenge in good faith and consistent with norms of scientific behavior?

- ○ Yes
- ○ No

What is motivating you to participate?

Beyond the open responses on the previous page, we're interested in learning more about a few specific aspects of the backgrounds and motivations of participants in the Challenge.

Your responses to these questions will only be used to help us understand the Fragile Families Challenge, and to design better challenges in the future.

There are many different reasons for wanting to participate in the Fragile Families Challenge. Which of the following best describe your reasons for participating? (Check all that apply.)

	Applies to me	Does not apply to me
General interest in topic	○	○

	Applies to me	Does not apply to me
Curiosity about the Challenge	○	○
Contributing to social science	○	○
To improve the lives of disadvantaged children	○	○
To make the best-performing model in the Challenge	○	○
For fun	○	○
Learning about cutting-edge research	○	○
Collaborating with colleagues/friends	○	○
Relevant to school or degree program	○	○
Collaborating with university researchers	○	○
Relevant to own research	○	○
Connecting with others who share my interests	○	○
Learning/practicing data analysis skills	○	○
Relevant to job	○	○
To work on a prestigious research project	○	○
To earn scholarly recognition	○	○
Required for coursework	○	○
To experience a mass scientific collaboration	○	○
To create the most interesting/innovative model in the Challenge	○	○
Collaborating with strangers	○	○
Contributing to data science	○	○

Which of the following best describes your academic background?

	Applies to me	Does not apply to me
Sociologist	○	○
Economist	○	○
Psychologist	○	○
Political scientist	○	○
Demographer	○	○
Data scientist	○	○
Computer scientist	○	○
No academic background	○	○
Other (please specify)	○	○

Which of the following best describes your professional background?

	Applies to me	Does not apply to me
Industry	○	○
Academia	○	○
Government	○	○

	Applies to me	Does not apply to me
Nonprofit	○	○
Undergraduate student	○	○
Graduate student	○	○
Not employed	○	○
Other (please specify)	○	○

Do you plan to participate in the Fragile Families Challenge individually, or as part of a group?

- ○ Individually
- ○ As part of a group

Are you participating in the Fragile Families Challenge as part of a class assignment?

- ○ Yes
- ○ No

Have you ever analyzed data from the Fragile Families and Child Wellbeing Study?

- ○ Yes
- ○ No

Have you ever published a paper using data from the Fragile Families and Child Wellbeing Study?

- ○ Yes
- ○ No

Terms and Conditions

You will be given access to the Fragile Families and Child Wellbeing Study (FFCWS) data for the Fragile Families Challenge. By completing this form, you agree to fulfill your responsibilities on this project according to the following guidelines.

- For each bolded statement, re-type the statement word for word in the text box.
- Additionally, please check all the boxes to agree with the sub-statements.
- Your signature at the bottom verifies your agreement with the entire document.

You may contact us at fragilefamilieschallenge@gmail.com with any questions.

The Fragile Families and Child Wellbeing Study is a dataset of real people who have selflessly opened up their lives to us for the last 15 years so that their experiences can contribute to scientific research. By participating in the Fragile Families Challenge, you

become a collaborator in this project. It is of the utmost importance that you respect the families in the data by using what they have told us responsibly.

I understand that harm could come to the survey respondents if their identities were made public. I will not do anything to harm the respondents. I agree not to attempt to identify individuals, families, households, or hospitals.

☐ Agree

Type the statement below to verify your agreement.

I agree not to attempt to identify individuals, families, households, or hospitals.

In the event that the identity of an individual, family, household, or hospital is discovered inadvertently, I will (a) make no use of this knowledge, (b) inform the Challenge organizers at fragilefamilieschallenge@gmail.com so that they can make changes to improve the security of the data, and (c) not inform any other persons of the discovered identity.

☐ Agree

Type the statement below to verify your agreement.

I agree to report any disclosure of participants or errors in data/documentation to the Challenge organizers through fragilefamilieschallenge@gmail.com.

I understand that each member of my research team must request their own copy of the data files. Files cannot be transferred. I will not at any time give, sell, show, disclose or otherwise disseminate the Fragile Families and Child Wellbeing Study data to anyone.

☐ Agree

Type the statement below to verify your agreement.

I understand that each member of my research team must request their own copy of the data files.

I will take steps to ensure that the data is safe-guarded, using protections such as password-protected access to all computers storing the data.

Type the statement below to verify your agreement.

I agree to store and process the data in a secure manner.

Once the Challenge is complete, I will destroy all the data files that I used during the Challenge. If I plan to continue research with this data, I will download a new version of the Fragile Families and Child Wellbeing Study data through the Princeton University Office of Population Research (OPR) Data Archive at: http://opr.princeton.edu/archive/.

☐ Agree

Type the statement below to verify your agreement.

I agree to destroy all copies of FFCWS data after the conclusion of this project.

Checking this box acknowledges that my name typed below constitutes my signature and indicates that I have read, understand, and agree to abide by the terms of this agreement.

☐ I have read, understand, and agree to abide by the terms of this agreement.

Signature

References

Aggarwal Charu C. 2005. "On ***k***-Anonymity and the Curse of Dimensionality." Pp. 901–909 in Proceedings of the 31st International Conference on Very Large Data Bases. Retrieved November 6, 2018 (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.3155&rep=rep1&type=pdf).

Barbaro Michael, and Zeller Tom Jr., 2006. "A Face Is Exposed for AOL Searcher No. 4417749." The New York Times. August 9, 2006.

Barocas Solon, and Selbst Andrew D.. 2016. "Big Data's Disparate Impact." California Law Review 104:671.

Beaulieu-Jones Brett K., Wu Zhiwei S., Williams Chris, and Greene Casey S.. 2017. "Privacy-preserving Generative Deep Neural Networks Support Clinical Data Sharing." bioRxiv. Retrieved November 6, 2018 (https://www.biorxiv.org/content/early/2017/07/05/159756).

Bennett James, and Lanning Stan. 2007. "The Netflix Prize." In Proceedings of International Conference on Knowledge Discovery and Data Mining Cup and Workshop. Retrieved November 6, 2018 (https://www.cs.uic.edu/~liub/KDD-cup-2007/NetflixPrize-description.pdf).

Berk Richard, Heidari Hoda, Jabbari Shahin, Kearns Michael, and Roth Aaron. 2017. "Fairness in Criminal Justice Risk Assessments: The State of the Art." arXiv. Retrieved November 6, 2018 (https://arxiv.org/abs/1703.09207).

Breiman Leo. 2001. "Statistical Modeling: The Two Cultures." Statistical Science 16(3):199–231.

Brickell Justin, and Shmatikov Vitaly. 2008. "The Cost of Privacy: Destruction of Data-mining Utility in Anonymized Data Publishing." Pp. 70–78 in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Retrieved November 6, 2018 (https://www.cs.cornell.edu/~shmat/shmat_kdd08.pdf).

Brier Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." Monthly Weather Review 78(1):1–3.

Carlson Marcia J., and Furstenberg Frank F. Jr., 2006. “The Prevalence and correlates of Multipartnered Fertility among Urban U.S. Parents.” Journal of Marriage and Family 68(3):718–32.

Chouldechova Alexandra, Benavides-Prado Diana, Fialko Oleksandr, and Vaithianathan Rhema. 2018. “A Case Study of Algorithm-assisted Decision Making in Child Maltreatment Hotline Screening Decisions.” Pp. 134–48 in Proceedings of the 1st Conference on Fairness, Accountability and Transparency, edited by Friedler SA and Wilson C. New York: PMLR.

Cox Lawrence H., Karr Alan F., and Kinney Satkartar K.. 2011. “Risk-utility Paradigms for Statistical Disclosure Limitation: How to Think, but Not How to Act.” International Statistical Review 79(2):160–83.

Crawford Jack L. H., Gentry Craig, Halevi Shai, Platt Daniel, and Shoup Victor. 2018. “Doing Real Work with FHE: The Case of Logistic Regression.” Cryptology ePrint Archive. Retrieved November 6, 2018 (https://eprint.iacr.org/2018/202.pdf).

Desmond Matthew, and Kimbro Rachel Tolbert. 2015. “Eviction’s Fallout: Housing, Hardship, and Health.” Social Forces 94(1):295–24.

Donnelly Louis, Garfinkel Irwin, Brooks-Gunn Jeanne, Wagner Brandon G., James Sarah, and McLanahan Sara. 2017. “Geography of Intergenerational Mobility and Child Development.” Proceedings of the National Academy of Sciences 114(35):9320–25.

Donoho David. 2017. “50 Years of Data Science.” Journal of Computational and Graphical Statistics 26(4):745–66.

Drechsler Jörg. 2011. Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation. Vol. 201. New York: Springer.

Duncan George T., Elliot Mark, and Salazar-González Juan Jose. 2011. Statistical Confidentiality: Principles and Practice. New York: Springer.

Dwork Cynthia. 2008. “Differential Privacy: A Survey of Results.” Pp. 1–19 in International Conference on Theory and Applications of Models of Computation. Berlin: Springer.

Dwork Cynthia, and Roth Aaron. 2014. “The Algorithmic Foundations of Differential Privacy.” Foundations and Trends in Theoretical Computer Science 9(3–4):211–407.

El Emam Khaled, Jonker Elizabeth, Arbuckle Luk, and Malin Bradley. 2011. “A Systematic Review of Re-identification Attacks on Health Data.” PLoS ONE 6(12):e28071. [PubMed: 22164229]

Emanuel Ezekiel J., Grady Christine C., Crouch Robert A., Lie Reidar K., Miller Franklin G., and Wendler David D.. 2008. The Oxford Textbook of Clinical Research Ethics. Oxford, UK: Oxford University Press.

Eubanks Virginia. 2018. Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor. New York: St. Martin’s.

Feuerverger Andrey, He Yu, and Khatri Shashi. 2012. “Statistical Significance of the Netflix Challenge.” Statistical Science 27(2):202–31.

Franceschi-Bicchierai Lorenzo. 2015. “Anonymous Hackers Officially Dox Hundreds of Alleged KKK Members.” Retrieved (https://motherboard.vice.com/en_us/article/kb7eyv/anonymous-hackers-officially-dox-hundreds-of-alleged-kkk-members).

Freese Jeremy, and Peterson David. 2017. “Replication in Social Science.” Annual Review of Sociology 43:147–65.

Gaboardi Marco, Honaker James, King Gary, Murtagh Jack, Nissim Kobbi, Ullman Jonathan, and Vadhan Salil. 2016. “Psi (Ψ): A Private Data Sharing Interface.” arXiv. Retrieved November 6, 2018 (https://arxiv.org/abs/1609.04340).

Gentry C 2009. “A Fully Homomorphic Encryption Scheme.” PhD dissertation, Stanford University.

Goroff Daniel L. 2015. “Balancing Privacy versus Accuracy in Research Protocols.” Science 347(6221):479–80. [PubMed: 25635075]

Hardt Moritz, and Blum Avrim. 2015. “The Ladder: A Reliable Leaderboard for Machine Learning Competitions.” Pp. 1006–14 in International Conference on Machine Learning. Retrieved November 6, 2018 (http://proceedings.mlr.press/v37/blum15.pdf).

Hardt Moritz, Ligett Katrina, and McSherry Frank. 2012. “A Simple and Practical Algorithm for Differentially Private Data Release.” Pp. 2339–47 in Advances in Neural Information Processing Systems. Retrieved November 6, 2018 (http://www.cs.huji.ac.il/~katrina/papers/mwem-nips.pdf).

Hofman Jake M., Sharma Amit, and Watts Duncan J.. 2017. “Prediction and Explanation in Social Systems.” Science 355(6324):486–88. [PubMed: 28154051]

Kahn David. 1996. The Codebreakers: The Comprehensive History of Secret Communication from Ancient Times to the Internet. New York: Simon & Schuster.

Karr Alan F., Kohnen Christine N., Oganian Anna, Reiter Jerome P., and Sanil Ashish P.. 2006. “A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality.” American Statistician 60(3):224–32.

Kim Jay J. 1986. “A Method for Limiting Disclosure In Microdata Based on Random Noise and Transformation.” Pp. 303–308 in Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Kinney Satkartar K., Reiter Jerome P., Reznek Arnold P., Miranda Javier, Jarmin Ron S., and Abowd John M.. 2011. “Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database.” International Statistical Review 79(3):362–84.

Kleinberg Jon, Lakkaraju Himabindu, Leskovec Jure, Ludwig Jens, and Mullainathan Sendhil. 2017. “Human Decisions and Machine Predictions.” Quarterly Journal of Economics 133(1):237–93. [PubMed: 29755141]

Kleinberg Jon, Ludwig Jens, Mullainathan Sendhil, and Obermeyer Ziad. 2015. “Prediction Policy Problems.” American Economic Review 105(5):491–95. [PubMed: 27199498]

Lambert Diane. 1993. “Measures of Disclosure Risk and Harm.” Journal of Official Statistics 9(2):313.

Malin Bradley, and Sweeney Latanya. 2004. “How (Not) to Protect Genomic Data Privacy in a Distributed Network: Using Trail Re-identification to Evaluate and Design Anonymity Protection Systems.” Journal of Biomedical Informatics 37(3):179–92. [PubMed: 15196482]

Mastroianni Anna, and Kahn Jeffrey. 2001. “Swinging on the Pendulum: Shifting Views of Justice in Human Subjects Research.” Hastings Center Report 31(3):21–28. [PubMed: 11478119]

McSherry Frank. 2009. “Privacy Integrated Queries: An Extensible Platform for Privacy-preserving Data Analysis.” Pp. 19–30 in Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. Retrieved November 6, 2018 (https://www.microsoft.com/en-us/research/wp-content/uploads/2010/09/pinq-CACM.pdf).

McSherry Frank, and Mironov Ilya. 2009. “Differentially Private Recommender Systems: Building Privacy into the Net.” Pp. 627–36 in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Retrieved November 6, 2018 (https://gnunet.org/sites/default/files/PrivateRecommender2009McSherry.pdf).

Menezes Alfred J., Van Oorschot Paul C., and Vanstone Scott A.. 1996. Handbook of Applied Cryptography. Boca Raton, FL: CRC Press.

Mullainathan Sendhil, and Spiess Jann. 2017. “Machine Learning: An Applied Econometric Approach.” Journal of Economic Perspectives 31(2):87–106.

Narayanan Arvind, Huey Joanna, and Felten Edward W.. 2016. “A Precautionary Approach to Big Data Privacy.” Pp. 357–85 in Data Protection on the Move: Current Developments in ICT and Privacy/Data Protection, edited by Gutwirth Serge, Leenes Ronald, and De Hert Paul. Houten, the Netherlands: Springer.

Narayanan Arvind, Shi Elaine, and Rubinstein Benjamin I. P.. 2011. “Link Prediction by De-anonymization: How We Won the Kaggle Social Network Challenge.” Pp. 1825–34 in International Joint Conference on Neural Networks. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Narayanan Arvind, and Shmatikov Vitaly. 2008. “Robust De-anonymization of Large Sparse Datasets.” Pp. 111–25 in Symposium on Security and Privacy. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Narayanan Arvind, and Shmatikov Vitaly. 2010. “Myths and Fallacies of Personally Identifiable Information.” Communications of the Association for Computing Machinery 53(6):24–26.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1978. “The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research.” Washington, DC: U.S. Government Printing Office.

National Research Council. 2004. Biotechnology Research in an Age of Terrorism. Washington, DC: National Academies Press.

National Research Council. 2014. “The Growth of Incarceration in the United States: Exploring Causes and Consequences.” Washington, DC: National Academies Press.

Nissenbaum Helen. 2009. Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford, CA: Stanford University Press.

Ohm Paul. 2010. “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization.” UCLA Law Review 57(6):1701.

O’Riordan Timothy, and Cameron James. 1994. Interpreting the Precautionary Principle. Vol. 2. London: Earthscan.

Phillips Whitney. 2015. This Is Why We Can’t Have Nice Things: Mapping the Relationship between Online Trolling and Mainstream Culture. Cambridge, MA: MIT Press.

Pilkauskas Natasha V. 2012. “Three-generation Family Households: Differences by Family Structure at Birth.” Journal of Marriage and Family 74(5):931–43. [PubMed: 24014117]

Reichman Nancy E., Teitler Julien O., Garfinkel Irwin, and McLanahan Sara S.. 2001. “Fragile Families: Sample and Design.” Children and Youth Services Review 23(4–5):303–26.

Reiter Jerome P. 2012. “Statistical Approaches to Protecting Confidentiality for Microdata and Their Effects on the Quality of Statistical Inferences.” Public Opinion Quarterly 76(1): 163–81.

Reiter Jerome P., and Kinney Satkartar K.. 2011. “Commentary: Sharing Confidential Data for Research Purposes: A Primer.” Epidemiology 22(5):632–35. [PubMed: 21811111]

Rubinstein Ira S., and Hartzog Woodrow. 2016. “Anonymization and Risk.” Washington Law Review 91:703.

Saez-Rodriguez Julio, Costello James C., Friend Stephen H., Kellen Michael R., Mangravite Lara, Meyer Pablo, Norman Thea, and Stolovitzky Gustavo. 2016. “Crowdsourcing Biomedical Research: Leveraging Communities as Innovation Engines.” Nature Reviews Genetics 17(8):470–86.

Salganik Matthew J. 2018. Bit by Bit: Social Research in the Digital Age. Princeton, NJ: Princeton University Press.

Salganik Matthew J., Lundberg Ian, Kindel Alexander T., and McLanahan Sara. 2019. “Introduction to the Special Collection on the Fragile Families Challenge.” Socius 5. doi:10.1177/2378023119871580.

Selgelid Michael J. 2009. “Governance of Dual-use Research: An Ethical Dilemma.” Bulletin of the World Health Organization 87(9):720–23. [PubMed: 19784453]

Shannon Claude E. 1949. “Communication Theory of Secrecy Systems.” Bell Labs Technical Journal 28(4):656–715.

Shmueli Galit. 2010. “To Explain or to Predict?” Statistical Science 25(3):289–310.

Shostack Adam. 2014. Threat Modeling: Designing for Security. New York: John Wiley.

Skinner Chris. 2012. “Statistical Disclosure Risk: Separating Potential and Harm.” International Statistical Review 80(3):349–68.

Sunstein Cass R. 2002. “Probability Neglect: Emotions, Worst Cases, and Law.” Yale Law Journal 112(1):61–107.

Sunstein Cass R. 2003. “Beyond the Precautionary Principle.” University of Pennsylvania Law Review 151(3):1003–58.

Sunstein Cass R. 2005. Laws of Fear: Beyond the Precautionary Principle. Vol. 6. Cambridge, UK: Cambridge University Press.

Sweeney Latanya. 2002a. “Achieving ***k***-Anonymity Privacy Protection Using Generalization and Suppression.” International Journal of Uncertainty, Fuzziness and Knowledge-based Systems 10(05):571–88.

Sweeney Latanya. 2002b. “***k***-Anonymity: A Model for Protecting Privacy.” International Journal of Uncertainty, Fuzziness and Knowledge-based Systems 10(05):557–70.

Sweeney Latanya. 2005. “Statement before the Pennsylvania House Select Committee on Information Security (House Resolution 351), ‘Recommendations to Identify and Combat Privacy Problems in the Commonwealth.’” Retrieved (https://dataprivacylab.org/dataprivacy/talks/Flick-05-10.html).

The Trustees of Princeton University. 2018. “Fragile Families & Child Wellbeing Study: Data and Documentation.” Retrieved November 6, 2018 (http://fragilefamilies.princeton.edu/documentation).

Watts Duncan J. 2014. “Common Sense and Sociological Explanations.” American Journal of Sociology 120(2):313–51.

Wildeman Christopher. 2009. “Parental Imprisonment, the Prison Boom, and the Concentration of Childhood Disadvantage.” Demography 46(2):265–80. [PubMed: 21305393]

Willenborg Leon, and de Waal Ton. 2001. Elements of Statistical Disclosure Control. Vol. 155. New York: Springer Science & Business Media.

Zimmer Michael. 2010. “‘But the Data Is Already Public’: On the Ethics of Research in Facebook.” Ethics and Information Technology 12(4):313–25.

Zook Matthew, Barocas Solon, boyd danah, Crawford Kate, Keller Emily, Gangadharan Seeta Peña, Goodman Alyssa, Hollander Rachelle, Koenig Barbara A., Metcalf Jacob, Narayanan Arvind, Nelson Alondra, and Pasquale Frank. 2017. “Ten Simple Rules for Responsible Big Data Research.” PLoS Computational Biology 13(3):e1005399. [PubMed: 28358831]

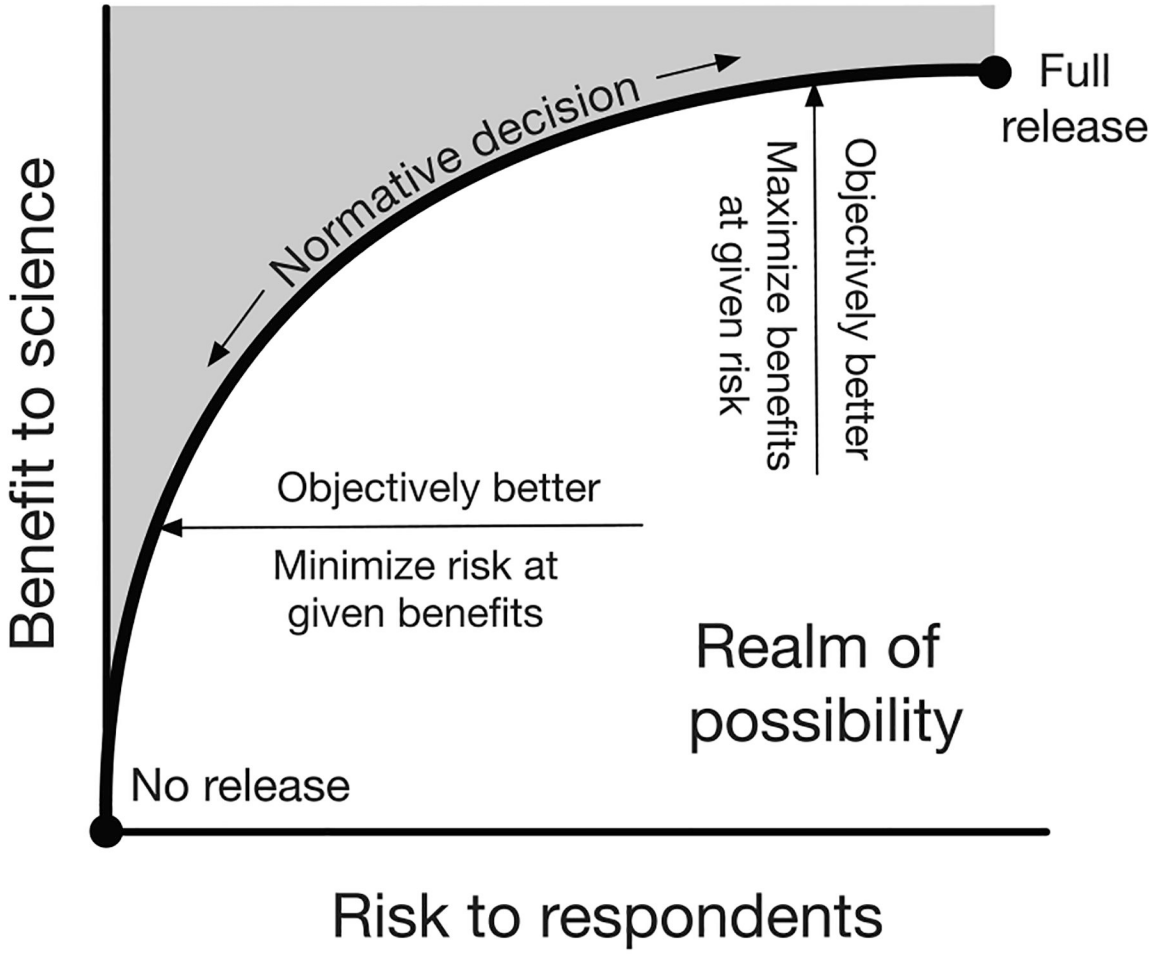

Figure 1.

Data access involves a fundamental tension between risk to respondents and benefits to science. If research requires releasing data, then those who manage the data face a tension between risk to respondents and benefits to science. It is objectively best to maximize benefits at a given level of risk and objectively best to minimize risk at a given level of benefits. On the frontier, the balance between risk and benefits becomes a normative question. The frontier is curved because we expect that, at low levels of risk, slight increases in risk might yield especially large benefits. For instance, moving from no release of data to release of a highly redacted form of the data might yield substantial benefits. At higher levels of risk, the returns to increased risk may be smaller. For instance, including respondents' addresses in the data release would substantially increase risk with only minimal benefits. We emphasize that this curve is merely a heuristic device; in realistic situations it is difficult, perhaps impossible, to define and quantify risks and benefits (Lambert 1993; Karr et al. 2006; Cox et al. 2011; Skinner 2012; Goroff 2015; Narayanan, Huey, and Felten 2016). Furthermore, many researchers are developing techniques, such as differential privacy (Dwork 2008), that try to improve the trade-offs.

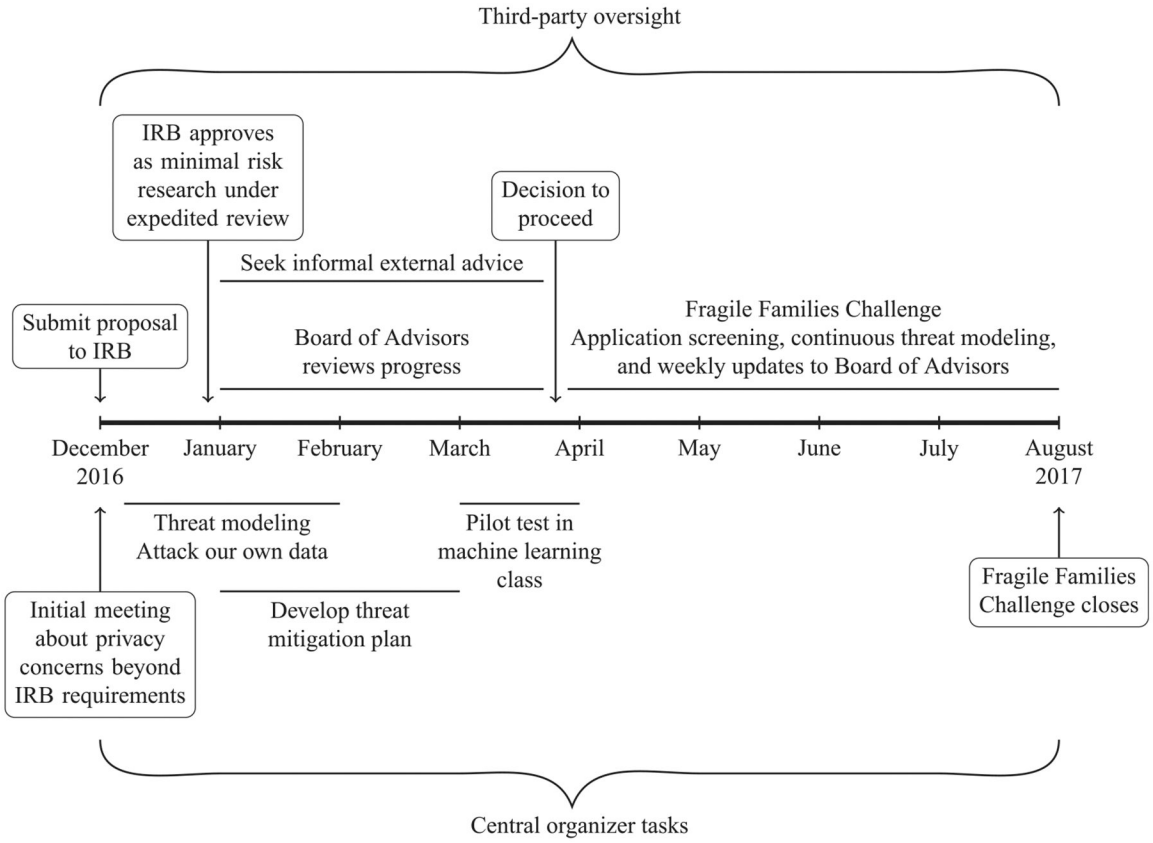

Figure 2.

Timeline of the privacy and ethics process for the Fragile Families Challenge. Boxed nodes represent events occurring at a specific point in time, such as the decision to proceed. Lines represent activities that occurred over the course of a period of time, such as seeking informal external advice.

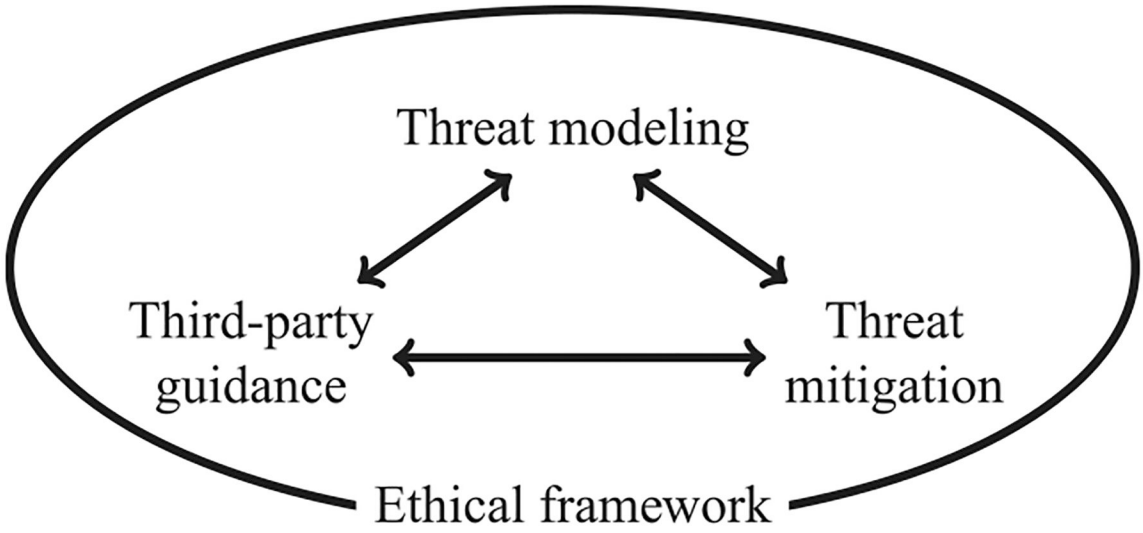

Figure 3.

Privacy and ethics audit for the Fragile Families Challenge. Our process involved (1) threat modeling to make precise our fear of reidentification, (2) threat mitigation to reduce risk, and (3) the guidance of third parties such as the Institutional Review Board at Princeton University and the Fragile Families Challenge Board of Advisors. The entire process was undertaken within an ethical framework, which we describe later in the paper. Although the article is written linearly, we emphasize that these steps were not taken linearly; we cycled through all the steps many times.

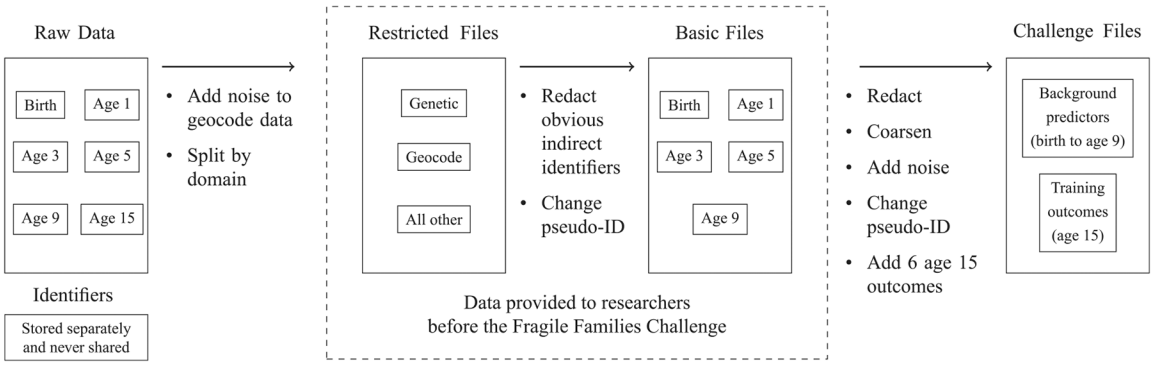

Figure 4.

Versions of the Fragile Families and Child Wellbeing Study data. There are several versions of the data which are made available to researchers depending on their particular needs. The raw data are used only by survey administrators and are stored in separate files to reduce the risk of a breach. For instance, no data file contains respondents' names and survey responses. All files given to researchers have names and other obvious identifiers removed and noise added to any data indicating place of residence. Among the files available to researchers, the restricted files provide the most information but are hardest to access. Researchers obtain restricted files through an intensive application and screening process. After this process, researchers are given only the portions of the restricted files needed for their particular projects. Most researchers' projects can be completed using only the basic files, for which one still must apply by proposing a research project. To create the Challenge files, we made modifications to the basic files.

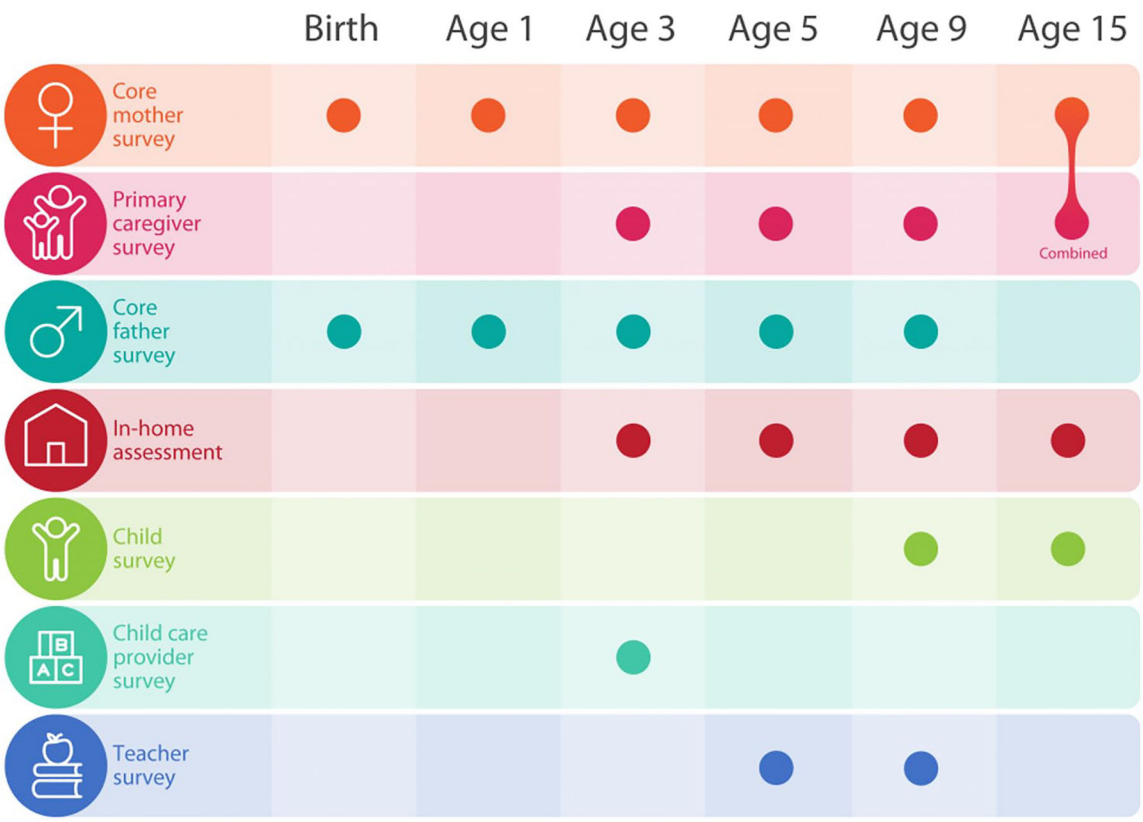

Figure 5.

Domains covered in the Fragile Families Study basic files. The number of substantive domains included makes the Fragile Families Study especially useful to social scientists. The number of domains also (1) facilitates reidentification because many possible auxiliary data sets may be used by an adversary and (2) increases the risk for harm in the event of reidentification because substantial information about respondents' private lives could become public.

Source: Fragile Families and Child Wellbeing Study.

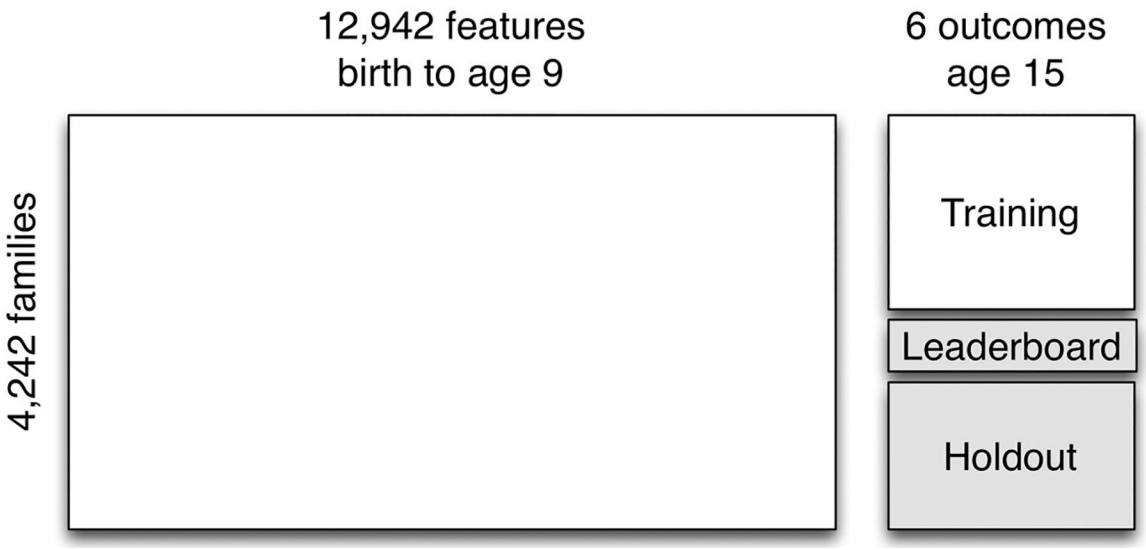

Figure 6.
Fragile Families Challenge data structure. Participants built models predicting the age 15 outcomes using data collected while the focal child was age 9 and younger. We provided participants with the data represented by the white boxes. Submissions were scored on the basis of their predictive performance (mean squared error) for the observations represented by the gray boxes, which were available only to organizers. The leaderboard set contained one-eighth of all observations and was used to provide instant feedback on submissions. The holdout set contained three-eighths of observations and was used to produce final scores for all submitted models at the end of the Challenge.

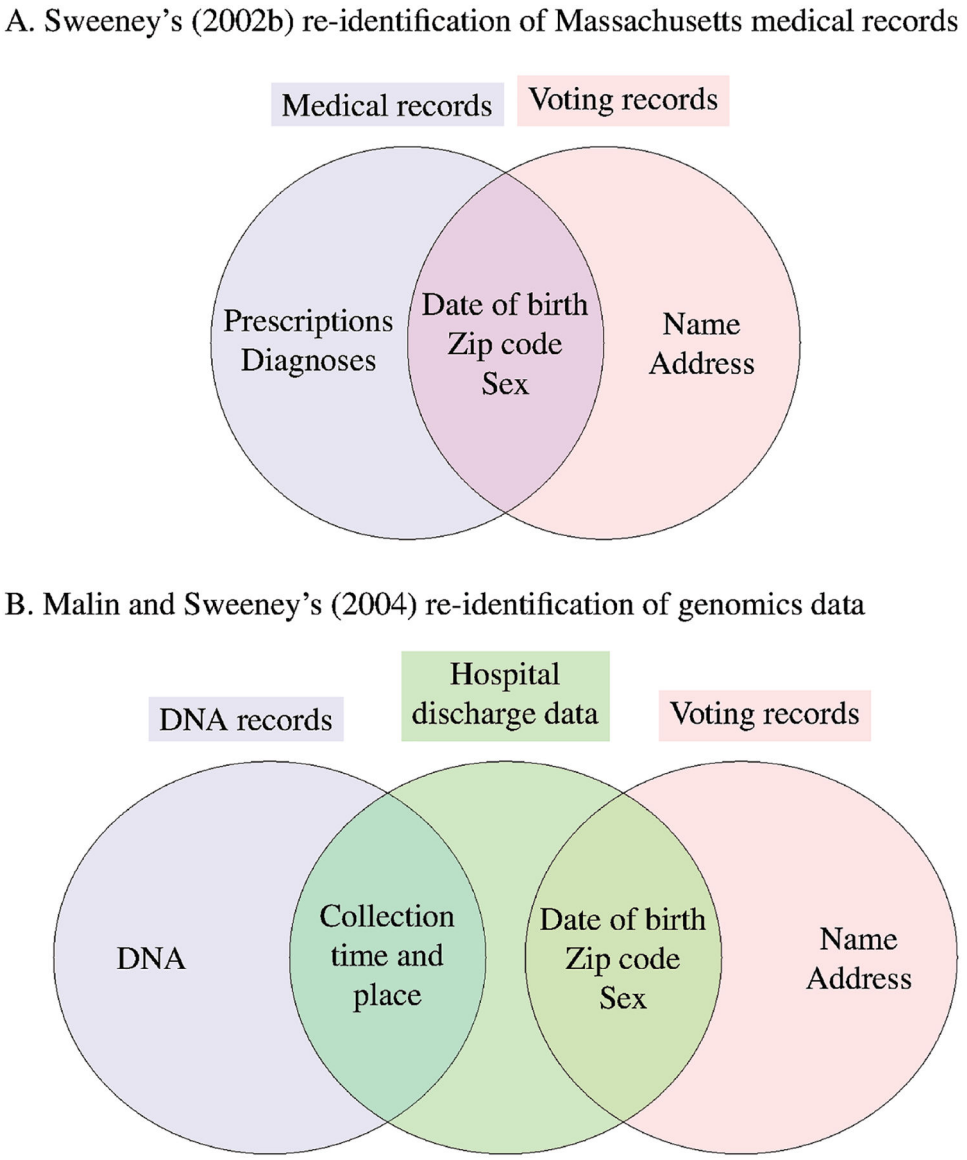

Figure 7.

Reidentification examples with deidentified data. In both examples, the adversary succeeded because key variables were available in both the deidentified data set (blue) and an identified auxiliary data set (red). In addition to merges between two data sets (Sweeney 2002b), reidentification can proceed through a chain of auxiliary data sets (Malin and Sweeney 2004).

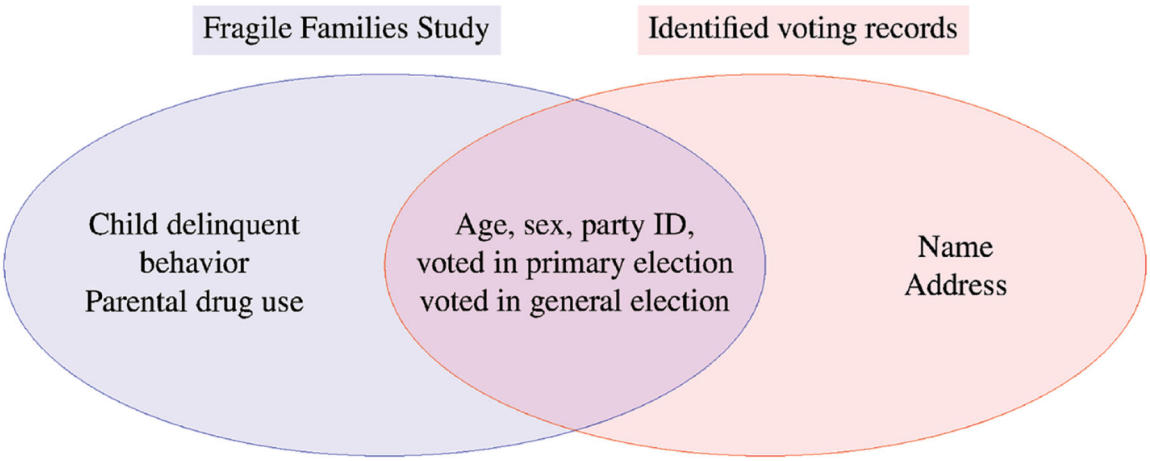

Figure 8.

Hypothetical example of reidentification attack of the Fragile Families Study. The Fragile Families Study does not contain information on voting, but if it did, these variables could be linked to an identified auxiliary data source: administrative voting records. After completing this linkage, an adversary could learn about potentially sensitive information, such as parental drug use and child delinquent behavior. This hypothetical attack, as well as our actual in-house attack, would be possible even if the data did not include any geographic information.

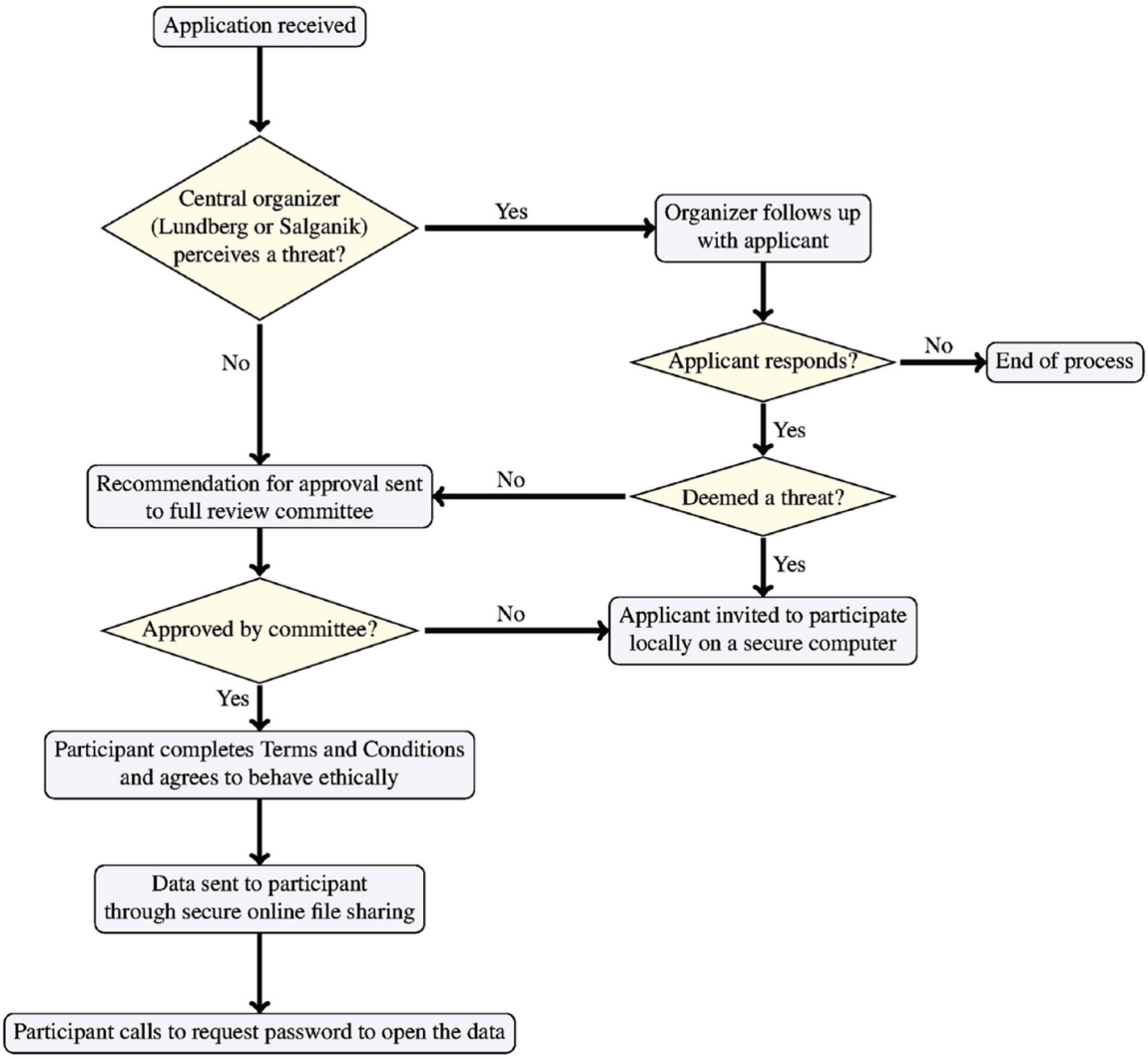

Figure 9.

Screening process for applications to the Fragile Families Challenge. Each potential participant completed an application describing his or her educational background, research experience, and motivations to participate in the study. We assessed these applications in terms of their ability to contribute to the goals of the Challenge and in terms of the risk that an applicant might try to reidentify respondents. Each application was reviewed by one of the lead organizers of the Challenge, sent to a review committee, and then approved 24 hours later if there were no objections. After approval, participants completed a set of terms and conditions, received a link to an encrypted file, and then called us for a password to open the file. For further details, see the main text. To review the application form and the terms and conditions, see the Appendix.

Table 1.

Five Main Threats and the Six Main Steps We Took to Mitigate Those Threats.

	Low Profile	Careful Language	Challenge Structure	Application Process	Ethical Appeal	Modifications to Data
Privacy researcher	✓	⊘	✓	✓	✓	
Nosy neighbor	⊘			✓		
Troll	⊘		✓	✓		✓
Journalist	✓	✓	✓	⊘	✓	✓
Cheater		✓	⊘		✓	✓

Note: Rows represent potential threats. Columns represent actions we took to mitigate the threats. Check marks indicate that we expected the action to be effective against the adversary. For each adversary (row), the circled check mark represents the action we felt was most effective against that adversary.