Volume Title: Economic Analysis of the Digital Economy

Volume Author/Editor: Avi Goldfarb, Shane M. Greenstein, and Catherine E. Tucker, editors

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-20684-X; 978-0-226-20684-4

Volume URL: http://www.nber.org/books/gree13-1

Conference Date: June 6–7, 2013

Publication Date: April 2015

Chapter Title: Estimation of Treatment Effects from Combined Data: Identification versus Data Security

Chapter Author(s): Tatiana Komarova, Denis Nekipelov, Evgeny Yakovlev

Chapter URL: http://www.nber.org/chapters/c12998

Chapter pages in book: (p. 279 – 308)

10

Estimation of Treatment Effects from Combined Data
Identification versus Data Security

Tatiana Komarova, Denis Nekipelov,
and Evgeny Yakovlev

10.1 Introduction

In policy analysis and decision making, it is instrumental in many areas to have access to individual data that may be considered sensitive or damaging when released publicly. For instance, a statistical analysis of the data from clinical studies that can include the information on the health status of their participants is crucial to study the effectiveness of medical procedures and treatments. In the financial industry, a statistical analysis of individual decisions combined with financial information, credit scores, and demographic data allows banks to evaluate risks associated with loans and mortgages. The resulting estimated statistical model will reflect the characteristics of individuals whose information was used in estimation. The policies based on this statistical model will also reflect the underlying individual data. The reality of the modern world is that the amount of publicly available (or searchable) individual information that comes from search traffic, social networks, and personal online file depositories (such as photo collections) is increasing on a daily basis. Thus, some of the variables in the data sets used

Tatiana Komarova is assistant professor of economics at the London School of Economics and Political Science. Denis Nekipelov is assistant professor of economics at the University of California, Berkeley. Evgeny Yakovlev is assistant professor at the New Economic School in Moscow, Russia.

We appreciate helpful comments from Philip Haile, Michael Jansson, Phillip Leslie, Aureo de Paula, Martin Pesendorfer, James Powell, Pasquale Schiraldi, John Sutton, and Elie Tamer. We also thank participants of the 2013 NBER conference "Economics of Digitization: An Agenda" for their feedback. For acknowledgments, sources of research support, and disclosure of the authors' material financial relationships, if any, please see http://www.nber.org/chapters/c12998.ack.

for policy analysis may be publicly observable.[1] Frequently, various bits of information regarding the same individual are contained in several separate data sets. Individual names or labels are most frequently absent from available data (either for the purposes of data anonymization or as an artifact of the data collection methodology). Each individual data set in this case may not pose a direct security threat to individuals. For instance, a collection of online search logs will not reveal any individual information unless one can attach the names of other identifying information to the generic identifiers attached to each unique user. However, if one can combine information from multiple sources, the combined array of data may pose a direct security threat to some or all individuals contained in the data. For instance, one data set may be a registry of HIV patients in which the names and locations of the patients are removed. Another data set may be the address book that contains names and addresses of people in a given area. Both these data sets individually do not disclose any sensitive information regarding concrete individuals. A combined data set will essentially attach names and addresses to the anonymous labels of patients in the registry and, thus, will disclose some sensitive individual information.

The path to digitization in a variety of markets with the simultaneous availability of the data from sources like social networks makes this scenario quite realistic. Clearly, from a policy perspective the prevention of a further increase in the availiability of such multiple sources is unrealistic. As a result, a feasible solution seems to be aimed at assuring some degree of anonymization as a possible security measure. At the same time, inferences and conclusions based on such multiple sources may be vital for making accurate policy decisions. Thus, a key agenda item in the design of methods and techniques for secure data storage and release is in finding a trade-off between keeping the data informative for policy-relevant statistical models and, at the same time, preventing an adversary from the reconstruction of sensitive information in the combined data set.

In this chapter we explore one question in this agenda. Our aim is to learn how one can evaluate the treatment effect when the treatment status of an individual may present sensitive information while the individual demographic characteristics are either publicly observable or may be inferred from some publicly observable characteristics. In such cases we are concerned with the *risk of disclosing sensitive individual information*. The questions that we address are, first, whether the point identification of treatment effects from the combined public and sensitive data is compatible with formal restrictions on the risk of the so-called partial disclosure. Second, we want to investigate how the public release of the estimated statistical model can lead to an increased risk of such a disclosure.

1. Reportedly, many businesses indeed rely on the combined data. See, for example, Wright (2010) and Bradley et al. (2010), among others.

In our empirical application we provide a concrete example of the analysis of treatment effects and propensity scores from two "anonymized" data sets. The data that we use come from the Russian Longitudinal Monitoring Survey (RLMS) that combines several questionnaires collected on a yearly basis. The respondents are surveyed on a variety of topics from employment to health. However, for anonymization purposes any identifying location information is removed from the data making it impossible to verify where exactly each respondent is located.

Due to the vast Soviet heritage, most people in Russia live in large apartment developments that include several blocks of multistory (usually five floors and up) apartment buildings connected together with common infrastructure, shops, schools, and medical facilities. With such a setup in place the life of each family becomes very visible to most of the neighbors. Our specific question of interest is the potential impact of the dominant religious affiliation in the neighborhood on the decision of parents to get their children checked up by a doctor in a given year as well as the decision of the parents to vaccinate their child with the age-prescribed vaccine.

Such an analysis is impossible without neighborhood identifiers. Neighborhood identifiers are made available to selected researchers upon a special agreement with the data curator (University of North Carolina and the Higher School of Economics in Moscow). This allows us to construct the benchmark where the neighborhood identification is known. Then we consider a realistic scenario where such an identification needs to be restored from the data. Using a record linkage technique adopted from the data mining literature, we reconstruct neighborhood affiliation using the individual demographic data. Our data linkage technique relies on observing data entries with infrequent attribute values. Accurate links for these entries may disclose individual location and then lead to the name disclosure based on the combination of the location and demographic data. We note that the goal of our work is not to demonstrate the vulnerability of anonymized personal data but to demonstrate a synthetic situation that reflects the component of the actual data-driven decision making and to show the privacy versus identification trade-off that arises in that situation. Further, we analyze how the estimates of the empirical model will be affected by the constraints on partial disclosure. We find that any such limitation leads to a loss of point identification in the model of interest. In other words, we find that there is a clear-cut trade-off between the restrictions imposed on partial disclosure and the point identification of the model using individual-level data.

Our analysis combines ideas from the data mining literature with those from the literature on statistical disclosure limitations, as well as the literature on model identification with corrupted or contaminated data. We provide a new approach to model identification from combined data sets as a limit in the sequence of statistical experiments.

A situation when the chosen data combination procedure provides a link

between at least one data entry in the data set with sensitive information (such as consumer choices, medical treatment, etc.) and auxiliary individual information from another data set with the probability exceeding the selected confidence threshold presents a case of a successful linkage attack and the so-called *individual disclosure*. The optimal structure of such attacks as well as the requirements in relation to the data release have been studied in the computer science literature. The structure of linkage attacks is based on the optimal record linkage results that have long been used in the analysis of databases and data mining. To some extent, these results were used in econometrics for combining data sets as described in Ridder and Moffitt (2007). In record linkage, one provides a (possibly) probabilistic rule that can match the records from one data set with the records from the other data set in an effort to link the data entries corresponding to the same individual. In several striking examples, computer scientists have shown that the simple removal of personal information such as names and Social Security numbers does not protect the data from individual disclosure. Sweeney (2002b) identified the medical records of William Weld, then governor of Massachusetts, by linking voter registration records to "anonymized" Massachusetts Group Insurance Commission (GIC) medical encounter data, which retained the birth date, sex, and zip code of the patient. Recent "depersonalized" data released for the Netflix prize challenge turned out to lead to a substantial privacy breach. As shown in Narayanan and Shmatikov (2008), using auxiliary information one can detect the identities of several Netflix users from the movie selection information and other data stored by Netflix.

Modern medical databases pose even larger threats to individual disclosure. A dramatic example of a large individual-level database is the data from genome-wide association studies (GWAS). The GWAS are devoted to an in-depth analysis of genetic origins of human health conditions and receptiveness to diseases, among other things. A common practice of such studies was to publish the data on the minor allele frequencies. The analysis of such data allows researchers to demonstrate the evidence of a genetic origin of the studied condition. However, there is a publicly available single nucleotide polymorphism (SNP) data set from the HapMap NIH project that consists of SNP data from four populations with about sixty individuals each. Homer et al. (2008) demonstrated that they could infer the presence of an individual with a known genotype in a mix of DNA samples from the reported averages of the minor allele frequencies using the HapMap data. To create the privacy breach, one can take an individual DNA sequence and then compare the nucleotide sequence of this individual with the reported averages of minor allele frequencies in the HapMap population and in the studied subsample. Provided that the entire list of reported allele frequencies can be very long, individual disclosure may occur with an extremely high probability. As a result, if a particular study is devoted to the analysis of a particular health condition or a disease, the discovery that a particular

individual belongs to the studied subsample means that this individual has that condition or that disease.

Samarati and Sweeney (1998), Sweeney (2002a, 2002b), LeFevre, DeWitt, and Ramakrishnan (2005), Aggarwal et al. (2005), LeFevre, DeWitt, and Ramakrishnan (2006), and Ciriani et al. (2007) developed and implemented the so-called k-anonymity approach to address the threats of linkage attacks. Intuitively, a database provides k-anonymity, for some number k, if every way of singling an individual out of the database returns records for at least k individuals. In other words, anyone whose information is stored in the database can be "confused" with k others. Several operational prototypes for maintaining k-anonymity have been offered for practical use. The data combination procedure will then respect the required boundary on the individual disclosure (disclosure of identities) risk if it only uses the links with at least k possible matches.

A different solution has been offered in the literature on synthetic data. Duncan and Lambert (1986), Duncan and Mukherjee (1991), Duncan and Pearson (1991), Fienberg (1994, 2001), Duncan et al. (2001), and Abowd and Woodcock (2001) show that synthetic data may be a useful tool in the analysis of particular distributional properties of the data such as tabulations, while guaranteeing a certain value for the measure of the individual disclosure risk (for instance, the probability of "singling out" some proportion of the population from the data). An interesting feature of the synthetic data is that they can be robust against stronger requirements for the risk of disclosure. Dwork and Nissim (2004) and Dwork (2006) introduced the notion of differential privacy that provides a probabilistic disclosure risk guarantee against the privacy breach associated with an arbitrary auxiliary data set. Abowd and Vilhuber (2008) demonstrate a striking result that the release of synthetic data is robust to differential privacy. As a result, one can use the synthetic data to enforce the constraints on the risk of disclosure by replacing the actual consumer data with the synthetic consumer data for a combination with an auxiliary individual data source.

In our chapter we focus on the threat of *partial disclosure*. Partial disclosure occurs if the released information such as statistical estimates obtained from the combined data sample reveals with high enough probability some sensitive characteristics of a group of individuals. We provide a formal definition of partial disclosure and show that generally one can control the risk of this disclosure, so the bounds on the partial disclosure risk are practically enforceable.

Although our identification approach is new, to understand the impact of the bounds on the individual disclosure risk we use ideas from the literature on partial identification of models with contaminated or corrupted data. Manski (2003), Horowitz et al. (2003), Horowitz and Manski (2006), and Magnac and Maurin (2008) have understood that many data modifications such as top-coding suppression of attributes and stratification lead to the

loss of point identification of parameters of interest. Consideration of the general setup in Molinari (2008) allows one to assess the impact of some data anonymization as a general misclassification problem. In this chapter we find the approach to the identification of the parameters of interest by constructing sets compatible with the chosen data combination procedure extremely useful. As we show in this chapter, the sizes of such identified sets for the propensity scores and the average treatment effect are directly proportional to the pessimistic measure of the partial disclosure risk. This is a powerful result that essentially states that there is a direct conflict between the informativeness of the data used in the consumer behavioral model and the security of individual data. An increase in the complexity and nonlinearity of the model can further worsen the trade-off.

In the chapter we associate the ability of a third party to recover sensitive information about consumers from the reported statistical estimates based on the combined data with the risk of partial disclosure. We argue that the estimated model *may itself be disclosive*. As a result, if this model is used to make (observable) policy decisions, some confidential information about consumers may become discoverable. Existing real-world examples of linkage attacks on the consumer data using the observable firm policies have been constructed for online advertising. In particular, Korolova (2010) gives examples of privacy breaches through micro ad targeting on Facebook.com. Facebook does not give advertisers direct access to user data. Instead, the advertiser interface allows them to create targeted advertising campaigns with a very granular set of targets. In other words, one can create a set of targets that will isolate a very small group of Facebook users (based on the location, friends, and likes). Korolova shows that certain users may be perfectly isolated from other users with a particularly detailed list of targets. Then, one can recover the "hidden" consumer attributes, such as age or sexual orientation, by constructing differential advertising campaigns such that a different version of the ad will be shown to the user depending on the value of the private attribute. Then the advertiser's tools allow the advertiser to observe which version of the ad was shown to the Facebook user.

When a company "customizes" its policy regarding individual users, for example, when a PPO gives its customers personalized recommendations regarding their daily routines and exercise or hospitals reassign specialty doctors based on the number of patients in need of specific procedures, then the observed policy results may disclose individual information. In other words, the disclosure may occur even when the company had no intention of disclosing customer information.

Security of individual data is not synonymous to privacy, as privacy may have subjective value for consumers (see Acquisti [2004]). Privacy is a complicated concept that frequently cannot be expressed as a formal guarantee against intruders' attacks. Considering personal information as a "good" valued by consumers leads to important insights in the economics of privacy. As seen in Varian (2009), this approach allowed the researchers to

analyze the release of private data in the context of the trade-off between the network effects created by the data release and the utility loss associated with this release. The network effect can be associated with the loss of competitive advantage of the owner of personal data, as discussed in Taylor (2004), Acquisti and Varian (2005), and Calzolari and Pavan (2006). Consider the setting where firms obtain a comparative advantage due to the possibility of offering prices that are based on past consumer behavior. Here, the subjective individual perception of privacy is important. This is clearly shown in both the lab experiments in Gross and Acquisti (2005), Acquisti and Grossklags (2008), as well as in the real-world environment in Acquisti, Friedman, and Telang (2006), Miller and Tucker (2009), and Goldfarb and Tucker (2010). Given all these findings, we believe that the disclosure protection plays a central role in the privacy discourse, as privacy protection is impossible without the data protection.

The rest of the chapter is organized as follows. Section 10.2 describes the analyzed treatment effects models, the availability of the data, and gives a description of data combination procedures employed in the chapter. Section 10.3 provides a notion of the identified values compatible with the data combination procedure for the propensity score and the average treatment effect. It looks at the properties of these values as the sizes of available data sets go to infinity. Section 10.4 introduces formal notions of partial disclosure and partial disclosure guarantees. It discusses the trade-off between the point identification of the true model parameters and partial disclosure limitations. Section 10.5 provides an empirical illustration.

10.2 Model Setup

In many practical settings the treatment status of an individual in the analyzed sample is a very sensitive piece of information, much more sensitive than the treatment outcome and/or the individual's demographics. For instance, in the evaluation of the effect of a particular drug, one may be concerned with the interference of this drug with other medications. Many anti-inflammatory medications may interfere with standard HIV treatments. To determine the effect of the interference one would evaluate how the HIV treatment status influences the effect of the studied anti-inflammatory drug. The fact that a particular person participates in the study of the anti-inflammatory drug does not necessarily present a very sensitive piece of information. However, the information that a particular person receives HIV treatment medications may be damaging.

We consider the problem of estimating the propensity score and the average treatment effect in cases when the treatment status is a sensitive (and potentially harmful) piece of information. Suppose that the response of an individual to the treatment is characterized by two potential outcomes $Y_1, Y_0 \in \mathcal{Y} \subset \mathbb{R}$, and the treatment status is characterized by $D \in \{0,1\}$. Outcome Y_1 corresponds to the individuals receiving the treatment and Y_0 cor-

responds to the nontreated individuals. Each individual is also characterized by the vector of individual-specific covariates $X \in \mathcal{X} \subset \mathbb{R}^p$ such as the demographic characteristics, income, and location.

Individuals are also described by vectors V and W containing a combination of real-valued and string-valued variables (such as Social Security numbers, names, addresses, etc.) that identify the individual but do not interfere with the treatment outcome. The realizations of V belong to the product space $\mathcal{V} = \boldsymbol{\mathcal{S}}^* \times \mathbb{R}^{\upsilon}$, where $\boldsymbol{\mathcal{S}}^*$ is a finite space of arbitrary (nonnumeric) nature. $\boldsymbol{\mathcal{S}}^*$, for instance, may be the space of combinations of all human names and dates of birth (where we impose some "reasonable" bound on the length of the name, e.g., thirty characters). The string combination {*'John','Smith', '01/01/1990'*} is an example of a point in this space. Each string in this combination can be converted into the digital binary format. Then the countability and finiteness of the space $\boldsymbol{\mathcal{S}}^*$ will follow from the countability of the set of all binary numbers of fixed length. We also assume that the space $\mathcal{V}$ is endowed with the distance. There are numerous examples of definitions of a distance over strings (e.g., see Wilson et al. 2006). We can then define the norm in $\boldsymbol{\mathcal{S}}^*$ as the distance between the given point in $\boldsymbol{\mathcal{S}}$ and a "generic" point corresponding to the most commonly observed set of attributes. We define the norm in $\mathcal{V}$ as the weighted sum of the defined norm in $\boldsymbol{\mathcal{S}}$ and the standard Euclidean norm in $\mathbb{R}^v$ and denote it $\| \|_v$. Similarly, we assume that W takes values in $\mathcal{W} = \boldsymbol{\mathcal{S}}^{**} \times \mathbb{R}^w$, where $\boldsymbol{\mathcal{S}}^{**}$ is also a finite space. The norm in $\mathcal{W}$ is defined as a weighted norm and denoted as $\| \|_w$. Spaces $\boldsymbol{\mathcal{S}}^*$ and $\boldsymbol{\mathcal{S}}^{**}$ may have common subspaces. For instance, they both may contain the first names of individuals. However, we do not require that such common elements indeed exist.

Random variables V and W are then defined by the probability space with a σ-finite probability measure defined on Borel subsets of $\mathcal{V}$ and $\mathcal{W}$.

We assume that the data-generating process creates N_y i.i.d. draws from the joint distribution of the random vector (Y, D, X, V, W). These draws form the (infeasible) "master" sample $\{y_i, d_i, x_i, \upsilon_i, w_i\}_{i=1}^{N_y}$. However, because either all the variables in this vector are not collected simultaneously or some of the variables are intentionally deleted, the data on the treatment status (treatment outcome) and individual-specific covariates are not contained in the same sample. One sample, containing N_y observations is the i.i.d. sample $\{x_i, \upsilon_i\}_{i=1}^{N_y}$ is in the *public domain*. In other words, individual researchers or research organizations can get access to this data set. The second data set is a subset of $N \leq N_y$ observations from the "master" data set and contains information regarding the treatment-related variables $\{y_j, d_j, w_j\}_{j=1}^{N}$.[2] This

2. Our analysis applies to other frameworks of split data sets. For instance, we could consider the case when x and y are contained in the same data subset, while d is observed only in the other data subset. We could also consider cases when some of the variables in x (but not all of them) are observed together with d. This is the situation we deal with in our empirical illustration. The important requirement in our analysis is that some of the relevant variables in x are not observed together with d.

data set is *private* in the sense that it is only available to the data curator (e.g., the hospital network) and cannot be acquired by external researchers or general public. We consider the case when, even for the data curator, there is no direct link between the private and the public data sets. In other words, the variables in υ_i and w_j do not provide immediate links between the two data sets. In our example of the HIV treatment status, we could consider cases where the data on the HIV treatment (or testing) are partially or fully anonymized (due to the requests by the patients) and there are only very few data attributes that allow the data curator to link the two data sets.

We impose the following assumptions on the elements of the model:

Assumption 1.

(a) The treatment outcomes satisfy the conditional unconfoundedness, that is, $(Y_1, Y_0) \perp D|X = x$.

(b) At least one element of X has a continuous distribution with the density strictly positive on its support.

We consider the propensity score $P(x) = E[D|X = x]$ and suppose that for some specified $0 < \delta < 1$ the knowledge that the propensity score exceeds $(1 - \delta)$—that is,

$$P(x) > 1 - \delta,$$

constitutes sensitive information. The next assumption states that there is a part of the population with the propensity score above the sensitivity threshold.

Assumption 2.

$$Pr(x : P(x) > 1 - \delta) > 0.$$

$\overline{P}$ will denote the average propensity score over the distribution of all individuals:

$$\overline{P} = E[P(x)].$$

We leave distributions of potential outcomes Y_1 and Y_0 conditional on X nonparametric with the observed outcome determined by

$$Y = DY_1 + (1 - D)Y_0.$$

In addition to the propensity score, we are interested in the value of the conditional average treatment effect

$$t_{ATE}(x) = E[Y_1 - Y_0|X = x],$$

or the average treatment effect conditional on individuals in a group described by some set of covariates $\mathcal{X}_0$:

$$t_{ATE}(\mathcal{X}_0) = E[Y_1 - Y_0|X \in \mathcal{X}_0],$$

as well as overall average treatment effect (ATE)

$$t_{ATE} = E[Y_1 - Y_0].$$

In this chapter we focus on the propensity score and the overall average treatment effect.

The evaluation of the propensity score and the treatment effects requires us to observe the treatment status and the outcome together with the covariates. A consistent estimator for the average treatment effect t_{ATE} could be constructed then by, first, evaluating the propensity score and then estimating the overall effect via the propensity score weighting:

$$t_{ATE} = E\left[\frac{DY}{P(X)} - \frac{(1-D)Y}{1-P(X)}\right]. \tag{2.1}$$

In our case, however, the treatment and its outcome are not observed together with the covariates. To deal with this challenge, we will use the information contained in the identifying vectors V and W to connect the information from the two split data sets and provide estimates for the propensity score and the ATE.

Provided that the data curator is interested in correctly estimating the treatment effect (to further use the findings to make potentially observable policy decisions, for example, by putting a warning label on the package of the studied drug), we assume that she will construct the linkage procedure that will correctly combine the two data sets with high probability.

We consider a two-step procedure that first uses the similarity of information contained in the identifiers and covariates to provide the links between the two data sets. Then, the effect of interest will be estimated from the reconstructed joint data set. To establish similarity between the two data sets, the researcher constructs vector-valued variables that exploit the numerical and string information contained in the variables. We assume that the researcher constructs variables $Z^d = Z^d(D, Y, W)$ and $Z^x = Z^x(X, V)$ (individual identifiers) that both belong to the space $\mathcal{Z} = \mathcal{S} \times \mathbb{R}^z$. The space $\mathcal{S}$ is a finite set of arbitrary nature such as a set of strings, corresponding to the string information contained in $\mathcal{S}^*$ and $\mathcal{S}^{**}$. We choose a distance in $\mathcal{S}$ constructed using one of commonly used distances defined on the strings $d_{\mathcal{S}}(\cdot,\cdot)$. Then the distance in $\mathcal{Z}$ is defined as a weighted combination of $d_{\mathcal{S}}$ and the standard Euclidean distance $d_z(Z^x, Z^d) = (\omega_s d_{\mathcal{S}}(z_s^x, z_s^d)^2 + \omega_z \|z_z^x - z_z^s\|^2)^{1/2}$, where $Z^x = (z_s^x, z_z^x)$ and $\omega_s, \omega_d > 0$.

Then we define the "null" element in $\mathcal{S}$ as the observed set of attributes that has the most number of components shared with the other observed sets of attributes and denote it $0_{\mathcal{S}}$. Then the norm in $\mathcal{Z}$ is defined as the distance from the null element: $\|Z\|_z = (\omega_s d_{\mathcal{S}}(z_s, 0_{\mathcal{S}})^2 + \omega_s \|z_z\|^2)^{1/2}$.

The construction of the variables Z^d and Z^x may exploit the fact that W and V can contain overlapping components, such as individuals' first names and the dates of birth. Then the corresponding components of the identifiers can be set equal to those characteristics. However, the identifiers may

also include a more remote similarity of the individual characteristics. For instance, V may contain the name of an individual and W may contain the race (but not contain the name). Then we can make one component of Z^d to take values from 0 to 4 corresponding to the individual in the private data set either having the race not recorded, or being black, white, Hispanic, or Asian.

Then, using the public data set we can construct a component of Z^x that will correspond to the guess regarding the race of an individual based on his name. This guess can be based on some simple classification rule, for example, whether the individual's name belongs to the list of top 500 Hispanic names in the US Census or if the name belongs to the top 500 names in a country that is dominated by a particular nationality. This classifier, for instance, will classify the name "Vladimir Putin" as the name of a white individual giving Z^x value 2, and it will classify the name "Kim Jong Il" as the name of an Asian individual giving Z^x value 4.

When the set of numeric and string characteristics used for combining two data sets is sufficiently large or it contains some potentially "hard to replicate" information such as the individual's full name, then if such a match occurs it very likely singles out the data of one person. We formalize this idea by expecting that if the identifiers take infrequent values (we model this situation as the case of identifiers having large norms), then the fact that the values of Z^d and Z^x are close implies that with high probability the two corresponding observations belong to the same individual. This probability is higher the more infrequent are the values of Z^d and Z^x. Our maintained assumptions regarding the distributions of constructed identifiers are listed below.

Assumption 3. We fix some $\bar{\alpha} \in (0,1)$ such that for any $\alpha \in (0,\bar{\alpha})$:

(a) (Proximity of identifiers) $Pr(d_z(Z^x,Z^d) < \alpha | X = x, D = d, Y = y, \|Z^d\|_z > 1/\alpha) \geq 1 - \alpha$.

(b) (Nonzero probability of extreme values)

$$\lim_{\alpha \to 0} Pr\left(\|Z^d\|_z > \frac{1}{\alpha} | D = d, Y = y\right) / \phi(\alpha) = 1$$

$$\lim_{\alpha \to 0} Pr\left(\|Z^x\|_z > \frac{1}{\alpha} | X = x\right) / \psi(\alpha) = 1$$

for some nondecreasing and positive functions $\phi(\cdot)$ and $\psi(\cdot)$.

(c) (Redundancy of identifiers in the combined data) There exists a sufficiently large M such that for all $\|Z^d\|_z \geq M$ and all $\|Z^x\|_z \geq M$

$$f(Y | D = d, X = x, Z^d = z^d, Z^x = z^x) = f(Y | D = d, X = x).$$

Assumption 3(a) reflects the idea that more reliable matches are provided by the pairs of identifiers whose values are infrequent. In other words, if, for example, in both public and private data sets collected in Durham, NC,

we found observations with an attribute "Denis Nekipelov," we expect them to belong to the same individual with a higher probability than if we found two attribute values "Jane Doe." Thus, the treatment status can be recovered more reliably for more unique individuals. We emphasize that infrequency of a particular identifier does not mean that the corresponding observation is an "outlier." In fact, if both public and private data sets contain very detailed individual information such as a combination of the full name and the address, most attribute values will be unique.

Assumption 3(b) requires that there are a sufficient number of observations with infrequent attribute values. This fact can actually be established empirically in each of the observed subsets and, thus, this assumption is testable.

Assumption 3(c) is the most important one for identification purposes. It implies that even for the extreme values of the identifiers and the observed covariates, the identifiers only served the purpose of data labels as soon as the "master" data set is recovered. There are two distinct arguments that allow us to use this assumption. First, in cases where the identifiers are high dimensional, infrequent attribute combinations do not have to correspond to unusual values of the variables. If both data sets contain, for instance, first and last names along with the dates of birth and the last four digits of the Social Security number of individuals, then a particular combination of all attributes can be can be extremely rare, even for individuals with common names. Second, even if the identifiers can contain model relevant information (e.g., we expect the restaurant choice of an individual labeled as "Vladimir Putin" to be different than the choice of an individual labeled as "Kim Jong Il"), we expect that information to be absorbed in the covariates. In other words, if the gender and the nationality of an individual may be information relevant for the model, than we include that information into the covariates.

We continue our analysis with the discussion of identification of the model from the combined data set.

In the remainder of the chapter we suppose that Assumptions 1–3 hold.

10.3 Identification of the Treatment Effect from the Combined Data

Provided that the variables are not contained in the same data set, the identification of the treatment effect parameter becomes impossible without having some approximation to the distribution of the data in the master sample. A way to link the observations in two data sets is to use the identifiers that we described in the previous section. The identifiers, on the other hand, are individual-level variables. Even though the data-generating process is characterized by the distribution over strings, such as names, we only recover the master data set correctly if we link the data of one concrete "John

Smith" in the two data sets. This means that the data combination is intrinsically a finite sample procedure. We represent the data combination procedure by the deterministic data combination rule $\mathcal{D}^N$ that for each pair of identifiers z_j^d and z_i^x returns a binary outcome

$$M_{ij} = \mathcal{D}^N(z_i^x, z_j^d),$$

which labels two observations as a match ($M_{ij} = 1$) if we think they belong to the same individual, and labels them as a nonmatch ($M_{ij} = 0$) if we think that the observations are unlikely to belong to the same individual or are simply uncertain about this. Although we can potentially consider many nonlinear data combination rules, in this chapter we focus on the set of data combination rules that are generated by our Assumption 3 (a). In particular, for some prespecified $\bar{\alpha} \in (0,1)$ we consider the data combination rule

$$\mathcal{D}^N = 1\{d_z(z_i^x, z_j^d) < \alpha_N, \|z_i^x\| > 1/\alpha_N\},$$

generated by a Cauchy sequence α_N such that $0 < \alpha_N < \bar{\alpha}$ and $\lim_{N\to\infty} \alpha_N = 0$. The goal of this sequence is to construct the set of thresholds that would isolate in the limit all of the infrequent observations. To guarantee that, such a sequence would have to satisfy the following two conditions. For infrequent observations, the probability of the correct match would be approaching one, as the probability of observing two identifiers taking very close values for two different individuals would be very small (proportional to the square of the probability of observing the infrequent attribute values). On the other hand, the conditional probability that the values of identifiers are close for a particular individual with infrequent values of the attributes would be of a larger order of magnitude (proportional to the probability of observing the attribute value). Thus, an appropriately scaled sequence of thresholds would be able to single out correct matches.

Let m_{ij} be the indicator of the event that the observation i from the public data set and the observation j from the private data set belong to the same individual. Given that we can make incorrect matches, M_{ij} is not necessarily equal to m_{ij}. However, we would want these two variables to be highly correlated, meaning that the data combination procedure that we use is good.

With our data combination procedure we will form the reconstructed master data set by taking the pairs of all observations from the public and the private data sets that we indicated as matches ($M_{ij} = 1$) and discard all other observations. We can consider more complicated rules for reconstructing the master sample. In particular, we can create multiple copies of the master sample by varying the threshold α_N and then we combine the information from those samples by downweighting the data sets that were constructed with higher threshold values.

The reconstructed master data set will have a small sample distribution, characterizing the joint distribution of outcomes and the covariates for all

observations that are identified as matches by the decision rule $\mathcal{D}^N$. We use $f^N_{\alpha_N}(y_i|d_j,x_i,z_i^x,z_j^d)$ to denote the conditional density of the outcome distribution with the decision rule applied to samples of size N. Provided that the decision rule does not perfectly identify the information from the same individual, density $f^N_{\alpha_N}(\cdot)$ will be a mixture of the "correct" distribution with the distribution of outcomes that were incorrectly identified as matches:

$$f^N_{\alpha_N}(y_j|d_j,x_i,z_i^x) = f_{Y|D,X}(y_j|d_j,x_i)Pr(m_{ij}=1|\mathcal{D}^N(z_i^x,z_j^d)=1) + f_{Y|D}(y_j|d_j)Pr(m_{ij}=0|\mathcal{D}^N(z_i^x,z_j^d)=1),$$

where we used the fact that identifiers are redundant once a correct match was made, as well as the fact that in the i.i.d. sample the observations have to be independent. Thus, if an incorrect match was made, the outcome should not be correlated with the treatment. By $E^N_{\alpha_N}[\cdot|d_j]$ we denote the conditional expectation with respect to the density product $f^N_{\alpha_N}(\cdot|d_j,x_i,z_i^x)f(x_i,z_i^x)$.

We can also introduce the propensity score implied by the finite sample distribution, which we denote $P^N_{\alpha_N}(\cdot)$. The finite sample propensity score is characterized by the mixture distribution combining the correct propensity score and the average propensity score

$$P^N_{\alpha_N}(x) = P(x)Pr(m_{ij}=1|x_i=x,\mathcal{D}^N(z_i^x,z_j^d)=1) + \bar{P}Pr(m_{ij}=0|x_i=x,\mathcal{D}^N(z_i^x,z_j^d)=1).$$

We can extend our data combination method by choosing sequences α_N depending on the value of x. Then the value of $Pr(m_{ij}=0|x_i=x,\mathcal{D}^N(z_i^x,z_j^d)=1)$ even in the limit will depend on x. We allow for such situations. In fact, later in the chapter we make use of this opportunity to choose differences threshold sequences for different values of x. To stress that we permit the threshold sequences to depend on x we denote a sequence of thresholds chosen for x as $\alpha_{N,x}$ (instead of α_N).

In the beginning of this section, we indicated that the estimation that requires combining the data based on the string-valued identifiers is an intrinsically finite sample procedure. As a result, we suggest the analysis of identification of this model as the limit of a sequence of data combination procedures. We allow for situations when the data curator could want to use several sequences $\alpha_{N,x}$ for some x and denote the collection of such sequences as $C_{0,x}$.

Definition 1. *By $\mathcal{P}^N$ we denote the set of all functions $p:\mathcal{X}\mapsto[0,1]$ that correspond to the set of finite sample propensity scores for all sequences $\alpha_{N,x}$ in $C_{0,x}$:*

$$\mathcal{P}^N = \bigcup_{\{\alpha_{N,x}\}\in C_{0,x}} \{P^N_{\alpha_{N,x}}(\cdot)\}.$$

We call $\mathcal{P}^N$ ***the N-identified set for the propensity score compatible with the data combination procedure with a threshold decision rule***.

By $\mathcal{T}^N$ *we denote the subset of* $\mathbb{R}$ *that corresponds to the set of treatment effects calculated as equation (2.1) for all sequences* $\alpha_{N,x}$ *in* $C_{0,x}$ *using the corresponding to* $\alpha_{N,x}$ *propensity score* $P^N_{\alpha_{N,x}}(\cdot)$*:*

$$\mathcal{T}^N = \bigcup_{\{\alpha_{N,x}\}\in C_{0,x}} E^N_{\alpha_{N,x}}\left[\frac{D_j Y_j}{P^N_{\alpha_{N,x}}(X_i)} - \frac{(1-D_j)Y_j}{1-P^N_{\alpha_{N,x}}(X_i)}\right].$$

We call $\mathcal{T}^N$ *the* ***N-identified set for the average treatment effect compatible with the data combination procedure with a threshold decision rule***.

Definition 2 below characterizes the identified set compatible with the data combination procedure as the set of all limits of the estimated treatment effects and the propensity scores under all possible threshold sequences that are bounded and converge to zero. Provided that the reconstructed master sample depends on the sample size, the set of treatment effect parameters that are compatible with the data combination procedure applied to random split samples of size N will depend on N. Provided that the small sample distribution in the sample of size N will always be a mixture of the correct joint distribution and the marginal outcome distribution for the outcomes that are misidentified as matches, the only way to attain the point identification is in the limit. Thus, we consider the concept of parameter identification in terms of the limiting behavior of the identified sets compatible with the data combination procedure constructed from the finite sample distributions as the sample size N approaches infinity.

Definition 2.

(a) We call $\mathcal{P}^\infty$ *the* ***identified set for the propensity score under the threshold decision rules*** *if* $\mathcal{P}^\infty$ *is the set of all partial pointwise limits of sequences of propensities score functions from the N-identified sets* $\mathcal{P}^N$*. That is, function* $f(\cdot) \in \mathcal{P}^\infty$ *if and only if for any x in the support of X,*

$$f(x) = \lim_{N_k\to\infty} f_{N_k}(x),$$

for some $f_{N_k}(\cdot) \in \mathcal{P}^{N_k}$.

(b) Similarly, we call $\mathcal{T}^\infty$ *the* ***identified set for the average treatment effect under the decision threshold rules*** *if* $\mathcal{T}^\infty$ *is the set of all partial limits of sequences of ATEs from the N-identified sets* $\mathcal{T}^N$*. That is,* $t \in \mathcal{T}^\infty$ *if*

$$t = \lim_{N_k\to\infty} t_{N_k},$$

for some $t_{N_k} \in \mathcal{T}^{N_k}$.

(c) The propensity score is point identified from the combined data if $\mathcal{P}^\infty = \{P(\cdot)\}$*. Otherwise, it is identified only up to a set compatible with the decision threshold rules.*

(d) The average treatment effect parameter is point identified from the combined data if the identified set is a singleton $\mathcal{T}^{\infty} = \{t_{ATE}\}$. Otherwise, it is identified only up to a set compatible with the decision threshold rules.

Our next idea will be based on the characterization of the sets for the average treatment effect parameter and the propensity score identified under the given threshold decision rule under Assumption 3. We start our analysis with the following lemma, that follows directly from the combination of Assumptions 3(b) and (c).

Lemma 1. *Under Assumption 3 the propensity score can be point identified from the observations with infrequent attribute values as follows:*

$$P(x) = E\left[D|X = x, d_z(Z^x, Z^d) < \alpha_{N,x}, \|Z^x\|_z > \frac{1}{\alpha_{N,x}}\right].$$

Also, the average treatment effect can be point identified from the observations with infrequent attribute values as follows:

$$t_{ATE} = E\left[\frac{DY}{P(X)} - \frac{(1-D)Y}{1-P(X)}\middle| d_z(Z^x, Z^d) < \alpha_{N,x}, \|Z^x\|_z > \frac{1}{\alpha_{N,x}}\right].$$

This lemma states that if we are able to correctly reconstruct the master data set only for the observations with infrequent values of the attributes, those observations are sufficient for correct identification of the components of interest. Two elements are crucial for these results. First, we need Assumption 3(c) to establish redundancy of identifiers for matches constructed for observations with infrequent values of those identifiers. Second, we need Assumption 3(b) to guarantee that there is a nonzero probability of observing individuals with those infrequent values of identifiers.

The biggest challenge in our analysis is to determine which Cauchy sequences have appropriate behavior to isolate the infrequent attribute values as $N \to \infty$ and guarantee that the probability of a mismatch, conditional on the observation being in the reconstructed master sample, approaches zero. We do so by an appropriate inversion of the probability of misidentifying a pair of observations as a match. We can provide the general result that delivers a fixed probability of a mismatch in the limiting reconstructed master sample.

Proposition 1. *Suppose that for $x \in \mathcal{X}$ the chosen sequence $\{\alpha_{N,x}\} \in C_{0,x}$ satisfies*

$$Pr(m_{ij} = 0|x_i = x, \mathcal{D}^N(Z_i^y, Z_j^d) = 1) \to \gamma(x)$$

for some $\gamma(x) \in [0,1]$ as $N \to \infty$. Then

$$P^N_{\alpha_{N,x}}(x) = E^N_{\alpha_{N,x}}[D_j|X_i = x] \to (1 - \gamma(x))P(x) + \gamma(x)\overline{P}, \tag{3.2}$$

and

$$T^N_{\alpha_{N,x}} = E^N_{\alpha_{N,x}}\left[\frac{D_jY_j}{P^N_{\alpha_{N,x}}(X_i)} - \frac{(1-D_j)Y_j}{1-P^N_{\alpha_{N,x}}(X_i)}\right] \to t_{ATE}$$

$$+ E\left[(E[Y_1] - E[Y\,|\,X, D=1]\overline{P})\frac{\gamma(X)}{(1-\gamma(X))P(X)+\gamma(X)\overline{P}}\right] \tag{3.3}$$

$$-E\left[(E[Y_0] - E[Y\,|\,X, D=0](1-\overline{P}))\frac{\gamma(X)}{1-(1-\gamma(X))P(X)-\gamma(X)\overline{P}}\right].$$

Proposition 1 states that if one controls the mismatch probability in the combined data set, then the propensity score recovered through such a procedure is a convex combination of the true propensity score and the expected fraction $\overline{P}$ of treated individuals. Thus, the propensity score recovered through the data combination procedure will be biased toward the expected fraction of treated individuals. Also, the resulting identified average treatment effect will be a sum of the true ATE and a nontrivial term. In other words, the presence of mismatched observations in the "limiting" reconstructed master data set biases the estimated ATE toward zero.

The formulated theorem is based on the premise that a sequence in $C_{0,x}$ that leads to the limiting probability of an incorrect match equal to $\gamma(x)$ exists. The proof of existence of fundamental sequences satisfying this property is given in Komarova, Nekipelov, and Yakovlev (2011). These sequences are determined from the behavior of functions $\phi(\cdot)$ and $\psi(\cdot)$. The result in that paper demonstrates that for each $\gamma(x) \in [0,1]$ we can find a Cauchy sequence that leads to the limiting mismatch probability equal to $\gamma(x)$.

Our next goal is to use one particular sequence that will make the mismatch probability approach zero in the limit.

THEOREM 1. ***(Point identification of the propensity score and the ATE).*** *Suppose that for each $x \in \mathcal{X}$ the chosen sequence $\{\alpha_{N,x}\} \in C_{0,x}$ satisfies*

$$\lim_{N\to\infty} Pr(m_{ij} = 0 | x_i = x, \mathcal{D}^N(Z^y_i, Z^d_j) = 1) = 0.$$

Then

$$P^N_{\alpha_{N,x}}(\cdot) \to P(\cdot)$$

pointwise everywhere on $\mathcal{X}$ and

$$T^N_{\alpha_{N,x}} \to t_{ATE}$$

as $N \to \infty$.

In other words, the propensity score and the treatment effect are point identified.

10.4 Inference of the Propensity Score and the Average Treatment Effect with Limited Partial Disclosure

The calculations of the propensity score and the treatment effect require the data curator to have a technique that would combine the two data sets with the available observation-identifying information. Our approach to data combination described above is based on constructing the threshold decision rule that identifies the observations as "a match" corresponding to the data on a single individual if the observed individual attributes are close in terms of the chosen distance. With this approach we can construct the sequences of thresholds that would lead to very high probabilities of correct matches for a part of the population that allows us to point identify the propensity score and the treatment-effect parameter.

If we provide a high-quality match, then we have a reliable link between the public information regarding the individual and this individual's treatment status. The release of the reconstructed master data set would then constitute an evident threat to the individual's privacy. However, even if the reconstructed master data set is not public, the release of the estimated propensity score and/or the value of the treatment effect itself *may pose a direct threat to the security of individual data*. To measure the risk of such a disclosure in the possible linkage attacks, we use a measure based on the notion of disclosure in Lambert (1993). We provide a formal definition for this measure.

Partial disclosure can occur if the released information that was obtained from the data may potentially reveal some sensitive characteristics of individual. In our case, the information we are concerned with are the propensity score and the treatment effect. In particular, in our case the sensitive characteristic of an individual is her treatment status, or how an individual with given characteristics is likely to receive a treatment.

Below we provide a formal definition of the risk of partial disclosure for the propensity score. The definition takes as given the following two parameters. One parameter is $1-\delta$ and it characterizes the sensitivity level of the information about the propensity score. Namely, the information that the propensity score of an individual is above $1-\delta$ is considered to be damaging. The other parameter is denoted as $\underline{\nu}$ and represents a tolerance level—specifically, $\underline{\nu}$ is the upper bound on the proportion of individuals for whom the damaging information that $P(x) > 1-\delta$ may be revealed.

Another important component of our definition of partial disclosure is how much information about the data combination procedure is revealed to the public by the data curator. We denote this information as $\mathcal{I}$. For instance, if the data curator reveals that $Pr(m_{ij} = 0 | x_i = x, \mathcal{D}^N(Z_i^y, Z_j^d) = 1) \to \gamma(x)$ for some $\gamma(x)$, then the public can determine that in the limit the released propensity score for an individual with characteristics x has the form $(1-\gamma(x))P(x) + \gamma(x)\overline{P}$. If, in addition, the data curator releases the value of

$Pr(m_{ij} = 0|x_i = x, \mathcal{D}^N(Z_i^y, Z_j^d) = 1)$ or the value of $\gamma(x)$, then the public can pin down the true propensity score $P(x)$[3] and, thus, obtain potentially damaging information if this propensity score is above $1 - \delta$.

DEFINITION 3. *Let $\mathcal{I}$ be the information about the data combination procedure released to the public by the data curator. Let $\delta \in (0,1)$ and $\underline{\nu} \in [0,1]$. Given $\mathcal{I}$, we say that a $(1 - \delta, \underline{\nu})$ bound guarantee is given for the risk of partial disclosure, if the proportion of individuals in the private data set for whom the public can determine with certainty that $P(x) > 1 - \delta$ does not exceed $\underline{\nu}$. The value of $\underline{\nu}$ is called the bound on the risk of partial disclosure.*

Setting $\underline{\nu}$ at $\underline{\nu} = 0$ means that we want to protect *all* the individuals in the private data set.

The idea behind our definition of partial disclosure is that one can use the released values of $P^N_{\alpha_{N,x}}$ (or $\lim_{N\to\infty} P^N_{\alpha_{N,x}}$) from the model to determine whether the probability of the positive treatment status exceeds the given threshold. If this is possible to determine with a high confidence level for some individuals, then this individual is identified as the one with "the high risk" of the positive treatment status. Such information can be extremely damaging.

In the following theorem we demonstrate that the release of the true propensity score is not compatible with a low disclosure risk.

THEOREM 2. *Suppose that*

$$\lim_{N\to\infty} Pr(m_{ij} = 0|x_i = x, \mathcal{D}^N(Z_i^x, Z_j^d) = 1) = 0 \text{ for } x \in \mathcal{X}. \tag{4.4}$$

If the data curator releases information (4.4), then for sufficiently large N the release of the propensity score $P^N_{\alpha_{N,x}}$ (or its limit) is not compatible with the bound on the risk of partial disclosure $\underline{\nu}$ for sufficiently small $\underline{\nu}$.

The formal result of Theorem 2 relies on Assumption 2, and Theorem 1 and is based on two elements. First, using the threshold decision rule we were able to construct the sequence of combined data sets where the finite-sample distribution of covariates approaches the true distribution. Second, from the estimated distribution we could improve our knowledge of the treatment status of individuals in the data. For some individuals the probability of the positive treatment status may be very high.

This result forces us to think about ways to avoid the situations where potentially very sensitive information may be learned about some individuals. The bound guarantee on the risk of partial disclosure essentially requires the data curator to keep a given proportion of incorrect matches in the data sets of any size. As discussed in Proposition 1, a fixed proportion of the incorrect matches leads to the calculated propensity score to be biased toward the proportion of treated individuals in the population, and also causes bias in the average treatment effect.

3. Note that the value $\overline{P}$ is known from the public data set.

THEOREM 3. *Suppose the value of* $\overline{P}$ *is publicly available, and* $\overline{P} < 1 - \delta$. *A* $(1 - \delta, 0)$ *bound guarantee for the risk of partial disclosure can be achieved if the data curator chooses* $\alpha_N(x)$ *in such a way that*

$$\gamma(x) = \lim_{N \to \infty} Pr(m_{ij} = 0 | x_i = x, \mathcal{D}^N(Z_i^x, Z_j^d) = 1) > 0 \text{ for all } x \in \mathcal{X}$$

and for individuals with $P(x) > 1 - \delta$ *the value of* $\gamma(x)$ *is chosen large enough to guarantee that*

$$\lim_{N \to \infty} P^N_{\alpha_{N,x}} = (1 - \gamma(x))P(x) + \gamma(x)\overline{P} < 1 - \delta.$$

We assume that the data curator provides information that the data were matched with an error and the matching error does not approach 0 *as* $N \to \infty$ *but does not provide the values of* $Pr(m_{ij} = 0 | x_i = x, \mathcal{D}^N(Z_i^x, Z_j^d) = 1)$ *or* $\gamma(x)$. *In this case, the behavior of the released propensity score and the treatment effect is as described in equations (3.2) and (3.3), and thus, the true propensity score and the true treatment effect are not identified.*

Note that in the framework of Theorem 3 for individuals with small $P(x)$ the data curator may want to choose a very small $\gamma(x) > 0$ whereas for individuals with large $P(x)$ the bias toward $\overline{P}$ has to be large enough.

REMARK 1. Continue to assume that $\overline{P} < 1 - \delta$. Note that if the released propensity score for an individual with x is strictly less than $\overline{P}$, then the public will be able to conclude that the true propensity score for this individual is strictly less than $\overline{P}$. If the released propensity score for an individual with x is strictly greater than $\overline{P}$, then the public will be able to conclude that the true propensity score for this individual is strictly greater than $\overline{P}$ but, under conditions of Theorem 3, will not know whether $P(x) > 1 - \delta$. If the released propensity score for an individual with x is equal to $\overline{P}$, then the public is unable to make any nontrivial conclusions about $P(x)$—that is, $P(x)$ can be any value from $[0,1]$.

We can consider other approaches the data curator may exploit regarding the release of the propensity score values and the information provided with this release. For instance, for some individuals with $P(x) < 1 - \delta$ she may choose $\gamma(x) = 0$ and provide information that *for some individuals* the data were matched without an error in the limit, but for the other individuals the matching error is strictly positive and does not approach 0 as $N \to \infty$ (given that she does not specify the values of $Pr(m_{ij} = 0 | x_i = x, \mathcal{D}^N(Z_i^x, Z_j^d) = 1)$ or $\gamma(x)$). In this case, the result of Theorem 3 continues to hold.

The next theorem gives a result on privacy protection when the data curator releases more information.

THEOREM 4. *Suppose the value of* $\overline{P}$ *is publicly available, and* $\overline{P} < 1 - \delta$. *A* $(1 - \delta, 0)$ *bound guarantee for the risk of partial disclosure can be achieved if the data curator chooses* $\alpha_N(x)$ *in such a way that*

$$Pr(m_{ij} = 0 | x_i = x, \mathcal{D}^N(Z_i^x, Z_j^d) = 1) \geq \overline{\gamma} \text{ for all } x \in \mathcal{X}$$

for all N, and for individuals with $P(x) > 1 - \delta$ *the value of* $Pr(m_{ij} = 0 | x_i = x, \mathcal{D}^N(Z_i^x, Z_j^d) = 1)$ *is chosen large enough to guarantee that*

$$P^N_{\alpha_{N,x}} = (1 - Pr(m_{ij} = 0 | x_i = x, \mathcal{D}^N(Z_i^x, Z_j^d) = 1))P(x)$$
$$+ Pr(m_{ij} = 0 | x_i = x, \mathcal{D}^N(Z_i^x, Z_j^d) = 1)\overline{P} < 1 - \delta$$

for all N. We assume that the data curator provides information that the data were matched with an error and the matching error is greater or equal than the known $\overline{\gamma}$ *but does not provide the values of* $Pr(m_{ij} = 0 | x_i = x, \mathcal{D}^N(Z_i^x, Z_j^d) = 1)$ *or* $\gamma(x)$*. In this case, the behavior of the released propensity score and the treatment effect is as described in equations (3.2) and (3.3), and thus, the true propensity score and the true treatment effect are not identified.*

To summarize, the fact that we want to impose a bound on the risk of disclosure leads us to the loss of the point identification of the true propensity score and the true average treatment effect. This means that the point identification of the econometric model from the combined data set is incompatible with the security of individual information. If the publicly observed policy is based on the combination of the nonpublic treatment status and the public information regarding the individual, then the treatment status of any individual cannot be learned from this policy only if it is based on a biased estimate for the propensity score and a biased treatment effect.

The next theorem considers the case when $\overline{P} > 1 - \delta$. It shows that in this case *any* release of point estimates of the propensity score from the treatment effect evaluation is not compatible with a low disclosure risk.

THEOREM 5. *Suppose the value of* $\overline{P}$ *is publicly available, and* $\overline{P} > 1 - \delta$*. Then the released propensity score will reveal all the individuals with* $P(x) > 1 - \delta$ *even if the data are combined with a positive (even very large) error. Let*

$$p^* = Pr(x : P(x) > 1 - \delta),$$

that is, p^* *is the proportion of individuals with the damaging information about the propensity score. Then a* $(1 - \delta, \underline{\nu})$ *bound guarantee cannot be attained for the risk of partial disclosure if* $\underline{\nu} \leq p^*$*.*

In the framework of Theorem 5 the release (or publicly observable use) of the propensity score is blatantly nonsecure. In other words, there will exist a sufficient number of individuals for whom we can learn their high propensity scores. To protect their privacy, *no* propensity scores should be released.

10.5 Does a Religious Affiliation Affect a Parent's Decision on Childhood Vaccination and Medical Checkups?

To illustrate our theoretical analysis, we want to bring our results to the real data.

Even though in the main body of this chapter we do not develop a formal theory of the statistical estimation of $P^N_{\alpha_{N,x}}(\cdot)$ or the true propensity score $P(\cdot)$ in a finite sample, in this section we want to illustrate an empirical procedure one could implement in practice.

The data come from the Russian Longitudinal Monitoring survey (RLMS).[4] The RLMS is a nationally representative annual survey that covers more than 4,000 households (the number of children varies between 1,900 and 3,682), from 1992 until 2011. The survey gathers information on a very broad set of questions, including demographic and household characteristics, health, religion, and so on. The survey covers 33 Russian regions—31 oblasts (krays, republics), and also Moscow and St. Petersburg. Islam is the dominant religion in two regions, and Orthodox Christianity is the dominant religion in the rest.

We combine our data from two parts of the RLMS—the survey for adults and the survey for children. The question that we want to answer can be informally stated as follows: Does the religion of family members affect the probability of a child getting regular medical checkups or to be vaccinated against tuberculosis? More specifically, we analyze whether (1) religious (Muslim or Orthodox Christian) families have their children seen by doctors or have their children vaccinated against tuberculosis with lower probability; and (2) families from neighborhoods with high percentages of religious people have their children seen by doctors with lower probability.

From the data set for children we extract the following individual characteristics for a child: the indicator for whether the child had a medical checkup in the last twelve (or three) months, the indicator for whether the child was vaccinated against tuberculosis, the indicator for whether the child lives in a city, and the child's age. We also have the following information on the child's family: the share of Orthodox Christian family members, the share of Muslim family members,[5] and the share of family members with a college degree. From other publicly available data sets we obtain the following information for the child's region: the share of Muslims and the gross regional product per capita. The summary statistics of all these variables are presented in table 10.1.

Our analysis focuses on the propensity scores that represent the probabil-

4. This survey is conducted by the Carolina Population Center at the University of Carolina at Chapel Hill, and by the Higher School of Economics in Moscow. Official Source name: "Russia Longitudinal Monitoring Survey, RLMS-HSE," conducted by Higher School of Economics and ZAO "Demoscope" together with Carolina Population Center, University of North Carolina at Chapel Hill and the Institute of Sociology RAS. (RLMS-HSE websites: http://www.cpc.unc.edu/projects/rlms-hse, http://www.hse.ru/org/hse/rlms).

5. Variables for the shares of Muslims and Orthodox Christians in a family are constructed based on the following definition of a Muslim (Orthodox Christian). We say that a person is a Muslim (Orthodox Christian) if the person (a) says that she believes in God, and (b) says that she is a Muslim (Orthodox Christian). There are people in the survey who said, for example, that they are Muslims, but at the same time said that they are not believers. We consider such people nonbelievers.

Table 10.1 Summary statistics of various variables for a child

Variable	Obs.	Mean	Std. Dev.	Min.	Max.
Child: Medical checkup in last 12 months?	33,924	0.69	0.46	0	1
Child: Medical checkup in last 3 months?	62,316	0.45	0.50	0	1
Child: Vaccinated (tuberculosis)?	49,464	0.96	0.19	0	1
Child: I (lives in a city)	73,100	0.38	0.49	0	1
Child: Age	73,100	7.19	4.09	0	18
Family: Share of Orthodox Christians	59,142	0.22	0.35	0	1
Family: Share of Muslims	59,142	0.06	0.23	0	1
Family: Share of those with college degree	66,314	0.26	0.37	0	1
Region: Share of Muslims	73,100	0.09	0.17	0	0.71
Region: Log group per capita	71,466	10.96	1.38	7.04	13.50

ity of the child getting regular checkups (being vaccinated against tuberculosis). In our model, the following information is considered to be sensitive: propensity scores are below a given threshold; the variable of the share of Orthodox Christian (or Muslim) family members has a negative marginal effect on the propensity score; the variable of the share of Orthodox Christians (or Muslims) in the child's neighborhood has a negative marginal effect on the propensity score.

The RLMS data set has a clustered structure as people are surveyed within small neighborhoods with a population of around 300 people (so-called census district; see Yakovlev [2012]). Thus, it is possible to construct characteristics of neighborhoods—in particular, on the shares of Orthodox Christians (or Muslims) in neighborhoods—by using the religion variable from the RLMS data set for adults[6] if one has information on neighborhood labels. Due to a vast Soviet heritage, the majority of people in Russia live in large communal developments that combine several multistory apartment buildings. These developments have common infrastructure, shops, and schools. High concentration in a relatively small area makes the life of each family very visible to all the neighbors. The neighborhoods are defined precisely by such developments. Neighborhood labels were publicly available till 2009 but then were deleted by the RLMS staff due to the privacy concerns.[7] In our study, we exploit the RLMS survey data from 1994 until 2009 because the neighborhood identifiers were publicly available in those years and, thus, one was able to consider the child's neighborhood and then use the religious affiliation variable from the adult data set to construct the data for religion in that particular neighborhood, and use the income variable from the adult data set to calculate the average logarithm of income

6. Thus, the variable for the shares of Muslims and Orthodox Christians in a neighborhood is constructed using the same principle as in the case of families.

7. Fortunately, we happened to have the data on neighborhood identifiers.

Table 10.2 Summary statistics of neighborhood characteristics

Variable	Obs.	Mean	Std. Dev.	Min.	Max.
Neighborhood: Share of Muslims	53,800	0.06	0.20	0	1
Neighborhood: Share of Orthodox	53,800	0.23	0.18	0	1
Neighborhood: Log(income)	58,578	6.25	1.86	0	10.9

in that particular neighborhood. The summary statistics of neighborhood characteristics are presented in table 10.2.

In order to answer the posed questions, we estimate the following probit regression

$$\begin{aligned} Pr(D_{it} = 1) = \Phi(&\alpha_1 share\ of\ Muslims\ in\ family_{it} \\ &+ \alpha_2 share\ of\ Orthodox\ Christians\ in\ family_{it} \\ &+ \beta_1 share\ of\ Muslims\ in\ neighborhood_{it} \\ &+ \beta_2 share\ of\ Orthodox\ Christians\ in\ neighborhood_{it} + \gamma' q_{it}), \end{aligned}$$

where D_{it} stands for the indicator of whether a child had a medical checkup within the last twelve (or three) months, or the indicator of whether a child has a vaccination against tuberculosis. The set of controls q_{it} contains child's characteristics of (age, I(live in city)), regional characteristics such as the GRP per capita and the share of Muslims in the region, family characteristics such as family income and the share of family members with a college degree, neighborhood characteristics (average income in neighborhood), and the year fixed effects. For notational simplicity, we write $Pr(D_{it} = 1)$ instead of $Pr(D_{it} = 1|religious\ characteristics_{it}, q_{it})$.

The estimation results are presented in table 10.3. Columns (2) and (4) in the table show the evidence that a higher percentage of Muslims in the family is associated with a lower chance of the child being regularly seen by a doctor. This holds for the sample of all children and for the subsample of children with health problems. Also, when the sample of all children is considered, a higher percentage of Muslims in the neighborhood has a negative marginal effect on the probability of the child being vaccinated against tuberculosis as well as being regularly seen by a doctor. The variables for the shares of Orthodox Christians are not significant.

The discussion below considers the sample of all children. The first two graphs in figure 10.1 are for the case when the dependent variable is the indicator for a checkup within the last twelve months. The last two graphs in that figure are for the case when the dependent variable is the indicator for a vaccination against tuberculosis. The large dot in the first graph in figure 10.1 shows the pair (–0.3416, –0.3314) of estimated coefficients for the share of Muslims in the family and the share of Muslims in the neighborhood from

Table 10.3 Probit regression estimation

	Sample: All children		Sample: Children with health problems
	Medical checkup in last 12 months?	Vaccinated against tuberculosis?	Medical checkup in last 3 months?
Child: Age	–0.0423	0.0685	–0.0438
	[0.0032]***	[0.0047]***	[0.0067]***
Child: I (live in city)	0.1704	–0.2062	0.0601
	[0.0313]***	[0.0441]***	[0.0543]
Family: Share of Muslims	–0.3314	–0.1506	–0.4193
	[0.1127]***	[0.1686]	[0.2515]*
Family: Share of Orthodox Christians	0.0478	–0.0936	–0.0244
	[0.0394]	[0.0604]	[0.0711]
Family: Average log(income)	0.0602	–0.0169	0.0437
	[0.0151]***	[0.0211]	[0.0303]
Family: Share of those with a college degree	0.0741	0.0296	0.1561
	[0.0367]**	[0.0571]	[0.0651]**
Region: Share of Muslims	–0.0129	–0.3195	0.2551
	[0.1421]	[0.2062]	[0.3075]
Region: Log GRP per capita	0.1838	–0.0412	–0.0858
	[0.0308]***	[0.0463]	[0.0544]
Neighborhood: Share of Muslims	–0.3416	–0.429	–0.4922
	[0.1757]*	[0.2319]*	[0.4512]
Neighborhood: Share of Orthodox	–0.105	–0.0169	–0.1603
	[0.0840]	[0.1272]	[0.1603]
Year fixed effects	yes	yes	yes
Constant	–2.0794	1.9472	–3.9003
	[0.3701]***	[0.4039]***	[103.6494]
Observations	10,780	17,413	2,902

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

column (2) in table 10.3. The large dot in the second graph in figure 10.1 shows the pair (–0.105, –0.3314) of estimated coefficients for the share of Orthodox Christians in the neighborhood and the share of Muslims in the neighborhood, respectively, from column (2) in table 10.3. The large dot in the third graph in figure 10.1 shows the pair (–0.1506, –0.429) of estimated coefficients for the share of Muslims in the family and the share of Muslims in the neighborhood from column (3) in table 10.3. The large dot in the fourth graph in figure 10.1 shows the pair (–0.105, –0.429) of estimated coefficients for the share of Orthodox Christians in the neighborhood and the share of Muslims in the neighborhood, respectively, from column (3) in table 10.3.

Finally, we analyze how the estimates of our parameters would change if

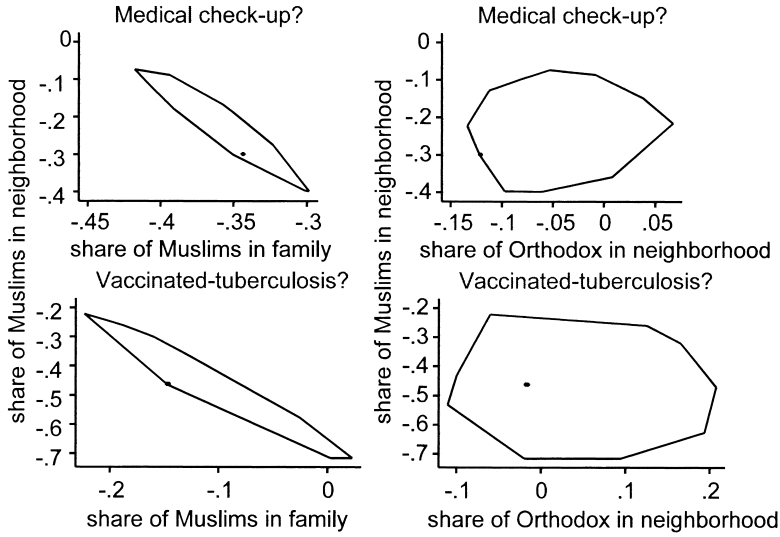

Fig. 10.1 Sets of estimates from 1,000 data sets combined using neighborhoods. Contour sets are for the cases of 2-anonymity

we enforce a bound on the risk of partial disclosure and consider the bound of 0.5—that is, $Pr(m_{ij} = 0|\mathcal{D}^N(Z_i^x, Z_j^d) = 1) \geq \overline{\gamma}$, where $\overline{\gamma} = 0.5$. This is the case of attaining 2-anonymity.

In order to attain 2-anonymity we conduct the following exercise. For every child in our sample we create two possible neighborhoods—one neighborhood is the true one, and the other one is drawn randomly from the empirical distribution of neighborhoods in the corresponding region. Such empirical distributions can be easily obtained from the publicly available data in RLMS. As a result, for every child we have two possible sets of values of neighborhood characteristics.

Then, ideally we would like to simulate all possible combined data sets but the number of these data sets is of exponential complexity, namely, of the rate 2^n. Instead of considering all possible combined data sets, we randomly simulate only 1,000 such data sets. For each simulated combined data set we conduct the probit estimation. Thus, we end up with a 1,000 different sets of estimated coefficients (as well as the propensity scores). The contour sets in the graphs in figure 10.1 are the convex hulls of the obtained estimates. Namely, the contour set in the first graph in figure 10.1 is the convex hull of the 1,000 pairs of estimated coefficients for the share of Muslims in the family and the share of Muslims in the neighborhood, respectively. The contour set in the second graph in figure 10.1 is the convex hull of the 1,000 pairs of estimated coefficients for the share of Orthodox Christians in the neighborhood and the share of Muslims in the neighborhood, respectively. Similarly for the other two graphs.

As can be seen, in the analysis of the probability of a medical checkup in the last twelve months, all the 1,000 coefficients corresponding to variables of the share of Muslims in the family and the share of Muslims in the neighborhood are negative.[8] If the data curator thinks that the release of these sets of estimates is not satisfactory with regard to partial disclosure guarantees, then she should increase the guarantee level by, for instance, attaining 3-anonymity.

As for the case of the probability of being vaccinated against tuberculosis, among the 1,000 coefficients corresponding to the share of Muslims in the family, there are some positive ones, even though all the 1,000 coefficients corresponding to the share of Muslims in the neighborhood are negative.[9] Again, the data curator may want to increase the guarantee level.

10.6 Conclusion

In this chapter we analyze how the combination of data from multiple anonymized sources can lead to the serious threats of the disclosure of individual information. While the anonymized data sets by themselves may pose no direct threat, such a threat may arise in the combined data. The main question that we address is whether statistical inference based on the information from all these data sets is possible without the risk of disclosure. We introduce the notion of *statistical partial disclosure* to characterize a situation when data combination allows an adversary to identify a certain individual characteristic with a small probability of misidentification. We focus our analysis on the estimation of treatment effects where the treatment status of an individual is sensitive and, thus, the possibility of the statistical recovery of this treatment status may be highly undesirable. We show that a variety of techniques from data mining literature can be used for reconstruction of the combined data sets with little to no auxiliary information. We also demonstrate that the point identification of the statistical model for the average treatment effects is incompatible with bounds imposed on the risk of statistical partial disclosure imposed to protect individual information. We illustrate our findings in the empirical study of the impact of religious affiliation of parents on the probability of a child's medical checkups and vaccination from tuberculosis using the individual-level data from Russia.

Statistical partial disclosure is becoming of central importance in the "big data" world. While many consumer companies have been routinely collecting private consumer data, the modern data-driven business paradigm calls for using these data in business decisions. A common example is the online ad-targeting technology where the consumer is exposed to the ads based on

8. These variables are significant in each of 1,000 cases (even though the confidence intervals are not depicted in the graphs).

9. The variable of the share of Muslims in the neighborhood is significant in each of 1,000 cases.

the past consumer behavior and the known consumer characteristics. The ad delivery is based on the estimator that would be used to predict the consumer click on the ad based on the historical behavior of the given consumer and other consumers similar in some sense to the consumer of interest. *Forbes* magazine published a story explaining how Target uses credit card information to identify the repeated purchases from the same customer, and using a variety of sources identifies the set of demographic characteristics. Then, based on the collected demographic information and the sets of products that the consumers purchased in the past, Target was able to identify the sets of purchased products that most likely lead to a customer (a female) being pregnant. Based on this prediction, Target sent out coupons for the baby section in the store. *Forbes* then proceeds with the anecdotal story of when Target customer service got a call from an angry father of a teenager stating that his daughter got the coupon. A week later the father called Target back with an apology, as his daughter had indeed turned out to be pregnant.

With further advancement in econometric and machine-learning methods, similar stories will emerge in a large variety of settings, from medical services (where people already get customized automatic medical advice based on their reported lifestyle, eating, and exercise habits) to real estate (where companies like Zillow give the homeowners automated recommendations for the timing of the house sale and purchase). We argue that confidentiality restrictions can go hand in hand with the big data tools to provide technologies that are both aimed at higher consumer welfare (leading to better consumer targeting) and provide formal privacy guarantees. We have studied some of these technologies in this chapter.

References

Abowd, J., and L. Vilhuber. 2008. "How Protective Are Synthetic Data?" *Privacy in Statistical Databases* 5262:239–46.

Abowd, J., and S.Woodcock. 2001. "Disclosure Limitation in Longitudinal Linked Data." In *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, edited by P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, 215–77. Amsterdam: North Holland.

Acquisti, A. 2004. "Privacy and Security of Personal Information." In *Economics of Information Security*, vol. 12, edited by L. Jean Camp and Stephen Lewis, 179–86. New York: Springer Science+Business Media.

Acquisti, A., A. Friedman, and R. Telang. 2006. "Is There a Cost to Privacy Breaches? An Event Study." *Proceedings of the Twenty-Seventh International Conference on Information Systems.* doi: 10.1.1.73.2942&rep=rep1&type=pdf.

Acquisti, A., and J. Grossklags. 2008. "What Can Behavioral Economics Teach Us about Privacy?" In *Digital Privacy: Theory, Technologies, and Practices*, edited by A. Acquisti, S. Gritzalis, S. DiVimercati, and C. Lambrinoudakis, 363–79. Boca Raton, FL: Auerbach Publications, Taylor & Francis Group.

Acquisti, A., and H. Varian. 2005. "Conditioning Prices on Purchase History." *Marketing Science* 33:367–81.

Aggarwal, G., T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. 2005. "Approximation Algorithms for k-anonymity." *Journal of Privacy Technology*, Paper no. 2005112001.

Bradley, C., L. Penberthy, K. Devers, and D. Holden. 2010. "Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future." *Health Services Research* 45 (5, pt. 2): 1468–88.

Calzolari, G., and A. Pavan. 2006. "On the Optimality of Privacy in Sequential Contracting." *Journal of Economic Theory* 130 (1): 168–204.

Ciriani, V., S. di Vimercati, S. Foresti, and P. Samarati. 2007. "k-Anonymity." In *Secure Data Management in Decentralized Systems*, vol. 33, edited by T. Yu and S. Jajodia. Berlin: Springer-Verlag.

Duncan, G., S. Fienberg, R. Krishnan, R. Padman, and S. Roehrig. 2001. "Disclosure Limitation Methods and Information Loss for Tabular Data." In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, edited by P. Doyle, 135–66. Amsterdam: North Holland.

Duncan, G., and D. Lambert. 1986. "Disclosure-Limited Data Dissemination." *Journal of the American Statistical Association* 81 (393): 10–18.

Duncan, G., and S. Mukherjee. 1991. "Microdata Disclosure Limitation in Statistical Databases: Query Size and Random Sample Query Control." In *Proceedings of IEEE Symposium on Security and Privacy*, 278–87.

Duncan, G., and R. Pearson. 1991. "Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future." *Statistical Science* 6 (3): 219–32.

Dwork, C. 2006. "Differential Privacy." In *Automata, Languages and Programming*, edited by M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, 1–12. Berlin: Springer-Verlag.

Dwork, C., and K. Nissim. 2004. "Privacy-Preserving Data Mining on Vertically Partitioned Databases." In *Advances in Cryptology–CRYPTO 2004*, edited by M. Franklin, 134–38. New York: Springer.

Fienberg, S. 1994. "Conflicts between the Needs for Access to Statistical Information and Demands for Confidentiality." *Journal of Official Statistics* 10:115.

———. 2001. "Statistical Perspectives on Confidentiality and Data Access in Public Health." *Statistics in Medicine* 20 (9–10): 1347–56.

Goldfarb, A., and C. Tucker. 2010. "Online Display Advertising: Targeting and Obtrusiveness." *Marketing Science* 30 (3): 389–404.

Gross, R., and A. Acquisti. 2005. "Information Revelation and Privacy in Online Social Networks." In *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*, edited by V. Atluri, S. di Vimercati, and R. Dingledine, 71–80. New York: Association for Computing Machinery.

Homer, N., S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. Pearson, D. Stephan, S. Nelson, and D. Craig. 2008. "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays." *PLoS Genetics* 4 (8): e1000167.

Horowitz, J., and C. Manski. 2006. "Identification and Estimation of Statistical Functionals Using Incomplete Data." *Journal of Econometrics* 132 (2): 445–59.

Horowitz, J., C. Manski, M. Ponomareva, and J. Stoye. 2003. "Computation of Bounds on Population Parameters When the Data are Incomplete." *Reliable Computing* 9 (6): 419–40.

Komarova, T., D. Nekipelov, and E. Yakovlev. 2011. "Identification, Data Combination and the Risk of Disclosure." CeMMAP Working Paper no. CWP39/11, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Korolova, A. 2010. "Privacy Violations Using Microtargeted Ads: A Case Study." In *IEEE International Workshop on Privacy Aspects of Data Mining (PADM'2010)*, 474–82, Washington, DC. doi:10.1109/ICDMW.2010.137.

Lambert, D. 1993. "Measures of Disclosure Risk and Harm." *Journal of Official Statistics* 9:313.

LeFevre, K., D. DeWitt, and R. Ramakrishnan. 2005. "Incognito: Efficient Full-Domain k-Anonymity." In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, edited by Fatma Ozcan, 49–60. Association for Computing Machinery.

———. 2006. "Mondrian Multidimensional k-anonymity." In ICDE'06 Proceedings of the 22nd International Conference on Data Engineering, 25. Institute of Electronics and Electronic Engineers.

Magnac, T., and E. Maurin. 2008. "Partial Identification in Monotone Binary Models: Discrete Regressors and Interval Data." *Review of Economic Studies* 75 (3): 835–64.

Manski, C. 2003. *Partial Identification of Probability Distributions*. Berlin: Springer-Verlag.

Miller, A., and C. Tucker. 2009. "Privacy Protection and Technology Diffusion: The Case of Electronic Medical Records." *Management Science* 55 (7): 1077–93.

Molinari, F. 2008. "Partial Identification of Probability Distributions with Misclassified Data." *Journal of Econometrics* 144 (1): 81–117.

Narayanan, A., and V. Shmatikov. 2008. "Robust De-Anonymization of Large Sparse Datasets." In *SP 2008 IEEE Symposium on Security and Privacy*, 111–125. Institute of Electronics and Electrical Engineers.

Ridder, G., and R. Moffitt. 2007. "The Econometrics of Data Combination." *Handbook of Econometrics* 6 (6b): 5469–547.

Samarati, P., and L. Sweeney. 1998. "Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression." Technical Report SRI-CSL-98-04, Computer Science Laboratory, SRI International.

Sweeney, L. 2002a. "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression." *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems* 10 (5): 571–88.

———. 2002b. "k-Anonymity: A Model for Protecting Privacy." *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems* 10 (5): 557–70.

Taylor, C. 2004. "Consumer Privacy and the Market for Customer Information." *RAND Journal of Economics* 35 (4): 631–50.

Varian, H. 2009. "Economic Aspects of Personal Privacy." In *Internet Policy and Economics*, edited by W. H. Lehr and L. M. Pupillo, 101–09. New York: Springer Science+Business Media.

Wilson, A., T. Graves, M. Hamada, and C. Reese. 2006. "Advances in Data Combination, Analysis and Collection for System Reliability Assessment." *Statistical Science* 21 (4): 514–31.

Wright, G. 2010. "Probabilistic Record Linkage in SAS®." Working Paper, Kaiser Permanente, Oakland, CA.

Yakovlev, E. 2012. "Peers and Alcohol: Evidence from Russia." CEFIR Working Paper no. 182, Center for Economic and Financial Research.