Estimation of Treatment Effects from Combined Data: Identification versus Data Security[1]

Tatiana Komarova,[2] **Denis Nekipelov**,[3] **Evgeny Yakovlev**,[4].

This version: May 2013

ABSTRACT

The security of sensitive individual data is a subject of undisputable importance. One of the major threats to sensitive data arises when one can link sensitive information and publicly available data. In this paper we demonstrate that even if the sensitive data are never publicly released, the point estimates from the empirical model estimated from the combined public and sensitive data may lead to a disclosure of individual information. Our theory builds on the work in Komarova, Nekipelov and Yakovlev (2012) where we analyze the individual disclosure that arises from the releases of marginal empirical distributions of individual data. The disclosure threat in that case is posed by the possibility of a linkage between the released marginal distributions. In this paper, we analyze a different type of disclosure. Namely, we use the notion of the risk of *statistical partial disclosure* to measure the threat from the inference on sensitive individual attributes from the released empirical model that uses the data combined from the public and private sources. As our main example we consider a treatment effect model in which the treatment status of an individual constitutes sensitive information.

JEL Classification: C35, C14, C25, C13.

Keywords: Data protection, model identification, data combination.

[1] We appreciate helpful comments from P. Haile, M. Jansson, A. de Paula, M. Pesendorfer, J. Powell., J. Sutton, and E. Tamer.

[2] Corresponding author, Department of Economics, London School of Economics and Political Science; e-mail: t.komarova@lse.ac.uk

[3] Department of Economics, UC-Berkeley; e-mail: `nekipelov@econ.berkeley.edu`.

[4] New Economic School; email: eyakovlev@nes.ru

1 Introduction

In policy analysis and decision making in many areas it is instrumental 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 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 models 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 the daily basis. Thus, some of the variables in the datasets used for policy analysis may be publicly observable.[1]

In this paper 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 identification of treatment effects from the combined public and sensitive data is compatible with formal restrictions on the risk of 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.*

In our empirical application we provide a concrete example of the analysis of treatment effects from the combined sensitive "anonymous" data and publicly observable individual characteristics. Such an analysis may be further used to make inferences regarding the future treatment assignments and other policy decisions. The results of these policy decisions may be publicly observable, for instance, involving the changes in staff and day-to-day operations. The data that we use comes from certain reviews

[1] Reportedly, many businesses indeed rely on the combined data. See, e.g. Wright (2010) and Bradley, Penberthy, Devers, and Holden (2010), among others.

on Yelp.com, where we focus on consumer ratings of the Health care businesses and where Yelp users rank those businesses based on their experiences. We use the review data for facilities located in Durham county, North Carolina. In particular, we want to evaluate the effect of a visit to an outpatient medical facility on the “mood” of individuals which is measured by changes in their Yelp ratings. In short, we want to answer the question: *Do the doctors make people happier?*

User information on Yelp.com does not, however, contain demographic or location information of users. Without controlling for this information, the inference solely based on the review data is prone to a selection bias: consumers who use the health-care facilities more frequently may be more prone to give a review. Moreover, a more frequent Yelp.com reviewer will be more likely to give a review of any Yelp business, not just the healthcare business. In order to model how a company may use the demographic data we collected the database of individual property tax records in Durham county. Using the record linkage technique adopted from the data mining literature, we merge the health service review data with the data on individual locations and property values which we use to control the sample selection. Our data linkage technique relies on observing data entries with infrequent attribute values (extracted from individual names and locations) in the two combined datasets. Accurate links for these entries may disclose consumer identities. We note that the goal of our work is not to demonstrate the vulnerability of online 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 consumer behavior 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 consumer-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 datasets 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 consumer dataset and auxiliary individual information with the probability exceeding the selected confidence threshold presents a case of a successful linkage attack. 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 been long used in the analysis of databases and data mining. To some extent, these results were used in econometrics for combination of datasets as described in Ridder and Moffitt (2007). In record linkage one provides a (possibly) probabilistic rule that can match the records from one dataset with the records from the other dataset 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 birthdate, sex, and zip code of the patient. Recent "de-personalized" 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 are the data from genome-wide association studies (GWAS). GWAS are devoted to an in-depth analysis of genetic origins of human health conditions and receptiveness to deceases, among other things. A common practice of such studies was to publish the data on the 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) dataset from the HapMap NIH project which consists of SNP data from 4 populations with about 60 individuals each. Homer, Szelinger, Redman, Duggan, Tembe, Muehling, Pearson, Stephan, Nelson, and Craig (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 decease, the discovery that a particular individual belongs to the studied subsample means that this individual has that condition or that decease.

Samarati and Sweeney (1998), Sweeney (2002b), Sweeney (2002a), LeFevre, DeWitt, and Ramakrishnan (2005), Aggarwal, Feder, Kenthapadi, Motwani, Panigrahy, Thomas, and Zhu (2005), LeFevre, DeWitt, and Ramakrishnan (2006), Ciriani, di Vimercati, Foresti, and Samarati (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 bound 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), and Fienberg (2001) Duncan, Fienberg, Krishnan, Padman, and Roehrig (2001), 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 disclosure risk. 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. 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 paper 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 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 literature on partial identification of models with contaminated or corrupted data. Manski (2003), Horowitz, Manski, Ponomareva, and Stoye (2003), Horowitz and Manski (2006), 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 paper 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 paper, the sizes of such identified sets for the propensity scores and the average treatment effect are directly proportional to the pessimistic measure of the 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. As a result, the combination of the company's internal data with the auxiliary public individual data is not compatible with the non-disclosure of individual identities. An increase in the complexity and nonlinearity of the model can further worsen the trade-off.

In the paper 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 the company "customizes" its policy regarding individual users, e.g. a PPO gives its customers personalized recommendations regarding their daily routines and exercise or hospitals re-assign specialty doctors based on the number of patients in need of specific procedures, then the observe 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), Calzolari and Pavan (2006). Consider the setting where firms obtain a comparative advantage due to the possibility of offering prices that are based on the 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 paper is organized as follows. Section 2 describes the analyzed treatment effects models, the availability of the data and gives a description of data combination procedures employed in the paper. Section 3 provides a notion of the identified sets, which are compatible with the data combination procedure, for the propensity score and the average treatment effect. It looks at the properties of these sets as the sizes of available data sets go to infinity. Section 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 5 provide an empirical illustration.

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 this individual's demographics. For instance, in 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 perhaps present a very sensitive piece of information. However, the fact that a particular person receives HIV treatment medications may be extremely sensitive.

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 corresponds to the non-treated 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 $\mathcal{V} = \mathcal{S}^* \times \mathbb{R}^v$, where $\mathcal{S}^*$ is a finite space of arbitrary (non-numeric) nature. $\mathcal{S}^*$, 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. 30 characters). The string combination $\{'John',' Smith',' 01/01/1990'\}$ in 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 $\mathcal{S}^*$ 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, Graves, Hamada, and Reese (2006)). We can then define the norm in $\mathcal{S}^*$ as the distance between the given point in $\mathcal{S}$ 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 $\mathcal{S}$ and the standard Euclidean norm in $\mathbb{R}^v$ and denote it $\|\|_{\mathcal{V}}$. Similarly, we assume that W takes values in $\mathcal{W} = \mathcal{S}^{**} \times \mathbb{R}^w$, where $\mathcal{S}^{**}$ is also a finite space. The norm in $\mathcal{W}$ is defined as a weighted norm and denoted as $\|\|_{\mathcal{W}}$. Spaces $\mathcal{S}^*$ and $\mathcal{S}^{**}$ 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 $\{y_i, d_i, x_i, v_i, w_i\}_{i=1}^{N_y}$. 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 and treatment outcome are not contained in the same sample. One sample, containing N_y observations is the i.i.d. sample $\{x_i, v_i\}_{i=1}^{N_y}$ is in the *public domain*. In other words, individual researchers or research organizations can get access to this dataset. The second dataset ia a subset of $N \leq N_y$ observations from the "master" dataset and contains information regarding the treatment status $\{y_j, d_j, w_j\}_{j=1}^{N}$. This dataset *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 datasets. In other words, the variables in v_i and w_j do not provide immediate links between the two datasets. In our example on the HIV treatment status, we could consider the cases where the data on 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 datasets.

Suppose that the true response model can be characterized by two potential outcomes Y_1 and Y_0 corresponding to the value of the treatment status. We impose the following assumptions on the elements of the model.

ASSUMPTION 1 *(i) The treatment outcomes satisfy the conditional unconfoundedness, i.e.* $(Y_1, Y_0) \perp D \,\big|X = x$

(ii) At least one element of X *has a continuous distribution with 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

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

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

$$\bar{P} = E\left[P(x)\right].$$

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

$$Y = D\,Y_1 + (1 - D)\,Y_0.$$

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

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

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

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

as well as overall average treatment effect (ATE)

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

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

Evaluation of the propensity score and the mentioned treatment effects requires us to observe the treatment status and 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 datasets and provide an estimate for 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, e.g. by putting a warning label on the package of the studied drug), we assume that she will construct the linkage procedure that will combine the two datasets.

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 datasets. Then, the effect of interest will be estimated from the reconstructed joint dataset. To establish similarity between the two datasets, 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 $Z^y = Z^d(X, V)$ (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 $d_z(Z^x, Z^d) = \big(\omega_s d_{\mathcal{S}}(z_s^x, z_s^d)^2 + \omega_z \|z_z^x - z_z^s\|^2\big)^{1/2}$, 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 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 variables 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 dataset either having the race not recorded, or being black, white, hispanic or asian.

Then, using the public dataset 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, e.g. 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 name 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.

If the set of numeric and string characteristics used for combining two datasets is sufficiently large or it can contain some potentially "hard to replicate" information such as an individual's full name, if such a match occurs it very likely singles out the data of one person. We formalize this idea by expecting that the probability of two observations with close values of identifiers Z^d and Z^x belong to the same individual is the higher, the more infrequent their values are (the larger is their norm, that we define as a distance from the "generic" set of attributes). Our maintained assumptions regarding the distribution of constructed identifiers are listed below.

ASSUMPTION 3 *We fix some $\underline{\alpha}, \bar{\alpha} \in (0,1)$ with $\underline{\alpha} < \bar{\alpha}$, then for any $\alpha \in (\underline{\alpha}, \bar{\alpha})$:*

(i) (Proximity of identifiers) $Pr\left(d_z(Z^x,\, Z^d) < \alpha \;\middle|\; X = x,\, D = d,\, Y = y,\, \|Z^d\|_z > \frac{1}{\alpha}\right) \geq 1 - \alpha$.

(ii) (Non-zero probability of extreme values)

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

for some non-decreasing and positive functions $\phi(\cdot)$ and $\psi(\cdot)$.

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

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

Assumption 3 (i) reflects the idea that more reliable matches are provided by the pairs of identifiers whose values are infrequent. In other words, if in both public and private datasets 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 datasets contain very detailed individual information such as a combination of the full name and the address, most attribute values will be unique.

Assumption (ii) requires that there are sufficiently many observations with infrequent attribute values. This fact can actually be established empirically in each of the observed subsets and, thus, this assumption is testable. The same is true for the Assumption (iv) as continuity of the of the observed marginal distributions can be directly observed and tested.

Assumption 3 (iii) 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" dataset 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 datasets 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 a 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 individual may be the 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 dataset.

In the remainder of the paper we suppose that Assumptions 1-3 hold.

3 Identification of the treatment effect from the combined data

Provided that the variables are not contained in the same dataset, identification of the treatment effect parameter becomes impossible without having some approximation to the distribution of the data in the "master" sample. The only way to link the observations in two datasets 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" dataset correctly if we link the data of one concrete 'John Smith' in the two datasets. This means that the data combination is an intrinsically small 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 label them as a "non-match" ($M_{ij} = 0$) if we think that the observations are unlikely to belong to the same individual or are simply uncertain. Although we can potentially consider many nonlinear data combination rules, in this paper we focus at the set of parametric data combination rules that are generated by our Assumption 3 (i). In particular for some pre-specified $\bar{\alpha} \in (0, 1)$ we consider a data combination rule

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

generated by a Cauchy sequence α_N such that $0 < \alpha_N < \bar{\alpha}$ and $\lim\limits_{N\to\infty} \alpha_N = 0$. The goal of this sequence is to construct the set of thresholds that in the limit would isolate all of the infrequent observations. For those observations, the probability of the correct match will be approaching one as the probability of observing two identifiers taking very close values for two different individuals will 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 will be of a larger order of magnitude (proportional to the probability of observing the attribute value). Thus, an appropriately scaled sequence of thresholds will be able to single out correct matches.

Let m_{ij} be the indicator of the event that the observation i from the public dataset and the observation j from the private dataset 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" dataset by taking the pairs of all observations from the public and the private datasets which 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 datasets that were constructed with higher threshold values.

The reconstructed master dataset 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 $f^N_{\alpha_N}(y_i\,|d_j, x_i, z^x_i, z^d_j)$ 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:

$$\begin{aligned} f^N_{\alpha_N}(y_j|d_j, x_i, z^x_i) = {} & f_{Y|D,X}(y_j|d_j, x_i) Pr(m_{ij} = 1 \mid \mathcal{D}^N(z^x_i, z^d_j) = 1) \\ & + f_{Y|X,Z^x}(y_j|x_i, z^x_i) Pr(m_{ij} = 0 \mid \mathcal{D}^N(z^x_i, z^d_j) = 1), \end{aligned}$$

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. So 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 $f^N_{\alpha_N}(\cdot|d_j, x_i, z^x_i)f(x_i, z^x_i)$.

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

$$\begin{aligned} P^N_{\alpha_N}(x) &= P(x)Pr(m_{ij} = 1 \mid x_i = x, \mathcal{D}^N(z^x_i, z^d_j) = 1) \\ &\quad + \bar{P}Pr(m_{ij} = 0 \mid x_i = x, \mathcal{D}^N(z^x_i, z^d_j) = 1). \end{aligned}$$

We can extend our data combination method by choosing sequences α_N depending on the value of x. Then the value of $Pr(m_{ij} = 0 \mid x_i = x, \mathcal{D}^N(z^x_i, z^d_j) = 1)$ even in the limit will depend on x. We allow for such situations. In fact, later in the paper 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 situation 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 $C_{0,x}$:*

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

We call $\mathcal{P}^N$ the 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 (2.1) for all sequences $\alpha_{N,x}$ in $C_{0,x}$ using the corresponding to $\alpha_{N,x}$

propensity score $P^N_{\alpha_{N,x}}(\cdot)$:

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

We call $\mathcal{T}^N$ the 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 chosen for the decision rule that are bounded and converge to zero. We note that 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 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 *(i) We call $\mathcal{P}^\infty$ the identified set for propensity score under the threshold decision rule if for the set of graphs of functions in $\mathcal{P}^\infty$ denoted as $G(\mathcal{P}^\infty)$ and the set of graphs of functions in $\mathcal{P}^N$ denoted as $G(\mathcal{P}^N)$ if*

$$\lim_{N\to\infty} d_H\left(G(\mathcal{P}^\infty), G(\mathcal{P}^N)\right) = 0,$$

where $d_H(\cdot,\cdot)$ stands for the Hausdorff distance.

(ii) Similarly, we call $\mathcal{T}^\infty$ the identified set for the average treatment effect under the decision threshold rules if

$$\lim_{N\to\infty} d_H\left(\mathcal{T}^\infty, \mathcal{T}^N\right) = 0.$$

(iii) 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 the decision threshold rules.

(iv) 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 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 (ii) and (iii).

LEMMA 1 *Under Assumption 3 the propensity score is point-identified from the observations with infrequent attribute values:*

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

Also, the average treatment effect is point-identified from the observations with infrequent attribute values:

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

This lemma states that if we are able to correctly reconstruct the "master" dataset 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 this results. First, we need Assumption 3 (iii) to establish redundancy of identifiers for matches constructed for observations with infrequent values of those identifiers. Second, we need Assumption 3 (ii) to guarantee that there is a non-zero 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 the mismatch, conditional on the observation being in the reconstructed master sample, approaches zero. We do so by an appropriate inversion of the probability of misidentification of the 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 $\{\alpha_{N,x}\} \in C_{0,x}$ satisfes*

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

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

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

and

$$\begin{aligned} T^N_{\alpha_{N,x}} = E^N_{\alpha_{N,x}}\left[\frac{D_j Y_j}{P^N_{\alpha_{N,x}}(X_i)} - \frac{(1-D_j)Y_j}{1-P^N_{\alpha_{N,x}}(X_i)}\right] \to t_{ATE} + \\ + E\left[\left(E[Y_1] - E[Y|X,\, D=1]\,\bar{P}\right)\frac{\gamma(X)}{(1-\gamma(X))\,P(X)+\gamma(X)\,\bar{P}}\right] - \\ - E\left[\left(E[Y_0] - E[Y|X,\, D=0]\,(1-\bar{P})\right)\frac{\gamma(X)}{1-(1-\gamma(X))\,P(X)-\gamma(X)\,\bar{P}}\right]. \end{aligned} \tag{3.3}$$

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

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).*** *There exists a sequence $\{\alpha_{N,x}\} \in C_{0,x}$ for which $\lim\limits_{N\to} Pr\left(m_{ij} = 0 \mid \mathcal{D}^N(Z_i^x, Z_j^d) = 1\right) = 0$ for $x \in \mathcal{X}$.*

In other words, for this sequence:

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

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.

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 datasets 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 which 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 dataset then would constitute an evident threat to individual's privacy. However, even if the reconstructed master dataset 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 partial 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 his or 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 $Pr\left(m_{ij} = 0 \mid x_i = x, \mathcal{D}^N(Z_i^y, Z_j^d) = 1\right) \to \gamma(x)$, then the public can determine that in the limit the released propensity score for an individual with characteristics x has the form $(1-\gamma(x))\,P(x) + \gamma(x)\,\bar{P}$. If, in addition, the data curator releases the value of $Pr\left(m_{ij} = 0 \mid x_i = x, \mathcal{D}^N(Z_i^y, Z_j^d) = 1\right)$ or the value of $\gamma(x)$, then the public can pin down the true propensity score $P(x)$[2] 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*

[2] Note that the value $\bar{P}$ is known from the public dataset.

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 dataset 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 dataset.

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 individual, 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 a release of the true propensity score is not compatible with a low disclosure risk.

THEOREM 2 *Suppose that*

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

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 datasets 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 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 the ways of avoiding the situations where potentially very sensitive information may be learned regarding 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 datasets of any size. As discussed in Proposition 1, a fixed proportion of the incorrect matches, leads to the the calculated propensity score to be biased towards the proportion of treated individuals in the population and also causes bias in the average treatment effect.

THEOREM 3 *Suppose the value of $\bar{P}$ is publicly available, and $\bar{P} < 1 - \delta$.*

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

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

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

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

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 $Pr\big(m_{ij} = 0 \mid \mathcal{D}^N(Z_i^x, Z_j^d) = 1\big)$ or$\gamma(x)$.

In this case, the behavior of the released propensity score and the treatment effect is as described in (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 towards $\bar{P}$ has to be large enough.

Remark 1 *Continue to assume that $\bar{P} < 1 - \delta$.*

Note that if the released propensity score for an individual with x is strictly less than $\bar{P}$, then the public will be able to conclude that the true propensity score for this individual is strictly less than $\bar{P}$.

If the released propensity score for an individual with x is strictly greater than $\bar{P}$, then the public will be able to conclude that the true propensity score for this individual is strictly greater than $\bar{P}$ 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 $\bar{P}$, then the public is unable to make any non-trivial 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 $Pr\left(m_{ij} = 0 \mid \mathcal{D}^N(Z_i^x, Z_j^d) = 1\right)$ 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 $\bar{P}$ is publicly available, and $\bar{P} < 1-\delta$.*

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

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

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

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

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 $\bar{\gamma}$ but does not provide the values of $Pr\left(m_{ij} = 0 \mid \mathcal{D}^N(Z_i^x, Z_j^d) = 1\right)$ or $\gamma(x)$.

In this case, the behavior of the released propensity score and the treatment effect is as described in (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 point identification of both the true propensity score and true average treatment effect. This means that point identification of the econometric model from the combined dataset is incompatible with the security of individual information. If the publicly observed policy is based on the combination of the non-public 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 $\bar{P} > 1 - \delta$. 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 $\bar{P}$ is publicly available, and $\bar{P} > 1 - \delta$.*

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 non-secure. In other words, there will exist sufficiently many individuals for whom we can learn their high propensity scores. To protect their privacy, *no* propensity scores whatsoever should be released.

5 Do doctors make people happier?

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

Even though in the main body of this paper 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 when only two split datasets are available, in this section we want to illustrate a reliable empirical procedure one could implement in practice.

The data that we use come from a combination of data from Yelp.com and the data from the property tax bills in Durham county, NC. The question that we want to answer can be informally formulated as: Does a visit to a doctor change the general attitudes of individuals towards rating businesses on Yelp.com?

Our analysis thus models the situation where there is an anonymous dataset of users on Yelp.com for whom we have information regarding their ratings of businesses as well as information on them specifically rating health-related businesses. We treat this dataset as if it were available only to a data curator. On the other hand, using public sources we can also collect demographic information for all individuals that potentially include the Yelp users.

The property tax data was extracted from the Durham county government website, tax administration record search (see http://www.ustaxdata.com/nc/durham/). Property tax bills are stored by the parcel numbers. Going over the list of all parcel numbers we collected data from property tax bills for years 2009/2010. In total we collected 104068 tax bills for year 2010 and 103445 tax bills for year 2009. Each bill contains information on taxable value of property, first and last names of the taxpayer and the location of the property (house number, street, and zip code). Then we merged the data between the years 2009 and 2010 by the parcel number and the property owner, removing the properties that change the owner from year to year. Property tax data allows us to assemble information on the name and location of individuals as well as an indicator of their wealth (as indicated by the taxable value of the property). Table 1 summarizes the distribution of taxable property values in the constructed dataset of tax bills.

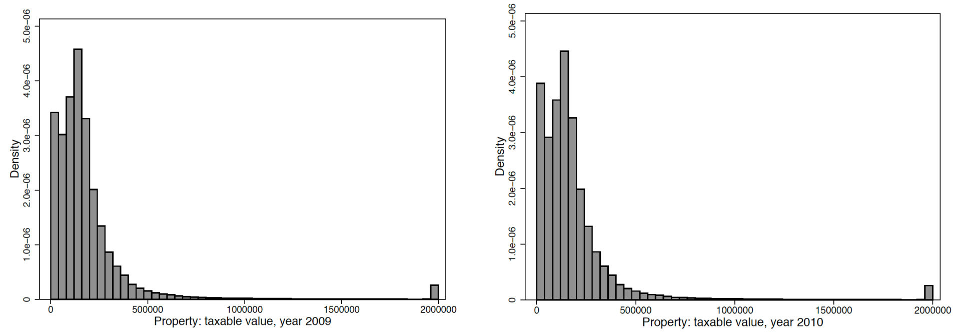
We demonstrate the distribution of taxable values of the properties on Figure 1. As we collect the entire dataset of the property tax bills, some of them are actually commercial properties. These are the outliers seen on the histogram.

We also collected the information from Yelp.com that covers all individuals who

Table 1: Summary statistics from property tax bills in Durham County, NC.

Variable	Obs	Mean	Std. Dev.	25%	50%	75%
year 2009-2010						
Property: taxable value	207513	261611.9	1723970	78375	140980	213373
year 2010						
Property: taxable value	104068	263216.1	1734340	78823.5	141490.5	214169.5

Figure 1: Empirical distribution of taxable property values in Durham county, NC



(a) (b)

Table 2: Summary statistics from Yelp.com for ratings of health services in Durham, NC

Variable	Obs	Mean	Std. Dev.	Min	Max
Rating	72	4.06	1.34	1	5
Category: fitness	72	0.17	0.38	0	1
Category: dentist	72	0.29	0.46	0	1
Category: physician	72	0.36	0.48	0	1
Category: hospital	72	0.04	0.20	0	1
Category: optometris	72	0.10	0.30	0	1
Category: urgent care	72	0.06	0.23	0	1
Appointment?	72	0.51	0.50	0	1
Kids friendly?	72	0.08	0.28	0	1

ever rated businesses in Durham, NC. Out of that information we found the most reliable business category on Yelp.com to be the ratings of restaurants where most user reviews are available among all the categories. The entire collected dataset was used mainly to assure that we find high-quality matches between Yelp users and the list of property owners in Durham county.

Then we focused specifically on the data of ratings of health care businesses. We used the indicator that a particular individual rated a health care business as an indication that this individual actually visited that business. We were able to extract reliable information from 59 yelp.com users who rated health care services in Durham. We focused on all of the publicly released ratings: yelp.com has a practice of filtering particular ratings that are believed to be unreliable. Even though, we collected the information from such ratings, we chose not to use it in our empirical analysis. The dataset from Yelp.com for the Durham, NC produces a total of 72 reviews for Durham health care businesses. We show the summary statistics for the constructed variables in Table 2.

As mentioned above, for data combination purposes we used the entire set of Yelp ratings in Durham, NC with a particular focus on the restaurant ratings as the largest

Table 3: Features of edit distance-based matches

# of matches	Freq.	Percent	# of yelp users
1 in yelp − > 1 in tax data	66	1.54	66
1− >2	92	2.19	46
2 − > 1	2	2.19	2
1 − > 3	72	1.68	24
1− > 4	36	0.84	9
1 − > 5	65	1.51	13
1 − > 6	114	2.65	19
1 − > 7	56	1.3	8
1 − > 8	88	2.05	11
1 − > 9	81	1.89	9
1− > 10 or more	3,623	84.35	97
Total	4,295	100	304

and most reliable rating category. We used a simple record linkage technique to combine two datasets. We constructed the individual identifiers using the rank cutoff rule combining the edit distance using the first and last name in the property tax dataset and the user name on Yelp.com, and the sum of ranks indicating that the taxpayer in the tax data is located in the same zip code as the rated business. Given this simple matching rule, we identified 397 individuals in the tax record data as positive matches. Fourteen people are uniquely identifiable in both databases. Table 3 shows the distribution of obtained matches. One-to-one matches correspond to the edit distance zero, one-to-two matches correspond to the edit distance one, etc.

The matched observations characterize the constructed merged dataset of Yelp reviews and the property tax bills. We were able to find the reviewers in Yelp and the property owners in the property tax bills for whom the the combined edit distance and the Euclidean distance between the numeric indicators (zip code and location of most frequent reviews) is equal to zero. We call this dataset the set of “one-to-one” matches. Based on reviewer first name we evaluate sex of reviewer and construct dummy variable indicating that the name of the individual in the taxpayer data has a name that belongs to the list of top 500 female names in the US from the Census data (as a proxi that the corresponding taxpayer is a female). We also constructed the indexes for other demographic indicators, but they did not improved the fit of our estimated rating model and we excluded them from our analysis.

We answer the posed question of the effect of a visit to a doctor on individual ratings of businesses on Yelp.com.

We focus on the analysis of the propensity score and the average treatment effect where the treatment status corresponds to an individual's visit to a doctor and the outcome corresponds to individual ratings. We find first, that at average after attending doctor a person has slightly (0.06 units or 0.05 SD) higher average rating than before (see column 1 of Table 4).

To visualize the heterogeneity of observed effects across individuals we calculated differences in mean of her/his rating before and after visiting doctor. We find a significant difference in before and after ratings: means in lower and upper quartile significantly different from zero (at 10% significant level). Those who belong to the upper quartile report at average increase in rating by 1.02 units whereas those who belong to the lower quartile reported at average decrease in rating by 1.14 units (see Table 5).

One of the caveats of OLS estimates is selection bias due to selection of those who are treated. For instance, males could visit doctors less frequently than females. As for the income, we can envision various scenarios. For example, one scenario would be that people with higher incomes may use the services of Yelp-listed businesses more frequently and, thus, in this case the effect of the income is positive. Another scenario would be that people with higher incomes are busier and, thus, are less likely to leave reviews and, therefore, the effect of the income is negative. This selection may result in a bias of the estimated ATE. The absence of the demographic characteristics, such as income or sex may result in inability to control for this section and so inability to get consistent estimates of treatment effects. Columns 2, 3 and 4 of 4 illustrate this point. To control for possible selection bias, we use a two-stage estimator for the subsample of data for which we have data on income.

Column 4 of Table 4 shows the evidence of selection showing a correlation between the income and the participation in the sample, as well as a correlation between the sex and the participation. Column 2 and column 3 exhibit the OLS and the two-stage estimates of the average treatment effect. After controlling for selection, the effect of visiting a doctor is much higher. According to the two-stage estimates, attending doctor increases the estimated rating coefficient to 0.353 units compared to 0.033 units obtained using OLS procedure. The large dot in the first graph in Figure 3 shows the pair $(0.353, -0.044)$ of estimated coefficients for the treatment effect and the income, respectively. The large dot in the first graph in Figure 3 shows the pair $(0.353, -0.044)$ of estimated coefficients for the treatment effect and the income, respectively, from columns 3 and 4 in Table 4. The large dot in the second graph in Figure 3 shows the pair $(0.353, -0.164)$ of estimated coefficients for the treatment effect and the male binary variable, respectively. The large dot in the third graph in Figure 3 shows the pair $(-0.164, -0.044)$ of estimated coefficients for the male binary variable and the income variable, respectively.

Finally, we analyze how the estimates of our parameters would change if we enforce a bound on the risk of partial disclosure and consider the bounds of 0.5 and 2/3 – that is, $Pr\left(m_{ij} = 0 \,|\, \mathcal{D}^N(Z_i^x, Z_j^d) = 1\right) \geq \bar{\gamma}$, where $\bar{\gamma} = 0.5$ (attaining k-anonymity with $k = 2$) or $\bar{\gamma} = 2/3$ (attaining k-anonymity with $k = 3$).

We do this by providing the following experiment. In order to attain 2-anonymity we erase some letters in the individuals' surnames to guarantee that for every yelp user from the one-to-one match database there are at least two good matches in the tax data. Then, ideally we would like to simulate all possible combined datasets but the number of these datasets is of exponential complexity, namely, of the rate 2^n. Instead of considering all possible combined datasets we randomly simulate only a 1000 of such datasets. For each simulated combined dataset we conduct the two-stage estimation. Thus, we end up with a 1000 of different estimated coefficients and the propensity scores. The smaller contour sets in the graphs in Figure 3 are the convex hulls of the obtained estimates. Namely, the smaller contour set in the first graph in Figure 3 is the convex hull of the 1000 pairs of estimated coefficients for the treatment effect and the income, respectively. Similarly, the smaller contour set in the second graph in Figure 3 is the convex hull of the 1000 pairs of estimated coefficients for the treatment effect and the male binary variable, respectively. These two graphs give us an idea about the range of the estimated treatment effect. Finally, in the third graph in Figure 3 the contour set that is bounded by 0 from the right is the convex hull of the 1000 pairs of estimated coefficients for the male binary variable and the income variable, respectively. Of course, the estimation in 1000 simulated datasets provide us with a 1000 of predicted propensity scores. In Figure 4 we consider only males and for each level of the log(income) we draw the range of the predicted propensity scores. An important question is which out of 1000 estimates of the propensity score, the treatment effect, the income coefficient and the male coefficient is the data curator going to release? A natural way, which is in accord with our theoretical analysis, is to average the 1000 propensity scores for each individual and obtain the average propensity score function. Then the data curator can choose and release the predicted propensity score function whose graph is the closest one to the average propensity score function as well as the corresponding to this propensity score function estimates of the treatment effect, the income effect and the male variable effect. As we can see in Figure 4, for males with log(income) around 12 all the predicted propensity scores are small and, thus, non-disclosure is guaranteed for such individuals. For the individuals with low or high log(income) some predicted propensity scores are close to 1. However, after averaging 1000 predicted values it may turn out that for these individuals this average is below the sensitivity threshold and, thus, for these individuals non-disclosure from the released propensity score is guaranteed too. If after averaging the released propensity score is above the sensitivity threshold for some individuals, then there are two things the data curator can do. First, the data curator can consider more than 1000 simulated combined datasets. If the partial disclosure is still not guaranteed then the data curator should increase the guarantee level by, for instance, attaining 3-anonymity.

In order to attain 3-anonymity we erase more letters in the individuals' surnames to achieve that for every yelp user from the one-to-one match database there are at least three good matches in the tax data. Then we conduct the analysis similar to the one for 2-anonymity. The estimation results for 3-anonymity are illustrated by the larger contour sets in the graphs in Figure 3. Figure 5 shows the range of estimates of the propensity score function for each log(income) level for males. The rest of the discussion is analogous to the case of 2-anonymity.

Table 4: Estimated treatment effects

	(1)	(2)	(3)	(4)
	OLS	OLS	Two-stage estimation	
	Rating	Rating	Rating	I(After visit)
I(After visit)	0.06	0.033	0.353	
	[0.015]***	[0.054]	[0.161]**	
log(property value)				-0.044
				[0.009]***
I(male)				-0.164
				[0.011]***
Observations	20723	2605	2605	2605

Column 1,2,4: SE in brackets; column 3: bootsrapped SE in brackets
* significant at 10%; ** significant at 5%; *** significant at 1%

Table 5: Quantile treatment effects

Variable	Obs	Mean	SD	Min	Max
Lower quartile					
Difference	57	-1.144	0.795	-4	-0.5
Upper quartile					
Difference	55	1.026	1.035	0.19	4

Mean difference test: t-stat =1.662

Figure 2: Distributions of Yelp.com ratings before and after a doctor visit

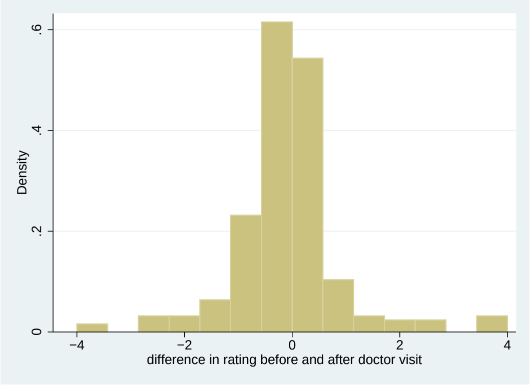


Figure 3: Sets of estimates from 1000 datasets combined from the given split datasets. Contour sets are for the cases of 2-anonymity and 3-anonymity.

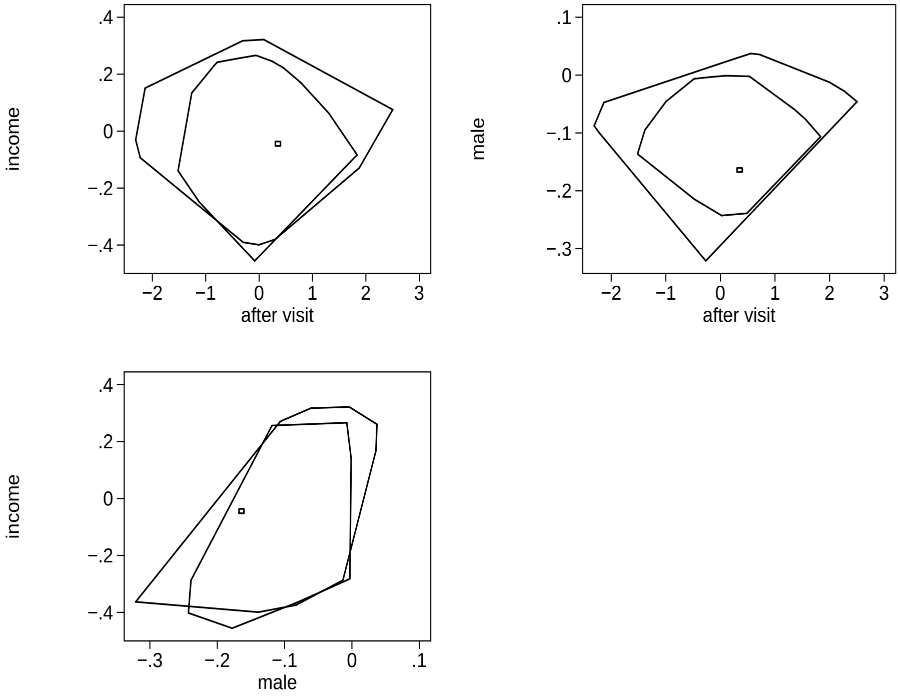


Figure 4: Range of predicted propensity scores for male in the case of 2-anonymity.

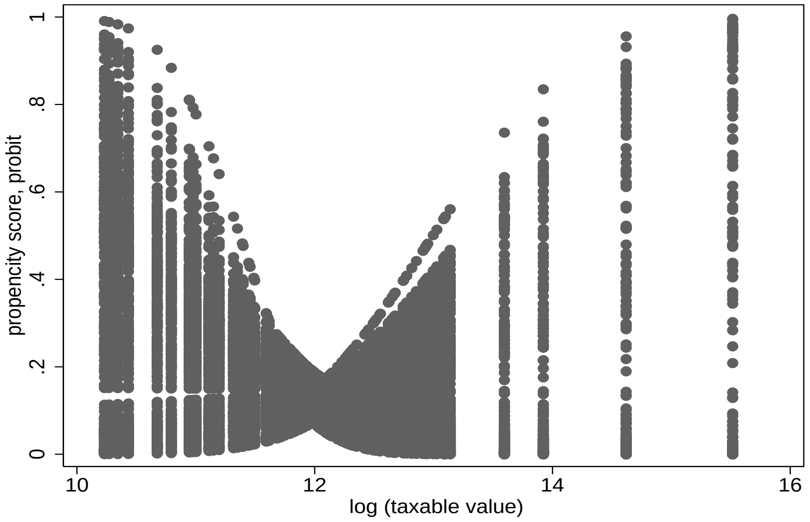
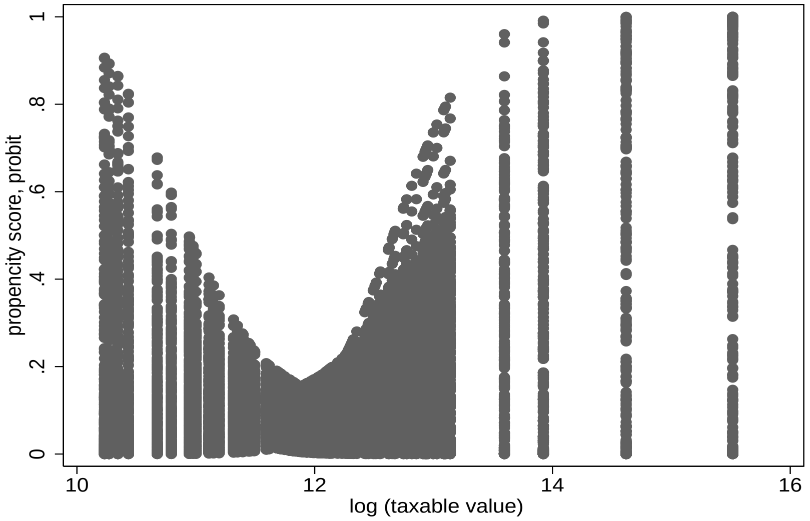


Figure 5: Range of predicted propensity scores for male in the case of 3-anonymity.



References

Abowd, J., and L. Vilhuber (2008): "How Protective Are Synthetic Data?," in *Privacy in Statistical Databases*, pp. 239–246. Springer.

Abowd, J., and S. Woodcock (2001): "Disclosure limitation in longitudinal linked data," *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, pp. 215–277.

Acquisti, A. (2004): "Privacy and security of personal information," *Economics of Information Security*, pp. 179–186.

Acquisti, A., A. Friedman, and R. Telang (2006): "Is there a cost to privacy breaches? An event study," in *Fifth Workshop on the Economics of Information Security*. Citeseer.

Acquisti, A., and J. Grossklags (2008): "What can behavioral economics teach us about privacy," *Digital Privacy: Theory, Technologies, and Practices*, pp. 363–377.

Acquisti, A., and H. Varian (2005): "Conditioning prices on purchase history," *Marketing Science*, pp. 367–381.

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*, 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(2)), 1468–1488.

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," *Secure Data Management in Decentralized Systems. Springer-Verlag.*

Duncan, G., S. Fienberg, R. Krishnan, R. Padman, and S. Roehrig (2001): "Disclosure limitation methods and information loss for tabular data," *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pp. 135–166.

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 Sizeand Random Sample Query Control," .

Duncan, G., and R. Pearson (1991): "Enhancing access to microdata while protecting confidentiality: Prospects for the future," *Statistical Science*, pp. 219–232.

Dwork, C. (2006): "Differential privacy," *Automata, languages and programming*, pp. 1–12.

Dwork, C., and K. Nissim (2004): "Privacy-preserving datamining on vertically partitioned databases," in *Advances in Cryptology–CRYPTO 2004*, pp. 134–138. Springer.

Fienberg, S. (1994): "Conflicts between the needs for access to statistical information and demands for confidentiality," *Journal of Official Statistics*, 10, 115–115.

——— (2001): "Statistical perspectives on confidentiality and data access in public health," *Statistics in medicine*, 20(9-10), 1347–1356.

Goldfarb, A., and C. Tucker (2010): "Online display advertising: Targeting and obtrusiveness," *Marketing Science*.

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*, pp. 71–80. ACM.

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), 445–459.

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), 419–440.

Komarova, T., D. Nekipelov, and E. Yakovlev (2011): "Identification, data combination and the risk of disclosure," *CeMMAP working papers*.

Korolova, A. (2010): "Privacy violations using microtargeted ads: A case study," in *IEEE International Workshop on Privacy Aspects of Data Mining (PADM'2010)*, pp. 474–482.

Lambert, D. (1993): "Measures of disclosure risk and harm," *Journal of Official Statistics*, 9, 313–313.

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*, pp. 49–60. ACM.

——— (2006): "Mondrian multidimensional k-anonymity," in *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference*, pp. 25–25. IEEE.

Magnac, T., and E. Maurin (2008): "Partial identification in monotone binary models: discrete regressors and interval data," *Review of Economic Studies*, 75(3), 835–864.

Manski, C. (2003): *Partial identification of probability distributions*. Springer Verlag.

Miller, A., and C. Tucker (2009): "Privacy protection and technology diffusion: The case of electronic medical records," *Management Science*, 55(7), 1077–1093.

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 *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pp. 111–125. IEEE.

RIDDER, G., AND R. MOFFITT (2007): "The econometrics of data combination," *Handbook of Econometrics*, 6, 5469–5547.

SAMARATI, P., AND L. SWEENEY (1998): "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," Discussion paper, Citeseer.

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

——— (2002b): "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, 10(5), 557–570.

TAYLOR, C. (2004): "Consumer privacy and the market for customer information," *RAND Journal of Economics*, pp. 631–650.

VARIAN, H. (2009): "Economic aspects of personal privacy," *Internet Policy and Economics*, pp. 101–109.

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), 514–531.

WRIGHT, G. (2010): "Probabilistic Record Linkage in SAS®," *Keiser Permanente, Oakland, CA.*