Identification, Data Combination and the Risk of Disclosure[1]

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

This version: December 21, 2016

ABSTRACT

It is commonplace that the data needed for econometric inference are not contained in a single source. In this paper we analyze the problem of parametric inference from combined individual-level data when data combination is based on personal and demographic identifiers such as name, age, or address. Our main question is the *identification* of the econometric model based on the combined data when the data do not contain exact individual identifiers and no parametric assumptions are imposed on the joint distribution of information that is common across the combined dataset. We demonstrate the conditions on the observable marginal distributions of data in individual datasets that can and cannot guarantee identification of the parameters of interest. We also note that the data combination procedure is essential in the semiparametric setting such as ours. Provided that the (non-parametric) data combination procedure can only be defined in finite samples, we introduce a new notion of identification based on the concept of limits of statistical experiments. Our results apply to the setting where the individual data used for inferences are sensitive and their combination may lead to a substantial increase in the data sensitivity or lead to a de-anonymization of the previously anonymized information. We demonstrate that the point identification of an econometric model from combined data is incompatible with restrictions on the risk of individual disclosure. If the data combination procedure guarantees a bound on the risk of individual disclosure, then the information available from the combined dataset allows one to identify the parameter of interest only partially, and the size of the identification region is inversely related to the upper bound guarantee for the disclosure risk. This result is new in the context of data combination as we notice that the quality of links that need to be used in the combined data to assure point identification may be much higher than the average link quality in the entire dataset, and thus point inference requires the use of the most sensitive subset of the data. Our results provide important insights into the ongoing discourse on the empirical analysis of merged administrative records as well as discussions on the disclosive nature of policies implemented by the data-driven companies (such as Internet services companies and medical companies using individual patient records for policy decisions).

JEL Classification: C35, C14, C25, C13.

Keywords: Data protection, model identification, data combination.

[1] *First version: December 2011.* Support from the NSF and STICERD is gratefully acknowledged. We appreciate helpful comments from Philip Haile, Michael Jansson, Charles Manski, Aureo de Paula, Martin Pesendorfer, James Powell, Catherine Tucker and Elie Tamer. We appreciate feedback from seminar participants at various universities.

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

[3] Department of Economics, University of Virginia, e-mail: denis@virginia.edu.

[4] New Economic School, Skolkovo, Moscow, Russia, e-mail: eyakovlev@nes.ru.

1 Introduction

Often, data combination is a vital step in a comprehensive analysis of industrial and government data and resulting policy decisions. Typical industrial data are contained in large, well-indexed databases and linking multiple datasets essentially reduces to finding the pairs of unique matching identifiers in disjoint databases. Examples of such databases include the supermarket inventory and scanner data that can be linked by the product UPCs, patient record and billing data that can be matched by name and social security number. Non-matches can occur, e.g., due to recording errors. Given that most industrial databases have a homogenous structure, prediction algorithms can be "trained" on a dataset of manually resolved linkage errors and then those algorithms can further be used for error control. These algorithms stem from the long-existing literature in Econometrics and Statistics on validation samples. Such procedures are on the list of routine daily tasks for database management companies and are applied in a variety of settings, from medical to tax and employment databases.[1]

A distinctive feature of data used in economic research is that the majority of utilized datasets are unique and, thus, any standardization of the data combination procedure may be problematic. Moreover, many distinct datasets that may need to be combined do not contain comprehensive unique identifiers either due to variation in data collection policies or because of the disclosure and privacy considerations. As a result, data combination tasks rarely reduce to a simple merger on unique identifiers with a subsequent error control. This means that in the combination of economic datasets, one may need to use not only the label-type information (such as the social security number, patient id or user name) but also some variables that have an economic and behavioral content and may be used in estimated models. In this case the error of data combination becomes heteroskedastic with an unknown distribution and does not satisfy the "mismatch-at-random" assumption that would otherwise allow one to mechanically correct the obtained estimates by incorporating a constant probability of an incorrect match.[2] In addition, economic datasets are usually more sensitive than typical industrial data and data curators may intentionally remove potentially identifying information from the data that further complicates combination of different datasets.

In this paper we introduce a novel framework for the parameter identifiability analysis from linked data when individual datasets used for combination do not contain unique individual identifiers. Our framework is suited to situations when only *partial* information regarding the quality of the links between the observations of separate datasets (e.g. upper and lower bounds on these probabilities) is available, and thus, it allows us to avoid making parametric assumptions regarding the joint distribution of combined variables or the joint distribution of additional variables utilized in a data combination procedure. This contrasts

[1] See, e.g. Wright [2010] and Bradley et al. [2010] among others.

[2] See, for instance, Lahiri and Larsen [2005]

many existing approaches that rely either on such parametric assumptions or assumptions about a known distribution of the data combination errors. This paper is an attempt to build a theoretical framework of how to think about parameter identification from combined data and to conform it with the tradition existing in the econometric literature of approaching the issue of identification from the population perspective.

Section 2 describes the problem of econometric inference and characterizes the structure of the data generating process.

In Section 3 we depict a class of data combination rules used in this paper. The data combination procedures suggested in this paper are based on *infrequent observations* of some numeric or string variables that are either available directly from the data or need to be constructed by the data curator. We formalize all the conditions that this procedure has to satisfy in order to give a meaningfully combined dataset. We prove that the accuracy of this procedure can be controlled and can vary from the "worst" (all the matches are incorrect) to the "best" (all the matches are correct) as the sizes of split data sets increase. We establish how exactly the control of its accuracy can be executed by a data curator.

Our framework naturally applies to the analysis of situations where the identifying information is intentionally removed from the data by the data curators to reduce the "sensitivity" of the data. In this case, an instance of a successful combination of two observations from two disjoined datasets means that the variables contain enough information to attribute these two observations to the same individual. This implies that the corresponding individual information can be de-anonymized, i.e. the *individual disclosure* can occur. We demonstrate the implications of the suggested data combination rules for individual identity disclosure. We introduce the notion of a bound on disclosure risk and show that there exist data combination rules that honor this bound.

In Section 4 we analyze the identifiability of the parameter of interest from combined data under restrictions on the information about the quality of the data combination rule that is released by the data curator to secondary users (researchers). Our approach to the identification analysis is novel as we notice that the data combination procedure in non-parametric settings can only be defined and implemented in the finite sample and not in the population. As a result, the identification analysis has to rely on the property of limits of sequences of data combination rules (as opposed to the property of the population distribution as in the standard literature on identification). This is a crucial aspect in our identification method as we provide a new approach to analyzing model identification from combined datasets as a limiting property in the sequence of statistical experiments.

Namely, we introduce the notion of the pseudo-identified set of model parameters from combined data through a limit of the set of parameters inferred from the combined data as the sizes of both datasets approach infinity. These sets and their limiting behavior depend on several factors: first, they depend on the properties of the data combination procedure; second, they depend on what kind of information about this procedure is provided

to the researcher by the data curator; and, finally, they could depend on the optimization criterion employed by researchers. We also study the tradeoff between disclosure limitation (defined by the probability that an individual disclosure can occur) and the quality of identification of the parameters of interest. To our knowledge, our paper is the first one to study such a tradeoff. This trade-off between the identifiability of the model and limitations on individual disclosure implies that whenever a non-zero disclosure restriction is imposed, the parameter in the model of interest based on the dataset combined from two separate datasets is not point identified. The analysis of pseudo-identified sets tells us what estimates e.g. a consumer behavior model can deliver under the constraints on the identity disclosure. We note that the goal of our work is not to demonstrate the vulnerability of online personal data but to provide a real example of the tradeoff between privacy and identification.

The importance of the risk of potential disclosure of confidential information is hard to overstate. With advances in data storage and collection technologies, issues and concerns regarding data security now generate front-page headlines. Private businesses and government entities are collecting and storing increasing amounts of confidential personal data. This data collection is accompanied by an unprecedented increase in publicly available (or searchable) individual information that comes from search traffic, social networks and personal online file depositories (such as photo collections), amongst other sources. One of the main messages of sections 2-4 is that if one of the data curator's objectives is to provide some privacy guarantees and prevent disclosure when conducting the task of combing the data, then the issues of model identification/estimation and the risk of disclosure should be analyzed jointly.

Sections 2-4 of the paper consider a scenario in which a data curator conducts the data combination procedure and the researcher is given a single combined dataset (with auxiliary variables that helped combine the data removed). This combined dataset is of course not guaranteed to contain all correct matches. Moreover, if the combined dataset is randomly selected from all possible constructed combined datasets with the data combination rule that honors the bound on the disclosure risk, there is a positive probability that all matches in this dataset will be incorrect. This scenario is likely to occur when a combined dataset is released into a public domain and thus the researcher does not bear the burden of assuring that an appropriate bound on the risk of disclosure has been imposed.

Section 5 contains an empirical application, where we illustrate a common situation where low resolution identifiers are removed from the dataset to protect privacy of individuals which then inhibits the linkage of this dataset with other data which can lead biased estimates in the models that do not use those additional linked variables. Our application uses the data from the Russian Longitudinal Monitoring Survey (RLMS) which is a comprehensive longitudinal survey of households in Russia. The survey is designed to be representative on the country level and the data are collected in over 50 geographical region.

In each region the survey households are typically clustered within small neighborhoods. The neighborhood identifiers along with demographic data turned out to be sufficient to single out individual households and de-anonymize them by linking the records with address databases. In light of this finding, the neighborhood identifiers were removed from the RLMS data distribution after year 2009.

In our empirical analysis we demonstrate that such an approach to privacy protection inhibits the inference of granular household-level decision models. Our main Economic question is the impact of religious affiliation of the household on the decision to allow the children in the household to complete schooling. We are also interested to find out whether in this decision females are withdrawn from schooling on average earlier than males. We note that in the absense of neighborhood identifiers, it will not be able to distinguish the group effects whithin local religious communities from the individual decision making within households.

In our dataset we do have access to the neighborhood identifiers that were subsequently suppressed. Using the neighborhood identifiers we can link the records in the RLMS with the religious census data collected by Rosstat (the Russian equivalent of the US Census Bureau). This allows us to estimate both the model that takes the group effects into account and the model that does not (i.e. the model that is feasible with the current RLMS data distribution). We find significant difference in the estimates obtained in the two models and then use our approach to construct the sets of parameters in the current data distribution that take into account the fact that the data was de-anonymized.

To relate this paper to other privacy frameworks, we want to note that we focus on the risk of individual disclosure as it describes the possibility of recovering the true identity of individuals in the anonymized dataset with sensitive individual information. However, even if the combined dataset is not publicly released, the estimated model may itself be disclosive in the sense that consumers' confidential information may become discoverable from the inference results based on the combined data. This situation may arise when there are no common identifiers in the combined data and only particular individuals may qualify to be included in the combined dataset. If the dataset is sufficiently small, a parametric model may give an accurate description of the individuals included in the dataset. We discuss this issue in more detail in Komarova et al. [2015] where we introduce the notion of a partial disclosure. In this paper we deal only with the identity disclosure.

The setup of this paper can be applied to situations when there are several independent data curators having access to separate datasets. Private firms and large government agencies collect large socio-economic datasets. The Internal Revenue Service, Social Security Administration and the US Census Bureau collect large comprehensive datasets that have large or complete overlaps over individuals whose data has been collected. Each of these agencies operate as independent data curators meaning that each of them has full control over their data, full exclusion rights over access to these data. Most existing data cura-

tors operate based on the vault storage model where the data is stored locally in a secure location and raw disaggregated data cannot be taken outside of the vault. Within their data management programs, each such a data owner allows researchers to access the data vault upon passing some clearance procedure. With this data analysis model there could be many researchers who can access many of such data vaults. However, provided that the raw data cannot be removed from the vault, neither of these researchers can combine individual data from two or more such vaults. Thus, this is the situation where each of the researchers knows the marginal distribution of the data in each of the vaults. However, none of the researchers knows the joint distribution of the data across the vaults and thus cannot estimate the model that contains the variables from multiple sources. Recently, several empirical researchers have been able to obtain permissions to merge separate administrative data sources. We note that while each data curator controls their own dataset, they also control the "sensitivity" of the variables contained in the dataset. For instance, some variables can be removed from the researcher's access based on the disclosure risk considerations. Such a risk cannot be controlled if the data from one source controlled by one data curator are combined with the data controlled by another data curator. Provided that the marginal data distributions from different sources are already known to the researchers the disclosure threat in this case comes precisely from the data combination.

Related literature.

Our paper is related to several strands in the computer science literature. One of them is on the optimal structures of linkage attacks as well as the requirements in relation to data releases. 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 have been used 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.[3] In several striking examples, computer scientists have shown that a simple removal of personal information such as names and social security numbers does not protect data from individual disclosure. For instance, 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.

In relation to the security of individual data, the computer science literature, e.g. Samarati and Sweeney [1998], Sweeney [2002a], Sweeney [2002b], LeFevre et al. [2005], Aggarwal et al. [2005], LeFevre et al. [2006], Ciriani et al. [2007], has developed and implemented the so-called k-anonymity approach. A database instance is said to provide k-anonymity, for some number k, if every way of singling an individual out of the database returns records for

[3]This is not what we are using in this paper as our data combination rule is deterministic.

at least k individuals. In other words, anyone whose information is stored in the database can be "confused" with k others. Under k-anonymity, a data combination procedure will respect the required bound on the disclosure risk. We describe it in Section 2.3 and use it in the empirical part. An alternative solution is in the use of synthetic data and a related notion of differential privacy, e.g. Dwork and Nissim [2004], Dwork [2006], Abowd and Vilhuber [2008], as well as Duncan and Lambert [1986], Duncan and Mukherjee [1991], Duncan and Pearson [1991], Fienberg [1994], and Fienberg [2001], Duncan et al. [2001], Abowd and Woodcock [2001], Kinney et al. [2011], Hu et al. [2014], among others.

We note that while the computer science literature has alluded to the point that data protection may lead to certain trade-offs in data analysis, data protection has never been considered in the context of model identification. For instance, a notion of "data utility" has been introduced that characterizes the accuracy of a statistical function that can be evaluated from the released data (e.g. see Lindell and Pinkas [2000], Karr et al. [2006], Brickell and Shmatikov [2008], Woo et al. [2009]), and it was found that existing data protection approaches lead to a decreasing quality of inference from the data measured in terms of this utility.

Our paper is also related to the literature on partial identification of models with contaminated or corrupted data, even though our identification approach is new. Manski [2003], Manski [2007] and Horowitz and Manski [1995] note that data errors or data modifications pose identification problems and generally result in only set identification of the parameter of interest. Manski and Tamer [2002] and Magnac and Maurin [2008] give examples where – for confidentiality or anonymity reasons – the data may be transformed into interval data or some attributes may be suppressed, leading to the loss of point identification of the 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. Cross and Manski [2002] and King [1997] study the ecological inference problem where a researcher needs to use the data from several distinct datasets to conduct inference on a population of interest. In ecological inference, several datasets usually of aggregate data are available. Making inferences about micro-units or individual behavior in this case is extremely difficult because variables that allow identification of units are not available. Cross and Manski [2002] show that the parameters of interest are only partially identified. We note that in our case the data contain individual observation on micro-units and there is a limited overlap between two datasets, making the inference problem dramatically different from ecological inference. Pacini [2016] considers estimation and inference on identified sets in linear regression models when the dependent variable is not observed together with covariates but some information available the conditional distribution of regressors conditional on another variable observed together with the outcome variable.

Our analysis relies on the data combination to estimate the econometric model of interest. A train of recent literature in statistics including Larsen [2005], Tancredi et al. [2011],

Chipperfield et al. [2011], Kim and Chambers [2012] establishes consistency for estimation of standard models, such as the regression model, when the data combination procedure is defined parametrically or it is based on exactly matching observations to combine the datasets based on one or more pre-defined variables. Our contribution to this literature is the development of identification properties of econometric models based on combined data for non-parametric data combination procedure when a deterministic a priori criterion for matching observations is not available.

Though less directly related to our analysis, there is also a literature within economics that considers privacy as something that may have a subjective value for consumers (see Acquisti [2004]) rather than 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 allows researchers to analyze the release of private data in the context of the tradeoff 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, a 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 et al. [2006], Miller and Tucker [2009] and Goldfarb and Tucker [2010]. Given all these findings, we believe that disclosure protection is a central theme in the privacy discourse, as privacy protection is impossible without the data protection.

2 Econometric model

2.1 Model and data structure

In this section, we formalize the empirical model based on the joint distribution of the observed outcome variable Y distributed on $\mathcal{Y} \subset \mathbb{R}^m$ and individual characteristics X distributed on $\mathcal{X} \subset \mathbb{R}^k$ that needs to be estimated from the individual level data. We assume that the parameter of interest is $\theta_0 \in \Theta \subset \mathbb{R}^l$, where Θ is a convex compact set.

We characterize the parameter of interest by a conditional moment restriction which, for instance, can describe the individual demand or decision:

$$E\left[\rho(Y, X, \theta_0) \,|\, X = x\right] = 0, \tag{2.1}$$

where $\rho(\cdot, \cdot, \cdot)$ is a known function with the values in $\mathbb{R}^p$. We assume that $\rho(\cdot, \cdot, \cdot)$ is continuous in θ and for almost all $x \in \mathcal{X}$,

$$E\left[\|\rho(Y, X; \theta)\| \,|\, X = x\right] < \infty \quad \text{for any } \theta \in \Theta.$$

We focus on a linear separable model for $\rho(\cdot,\cdot,\cdot)$ as our lead example, which can be directly extended to monotone nonlinear models.

In a typical Internet environment the outcome variable may reflect individual consumer choices by characterizing purchases in an online store, specific messages on a discussion board, comments on a rating website, or a profile on a social networking website. Consumer characteristics are relevant socio-demographic characteristics such as location, demographic characteristics, and social links with other individuals. We assume that if the true joint distribution of (Y, X) were available, one would be able to point identify parameter θ_0 from the condition (2.1). Formally we write this as the following assumption.

ASSUMPTION 1. *Parameter θ_0 is uniquely determined from the moment equation (2.1) and the conditional distribution of $Y|X$.*

As an empirical illustration, in Section 5 we estimate a model the relates the household decision to withdraw the child from the schooling to the religious affiliation of the household and the child's gender as well as the characteristics of the neighborhood. The outcome variable corresponds to the number of completed year of schooling by each child in the household. In this context, we are interested in separating the household-level effect from the neighborhood effect. However, the latest distribution of the survey that we use does not have neighborhood identifiers.

In our data, however, we were able to trace back the original neighborhood identifiers that were used in the early distributions of the survey that we used. That allowed us to construct a (now infeasible) merged dataset that combines individual household characteristics with the neighborhood characteristics that we take from the Russian Census data. In this case Y corresponds to the set of household-level variables and X corresponds to the set of neighborhood variables. The current distribution of the survey does not contain the neighborhood-level variables.

As a result, the variables of interest Y and X are not observed jointly. One can only separately observe the dataset containing the values of Y and the dataset containing the values of X for subsets of the same population.

The following assumption formalizes the idea of the data sample broken into two separate datasets.

ASSUMPTION 2. *(i) The population is characterized by the joint distribution of random vectors (Y, W, X, V) distributed on $\mathcal{Y} \times \mathcal{W} \times \mathcal{X} \times \mathcal{V} \subset \mathbb{R}^m \times \mathbb{R}^q \times \mathbb{R}^k \times \mathbb{R}^r$.*

(ii) The (infeasible) data sample $\{y_i, w_i, x_i, v_i\}_{i=1}^{N_0}$ is a random sample from the population distribution of the data.

(iii) The observable data is formed by two independently created random data subsamples from the sample of size N_0 such that the first data subsample is $\mathcal{D}^{yw} = \{y_j, w_j\}_{j=1}^{N^y}$

and the second subsample is $\mathcal{D}^{xv} = \{x_i, v_i\}_{i=1}^{N^x}$.[4]

(iv) Any individual in $\mathcal{D}^{yw}$ *is present in* $\mathcal{D}^{xv}$*. In other words, for each* (y_j, w_j) *in* $\mathcal{D}^{yw}$ *there exists* (x_i, v_i) *in* $\mathcal{D}^{xv}$ *such that* (y_j, w_j) *and* (x_i, v_i) *correspond to the same individual.*[5]

Assumption 2 characterizes the observable variables as independently drawn subsamples of the infeasible "master" dataset. This means that without any additional information, one can only re-construct distributions $F_{X,V}$ of (X, V) and $F_{Y,W}$ of (Y, W) but this is not enough to learn the joint distribution $F_{Y,X}$ of (Y, X), even though one can use the Fréchet sharp bounds on $F_{Y,X}$ in terms of the marginal distributions F_Y and F_X, or on $F_{Y,W,X,V}$ in terms of the distributions $F_{Y,W}$ and $F_{X,V}$.

EXAMPLE 1. *For linear models, without any additional information identification with split sample data comes down to computing Fréchet bounds. For example, in a bivariate linear regression of random variable* Y *on random variable* X *with* $Var[X] > 0$*, the slope coefficient can be expressed as* $b_0 = \frac{cov(Y,X)}{Var[X]}$*. Because the joint distribution of* Y *and* X *is unknown,* $cov(Y, X)$ *cannot be calculated even if the marginal distributions of* Y *and* X *are available.*

As a result, the only information that allows to draw conclusions about the joint moments of the regressor and the outcome can be summarized by the Cauchy-Schwartz inequality $|cov(Y, X)| \leq \sqrt{Var[Y]}\sqrt{Var[X]}$*, which gives the sharp bounds on* $cov(Y, X)$*. Therefore, we can determine the slope coefficient only up to a set:*

$$-\sqrt{\frac{Var[Y]}{Var[X]}} \leq b_0 \leq \sqrt{\frac{Var[Y]}{Var[X]}}.$$

As we can see, the bounds on b_0 *are extremely wide, especially when there is not much variation in the regressor. Moreover, we cannot even identify the direction of the relationship between the regressor and the outcome, which is of interest in many economic applications.* □

The information contained in vectors V and W is not necessarily immediately useful for the econometric model that is being estimated. However, this information can help us to construct measures of similarity between observations y_j in dataset $\mathcal{D}^{yw}$ and observations x_i in dataset $\mathcal{D}^{xv}$. Random vectors W and V are very likely to be highly correlated for a given

[4]Our analysis applies to other frameworks of split datasets. For instance, we could consider the case when some of the variables in x (but not all of them) are observed together with y. This is the situation we deal with in our empirical illustration. The important requirement in our analysis is that at least some of the relevant variables in x are not observed together with y.

[5]This part of the assumption can be relaxed by allowing $\mathcal{D}^{yw}$ and $\mathcal{D}^{xv}$ to overlap rather than the former to be nested in the latter. In this case, we would replace the requirement $N^y \to \infty$ later in the paper with the requirement that the size of the overlap goes to infinity. When there is no common individual between $\mathcal{D}^{yw}$ and $\mathcal{D}^{xv}$, then the procedures suggested in this paper will not work but some ecological inference methods can be used – see e.g., Cross and Manski [2002] and King [1997], among others.

individual but uncorrelated across different individuals. In our empirical example, the main survey dataset that we use contains the identifier for a large geographic region in Russia (equivalent to the US State in terms of scale) as well as the identifier for the neighborhood. The Russian Census data can be obtained on the level of the neighborhood. Thus, if the neighborhood identifiers are available, then W and V are the perfectly matching placing a given household in its neighborhood. However, the neighborhood identifiers have been removed in the recent distributions of our survey. Therefore, V are larger geographic region identifiers and W (corresponding to the identifier in the Russian Census) contains both the neighborhood identifier and the larger region identifier. In principle, we can expand the set of variables in V and W, for instance, including household demographics, income, property and health data in V and including the neighborhood averages contained in the Russian Census as a part of W. Then we can use a weighted Euclidean distance between V and W as a measure of similarity. This measure of similarity will be used to combine observations in the two datasets.

2.2 Identifiers and decisions rules for data combination

Our data linkage procedure is based on comparing the value of an identifier Z^y constructed for each observation in the main dataset with the value of an identifier Z^x constructed for each observation in the auxiliary dataset. These identifiers are random vectors that can consist of both numerical and string variables. $Z^x = Z^x(X, V)$ is a multivariate function of X and some auxiliary random vector V observed together with X, whereas $Z^y = Z^y(W)$ is a multivariate function of an auxiliary random vector W observed together with Y. Thus, identifiers constructed in the dataset of outcome variables Y are assumed not to be determined by the values of those outcomes. We suppose that these identifiers Z^y and Z^x are constructed in such a way that they have the same dimension and the same support. Our combination rule is based on comparing the values of z_i^y and z_i^x for each $j = 1, \ldots, N^y$ and each $i = 1, \ldots, N^x$.

Namely, we describe the linkage procedure employed by the data curator by means of a binary decision rule $\mathcal{D}_N(y_j, z_j^y, x_i, z_i^x)$, where $N \equiv (N^y, N^x)$, such as

$$\mathcal{D}_N(y_j, z_j^y, x_i, z_i^x) = \begin{cases} 1, & \text{if } z_j^y \text{ and } z_i^x \text{ satisfy certain conditions,} \\ 0, & \text{otherwise.} \end{cases}$$

If $\mathcal{D}_N(y_j, z_j^y, x_i, z_i^x) = 1$, this means that observations j from the main dataset and i from the auxiliary one can potentially be linked. If $\mathcal{D}_N(y_j, z_j^y, x_i, z_i^x) = 0$, then we do not consider j and i to be a possible match. Conditions in the definition of $\mathcal{D}_N(y_j, z_j^y, x_i, z_i^x)$ are chosen by the data curator and in general depend on N, features of the data and objectives on the non-disclosure guarantees discussed later in the paper. A specific feature of such a decision rule is that these conditions do not depend on the values of y_j and x_i and only depend on the values of z_j^y and z_i^x.

Decisions rules used in this paper are based on a chosen distance between z_j^y and z_i^x. Without a loss of generality, suppose that $Z^y = (Z^{y,n}, Z^{y,s})$ and $Z^x = (Z^{x,n}, Z^{x,s})$, where $Z^{y,n}$ and $Z^{x,n}$ are random subvectors of the same dimension that contain all the numeric variables in Z^y and Z^x, respectively, and $Z^{y,s}$ and $Z^{x,s}$ are random subvectors of the same dimension that contain all the string variables in Z^y and Z^x. Then we can define a distance $d(z_j^y, d_i^x)$ between z_j^y and z_i^x as

$$d(z_j^y, d_i^x) = \omega_n \|z_j^{y,n} - z_i^{x,n}\|_E + \omega_s \|z_j^{y,s} - z_i^{x,s}\|_S,$$

where $\|\cdot\|_E$ denotes the Euclidean distance, $\|\cdot\|_S$ stands for a distance between strings (e.g., the edit distance), and $\omega_n, \omega_s \geq 0$ are weights. Below we give some examples of decision rules.

Notation. *Let m_{ij} be the indicator of the event that j and i are the same individual.*

EXAMPLE 2. *A decision rule can be chosen as*

$$\mathcal{D}_N(y_j, z_j^y, x_i, z_i^x) = 1\left\{d(z_j^y, z_i^x) < \alpha_N\right\}. \tag{2.2}$$

The properties of this decision rule – such as the behavior of probabilities of making linkage errors as $N^y, N^x \to \infty$, – would depend on the behavior of the sequence of thresholds $\{\alpha_N\}$ and the properties of the joint distribution of (Y, Z^y, X, Z^x).

Suppose that Z^y and Z^x contain a common variable (e.g., a binary variable for gender). It is clear that in this case j and i can be a potential match only if the values of this variable coincide. Let us denote this variable as $Z^{y,g}$ in the main dataset and as $Z^{x,g}$ in the auxiliary dataset. Then the distance for the decision rule (2.2) can be defined as

$$d(z_j^y, z_i^x) = \begin{cases} \omega_N \|z_j^{y,n} - z_i^{x,n}\|_E + \omega_s \|z_j^{y,s} - z_i^{x,s}\|_S, & \text{if } z_j^{y,g} = z_i^{x,g} \\ \infty, & \text{otherwise.} \end{cases}$$

This idea can be extended to any situation when data linkage is partly based on the values of discrete variables whose values must coincide exactly for the same individual. □

We focus on two types of data combination procedures. Procedures of the first type look only at *observations with infrequent values* of z_i^x. To the best of our knowledge, this paper offers the first formal analysis of the record linkage based on infrequent observations. Procedures of the second type employ decision rules that satisfy the property of *k-anonymity* suggested in the computer science literature.

2.3 Data combination from observations with infrequent values

Let us define the norm of z_i^x as $\|z_i^x\| = \omega_n \|z_i^{x,n}\|_E + \omega_s \|z_i^{x,s}\|_S$. Analogously, the norm of z_j^y is $\|z_j^y\| = \omega_n \|z_j^{y,n}\|_E + \omega_s \|z_j^{y,s}\|_S$. By infrequent attributes we mean the values of identifiers

in the tails.

We suppose that all the variables in Z^x and Z^y are either discrete or continuous with respect to the Lebesgue measure. For technical simplicity, we also suppose that at least one variable in Z^x (and, analogously, in Z^y) is continuous with respect to the Lebesgue measure, which implies that the norms $\|Z^x\|$ and $\|Z^y\|$ are continuous with respect to the Lebesgue measure too.

ASSUMPTION 3. *There exists $\bar{\alpha} > 0$ such that for any $0 < \alpha < \bar{\alpha}$ the following hold:*

(i) (Proximity of identifiers with extreme values)

$$Pr\left(d(Z^y, Z^x) < \alpha \mid X = x, Y = y, \|Z^x\| > \frac{1}{\alpha}\right) \geq 1 - \alpha.$$

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

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

for some non-decreasing and positive at $\alpha > 0$ functions $\phi(\cdot)$ and $\psi(\cdot)$.

(iii) (Redundancy of identifiers in the full data)[6]

$$F_{Y|X,Z^x,Z^y}(y \,|\, X = x, Z^x = z^x, Z^y = z^y) = F_{Y|X}(y \,|\, X = x),$$

where $F_{Y|X,Z^x,Z^y}$ denotes the conditional CDF of Y conditional on X, Z^x and Z^y, and $F_{Y|X}$ denotes the conditional CDF of Y conditional on X.

(iv) (Uniform conditional decay of the tails of identifiers' densities) There exist positive at large z functions $g_1(\cdot)$ and $g_2(\cdot)$ such that

$$\lim_{z \to \infty} \sup_{x} \left| \frac{f_{\|Z^x\||X}(z|X = x)}{g_1(z)} - 1 \right| = 0,$$

$$\lim_{z \to \infty} \sup_{y} \left| \frac{f_{\|Z^y\||Y}(z|Y = y)}{g_2(z)} - 1 \right| = 0,$$

where $f_{\|Z^x\||X}$ denotes the conditional density of $\|Z^x\|$ conditional on X, and $f_{\|Z^y\||Y}$ denotes the conditional density of $\|Z^y\|$ conditional on Y.

[6]It is enough to just impose Assumption 3(iii) under the event described in Assumption 3(i). In this case, we can generalize the definition of $Z^y = Z^y(W)$ to a function $Z^y = Z^y(Y, W)$ that can depend on Y and requiring the redundancy (conditional independence) only in those infrequence events. This kind of an extension can be important in some applications where Y directly contains some information on X, which should help link the two data sets.

Assumption 3 implies that the ordering of the values of $\|Z^y\|$ and $\|Z^x\|$ is meaningful and that the tails of the distributions of $\|Z^x\|$ and $\|Z^y\|$ contains extreme values. If we considered a situation when all the variables in Z^y and Z^x were discrete, this would mean that at least one of these variables has a denumerable support. Ridder and Moffitt [2007] overview cases where *a priori* available numeric identifiers Z^y and Z^x are jointly normally distributed random variables, but we avoid making such specific distributional assumptions.

Assumption 3 (i) states that for infrequent observations – those for which the values of $\|Z^x\|$ are in the tail of the distribution $f_{\|Z^x\||X,Y}$ – the values of Z^y and Z^x are very close, and that they become arbitrarily close as the mass of the tails approaches 0.

Functions $\phi(\cdot)$ and $\psi(\cdot)$ in Assumption 3 (ii) characterize the decay of the marginal distributions of $\|Z^x\|$ and $\|Z^y\|$ at the tail values. The assumptions on these functions imply that

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

and therefore $\phi(\cdot)$ and $\psi(\cdot)$ can be estimated from the split datasets. Moreover, our assumption on the existence of densities for the distributions of $\|Z^x\||X$ and $\|Z^y\||Y$ implies that without a loss of generality, functions $\phi(\cdot)$ and $\psi(\cdot)$ are absolutely continuous.

Assumption 3 (iii) states that for a pair of correctly matched observations from the two databases, their values of identifiers Z^x and Z^y do not add any information regarding the distribution of the outcome Y conditional on X. In other words, if the datasets are already correctly combined, the constructed identifiers only label observations and do not improve any knowledge about the economic model that is being estimated. For instance, if the data combination is based on the names of individuals, then once we extract all model-relevant information from the name (for instance, whether a specific individual is likely to be male or female, or white, black or hispanic) and combine the information from the two databases, the name itself will not be important for the model and will only play the role of a label for a particular observation. Assumption 3 (iii) can be violated, for example, if Z^x and Z^y are proxies for a random vector Z:

$$Z^x = Z + u_x, \quad Z^y = Z + u_y,$$

and measurement errors u_x and u_y are not independent of X and Y.

Function $g_1(\cdot)$ ($g_2(\cdot)$) in Assumption 3 (iv) describes the uniform over x (over y) rate of the conditional density of $\|Z^x\|$ conditional on X ($\|Z^y\|$ conditional on Y) for extreme values of $\|Z^x\|$ ($\|Z^y\|$). If Assumption 3 (iv) holds, then necessarily

$$\lim_{z\to\infty} \frac{\phi'\left(\frac{1}{z}\right)}{z^2 g_1(z)} = 1, \quad \lim_{z\to\infty} \frac{\psi'\left(\frac{1}{z}\right)}{z^2 g_2(z)} = 1.$$

We recognize that Assumption 3 puts restrictions on the behavior of infrequent (tail) realizations of identifiers Z^x and Z^y. Specifically, we expect that conditional on $\|Z^x\|$ taking a high value, the values of identifiers constructed from two datasets must be close. We illustrate this assumption with our empirical application, where different geographic regions in Russia have different population density. As a result, linking the regional data with the household level neighborhood data will lead to higher quality matches in less populated regions.

REMARK 1. *Assumption 3 (iii) can be relaxed to allow for situations when matching is based on income, health or demographic characteristics that would also be included among the regressors. But weakening of Assumption 3 (iii) has to be done together with imposing stricter requirements on the distance function $d(\cdot,\cdot)$.*

Suppose that $Z^y = (\tilde{Z}^y, \tilde{\tilde{Z}}^y)$, $Z^x = (\tilde{Z}^x, \tilde{\tilde{Z}}^x)$ and $X = (\tilde{X}, \tilde{\tilde{X}})$, where $\tilde{Z}^x = \tilde{X}$, and $\tilde{Z}^y$ in the main dataset and $\tilde{X}$ in the auxiliary dataset contain common variables (e.g., discrete variables for age and gender). Suppose that the distance for the decision rule is defined in such a way that

$$d(z_j^y, z_i^x) = \infty \quad \text{if} \quad \tilde{z}_j^y \neq \tilde{x}_i$$

– that is, individuals j and i with different observations for age or gender cannot possibly be matched. Then instead of assumption 3 (iii) we can impose the following weaker restriction:

$$F_{Y|X,\tilde{\tilde{Z}}^x,\tilde{\tilde{Z}}^y}(y \,|\, X = x, \tilde{\tilde{Z}}^x = \tilde{\tilde{z}}^x, \tilde{\tilde{Z}}^y = \tilde{\tilde{z}}^y) = F_{Y|X}(y \,|\, X = x).$$

REMARK 2 (k-anonymity)**.** *The description of k-anonymity approach can be found Samarati and Sweeney [1998], Sweeney [2002a], Sweeney [2002b], among others. We describe it here with the purpose of illustrating how the k-anonymity rule would translate into the properties of the decision rule.*

Given the binary decision rule $\mathcal{D}_N(y_j, z_j^y, x_i, z_i^x)$ in (3.5), we say that the k-anonymity property is implemented if for each observation j in the main dataset, $j = 1, \ldots, N^y$, one of the following conditions hold:

either

a) $\mathcal{D}_N(y_j, z_j^y, x_i, z_i^x) = 0$ for all $i = 1, \ldots, N^x$; that is, j cannot be combined with any individual i in the auxiliary dataset;

or

b) $\sum_{i=1}^{N^x} \mathcal{D}_N(y_j, z_j^y, x_i, z_i^x) \geq k$; that is, for j there are at least k equally good matches in the auxiliary dataset.

Under the rule of k-anonymity, for any j from $\mathcal{D}^y$ and any i from $\mathcal{D}^x$,

$$Pr\left(m_{ij}=1 \,\middle|\, \mathcal{D}_N(y_j,z_j^y,x_i,z_i^x)=1,\mathcal{D}^y,\mathcal{D}^x\right)=\begin{cases}0, & \text{if } \sum_{l=1}^{N^x}\mathcal{D}_N(y_j,z_j^y,x_l,z_l^x)=0,\\ \frac{1}{\sum_{l=1}^{N^x}\mathcal{D}_N(y_j,z_j^y,x_l,z_l^x)}, & \text{otherwise.}\end{cases}$$

Clealry, it always holds that

$$Pr\left(m_{ij}=1 \,\middle|\, \mathcal{D}_N(y_j,z_j^y,x_i,z_i^x)=1,\mathcal{D}^y,\mathcal{D}^x\right)\leq\frac{1}{k}. \tag{2.3}$$

The binary decision rule for k-anonymity does not have to be based on infrequent observations and can use much more general ideas. One only has to guarantee that (2.3) holds.

3 Implementation of data combination and implications for identity disclosure

In this section, we characterize in more detail the class of data combination procedures that we use in this paper, introduce the formal notion of identity disclosure and characterize a subclass of data combination procedures that are compatible with a bound for the risk of the identity disclosure. We suppose henceforth that Assumptions 1-3 hold.

3.1 Implementation of data combination

In our model, the realizations of random variables Y and X are contained in disjoint datasets. After constructing identifiers Z^y and Z^x, we directly observe the empirical distributions of (Y,Z^y) and (X,Z^x). Even though these two distributions provide some information about the joint distribution of (Y,X), such as Fréchet bounds, they do not fully characterize it if no data combination whatsoever is conducted, and thus, there are many joint distributions of (Y,X) (or, more generally, joint distributions of (Y,Z^y,X,Z^x)) consistent with the observed distributions of (Y,Z^y) and (X,Z^x).[7]

We can hope to identify the econometric model only if the two datasets are combined for at least some observations and thus, more information becomes available about the dependence structure between vectors (Y,Z^y) and (X,Z^x), from which we can consequently obtain more information about the dependence structure between Y and X. The best case scenario from the identification point of view occurs if our data combination procedure allows us to learn the copula describing the true joint distribution of (Y,Z^y,X,Z^x) as a function of two separate distributions of (Y,Z^y) and (X,Z^x). This would automatically give

[7]This means that we would have to consider all such compatible joint distributions of (Y,X) when trying to determine the parameter of interest using (2.1). Intuitively, any compatible joint distribution of (Y,X) would give us a different value of the parameter of interest, which means that the parameter of interest can only be determined up to a set. Thus, the econometric model of interest is not identified from the available information about the distributions of (Y,Z^y) and (X,Z^x).

us the copula describing the true joint distribution of (Y, X) as a function of the marginal distributions of Y and X, and then we would be able to point identify θ_0 using (2.1). Whether this scenario will occur clearly depends on the quality of the data combination procedure.

Now let us describe data combination procedures in more detail. Once the identifiers Z^y and Z^x are constructed, we have the following two split data sets:

$$\mathcal{D}^y = \{y_j, z_j^y\}_{j=1}^{N^y}, \quad \mathcal{D}^x = \{x_i, z_i^x\}_{i=1}^{N^x}. \tag{3.4}$$

Provided that the indexes of matching entries are not known in advance, the entries with the same index i and j do not necessarily belong to the same individual.

We base our decision rule on the postulated properties in Assumption 3:

$$\mathcal{D}_N(y_j, z_j^y, x_i, z_i^x) = 1\left\{d(z_j^y, z_i^x) < \alpha_N, \|z_i^x\| > 1/\alpha_N\right\}, \tag{3.5}$$

for a chosen α_N such that $0 < \alpha_N < \bar{\alpha}$. We notice that for each rate $r_N \to \infty$ there is a whole class of data combination rules $\mathcal{D}_N(y_j, z_j^y, x_i, z_i^x)$ corresponding to all threshold sequences for which $\alpha_N r_N$ converges to a non-zero value as $N^y, N^x \to \infty$. As is clear from our results later in this section, the rate r_N is what determines the asymptotic properties of the data combination procedure. Provided that the focus of this paper is on identification rather than estimation in the context of data combination, in the remainder of the paper, our discussion about a data combination rule refers the whole class of data combination rules characterized by the threshold sequences with a given rate.

Consider an observation i from $\mathcal{D}^x$ such that $\|z_i^x\| \geq 1/\alpha_N$. If we find a data entry j from the dataset $\mathcal{D}^y$ such that $d(z_j^y, z_i^x) < \alpha_N$, then we consider i and j as a potential match. In other words, if identifiers z_i^x and z_j^y are both large and are close, then we consider (x_i, z_i^x) and (y_j, z_j^y) as observations possibly corresponding to the same individual. This seems to be a good strategy when α_N is small because, according to Assumption 3, when the pair (Z^x, Z^y) is drawn from their true joint distribution, the conditional probability of Z^x and Z^y taking proximate values when Z^x is large in the absolute value is close to 1. Even though the decision rule is independent of the values of x_i and y_j, the probability $Pr\left(m_{ij} = 1 \mid \mathcal{D}_N(y_j, x_i, z_j^y, z_i^x), x_i = x, y_j = y, \mathcal{D}^x, \mathcal{D}^y\right)$ for a finite $N = (N^x, N^y)$ can depend on these values (and also depend on the sizes of datasets $\mathcal{D}^x$ and $\mathcal{D}^y$) and therefore can differ across pairs of i and j.

Using the combination rule $\mathcal{D}_N(\cdot)$, for each $j \in \{1, \ldots, N^y\}$ from the database $\mathcal{D}^y$ we try to find an observation i from the database $\mathcal{D}^x$ that satisfies our matching criteria and thus presents a potential match for j. We can then add the "long" vector (y_j, z_j^y, x_i, z_i^x) to our combined dataset if neither (y_j, z_j^y) for this specific j nor (x_i, z_i^x) for this specific i enter the combined dataset as subvectors of other "long" observations. In other words, if there are

several possible matches i from $\mathcal{D}^x$ for some j in $\mathcal{D}^y$(or several possible matches j from $\mathcal{D}^y$ for some i in $\mathcal{D}^x$), we can put only one of them in our combined dataset. Mathematically, each combined dataset $\mathcal{G}_N$ can be described by an $N^y \times N^x$ matrix $\{d_{ji}, j = 1, \ldots, N^y; i = 1, \ldots, N^x\}$ of zeros and ones, which satisfies the following conditions:

(a) $d_{ji} = 1$ if observations (y_j, z_j^y) and (x_i, z_i^x) are matched; $d_{ji} = 0$ otherwise.

(b) For each $j = 1, \ldots, N^y$, $\sum_{i=1}^{N^x} d_{ji} \leq 1$ (i.e., each j can be added to our combined dataset with at most one i).

(c) For each $i = 1, \ldots, N^x$, $\sum_{j=1}^{N^y} d_{ji} \leq 1$ (i.e., each i can be added to our combined dataset with at most one j).

Because some j in $\mathcal{D}^y$ or some i in $\mathcal{D}^x$ can have several possible matches, several different combined datasets $\mathcal{G}_N$ can be constructed. The data curator decides which one of these combined datasets to use (e.g., it can be chosen randomly, or the data curator could choose a different selection principle). Once the data curator chooses some $\mathcal{G}_N$, from this combined dataset she deletes the data on z_j^y and z_i^x leaving only that data on linked pairs (y_j, x_i). This reduced dataset $\mathcal{G}_N^{xy}$ is released to the public along with some information about the properties of identifiers. This information is used by the researchers to conduct the identification analysis. Even though the dataset $\mathcal{D}^{xv} = \{(x_i, v_i)\}$ is publicly available and, thus, the researcher can potentially construct some identifiers (possibly similar to z_i^x) from that dataset, the researcher is not given any data on w_j and thus would not be able to construct identifiers similar to z_j^y (or any other identifiers for observations y_j).

Our identification approach in section 4 will take into account all possible combined datasets and take into account the probabilities of making data combination errors.

Consider an observation i from $\mathcal{D}^x$ such that $\|z_i^x\| \geq 1/\alpha_N$. Two kinds of errors can be made when finding entry i's counterpart in the dataset $\mathcal{D}^y$.

(1) Data combination errors of the first kind occur when the decision rule links an observation j from $\mathcal{D}^y$ to i, but in fact j and i do not correspond to the same individual. For the two given split datasets, the probability of the error of this kind is

$$Pr\left(d(z_j^y, z_i^x) < \alpha_N \;\middle|\; \|z_i^x\| > 1/\alpha_N, x_i = x, y_j = y, m_{ij} = 0, \mathcal{D}^y, \mathcal{D}^x\right),$$

or

$$Pr\left(d(\widetilde{Z^y}, Z^x) < \alpha_N \;\middle|\; \|Z^x\| > 1/\alpha_N, X = x, \widetilde{Y} = y\right),$$

where (X, Z^x) and $(\widetilde{Y}, \widetilde{Z^y})$ are independent random vectors with the distributions F_{X,Z^x} and F_{Y,Z^y}, respectively.

(2) Data combination errors of the second kind occur when observations j and i do belong to the same individual but the procedure does not identify these two observations

as a potential match (we still consider i such that $\|z_i^x\| \geq 1/\alpha_N$). For the two given split datasets, the probability of the error of this kind is

$$Pr\left(d(z_j^y, z_i^x) \geq \alpha_N \;\middle|\; \|z_i^x\| > 1/\alpha_N, x_i = x, y_j = y, m_{ij} = 1, \mathcal{D}^y, \mathcal{D}^x\right),$$

or

$$Pr\left(d(Z^y, Z^x) \geq \alpha_N \;\middle|\; \|Z^x\| > 1/\alpha_N, X = x, Y = y\right), \tag{3.6}$$

where (Y, X, Z^x, Z^y) is distributed with F_{Y,X,Z^x,Z^y}. Assumption 3 guarantees that (3.6) converges to 0 as $\alpha_N \to 0$.

While the second kind of error vanishes as one considers increasingly infrequent values, the behavior of the probability of the first kind of error depends on the rate of α_N and can be controlled by the data curator. As we establish later in this section, this rate can be chosen e.g. in such a way that the probability of the first kind of error will be separated away from 0 even for arbitrarily large split datasets.

3.2 Risk of disclosure

What we notice so far is that given that there is no readily available completely reliable similarity metric between the two databases we rely on the probabilistic properties of the data. As a result, in estimation we have to resort to only using the pairs of combined observations. If correct matches are made with a sufficiently high probability, this may pose a potential problem if one of the two datasets contains sensitive individual-level information. The only way to avoid such an information leakage is to control the accuracy of utilized data combination procedures. In particular, we consider controlling the error of the first kind.

For technical convenience, in the remainder of the paper we consider the case when Z^y and Z^x are random variables, and the distance $d(Z^y, Z^x)$ is defined as $|Z^y - Z^x|$. The the decision rule is

$$\mathcal{D}_N(y_j, z_j^y, x_i, z_i^x) = 1\left\{|z_j^y - z_i^x| < \alpha_N, |z_i^x| > 1/\alpha_N\right\}. \tag{3.7}$$

Propositions 1 and 2, which appear later in this section, give conditions on the sequence of α_N, $\alpha_N \to 0$, that are sufficient to guarantee that the probability of the error of the first kind vanishes as $N^y \to \infty$. Proposition 3 give conditions on α_N, $\alpha_N \to 0$, under which the probability of the error of the first kind is separated away from 0 as $N^y \to \infty$.

For given split datasets $\mathcal{D}^y$ of size N^y and $\mathcal{D}^x$ of size N^x as in (3.4), and given y and x, consider the conditional probability

$$p_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y) = Pr\left(m_{ij} = 1 \mid x_i = x, y_j = y, |z_i^x| > \frac{1}{\alpha_N}, |z_j^y - z_i^x| < \alpha_N, \mathcal{D}^x, \mathcal{D}^y\right) \tag{3.8}$$

of a successful match of (y_j, z_j^y) from $\mathcal{D}^y$ with (x_i, z_i^x) from $\mathcal{D}^x$.

According to our discussion, potential privacy threats occur when one establishes that a particular combined data pair (y_j, x_i, z_j^y, z_i^x) is correct with a high probability. This is the idea that we use to define the notion of the risk of the identity disclosure. Our definition of the risk of disclosure in possible linkage attacks is similar to the definition of the pessimistic disclosure risk in Lambert [1993]. We formalize the pessimistic disclosure risk by considering the maximum probability of a successful linkage attack over all individuals in a database.

Since by Assumption 2 (iv), $N^x \geq N^y$, all of our asymptotic results will be formulated as the ones obtained when $N^y \to \infty$ since this also implies that $N^x \to \infty$.

DEFINITION 1. *A bound guarantee is given for the risk of disclosure if*

$$\sup_{x,y} \sup_{\mathcal{D}^x, \mathcal{D}^y} \sup_{i,j} p_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y) < 1$$

for all N, and there exists $0 < \underline{\gamma} \leq 1$ such that

$$\sup_{x,y} \limsup_{N^y \to \infty} \sup_{\mathcal{D}^x, \mathcal{D}^y} \sup_{i,j} p_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y) \leq 1 - \underline{\gamma}. \tag{3.9}$$

The value of $\underline{\gamma}$ is called a bound on the disclosure risk.

Our definition of the disclosure guarantee requires, first of all, that for any two finite datasets $\mathcal{D}^y$ and $\mathcal{D}^x$ and any matched pair, the value of $p_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y)$ is strictly less than one. In other words, there is always a positive probability of making a linkage mistake. However, even if probabilities $p_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y)$ are strictly less than 1, they may turn out to be very high when N^y is sufficiently large and α_N is sufficiently small. Our definition of the disclosure guarantee requires that such situations do not arise. The value of $\underline{\gamma}$ is the extent of the non-disclosure risk guarantee.

An important practical question is whether there exist (the classes of the) decision rules that guarantee a specified bound on the disclosure risk. Below we present results that indicate, first, that for a given bound on the disclosure risk we can find sequences of thresholds such that the corresponding decision rules honor this bound, and second, that the rates of convergence for these sequences depend on the tail behaviour of identifiers used in the data combination procedure. Propositions 1 and 3 give general results.

PROPOSITION 1. *Suppose Assumptions 2 and 3 hold. Suppose that for given non-decreasing and positive for $\alpha \in (0, \bar{\alpha})$ functions $\phi(\cdot)$ and $\psi(\cdot)$ the sequence of $\alpha_N \to 0$ (as $N^y \to \infty$) is chosen in such a way that*

$$\frac{N^x}{\phi(\alpha_N)} \int_{\frac{1}{\alpha_N}}^{\infty} \left(\psi\left(\frac{1}{z - \alpha_N} \right) - \psi\left(\frac{1}{z + \alpha_N} \right) \right) \frac{\phi'\left(\frac{1}{z}\right)}{z^2} \, dz \to 0 \tag{3.10}$$

as $N^y \to \infty$. *Then*

$$\inf_{x\in\mathcal{X},y\in\mathcal{Y}} \inf_{\mathcal{D}^x,\mathcal{D}^y} \inf_{i,j} p_{ij}^N(x,y,\mathcal{D}^x,\mathcal{D}^y) \to 1 \quad \text{as } N^y \to \infty.$$

The result of Proposition 1 implies the following result in Proposition 2.

PROPOSITION 2. ***(Absence of non-disclosure risk guarantee).*** *Suppose the conditions in Proposition 1 hold.*

Then non-disclosure is not guaranteed.

The result of Proposition 1 is stronger than that of Proposition 2 and will provide an important link between the absence of non-disclosure risk guarantees and the point identification of the parameter of interest discussed in Theorem 1.

The next proposition describes instances in which non-disclosure can be guaranteed.

PROPOSITION 3. ***(Non-disclosure risk guarantee).*** *Suppose Assumptions 2 and 3 hold. Suppose that for given non-decreasing and positive for* $\alpha \in (0,\bar{\alpha})$ *functions* $\phi(\cdot)$ *and* $\psi(\cdot)$ *the sequence of* $\alpha_N \to 0$ *(as* $N^y \to \infty$*) is chosen in such a way that*

$$\liminf_{N^y\to\infty} \frac{N^x}{\phi(\alpha_N)} \int_{\frac{1}{\alpha_N}}^{\infty} \left(\psi\left(\frac{1}{z-\alpha_N}\right) - \psi\left(\frac{1}{z+\alpha_N}\right)\right) \frac{\phi'\left(\frac{1}{z}\right)}{z^2}\, dz > 0. \tag{3.11}$$

Then non-disclosure is guaranteed.

The proofs of propositions 1–3 are in the Appendix.

Propositions 2 and 3 demonstrate that the compliance of the decision rule generated by a particular threshold sequence with a given bound guarantee for the disclosure risk depends on the rate at which the threshold sequence converges towards zero as the sizes of $\mathcal{D}^y$ and $\mathcal{D}^x$ increase.

The decision rules that we constructed are well-defined and there exists a non-empty class of sequences of thresholds that can be used for data combination and that guarantee the avoidance of identity disclosure with a given probability. The rate of these sequences depends on the tail behaviour of the identifiers' distributions. A more detailed discussion of the choice of threshold sequences can be found in the Online supplement.

4 Analysis of identifiability with combined data

In the previous section we described the decision rule that can be used for combining data and its implications for potential identity disclosure. In this section, we characterize the identification of the econometric model from the combined dataset constructed using the proposed data combination procedure. We also show the implications of the bound on the disclosure risk for identification.

We emphasize that the structure of our identification argument is non-standard. In fact, the most common identification argument in the econometrics literature is based on finding a mapping between the population distribution of the data and parameters of interest. If the data distribution leads to a single parameter value, this parameter is called point identified. However, as we explained in the previous section, the population distribution of the immediately available data in our case is not informative, because it consists of two unrelated marginal distributions corresponding to population distributions generating split samples $\mathcal{D}^y$ and $\mathcal{D}^x$. Combination of these two samples and construction of a combined subsample is only possible when these samples are finite. In other words, knowing the probability that a given household may reside in a certain neighborhood is not informative to us. For correct inference we need to make sure that a combined observation contains the split pieces of information regarding the same household and does not misassign a household to a different neighborhood within the same region. As a result, our identification argument is based on the analysis of the limiting behavior of identified sets of parameters that are obtained by applying the (finite sample) data combination procedure to samples of an increasing size.

The proposition below brings together the conditional moment restriction (2.1) describing the model and our threshold-based data combination procedure. This proposition establishes that if there is a "sufficient" number of data entries which we *correctly* identify as matched observations, then there is "enough" knowledge about the joint distribution of (Y, X) to point identify and estimate the model of interest.

PROPOSITION 4. *Suppose Assumption 3 holds. For any $\theta \in \Theta$ and any $\alpha \in (0, \bar{\alpha})$,*

$$E\left[\rho(Y, X; \theta) \;\middle|\; X = x, |Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\right] = E\left[\rho(Y, X; \theta) \;\middle|\; X = x\right]. \tag{4.12}$$

The proof of this proposition is in the Appendix.

The result in Proposition 4 is quite intuitive. Record linkage is based on Z^x and Z^y, which are by Assumption 3 are unrelated to Y and hence to $\rho(Y, X, \theta)$ given X. This immediately makes $E[\rho(Y, X, \theta) \,|\, X] = E[\rho(Y, X, \theta) \,|\, X, G(Z^x, Z^y)]$ for any function G, so we can in particular define G to indicate a high probability of correctly matched data. In short, we can identify the parameters in the model just using a subpopulation with relatively infrequent characteristics because are the observations that are very likely to be correctly matched, because information used for matching is by assumption conditionally independent of the model.

Thus, if the joint distribution of Y and X is known when the constructed identifiers are compatible with the data combination rule ($\left\{|Z^x| > \frac{1}{\alpha}, |Z^x - Z^y| < \alpha\right\}$), then, also employing Assumption 1, one can conclude that θ_0 can be identified and estimated from

the moment equation

$$E\left[\rho(Y,X;\theta_0) \mid X=x, |Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\right] = 0 \tag{4.13}$$

using only observations from the combined dataset. This is true even for extremely small $\alpha > 0$. Using this approach, we effectively ignore a large portion of observations of covariates and concentrate only on observations with extreme values of identifiers. For the population analysis based on (4.13) it does not matter how small the event $\{|Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\}$ is because Assumption 3(i)-3(ii) guarantee that its probability is strictly positive. In a sample, if the set of infrequent observation turns out to be very small, our recommendation to researchers would be to try increasing the dimension of Z^X and Z^y by employing more information contained in auxiliary vectors V and W.

A useful implication of Proposition 4 is that

$$\lim_{\alpha \downarrow 0} E\left[\rho(Y,X;\theta) \mid X=x, |Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\right] = E\left[\rho(Y,X;\theta) \mid X=x\right].$$

EXAMPLE 3. *Here we continue Example 1 and illustrate the identification approach based on infrequent data attributes in a bivariate linear model. Let Y and X be two scalar random variables, and $Var[X] > 0$. Suppose the model of interest is characterized by the conditional mean restriction*

$$E\left[Y - a_0 - b_0 X \mid X = x\right] = 0,$$

where $\theta_0 = (a_0, b_0)$ is the parameter of interest. If the joint distribution of (Y, X) was known, then applying the least squares approach, we would find θ_0 from the following system of equations for unconditional means implied by the conditional mean restriction:

$$\begin{aligned} 0 &= E[Y - a_0 - b_0 X] \\ 0 &= E[X(Y - a_0 - b_0 X)]. \end{aligned}$$

This system gives $b_0 = \frac{Cov(X,Y)}{Var[X]}$ and $a_0 = E[Y] - b_0 E[X]$.

When using infrequent observations only, we can apply Proposition 4 and identify θ_0 from the "trimmed" moments. The solution can be expressed as

$$\begin{aligned} b_0 &= \frac{E\left[X^* Y^*\right] E[\mathbf{1}\{|Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\}] - E[X^*]E[Y^*]}{E\left[X^{*2}\right] E[\mathbf{1}\{|Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\}] - \left(E[X^*]\right)^2}, \\ a_0 &= \frac{E[Y^*] - b_0 E[X^*]}{E[\mathbf{1}\{|Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\}]^{1/2}}, \end{aligned}$$

where $X^* = \frac{X\mathbf{1}\{|Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\}}{E[\mathbf{1}\{|Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\}]^{1/2}}$ and $Y^* = \frac{Y\mathbf{1}\{|Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\}}{E[\mathbf{1}\{|Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\}]^{1/2}}$. □

It is worth noting that observations with more common values of identifiers (not sufficiently far in the tail of the distribution) have a higher probability of resulting in false matches and are thus less reliable for the purpose of model identification.

Our next step is to introduce a notion of the pseudo-identified set based on the combined data. This notion incorporates several features. First, it takes into account the result of Proposition 4, which tells us that the information obtained from the correctly linked data is enough to point identify the model. Second, it takes into consideration the fact that it is possible to make some incorrect matches, and the extent to which the data are mismatched determines how much we can learn about the model. Third, it takes into account the fact that the data combination procedure is a finite-sample technique and identification must therefore be treated as a limiting property as the sizes of both datasets increase. We start with a discussion of the second feature and then conclude this section with a discussion of the third feature.

As before, $\mathcal{G}_N$ denotes some combined dataset of (y_j, z_j^y, x_i, z_i^x) constructed from $\mathcal{D}^x$ of size N^x and $\mathcal{D}^y$ of size N^y by means of a chosen data combination procedure. The joint density of observations (y_j, z_j^y, x_i, z_i^x) in $\mathcal{G}_N$ can be expressed in terms of the true joint density of the random vector (Y, Z^y, X, Z^x) and the marginal densities of (Y, Z^y) and (X, Z^x):

$$f_{Y,Z^y,X,Z^x}(y_j, z_j^y, x_i, z_i^x)1(m_{ij} = 1) + f_{Y,Z^y}(y_j, z_j^y)f_{X,Z^x}(x_i, z_i^x)1(m_{ij} = 0).$$

In other words, if j and i correspond to the same individual, then (y_j, z_j^y, x_i, z_i^x) is a drawing from the distribution f_{Y,Z^y,X,Z^x}, whereas if j and i do not correspond to the same individual, then the subvector (y_j, z_j^y) and the subvector (x_i, z_i^x) are independent and are drawn from the marginal distributions f_{Y,Z^y} and f_{X,Z^x} respectively.

For a given value $y \in Y$ and a given value $x \in X$, let $\pi^N(y, x, \mathcal{G}_N)$ denote the proportion of incorrect matches in the set

$$S_{yx}(\mathcal{G}_N) = \{(y_j, z_j^y), (x_i, z_i^x) : y_j = y, x_i = x,\ (y_j, z_j^y, x_i, z_i^x) \in \mathcal{G}_N\}.$$

If this set is empty, then $\pi^N(y, x, \mathcal{G}_N)$ is not defined.

By $\pi^N\left(y, x, \{y_j, z_j^y\}_{j=1}^{N^y}, \{x_i, z_i^x\}_{i=1}^{N^x}\right)$ let us denote the average proportion of incorrect matches across *all possible combined datasets* $\mathcal{G}_N$ that can be obtained from $\mathcal{D}^y$ and $\mathcal{D}^x$ according to the chosen data combination. Then we find that

$$\pi^N\left(y, x, \{y_j, z_j^y\}_{j=1}^{N^y}, \{x_i, z_i^x\}_{i=1}^{N^x}\right) = \frac{\sum\limits_{\mathcal{G}_N} \pi^N(y, x, \mathcal{G}_N)1\left(S_{yx}(\mathcal{G}_N) \neq \emptyset\right)}{\sum\limits_{\mathcal{G}_N} 1\left(S_{yx}(\mathcal{G}_N) \neq \emptyset\right)} \quad \text{if} \quad \sum_{\mathcal{G}_N} 1\left(S_{yx}(\mathcal{G}_N) \neq \emptyset\right) > 0.$$

This value is not defined otherwise (that is, if (y_j, z_j^y) and (x_i, z_i^x) with $y_j = y$, $x_i = x$ are

never combined). Define $\pi^N(y,x)$ as the mean of $\pi^N\left(y,x,\{y_j,z_j^y\}_{j=1}^{N^y},\{x_i,z_i^x\}_{i=1}^{N^x}\right)$ over all possible datasets of N^y observations of (y_j,z_j^y) and all possible datasets of N^x observations of (x_i,z_i^x) that contain y_j that coincide with y and x_i that coincide with x. It is assumed that these datasets originated from split datasets $\{y_j,w_j\}_{j=1}^{N^y}$ and $\{x_i,v_i\}_{i=1}^{N^x}$ that satisfy Assumption 2.

Next, we define the distribution density for an observation in a "generic" combined dataset of size $N=(N^x,N^y)$:

$$f^N_{Y,Z^y,X,Z^x}(y_j,z_j^y,x_i,z_i^x)=(1-\pi^N(y_j,x_i))f_{Y,Z^y,X,Z^x}(y_j,z_j^y,x_i,z_i^x)+\pi^N(y_j,x_i)f_{Y,Z^y}(y_j,z_j^y)f_{X,Z^x}(x_i,z_i^x)$$

for any pairs (y_j,z_j^y) and (x_i,z_i^x) with $\mathcal{D}_N(y_j,z_j^y,x_i,z_i^x)=1$. Using this density we can define the expectation with respect to the distribution of the data in the combined dataset and denote it $E^N[\cdot]$.

In light of the result in (4.13), we want to consider $E^N\left[\rho(y,x;\theta)\,\middle|\,X=x\right]$ and analyze how close this conditional mean is to 0, and how close it gets to 0 as $\alpha_N\to 0$. If, for instance, $\pi^N(y,x)$ approaches 0 almost everywhere, then in the limit we expect this conditional mean to coincide with the left-hand side in (4.13), and thus, take the value of 0 if and only if $\theta=\theta_0$. Intuitively, the situation is going to be completely different if even for arbitrarily small thresholds the values of $\pi^N(y,x)$ will be separated away from 0 for a positive measure of (y,x) .

We want to introduce a distance $r(\cdot)$ that measures the proximity of the conditional moment vector $E^N[\rho(y_j,x_i;\theta)\mid x_i=x]$ to 0. We want this distance to take only non-negative values and satisfy the following condition in the special case when $\pi^N(y,x)$ is equal to 0 a.e.:

$$r\left(E\left[\rho(Y,X;\theta)\,\middle|\,X=x\right]\right)=0\quad\Longrightarrow\quad\theta=\theta_0 \tag{4.14}$$

The distance function $r(\cdot)$ can be constructed, for instance, by using the idea behind the generalized method of moments. We consider

$$r\left(E^N\left[\rho(y_j,x_i;\theta)\,\middle|\,x_i=x\right]\right)=g^N(\theta)'W_0g^N(\theta),$$

where

$$g^N(\theta)=E_X\left[h(x)E^N[\rho(y_j,x_i;\theta)|x_i=x]\right]=E^N\left[h(x_i)\rho(y_j,x_i;\theta)\right],$$

with a $J\times J$ positive definite matrix W_0, and a chosen (nonlinear) $J\times p$, $J\geq k$ instrument $h(\cdot)$ such that

$$E\left[\sup_{\theta\in\Theta}\|h(X)\rho(Y,X;\theta)\|\right]<\infty,\quad E^*\left[\sup_{\theta\in\Theta}\|h(X)\rho(\widetilde{Y},X;\theta)\|\right]<\infty \tag{4.15}$$

where $E_X[\cdot]$ denotes the expectation over the distribution of X, and E^* denotes the expectation taken over the distribution $f_Y(\widetilde{y})f_X(x)$.

Condition (4.14) is satisfied if and only if for $\pi^N(y,x) = 0$ a.e.,

$$E\left[h(X)\rho(Y,X;\theta)\right] = 0 \quad \Longrightarrow \quad \theta = \theta_0.$$

In rare situations this condition can be violated for some choices of instruments $h(\cdot)$[8], so $h(\cdot)$ has to be chosen in a way to guarantee that it holds. Here and thereafter we suppose that (4.14) is satisfied.

For a given N and a known $\pi^N(y,x)$, the minimizer (or the set of minimizers) of $r\left(E^N\left[\rho(y_j,x_i;\theta) \mid x_i = x\right]\right)$ is the best approximation of θ_0 under the chosen $r(\cdot)$. The important question, of course, is how much is known (or, told by the data curator) to the researcher about the sequences of $\pi^N(y,x)$.

Let Π^N denote the information available to the researcher about the proportions $\pi^N(\cdot,\cdot)$. We can interpret Π^N as the set of all functions $\pi^N(\cdot,\cdot)$ that are possible under the available to the researcher information about the data combination procedure. For instance, the data curator could provide the researcher with the information that any value of $\pi^N(y,x)$ is between some known π_1 and π_2. Then any measurable function $\pi^N(\cdot,\cdot)$ taking values between π_1 and π_2 has to be considered in Π^N. The empirical evidence thus generates a set of values for θ approximating θ_0. We call it the N-identified set and denote it as Θ_N:

$$\Theta_N = \bigcup_{\pi^N \in \Pi^N} \underset{\theta\in\Theta}{\text{Argmin}}\ r\left(E^N\left[\rho(y_j,x_i;\theta) \mid x_i = x\right]\right). \tag{4.16}$$

The next step is to consider the behavior of sets Θ_N as $N \to \infty$, which, of course, depends on the behavior of Π^N as $N \to \infty$.

Let Π^∞ denote the set of possible uniform over all $y \in \mathcal{Y}$ and over all $x \in \mathcal{X}$ limits of elements in Π^N. That is, Π^∞ is the set of $\pi(\cdot,\cdot)$ such that for each N, there exists $\pi^N(\cdot,\cdot) \in \Pi^N$ such that $\sup_{y\in\mathcal{Y},x\in\mathcal{X}} |\pi^N(y,x) - \pi(y,x)| \to 0$.

The fact that the data combination procedure does not depend on the values of y and x (even though the probability of the match being correct may depend on y and x) implies that Π^∞ is a set of some constant values π. Suppose that this is known to the researcher.

Proposition 5 below shows that in this situation the following set Θ_∞ is a limit of the sequence of N-identified sets Θ_N:

$$\Theta_\infty = \bigcup_{\pi\in\Pi^\infty} \underset{\theta\in\Theta}{\text{Argmin}}\ r\left((1-\pi)E\left[\rho(Y,X;\theta) \mid X = x\right] + \pi E^*\left[\rho(\widetilde{Y},\, X;\theta)|X = x\right]\right), \tag{4.17}$$

[8]Dominguez and Lobato [2004] give examples of situations when the selected unconditional moment restrictions may hold for several parameter values even if the conditional restrictions from the are obtained hold only for one value.

where

$$r\left((1-\pi)E\left[\rho(Y,X;\theta)\,\middle|\,X=x\right]+\pi E^{*}\left[\rho(\widetilde{Y},\,X;\theta)|X=x\right]\right)=g_{\pi}(\theta)'W_{0}g_{\pi}(\theta)$$

with

$$\begin{aligned}g_{\pi}(\theta)&=E_{X}\left[h(x)\left((1-\pi)E\left[\rho(Y,X;\theta)\,\middle|\,X=x\right]+\pi E^{*}\left[\rho(\widetilde{Y},\,X;\theta)|X=x\right]\right)\right]\\&=(1-\pi)E\left[h(X)\rho(Y,X;\theta)\right]+\pi E^{*}\left[h(X)\rho(\widetilde{Y},\,X;\theta)\right].\end{aligned}$$

PROPOSITION 5. *Suppose that Π^{∞} consists of constant values and for any $\pi\in\Pi^{\infty}$ there exists $\pi^{N}(\cdot,\cdot)\in\Pi^{N}$ such that*

$$\sup_{y\in Y,x\in X}|\pi^{N}(y,x)-\pi|\to 0\ \textit{as}\ N^{y}\to\infty. \tag{4.18}$$

Also suppose that for any $\pi\in\Pi^{\infty}$ the function $g_{\pi}(\theta)'W_{0}g_{\pi}(\theta)$ has a unique minimizer. Consider Θ_{N} defined as in (4.16) and Θ_{∞} defined as in (4.17). Then for any $\theta\in\Theta_{\infty}$ there exists a sequence $\{\theta_{N}\}$, $\theta_{N}\in\Theta_{N}$, such that $\theta_{N}\to\theta$ as $N^{y}\to\infty$.

The proof of this proposition is in the Online supplement.

Proposition 5 can be rewritten in terms of the distances between sets Π^{∞} and Π^{N} and sets Θ_{∞} and Θ_{N}:

$$\begin{aligned}d(\Pi^{\infty},\Pi^{N})&=\sup_{\pi\in\Pi^{\infty}}\inf_{\pi^{N}\in\Pi^{N}}\sup_{y\in Y,x\in X}|\pi^{N}(y,x)-\pi|\\d(\Theta_{\infty},\Theta_{N})&=\sup_{\theta\in\Theta_{\infty}}\inf_{\theta_{N}\in\Theta_{N}}\|\theta_{N}-\theta\|.\end{aligned}$$

Indeed, the definition of Π^{∞} gives that $d(\Pi^{\infty},\Pi^{N})\to 0$ as $N^{y}\to\infty$. Proposition 5 establishes that this condition together with the condition on the uniqueness of the minimizer of $g_{\pi}(\theta)'W_{0}g_{\pi}(\theta)$ for each $\pi\in\Pi^{\infty}$ gives that $d(\Theta_{\infty},\Theta_{N})\to 0$ as $N^{y}\to\infty$.

DEFINITION 2. *Θ_{∞} is what we call the pseudo-identified set or the set of parameter values identified from infrequent attribute values.*

Obviously, the size of Θ_{∞} depends on the information set Π^{∞} because Θ_{∞} generally becomes larger if Π^{∞} becomes a larger interval.

The definition below provides notions of point identification and partial pseudo-identification.

DEFINITION 3. *We say that parameter θ_{0} is point identified (partially pseudo-identified) from infrequent attribute values if $\Theta_{\infty}=\{\theta_{0}\}$ ($\Theta_{\infty}\neq\{\theta_{0}\}$).*

Whether the model is point identified depends on the properties of the model, the distribution of the data, and the matching procedure. Definition 3 implies that if θ_{0} is

point identified, then at infinity we can construct only one combined data subset using a chosen matching decision rule and that all the matches are correct ($\Pi^{\infty} = \{0\}$). If for a chosen $h(\cdot)$ in the definition of the distance $r(\cdot)$ parameter θ_0 is point identified in the sense of Definition 3, then θ_0 is point identified under any other choice of function $h(\cdot)$ that satisfies (4.14), and (4.15).

If the parameter of interest is only partially pseudo-identified from infrequent attribute values, then Θ_{∞} is the best approximation to θ_0 in the limit in terms of the distance $r(\cdot)$ under a chosen $h(\cdot)$. In this case, Θ_{∞} is sensitive to the choice of $h(\cdot)$ and W_0 and in general will be different for different $r(\cdot)$ satisfying (4.14) and (4.15). In the case of partial pseudo-identification, $0 \in \Pi^{\infty}$ implies that $\theta_0 \in \Theta_0$, but otherwise θ_0 does not necessarily belong to Θ_0.

Our next step is to analyze identification from combined data sets obtained using a decision rule that honors a particular bound on the risk of individual disclosure. Having the bound on the risk of individual disclosure does not mean that making a correct match in a particular dataset is impossible. What it implies is that there will be multiple versions of a combined dataset. One of these versions can correspond to the "true" dataset for which $d_{ji} = m_{ij}$ (using the notation from Section 3). However, as is clear from our discussion before, in addition to this dataset we can also construct combined datasets with varying fractions of incorrect matches. This implies that for any x and y, and any $\mathcal{D}^x = \{x_i, z_i^x\}_{i=1}^{N^x}$ that contains x as one of the values x_i, and any $\mathcal{D}^y = \{y_j, z_j^y\}_{j=1}^{N^y}$ that contains y as one of the values y_j, we have that if $\inf_{i,j} \pi^N\left(y_j = y, x_i = x, \{y_j, z_j^y\}_{j=1}^{N^y}, \{x_i, z_i^x\}_{i=1}^{N^x}\right) > 0$ if $\pi^N\left(y_j = y, x_i = x, \{y_j, z_j^y\}_{j=1}^{N^y}, \{x_i, z_i^x\}_{i=1}^{N^x}\right)$ is defined.

Condition (3.9) in the definition of the disclosure risk implies that

$$\inf_{x,y} \liminf_{N^y \to \infty} \pi^N(y, x) \geq \underline{\gamma}.$$

Taking into account Assumptions 3 (i)-(ii) for $\alpha_N \to 0$ and the property of our data combination procedure – namely, that the values of y_j and x_i are not taken into account in matching (y_j, z_j^y) with (x_i, z_i^x) and it only matters whether identifiers satisfy conditions $|z_i^x - z_j^y| < \alpha_N$ and $|z_i^x| > 1/\alpha_N$, – we obtain that the limit of $\pi^N(y, x)$ does not depend on the value of y and x. Denote this limit as π. Uniformity over $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ in Assumptions 3 (i)-(ii) imply that π is the uniform limit of $\pi^N(y, x)$:

$$\sup_{y \in \mathcal{Y}, x \in \mathcal{X}} |\pi^N(y, x) - \pi| \to 0 \quad \text{as } N^y \to \infty.$$

If the only information released by the data curator about the disclosure risk is a bound $\underline{\gamma}$, then the researcher can only infer that $\pi \geq \underline{\gamma}$, that is, $\Pi^{\infty} = [\underline{\gamma}, 1]$. This fact will allow us to establish results on point (partial pseudo-) identification of θ_0 in Theorem 1 (Theorem 2).

Theorems 1 and 2 below link point identification and partial pseudo-identification with the risk of disclosure.

THEOREM 1. ***(Point identification of*** θ_0***).*** *Suppose Assumptions 1-3 hold. Let* $\alpha_N \to 0$ *as* $N^y \to \infty$ *in such a way that*

$$\inf_{x\in\mathcal{X},x\in\mathcal{Y}} \inf_{\mathcal{D}^x,\mathcal{D}^y} \inf_{i,j} p_{ij}^N(x,y,\mathcal{D}^x,\mathcal{D}^y) \to 1 \quad \text{as } N^y \to \infty.$$

Then θ_0 *is point identified from matches of infrequent values of the attributes.*

Proof. Condition

$$\lim_{N^y\to\infty} \inf_{x\in\mathcal{X},y\in\mathcal{Y}} \inf_{\mathcal{D}^x,\mathcal{D}^y} \inf_{i,j} p_{ij}^N(x,y,\mathcal{D}^x,\mathcal{D}^y) = 1$$

can equivalently be written as

$$\lim_{N^y\to\infty} \sup_{x\in\mathcal{X},y\in\mathcal{Y}} \sup_{\mathcal{D}^x,\mathcal{D}^y} \sup_{i,j} \left(1 - p_{ij}^N(x,y,\mathcal{D}^x,\mathcal{D}^y)\right) = 0,$$

which means that for any $\varepsilon > 0$, when N^x and N^y are large enough, $\sup_{x\in\mathcal{X},y\in\mathcal{Y}} \pi^N(y,x) < \varepsilon$. Since $\varepsilon > 0$ can be chosen arbitrarily small, we obtain that

$$\lim_{N^y\to\infty} \sup_{x\in\mathcal{X},y\in\mathcal{Y}} \pi^N(y,x) = 0.$$

From here we can conclude that $\Pi^\infty = \{0\}$, and hence, $\Theta_\infty = \{\theta_0\}$. □

As we can see, Theorem 1 provides the identification result when there is no bound imposed on disclosure risk. The rates of the sequences of thresholds for which the condition of this theorem is satisfied are established in Section 3.

Theorem 2 gives a partial pseudo-identification result when data combination rules are restricted to those that honor a given bound on the disclosure risk and follows from our discussion earlier in this section.

THEOREM 2. ***(Absence of point identification of*** θ_0***).*** *Suppose Assumptions 1-3 hold. Let* $\alpha_N \to 0$ *as* $N \to \infty$ *in such a way that there is a bound* $\underline{\gamma} > 0$ *imposed on the disclosure risk. Then* θ_0 *is only partially pseudo-identified from the combined dataset which is constructed by applying the data combination rules that honor the bound* $\underline{\gamma} > 0$.

Proof. As discussed earlier in this section, in this case $\Pi^\infty = [\underline{\gamma}, 1]$, and thus,

$$\Theta_\infty = \bigcup_{\pi\in[\underline{\gamma},1]} \operatorname*{Argmin}_{\theta\in\Theta} r\left(\pi E\left[\rho(Y,X;\theta) \,\middle|\, X = x\right] + (1-\pi)E^*\left[\rho(\widetilde{Y},\, X;\theta)|X = x\right]\right).$$

In general, $r\left(\pi E\left[\rho(Y,X;\theta) \,\middle|\, X = x\right] + (1-\pi)E^*\left[\rho(\widetilde{Y},\, X;\theta)|X = x\right]\right)$ is minimized at different values for different π meaning that generally Θ_∞ is not a singleton. □

Using the result of Theorem 2, we are able to provide a clear characterization of the identified set in the linear case.

COROLLARY 1. *Consider a linear model with θ_0 defined by*

$$E[Y - X'\theta_0 | X = x] = 0,$$

where $E[XX']$ has full rank. Suppose Assumptions 2 and 3 hold, and there is a bound $\underline{\gamma} > 0$ on the disclosure risk. Then θ_0 is only partially pseudo-identified from matches on infrequent attribute values, and, under the distance $r(\cdot)$ chosen in the spirit of least squares, the pseudo-identified set is the following collection of convex combinations of parameters θ_0 and θ_1:

$$\Theta_\infty = \{\theta_\pi, \pi \in [\underline{\gamma}, 1] \ : \ \theta_\pi = (1-\pi)\theta_0 + \pi\theta_1\},$$

where θ_1 is the parameter obtained under the complete independence of X and Y.

The proof of Corollary 1 is in the Appendix.

Note that $\theta_0 = E_X[XX']^{-1}E[XY]$. The matrix $E[XX']$ can be found from the marginal distribution of X (we write $E_X[]$ to emphasize this fact) and, thus, is identified without any matching procedure. The value of $E[XY]$, however, can be found only if the joint distribution of (Y, X) is known in the limit – that is, only if there is no non-disclosure guarantee.

When we consider *independent* X and Y with distributions f_X and f_Y, we have $E^*[X(Y - X'\theta)] = 0$. Solving the last equation, we obtain

$$\theta_1 = E_X[XX']^{-1}E_X[X]E_Y[Y], \tag{4.19}$$

which can be found from split samples without using any matching methodology. When the combined data contain a positive proportion of incorrect matches in the limit, the resulting value of θ is a mixture of two values obtained in two extreme situations: θ_0 when $\pi = 0$, and θ_1 when $\pi = 1$.

The next example illustrates that the pseudo-identified set Θ_∞, even if $\theta_0 \notin \Theta_\infty$ is informative about the true parameter value of θ_0.

EXAMPLE 4. *As a special case, consider a bivariate linear regression model*

$$E[Y - a_0 - b_0 X | X = x] = 0,$$

where $Var[X] > 0$. Using our previous calculations, we obtain that the pseudo-identified set for the slope coefficient is

$$\{b_\pi \ : \ b_\pi = (1-\pi)b_0, \ \ \pi \in [\underline{\gamma}, 1]\}$$

because $b_1 = 0$. Here we can see that we are able to learn the sign of b_0, and in addition to the sign, we can conclude that $|b_0| \geq \frac{b_\pi}{1-\underline{\gamma}}$. This result is much more than we were able to learn about b_0 in Example 1.

The pseudo-identified set for the intercept is

$$\{a_\pi \ : \ a_\pi = (1-\pi)a_0 + \pi E_Y[Y], \ \ \pi \in [\underline{\gamma}, 1]\} = \{a_\pi \ : \ a_\pi = E_Y[Y] - (1-\pi)b_0 E_X[X], \ \ \pi \in [\underline{\gamma}, 1]\}.$$

Thus far, we have shown that using a high quality data combination rule that selects observations with infrequent values of some attributes allows us to point identify the parameters of the econometric model. However, given that we may be using a small subset of individuals to estimate the model, the obtained estimates may reveal sensitive information on those individuals. To prevent this, the data curator can decide to conduct the data linkage in a way that guarantees a bound on the risk of disclosure. As we have seen however, in this case it is generally not possible to point identify the parameter of interest, and the pseudo-identified set that can be obtained from the data does not generally contain the true parameter value.

4.1 Computational illustration

Experiment 1 Consider the bivariate regression model

$$Y = \alpha_0 + \beta_0 X + \epsilon,$$

where $X \sim \mathcal{N}(0,1)$, $\epsilon \sim \mathcal{N}(0,1)$ and $(\alpha_0, \beta_0) = (1,1)$. Consider the original identifiers $V = W$ that are both generated from Poisson distribution with parameter $\lambda = 10$ independently of X and ε. So in the "raw" design all matches are one-to-one.

We then split a sample into subsamples of the observations of Y in the first one and X in the second one.

Our goal now is to construct identifiers Z^x and Z^y and construct approximations for the joint distributions $F^N(Y, X)$. To do so in each simulation draw we draw the "raw" sample of size N.

Then we consider sample $\{V_i\}_{i=1}^N$ and split the interval $[\min_i V_i, \max_i V_i]$ into the segments of the same length. We consider three designs for this:

(A) The bin most to the right contains a single observation.

(B) The bin most to the right contains at least two observations.

(C) The bin most to the right contains at least three observations.

Then we set $Z_i^x = Z_i^y$ and equal to the average V in the bin that contains V_i.

By an infrequent event we understand the event when observation i is in the bin most to the right. Thus, the linkage of data will be based on observations in that bin only. We will

say that the values of quasi-identifiers Z_i^x and Z_j^y are close if $Z_i^x = Z_j^y$. This corresponds to considering the threshold $\alpha_N = \frac{1}{\min_i\{V_i : V_i \text{ is in the bin most to the right}\}-0.5}$ and then using the linkage rule based on $\mathbf{1}\{|Z_i^x| > \frac{1}{\alpha_N}, |Z_i^x - Z_j^y| < \alpha_N\}$, as described in the main text of the paper.

Then in each Monte Carlo draw $m = 1, \ldots, M$, we construct the empirical joint distribution $F_m^N(y, x) = \frac{1}{N_m} \sum \mathbf{1}\{Y_j \leq y\}\mathbf{1}\{X_i \leq x\}$ for the matched sample of pairs, and empirical marginal distributions $F_{m,Y}^N(y) = \frac{1}{N} \sum_{j=1}^N \mathbf{1}\{Y_j \leq y\}$ and $F_{m,X}^N(x) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{X_i \leq x\}$ from split datasets. The approximate joint and marginal distributions of interest are computed from the Monte Carlo sample as simple averages

$$\widehat{F}^N(y, x) = \frac{1}{M} \sum_{m=1}^M F_m^N(y, x), \quad \widehat{F}_Y^N(y) = \frac{1}{M} \sum_{m=1}^M F_{m,Y}^N(y), \quad \widehat{F}_X^N(x) = \frac{1}{M} \sum_{m=1}^M F_{m,X}^N(x),$$

respectively. Then we construct the moment vector with two equations

$$\frac{\int\int y\, x\, \widehat{F}^N(dy,\, dx) - \int y\, \widehat{F}_Y^N(dy) \int x\, \widehat{F}_X^N(dx)}{\int x^2\, \widehat{F}_X^N(dx) - \left(\int x\, \widehat{F}_X^N(dx)\right)^2} = \tilde{\beta},$$

and

$$\int y\, \widehat{F}_Y^N(dy) - \tilde{\beta} \int x\, \widehat{F}_X^N(dx) = \tilde{\alpha}.$$

Then we solve this system of equations for $\tilde{\alpha}$ and $\tilde{\beta}$.

In order to construct pseudo-identified sets, in each scenario we proceed in the following way:

(A) For each simulation we construct only one bin that is most to the right – the bin with a single observation.[9] This corresponds to the case of the lower bound guarantee $\underline{\gamma} = 0$ $\Pi^N = \{0\}$. This is the case when the identity disclosure is not guaranteed

(B) For each simulation we consider a series of bins that are most to the right – starting from the case when that bin contain only two observations (this corresponds to the case $\gamma = \frac{1}{2}$) and ending with the case when that bin contains all the observations (this corresponds to the case $\gamma = 1$). Overall, this described the situation of the lower bound guarantee $\underline{\gamma} = \frac{1}{2}$ and $\Pi^N = [\frac{1}{2}, 1]$.

(C) For each simulation we consider a series of bins that are most to the right – starting from the case when that bin contain only three observations (this corresponds to the case $\gamma = \frac{2}{3}$) and ending with the case when that bin contains all the observations (this corresponds to the case $\gamma = 1$). Overall, this described the situation of the lower bound guarantee $\underline{\gamma} = \frac{2}{3}$ and $\Pi^N = [\frac{2}{3}, 1]$.

[9] If we draw a sample that contains several observations with $\max_i V_i$, then we re-draw this sample until we have only one observation with the maximum value of the variable V

The results of the experiment are illustrated in Figure 1. The left panel in Figure 1 shows the N-pseudo-identified sets in scenarios (A), (B) and (C), respectively, obtained for $N = 1000$ (with $M = 200$ simulations). Thus, the left panel in Figure 1 looks at the situation from the perspective of the data curator (primary user of the dataset).

The right panel in Figure 1 describes what a secondary user (that is, a researcher) can learn about the true parameter in all three scenarios from just *one* combined dataset released to her, where the proportion of correct matches is between 0 and $1-\underline{\gamma}$ in scenarios (B) and (C) (we choose this proportion randomly on $[0, 1-\underline{\gamma}]$). As discussed in Example 4, we can learn the sign of b_0, which in our example we learn to be positive, and then, in addition to the sign, we can conclude make a conclusion about the range that $b_0 \geq \frac{b_\pi}{1-\underline{\gamma}}$, and then find a respective range for α_0. Those ranges are illustrated in the second and third graphs in the right panel in Figure 1, for scenarios (B) and (C), respectively (in that figure the proportion of correct matches is somewhere in the middle of $[0, 1-\underline{\gamma}]$).

Experiment 2 Now we analyze the extension of the regression model to the case of instrumental variables. We consider the regression model

$$Y = \alpha_0 + \beta_0 X^* + \epsilon,$$

where X^* is not observed. What is observed is its error-ridden version $X = X^* + .1\xi$, where $X^* \sim N(0,1)$, $(\epsilon, \xi)^T \sim N(0, I_2)$, and $(\epsilon, \xi)^T \perp X^*$. We want to use the IV estimator with the excluded instrument $Z^1 = 0.8X^* + 0.1u$, where $u \sim N(0,1)$, $u \perp \epsilon, \xi, X^*$. Let $(\alpha_0, \beta_0) = (1, 1)$. Just as in Experiment 1, we consider the original identifiers $V = W$ that are both generated from Poisson distribution with parameter $\lambda = 10$ independently of anything in the model. So in the "raw" design all matches are one-to-one.

We then split a sample into subsamples of the observations of Y in the first one and (X, Z) in the second one.

Our goal now is to construct the quasi-identifiers Z^z and Z^y and construct approximations for the joint distribution $F^N(Y, Z)$. We consider the same three scenarios (A), (B), (C) as in Experiment 1 and construct blocks and quasi-identifiers in the same way as there.

In each Monte Carlo draw m we construct the empirical distribution function $F^N_m(y, z) = \frac{1}{N_{obs}} \sum \mathbf{1}\{Y_j \leq y\}\mathbf{1}\{Z_i \leq z\}$ for the linked sample of pairs of Y and Z, and empirical distributions $F^N_{m,Y}(y) = \frac{1}{N} \sum_{j=1}^N \mathbf{1}\{Y_j \leq y\}$ and $F^N_{m,XZ}(x, z) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{X_i \leq x\}\mathbf{1}\{Z_i \leq z\}$ from split datasets. The approximate distributions of interest are computed from the Monte Carlo sample as simple averages

$$\widehat{F}^N(y, z) = \frac{1}{M} \sum_{m=1}^M F^N_m(y, z), \quad \widehat{F}^N_Y(y) = \frac{1}{M} \sum_{m=1}^M F^N_{m,Y}(y), \quad \widehat{F}^N_{XZ}(x, z) = \frac{1}{M} \sum_{m=1}^M F^N_{m,XZ}(x, z).$$

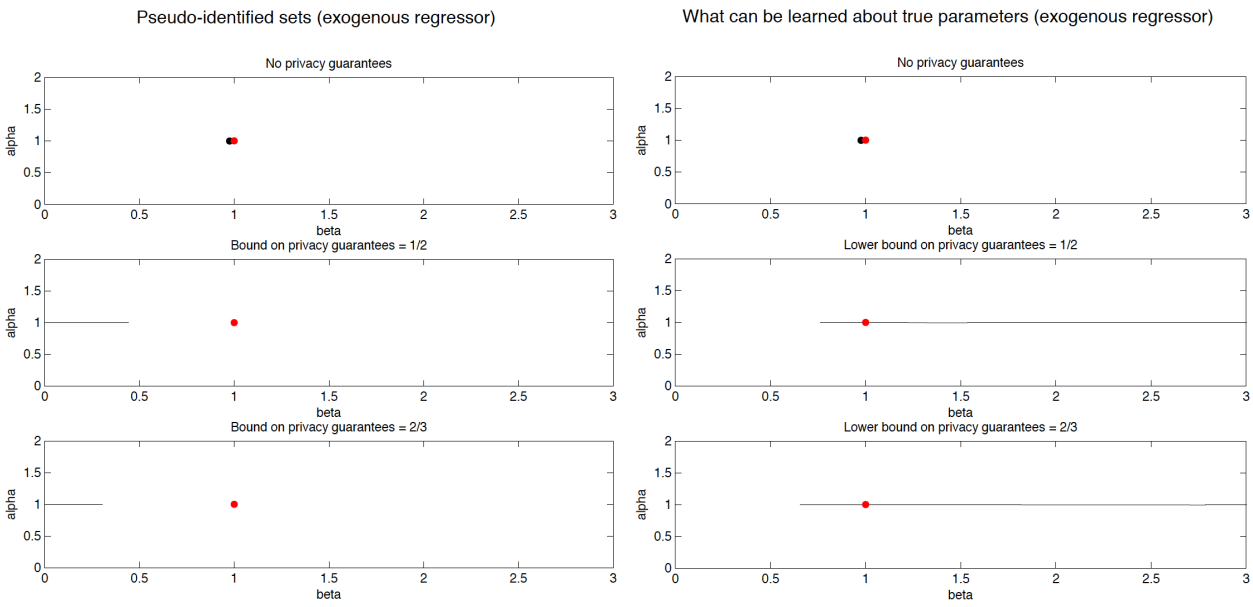

Figure 1. Monte Carlo simulation in the case of an exogenous regressor. On the left-hand side – pseudo-identified sets. On the right-hand side – what can be learned about the true parameter.

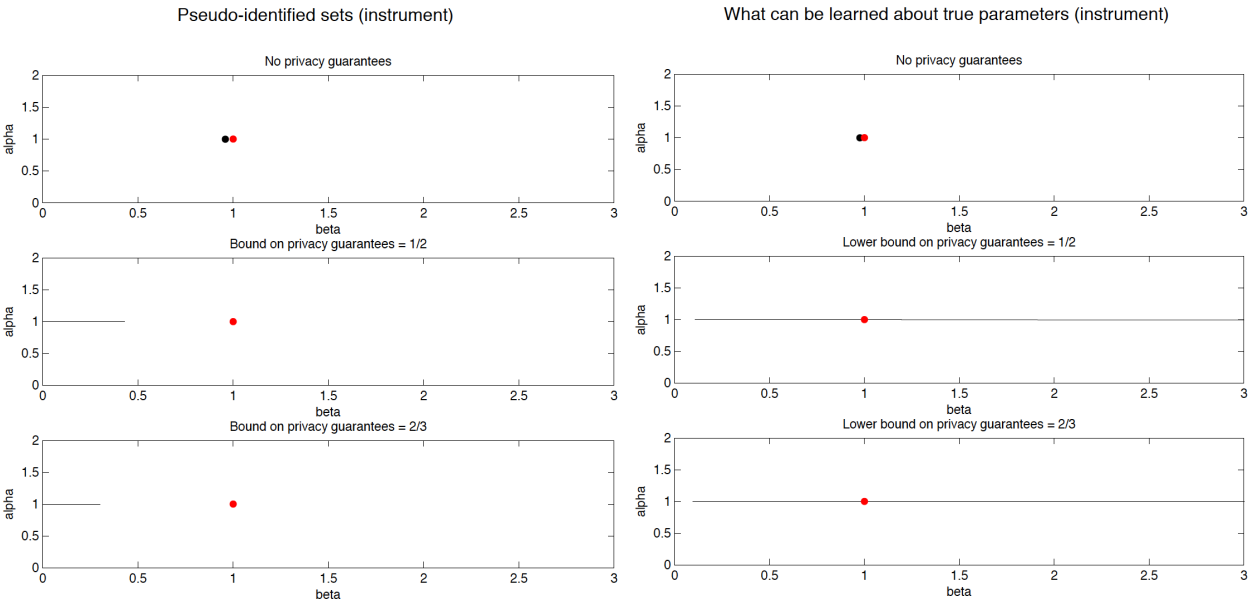

Figure 2. Monte Carlo simulation in the case of an endogenous regressor and a strong instrument. On the left-hand side – pseudo-identified sets. On the right-hand side – what can be learned about the true parameter.

Then we construct the moment vector with two equations

$$\frac{\int\int y\,z\,\widehat{F}^N(dy,\,dz) - \int y\,\widehat{F}^N_Y(dy)\int\int x\,\widehat{F}^N_{XZ}(dx,\,dz)}{\int\int x\,z\,\widehat{F}^N_{XZ}(dx,\,dz) - \int\int x\,\widehat{F}^N_{XZ}(dx,\,dz)\int\int z\,\widehat{F}^N_{XZ}(dx,\,dz)} = \tilde{\beta},$$

and

$$\int y\, \widehat{F}_Y^N(dy) - \tilde{\beta} \int \int x\, \widehat{F}_{XZ}^N(dx,\, dz) = \tilde{\alpha}.$$

Then solve this system of equations for $\tilde{\alpha}$ and $\tilde{\beta}$.

In order to construct pseudo-identified sets, in each scenario we proceed just as we did in Experiment 1. The results of the experiment are illustrated in Figure 2. The left panel of Figure 2 shows the N-pseudo-identified sets in scenarios (A), (B) and (C), respectively, obtained for $N = 1000$ (with $M = 200$ simulations). Thus, the left panel of Figure 2 looks at the situation from the perspective of the data curator.

The right panel of Figure 2 describes what a a researcher as a secondary user of the data can learn about the true parameter in all three scenarios from just *one* combined dataset released to her, where the proportion of correct matches is between 0 and $1-\underline{\gamma}$ in scenarios (B) and (C) (we choose this proportion randomly on $[0, 1-\underline{\gamma}]$). Analogously to Experiment 1, we can learn the sign of b_0, which in our example we learn to be positive, and then, in addition to the sign, we can conclude make a conclusion about the range that $b_0 \geq \frac{b_{\pi,IV}}{1-\underline{\gamma}}$, and then find a respective range for α_0. Those ranges are illustrated in the second and third graphs on the right panel in Figure 2, for scenarios (B) and (C), respectively (in that figure the proportion of correct matches is close to zero).

5 Empirical example

In our theoretical analysis we focus on the tradeoff between the quality of identification of the empirical model from the combined Economic data and the potential privacy threats that arise from data linkage. If it is possible to identify the model of interest, that means that there exist "high quality" links between the combined datasets.

In the context where one or both of the combined datasets contain sensitive information, the combined records can be significantly more sensitive. We illustrate this idea and demonstrate the impact of the data security constraints on the identification of the econometric model using the example of gender-based discrimination.

The anecdotal evidence from the recent news publications indicates that a common practice in the Christian Orthodox religious communities in Central Russia and in the Muslim communities of the Caucasus republics of Russia it is a commonplace practice to withdraw children from schooling (which is mandatory in Russia) and common preventive medical procedures (such as vaccinations). The press reports that this practice is disproportionally applied to females. Our goal is to empirically study the presence of this practice.

The clear difficulty that would arise if we were to use aggregate data to address these questions is that there exist group effects (that are correlated with the religious affiliation) not accounting for which can significantly bias the analysis. As a result, for analysis we need to combine the data that contains family-level demographics with the religious census.

Our main source of the household-level characteristics is the Russian Longitudinal Mon-

itoring Survey (RLMS).[10] The RLMS is a nationally representative annual longitudinal survey that covers more than 4,000 households (that include between 1900 and 3682 children), starting from 1992 and the last available years is 2014. RLMS provides a survey of a broad set of variables, including a variety of individual demographic characteristics, health information, employment and income data. The data are collected from 33 Russian regions, which include 31 large regions (equivalents of counties in the United States), as well as two largest cities of Moscow and St. Petersburg. The religious census (conducted by Rosstat, the equivalent of the Census Bureau in Russia) indicates that 2 out of 33 regions are dominated by individuals who identify themselves as Muslim, while in the remaining regions the majority of the population identify as Orthodox Christians.

Due to extremely low population mobility in Russia, the group effects are localized geographically. In the context of the RLMS data this has been documented in Yakovlev [2017], where the group effects were associated with the "neighborhood" effects indicating the common component in the behavior and characteristics of households from the same geographical area. To identify the neighborhood effects we use the RLMS data on the neighborhood identifiers, that were available for researchers in initial years the survey was conducted (also referred to as rounds). These identifiers are available up to year 2009 while in the later rounds these identifiers were withheld due to privacy considerations. The RLMS covers households within clusters of small neighborhoods (referred to as census districts by Rosstat). The information on these small neighborhood identifiers allows one combine the data from the RLMS with the data from Rosstat that contains neighborhood characteristics, such as the predominant religious affiliation.

The empirical question that we analyze is the impact of the religious affiliation of a family on the number of completed classes of mandatory schooling. We are particularly interested in whether the number of completed grades differs for males and females. In other words, how likely it is that females may be withheld from school by their parents for religious reasons.

From the perspective of the privacy analysis, our goal is to see how the obfuscation of small neighborhood identifiers can impact the identification of the causal effect of interest. First, we consider the *status quo* situation where the actual neighborhoods are aggregated to regions and thus eah neighborhood can be confused with 10-15 other neighborhoods within the same region corresponding to k-anonymity with k equal to the total number of neighborhoods in the region. Then we consider a hypothetical situation of of $k = 2$-anonymity: we combine the data from individual neighborhoods into pairs of neighborhoods within the same region.

[10]This survey is conducted by the Carolina Population Center at the University of North Carolina at Chapel Hill, and by the Higher School of Economics in Moscow. The official source name is "Russian Longitudinal Monitoring survey, RLMS-HSE," conducted by Higher School of Economics and ZAO *Demoscope* jointly with Carolina Population Center, University of North Carolina at Chapel Hill and the Institute of Sociology RAS.

In the context of model specification, we want to determine the importance of the neighborhood effects and determine to which extent the observed disparities in schooling are affected by differential treatment of males and females in families as opposed to just reflect the difference in school attendance across different neighborhoods. To do this, we estimate two empirical models where the unit of observation is each child. The first empirical model analyzes the number of completed grades in school as a function of child's gender and age, religious affiliation of the family as well as demographic characteristics of the family. The second model adds neighborhood characteristics to the first model.

After we estimated the "infeasible" model that uses the neighborhood identifiers, we proceed to implement the "feasible" procedure. This application combines (i) the actual data combination procedure; (ii) the empirical characterization of functions and parameters used in the theorems and assumptions. For our data coming from the RLMS this imply the illustration of the actual situation where the data curator suppressed the ability of researchers for data linkage for privacy considerations.

For each household in the data we take available demographic information: size of the household, number of children, age of the head of the household, gender of the head of household, income, education, occupation. We take each region (oblast, republic, etc.) at a time and perform clustering of households using the demographic variables and known average neighborhood demographics that are de-linked from the individual demographic data. The distance function that we use for such a clustering is

$$d(x,y) = \sqrt{\sum_{k=1}^{K}(x_k - y_k)^2/\sigma_k^2},$$

where K is the number of demographic variables used for clustering and σ_k^2 is the overall sample variance of a given variable. The points are selected into a given cluster simply by verifying the distance between a given household and the nearest average characteristics of the neighborhood is smaller than the pre-specified threshold $1/\alpha_N$. We restrict the number of clusters to be smaller than the ("infeasibly known" to us) the maximum number of neighborhoods in a region. Each recovered cluster will be associated with the "inferred neighborhood."

Set constant α_N such that the number of households over all clusters for which the minimum over all neighborhoods included in the region $d(x, x^c) < 1/\alpha_N$ constitutes fraction $\theta = .9$ of the samples. We use this data-driven definition of α_N to construct a "scale-free" measure of frequency and proximity of observations. Our notion of "infrequent" observations is slightly changed from the theoretical definition and we now focus on the set of observations that are the closest to the mean characteristics of a given neighborhood. To do this in practice, we drop all the points from each cluster for each $d(x, x^c) > 1/\alpha_N$ and for each remaining household, call its cluster identifier the "inferred neighborhood." Now we

re-run our main two models but using the subsample of points that satisfy $d(x, x^c) < 1/\alpha_N$ and using their "inferred" neighborhoods instead of the true neighborhoods.

Then we consider the case of $k = 2$-anonymity. To do that take the neighborhood identifiers and randomly select pairs of neighborhoods within each region without replacement and create a new identifier that now corresponds to the pair (rather than the individual neighborhoods). If there is a neighborhood left without a pair, we randomly join it with any of the selected pairs within the region.

After that we again isolate the clusters of the neighborhoods within each aggregated neighborhood using our distance criterion. Then we estimate our two specifications of the econometric model using the data across all the clusters.

Estimation results for each model are presented in Table 1. In our models the religious affiliations are dummies for each household, Income stands for the household income, Share College is the share of the individuals in the household who have a college degree and City is the dummy indicating that a given neighborhood is within a city. We notice that in the model that does not take the group effects into account the estimates indicate the potential adverse effect of both considered religious denominations on the school completion by women. The effect appears to be stronger for Muslim families where a female child has a lower average number of school grades completed. The model that does take the group effects into account shows a different picture. While the significant effect of religious affiliation of the household on the schooling of females is still present for Orthodox Christian households, it disappears for households that identify as Muslim. This can partly be explained by the lower overall school completion rates in the traditionally Muslim parts of Russia. Our estimate, therefore, indicate that it is important to have neighborhood identification of households to estimate the true causal effect in the Econometric model.

In Table 1 the unit of observation is a child in a given round of the RLMS. The dependent variable is the number of grade completed. Models 1,3 and 5 are estimated without accounting for the neighbothood identifiers. Models 2,4 and 6 account for neighborhoods. Model 2 uses actual neighborhoods, while Models 4 and 6 use neighborhoods constructed from matches between the individual data and actual neighborhoods. The sample in Models 3 and 4 is restricted to the observations where the weighted Euclidian distance between the average neighborhood characteristis and individual households is smaller than the threshold that eliminates 10% of the households overall that are too far from the distance. The sample in Models 5 and 6 is restricted to the observations where the weighted Euclidian distance between the average neighborhood characteristis and individual households is smaller than the threshold that eliminates 10% of the households overall that are too far from the distance, but instead the data on grouped neighborhoods are available that preserve 2-anonymity.

The analysis of the results for the data combination procedure with the actual data and the case of 2-anonymity demonstrates that the previously observed pattern where we

Table 1. The impact of religious affiliation, family and neighborhood characteristics on school completion

Variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Female × Muslim	-0.2464 ***	-0.0491	-0.1728***	-0.1720	-0.1333**	0.0293
	(0.051)	(0.104)	(0.066)	(0.131)	(0.064)	(0.126)
Female × Orthodox	-0.1 ***	-0.0883 **	-0.0699	-0.0577	-0.0949**	-0.0998**
	(0.033)	(0.037)	(0.043)	(0.046)	(0.042)	(0.045)
Female	0.1385 ***	0.1279 ***	0.1403***	0.1351***	0.1272***	0.1143***
	(0.018)	(0.019)	(0.023)	(0.024)	(0.022)	(0.023)
Orthodox	0.0023	0.0149	0.0174	0.0060	0.0268	0.0316
	(0.023)	(0.026)	[0.032)	(0.030)	(0.032)	
Muslim	0.2326 ***	-0.0246	0.3302***	0.0400	0.3319***	-0.0063
	(0.036)	(0.074)	(0.045)	(0.091)	(0.045)	(0.092)
Share College	0.1566 ***	0.2163 ***	.2052***	0.2391***	0.1901***	0.2206***
	(0.023)	(0.024)	(0.029)	(0.029)	(0.029)	(0.029)
log(Income)	0.0216 ***	0.0146 ***	0.0244***	0.0162***	0.0249***	0.0171***
	(0.004)	(0.004)	0.004)	(0.005)	(0.004)	(0.004)
City	-0.0452 ***	-0.0178	-0.0640***	-0.0325	-0.0779***	-0.0466**
	(0.017)	(0.018)	0.022)	(0.022)	(0.021)	(0.021)
Child's Age	1.0449 ***	1.0464 ***	1.0626***	1.0620***	1.0621***	1.0617***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Female × Share College	-0.0926 ***	-0.1079 ***	-0.1181***	-0.1220***	-0.1042**	-0.1043***
	(0.032)	(0.033)	(0.042)	(0.041)	(0.040)	(0.040)
Female × log(Income)	-0.0126 **	-0.013 **	-0.0127**	-0.0125**	-0.0089	-0.0092
	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
Female × City	-0.0257	-0.0208	-0.0195	-0.0154	-0.0208	-0.0159
	(0.025)	(0.026)	(0.031)	(0.031)	(0.031)	(0.031)
Female × Muslim neighb		-0.2171 *		-0.0054		-0.1986
		(0.120)		(0.151)		(0.144)
Female × Orthodox neighb		-0.0338		-0.0332		0.0239
		(0.075)		(0.092)		(0.090)
Muslim neighb		0.3607 ***		0.3717***		0.4173***
		(0.086)		(0.106)		(0.106)
Orthodox neighb		0.0868 *		0.1096*		0.0510
		(0.052)		(0.065)		(0.062)
log(Avg. Neighb Income)		-0.063 ***		-0.0448***		-0.0437***
		(0.004)		(0.005)		(0.005)
Constant	-7.6509 ***	-7.6561 ***	-7.7849***	-7.7694***	-7.7669***	-7.7581***
	(0.035)	(0.037)	(0.047)	(0.047)	(0.045)	(0.045)
Observations	13,580	12,349	8,218	8,216	8,742	8,741
$\mathbf{R}^2$	0.898	0.899	0.896	0.898	0.895	0.897

Robust standard errors are given in parentheses. *** indicates the significance of a given variable on 1% significance level, ** on 5% level, and * on a 10% level.

observe a significant negative effect of a muslim family on the years of schooling for female without controling for neighborhood religios affiliation and do not observe this effect in case where we control for the dominating religion of the neighborhood remains in place. However, we also observe that the negative significant effect of the orthodox families (observed even controlling for the neighborhood effect in the "infeasible" estimation) vanishes

with the use of the actual data and only becomes significant in case of 2-anonymity. This clearly indicates that privacy constraints can significantly affect the model estimates (and, therefore, the policy implications).

We note also, that the effects of several other demographic characteristics (unrelated to religion) remain consistent through the model and preserve both the sign and the general magnitude.

Recall that we construct an adaptive clustering procedure where the threshold α_N used in our theoretical analysis is chosen such that a fraction θ of households overall are dropped from the sample for being "too far" from any average neighborhood characteristics. While Table 1 reports the results for $\theta = .9$, we also analyze the cases of $\theta = .5$ and $\theta = .1$.

To illustrate the performance of our data combination procedure, we report the empirical analog of the parameter π_N corresponding to the expected fraction of the correctly identified mathed observations over a distribution of combined datasets. To do this for each inferred neighborhood we count the number of households that indeed belong to the same neighborhood. On figures 3 -5 we illustrate the impact of the stringency of the data combination constraint and the degree of anonymity of the data on the quality of matches by showing the distribution of the number of correct matches across neighborhoods.

We note that in general, with the actual data the modal number of correct matches is 1 per per neighborhood. This pattern is generally preserved for all choices of the fraction of dropped observations. However, in the dataset with 2-anonymity, the modal number of correctly identified matches varies between 2 and 3.

6 Conclusion

In this paper we analyze an important problem of identification of econometric model from the split sample data without common numeric variables. Data combination with combined string an numeric variables requires the measures of proximity between strings, which we borrow from the data mining literature. Model identification from combined data cannot be established using the traditional machinery as the population distributions only characterize the marginal distribution of the data in the split samples without providing the guidance regarding the joint data distribution. As a result, we need to embed the data combination procedure (which is an intrinsically finite sample procedure) into the identification argument. Then the model identification can be defined in terms of the limit of the sequence of parameters inferred from the samples with increasing sizes. We discover, however, that in order to provide identification, one needs to establish some strong links between the two databases. The presence of these links means that the identities of the corresponding individuals will be disclosed with a very high probability.

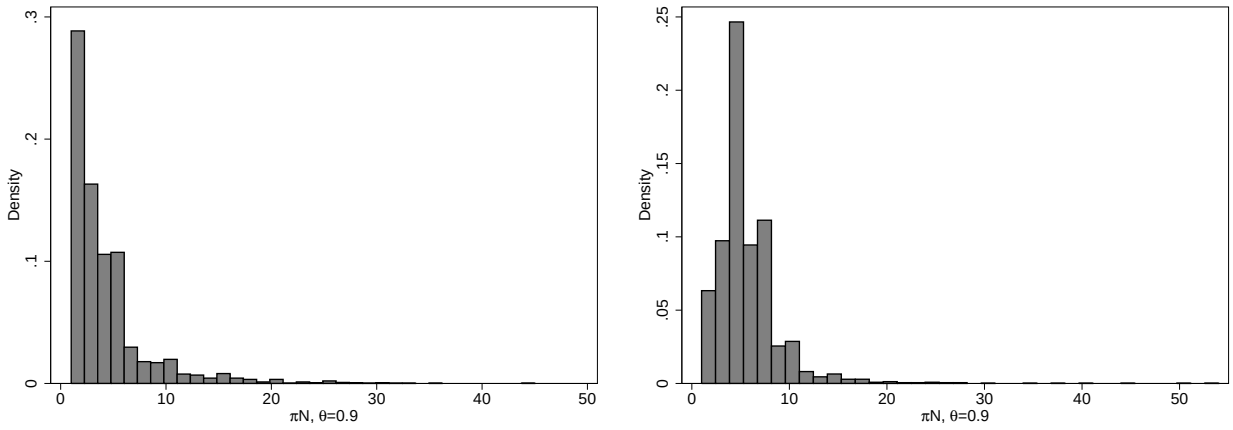

Figure 3. Empirical π_N for the data combination procedure ($\theta = .9$). On the left: with the actual data. On the right: with 2-anonymity.

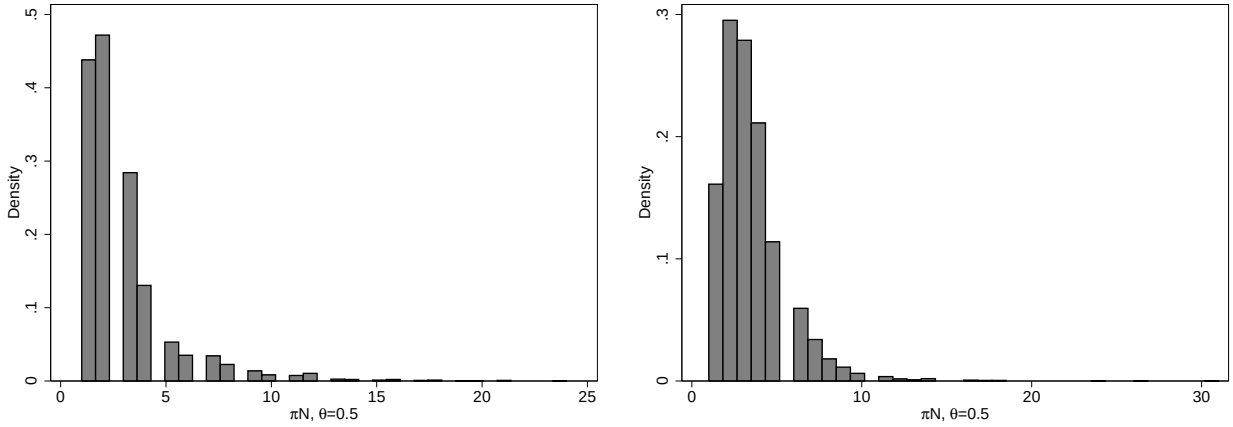

Figure 4. Empirical π_N for the data combination procedure ($\theta = .5$). On the left: with the actual data. On the right: with 2-anonymity.

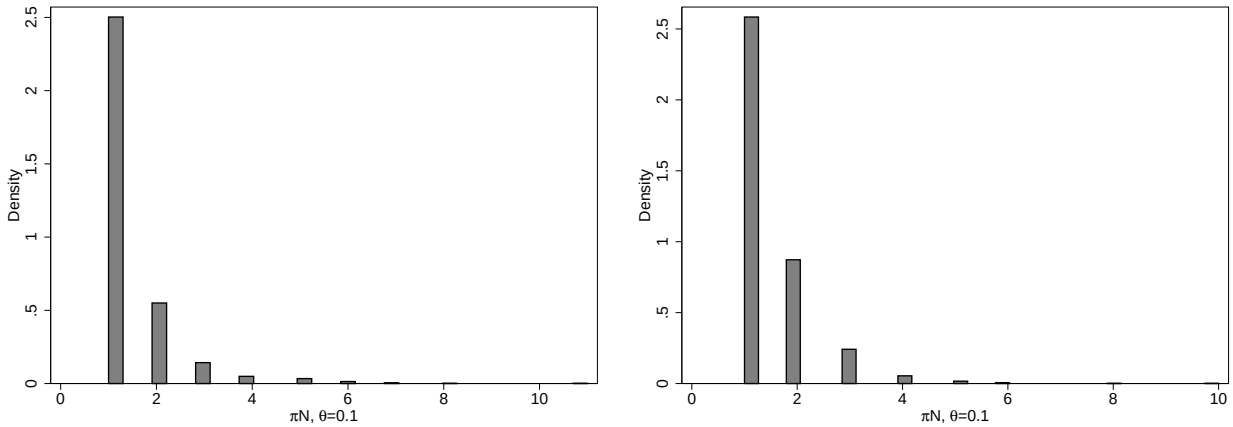

Figure 5. Empirical π_N for the data combination procedure ($\theta = .1$). On the left: with the actual data. On the right: with 2-anonymity.

References

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

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

Acquisti, A. (2004): "Privacy and security of personal information," *Economics of Information Security*, 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*, 363–377.

Acquisti, A. and H. Varian (2005): "Conditioning prices on purchase history," *Marketing Science*, 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, 1468–1488.

Brickell, J. and V. Shmatikov (2008): "The cost of privacy: destruction of data-mining utility in anonymized data publishing," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 70–78.

Calzolari, G. and A. Pavan (2006): "On the optimality of privacy in sequential contracting," *Journal of Economic Theory*, 130, 168–204.

Chipperfield, J. O., G. Bishop, P. D. Campbell, et al. (2011): "Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data," .

Ciriani, V., S. di Vimercati, S. Foresti, and P. Samarati (2007): "k-Anonymity," *Secure Data Management in Decentralized Systems. Springer-Verlag.*

Cross, P. and C. Manski (2002): "Regressions, Short and Long," *Econometrica*, 70, 357–368.

Dominguez, M. and I. Lobato (2004): "Consistent estimation of models defined by conditional moment restrictions," *Econometrica*, 72, 1601–1615.

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*, 135–166.

Duncan, G. and D. Lambert (1986): "Disclosure-limited data dissemination," *Journal of the American statistical association*, 81, 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*, 219–232.

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

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

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, 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*, ACM, 71–80.

Horowitz, J. and C. Manski (1995): "Identification and robustness with contaminated and corrupted data," *Econometrica*, 63, 281–302.

Hu, J., J. P. Reiter, and Q. Wang (2014): "Disclosure risk evaluation for fully synthetic data," in *Privacy in Statistical Databases*, ed. by J. Domingo-Ferrer, Heidelberg: Springer.

Karr, A. F., C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil (2006): "A framework for evaluating the utility of data altered to protect confidentiality," *The American Statistician*, 60, 224–232.

Kim, G. and R. Chambers (2012): "Regression analysis under incomplete linkage," *Computational Statistics & Data Analysis*, 56, 2756–2770.

King, G. (1997): *A solution to the ecological inference problem: reconstructing individual behavior from aggregate data*, Princeton University Press.

Kinney, S. K., J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd (2011): "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database," *International Statistical Review*, 79, 362–384.

Komarova, T., D. Nekipelov, and E. Yakovlev (2015): "Estimation of Treatment Effects from Combined Data: Identification versus Data Security," in *Economic Analysis of the Digital Economy*, ed. by A. Goldfarb, S. Greenstein, and C. Tucker, Chicago: The University of Chicago Press.

Lahiri, P. and M. Larsen (2005): "Regression analysis with linked data," *Journal of the American statistical association*, 100, 222–230.

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

Larsen, M. D. (2005): "Hierarchical bayesian record linkage theory," *Iowa State University, Statistics*.

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*, ACM, 49–60.

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

Lindell, Y. and B. Pinkas (2000): "Privacy preserving data mining," in *Advances in CryptologyCRYPTO 2000*, Springer, 36–54.

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

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

——— (2007): *Identification for prediction and decision*, Harvard University Press.

Manski, C. and E. Tamer (2002): "Inference on regressions with interval data on a regressor or outcome," *Econometrica*, 70, 519–546.

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

Molinari, F. (2008): "Partial identification of probability distributions with misclassified data," *Journal of Econometrics*, 144, 81–117.

Pacini, D. (2016): "Two-sample least squares projection," *Econometric Reviews*, forthcoming.

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," Tech. rep., Citeseer.

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

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

Tancredi, A., B. Liseo, et al. (2011): "A hierarchical Bayesian approach to record linkage and population size problems," *The Annals of Applied Statistics*, 5, 1553–1585.

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

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

Woo, M., J. P. Reiter, A. Oganian, and A. F. Karr (2009): "Global measures of data utility for microdata masked for disclosure limitation," *Journal of Privacy and Confidentiality*, 1, 111–124.

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

Yakovlev, E. (2017): "Demand for Alcohol Consumption in Russia and Its Implication for Mortality," *American Economic Journal: Applied Economics*, forthcoming.

A Appendix

Proof of Proposition 1. Probability $p_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y)$ in (3.8) is equal to

$$\frac{\overline{p}_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y)\, p_{ij}(x, y, \mathcal{D}^x, \mathcal{D}^y)}{\overline{p}_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y)\, p_{ij}(x, y, \mathcal{D}^x, \mathcal{D}^y) + \underline{p}_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y)\ (1 - p_{ij}(x, y, \mathcal{D}^x, \mathcal{D}^y))}, \tag{A.20}$$

where

$$p_{ij}(x,y,\mathcal{D}^x,\mathcal{D}^y) = Pr(m_{ij}=1 \,|\, x_i=x, y_j=y, \mathcal{D}^x, \mathcal{D}^y),$$
$$\overline{p}_{ij}^N(x,y,\mathcal{D}^x,\mathcal{D}^y) = Pr\left(|z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N} \,|\, m_{ij}=1, x_i=x, y_j=y, \mathcal{D}^x, \mathcal{D}^y\right),$$
$$\underline{p}_{ij}^N(x,y,\mathcal{D}^x,\mathcal{D}^y) = Pr\left(|z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N} \,|\, m_{ij}=0, x_i=x, y_j=y, \mathcal{D}^x, \mathcal{D}^y\right).$$

Note that $Pr(m_{ij}=1 \,|\, x_i=x, y_j=y, \mathcal{D}^x, \mathcal{D}^y) = \frac{1}{N^x}$. By Assumption 3, for $\alpha_N \in (0, \bar{\alpha})$,

$$\inf_{\mathcal{D}^x,\mathcal{D}^y} \inf_{i,j} \overline{p}_{ij}^N(x,y,\mathcal{D}^x,\mathcal{D}^y) \geq (1-\alpha_N)(\phi(\alpha_N) + o(\phi(\alpha_N))).$$

Therefore, $\inf_{\mathcal{D}^x,\mathcal{D}^y} \inf_{i,j} p_{ij}^N(x,y,\mathcal{D}^x,\mathcal{D}^y)$ is bounded from below by

$$\frac{(1-\alpha_N)(\phi(\alpha_N) + o(\phi(\alpha_N)))\frac{1}{N^x}}{(1-\alpha_N)(\phi(\alpha_N) + o(\phi(\alpha_N)))\frac{1}{N^x} + \sup_{\mathcal{D}^x,\mathcal{D}^y} \sup_{i,j} \underline{p}_{ij}^N(x,y,\mathcal{D}^x,\mathcal{D}^y)}.$$

The last ratio will converge to 1 as $N^y \to \infty$ if $\frac{N^x}{\phi(\alpha_N)} \sup_{\mathcal{D}^x,\mathcal{D}^y} \sup_{i,j} \underline{p}_{ij}^N(x,y,\mathcal{D}^x,\mathcal{D}^y)$ converges to 0. Note that

$$\underline{p}_{ij}^N(x,y,\mathcal{D}^x,\mathcal{D}^y) = \int\limits_{|z_i^x|>\frac{1}{\alpha_N}} \int_{z_i^x-\alpha_N}^{z_i^x+\alpha_N} f_{Z^y|Y}(z_j^y|y_j=y) f_{Z^x|X}(z_i^x|x_i=x)\, dz_j^y\, dz_i^x.$$

From Assumption 3, for small α_N,

$$\underline{p}_{ij}^N(x,y,\mathcal{D}^x,\mathcal{D}^y) = \int\limits_{|z_i^x|>\frac{1}{\alpha_N}} \left(\psi\left(\frac{1}{|z_i^x|-\alpha_N}\right) - \psi\left(\frac{1}{|z_i^x|+\alpha_N}\right)\right)(1+o_y(1)) g_1(|z_i^x|)(1+o_{xz^x}(1))\, dz_i^x,$$

where $\sup_{|z_i^x|>\frac{1}{\alpha_N}} \sup_{x_i\in\mathcal{X}} |o_{xz^x}(1)| \to 0$ and $\sup_{y_i\in\mathcal{Y}} |o_y(1)| \to 0$ as $\alpha_N \to 0$. Thus, for any x and y,

$$\frac{N^x}{\phi(\alpha_N)} \sup_{\mathcal{D}^x,\mathcal{D}^y} \sup_{i,j} \underline{p}_{ij}^N(x,y,\mathcal{D}^x,\mathcal{D}^y) \leq \frac{N^x}{\phi(\alpha_N)} \int\limits_{|z|>\frac{1}{\alpha_N}} \left(\psi\left(\frac{1}{|z|-\alpha_N}\right) - \psi\left(\frac{1}{|z|+\alpha_N}\right)\right) g_1(|z|)\, dz +$$

$$+ \left(\sup_{|z_i^x|>\frac{1}{\alpha_N}} \sup_{x_i\in\mathcal{X}} |o_{xz^x}(1)| + \sup_{y_i\in\mathcal{Y}} |o_y(1)|\right) \frac{N^x}{\phi(\alpha_N)} \int\limits_{|z|>\frac{1}{\alpha_N}} \left(\psi\left(\frac{1}{|z|-\alpha_N}\right) - \psi\left(\frac{1}{|z|+\alpha_N}\right)\right) g_1(|z|)\, dz.$$

Taking into account the relationship between $g_1(z)$ and $\phi\left(\frac{1}{z}\right)$, we obtain the result in the proposition. □

Proof of Proposition 2. This result of this proposition obviously follows from Proposition 1 because $\sup_{\mathcal{D}^x,\mathcal{D}^y} \sup_{i,j} p_{ij}^N(x,y,\mathcal{D}^x,\mathcal{D}^y) \geq \inf_{\mathcal{D}^x,\mathcal{D}^y} \inf_{i,j} p_{ij}^N(x,y,\mathcal{D}^x,\mathcal{D}^y)$. □

Proof of Proposition 3. From (A.20), using Assumption 3 obtain that for $\alpha_N \in (0, \bar{\alpha})$

$$p_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y) \leq \frac{1}{1 + \frac{N^x}{\phi(\alpha_N) + o_{xy}(\phi(\alpha_N))} \left(1 - \frac{1}{N^x}\right) \underline{p}_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y)},$$

and, thus,

$$\sup_{\mathcal{D}^x, \mathcal{D}^y} \sup_{i,j} p_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y) \leq \frac{1}{1 + \frac{N^x}{\phi(\alpha_N) + o_{xy}(\phi(\alpha_N))} \left(1 - \frac{1}{N^x}\right) \inf_{\mathcal{D}^x, \mathcal{D}^y} \inf_{i,j} \underline{p}_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y)}.$$

Now obtain that $\sup_{x,y} \sup_{\mathcal{D}^x, \mathcal{D}^y} \sup_{i,j} p_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y)$ will be bounded away from 1 as $N^y \to \infty$ if

$$\frac{N^x}{\phi(\alpha_N)} \inf_{x,y} \inf_{\mathcal{D}^x, \mathcal{D}^y} \inf_{i,j} \underline{p}_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y)$$

is bounded away from 0 as $N^y \to \infty$, that is, if

$$\liminf_{N^y \to \infty} \frac{N^x}{\phi(\alpha_N)} \inf_{x,y} \inf_{\mathcal{D}^x, \mathcal{D}^y} \inf_{i,j} \underline{p}_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y) > 0. \tag{A.21}$$

Using the expression for $\underline{p}_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y)$ from the proof of Proposition 1, for small α_N obtain

$$\begin{aligned} \frac{N^x}{\phi(\alpha_N)} \underline{p}_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y) \geq & \left(1 - \sup_{|z_i^x| > \frac{1}{\alpha_N}} \sup_{x_i \in \mathcal{X}} |o_{xz^x}(1)| - \sup_{y_i \in \mathcal{Y}} |o_y(1)|\right) \times \\ & \frac{N^x}{\phi(\alpha_N)} \int_{|z_i^x| > \frac{1}{\alpha_N}} \left(\psi\left(\frac{1}{|z_i^x| - \alpha_N}\right) - \psi\left(\frac{1}{|z_i^x| + \alpha_N}\right)\right) g_1(|z_i^x|)\, dz_i^x. \end{aligned}$$

Clearly, the expression on the right-hand side of the last inequality is also a lower bound for $\frac{N^x}{\phi(\alpha_N)} \inf_{x,y} \inf_{\mathcal{D}^x, \mathcal{D}^y} \inf_{i,j} \underline{p}_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y)$. Taking into account the relationship between $g_1(z)$ and $\phi\left(\frac{1}{z}\right)$, and the fact that $\sup_{|z_i^x| > \frac{1}{\alpha_N}} \sup_{x_i \in \mathcal{X}} |o_{xz^x}(1)| \to 0$ and $\sup_{y_i \in \mathcal{Y}} |o_y(1)| \to 0$ as $\alpha_N \to 0$, we obtain that the condition (3.11) guarantees then that (A.21) holds. □

Proof of Proposition 4. Using Assumption 3 (iii) and the law of iterated expectations,

$$\begin{aligned} & E\left[1\left(|Z^x| > \frac{1}{\alpha}, |Z^x - Z^y| < \alpha\right) \rho(Y, X; \theta) \mid X = x\right] = \\ & E\left[E\left[1\left(|Z^x| > \frac{1}{\alpha}, |Z^x - Z^y| < \alpha\right) \rho(Y, X; \theta) \mid X = x, Z^x = z^x, Z^y = z^y\right] \mid X = x\right] = \\ & E\left[1\left(|Z^x| > \frac{1}{\alpha}, |Z^x - Z^y| < \alpha\right) E\left[\rho(Y, X; \theta) \mid X = x, Z^x = z^x, Z^y = z^y\right] \mid X = x\right] = \\ & E\left[1\left(|Z^x| > \frac{1}{\alpha}, |Z^x - Z^y| < \alpha\right) E\left[\rho(Y, X; \theta) \mid X = x\right] \mid X = x\right] = \\ & E\left[1\left(|Z^x| > \frac{1}{\alpha}, |Z^x - Z^y| < \alpha\right) \mid X = x\right] \cdot E\left[\rho(Y, X; \theta) \mid X = x\right]. \end{aligned}$$

By Assumption 3 (i) and (ii),

$$E\left[1\left(|Z^x| > \frac{1}{\alpha}, |Z^x - Z^y| < \alpha\right) \mid X = x\right] > 0.$$

This implies

$$\frac{E\left[1\left(|Z^x| > \frac{1}{\alpha}, |Z^x - Z^y| < \alpha\right)\rho(Y, X; \theta) \mid X = x\right]}{E\left[1\left(|Z^x| > \frac{1}{\alpha}, |Z^x - Z^y| < \alpha\right) \mid X = x\right]} = E\left[\rho(Y, X; \theta) \mid X = x\right],$$

which is equivalent to (4.12). □

Proof of Corollary 1. Here $\rho(Y, X, \theta) = Y - X'\theta$. From the conditional moment restriction we obtain that $E\left[X(Y - X'\theta_0)\right] = 0$ and, thus, $\theta_0 = E_X[XX']^{-1}E[XY]$. When $\widetilde{Y}$ is drawn from $f_Y(\cdot)$ independently of X, then $E^*\left[X(\widetilde{Y} - X'\theta_1)\right] = 0$ gives $\theta_1 = E_X[XX']^{-1}E_X[X]E_Y[\widetilde{Y}]$.

As established in Theorem 2, the identified set is

$$\Theta_\infty = \bigcup_{\pi \in [\underline{\gamma}, 1]} \underset{\theta \in \Theta}{\text{Argmin}}\ r\left(\pi E\left[\rho(Y, X; \theta) \,\middle|\, X = x\right] + (1 - \pi)E^*\left[\rho(\widetilde{Y}, X; \theta) | X = x\right]\right).$$

Here $\rho(Y, X, \theta) = Y - X'\theta$. In the spirit of least squares, let us choose instruments $h(X) = X$ and consider the distance

$$r\left(\pi E\left[\rho(Y, X; \theta) \,\middle|\, X = x\right] + (1 - \pi)E^*\left[\rho(\widetilde{Y}, X; \theta) | X = x\right]\right) = g_\pi(\theta)' g_\pi(\theta),$$

where

$$g_\pi(\theta) = (1 - \pi)E[X(Y - X'\theta)] + \pi E^*[X(\widetilde{Y} - X'\theta)].$$

Note that

$$\begin{aligned} g_\pi(\theta) &= (1 - \pi)E[XY] - (1 - \pi)E_X[XX']\theta + \pi E_X[X]E_Y[\widetilde{Y}] - \pi E_X[XX']\theta \\ &= (1 - \pi)E[XY] + \pi E_X[X]E_Y[Y] - E_X[XX']\theta \\ &= E_X[XX']\left((1 - \pi)E_X[XX']^{-1}E[XY] + \pi E_X[XX']^{-1}E_X[X]E_Y[Y] - \theta\right) \\ &= E_X[XX']\left((1 - \pi)\theta_0 + \pi\theta_1 - \theta\right). \end{aligned}$$

Clearly, $g_\pi(\theta)' g_\pi(\theta)$ takes the value of 0 if and only if $g_\pi(\theta)$ takes the value of 0, which happens if and only if $\theta = (1 - \pi)\theta_0 + \pi\theta_1$. Thus for each $\pi \in [\underline{\gamma}, 1]$,

$$\theta_\pi = (1 - \pi)\theta_0 + \pi\theta_1$$

is the unique minimizer of $r\left(\pi E\left[\rho(Y, X; \theta) \,\middle|\, X = x\right] + (1 - \pi)E^*\left[\rho(\widetilde{Y}, X; \theta) | X = x\right]\right)$. Therefore,

$$\Theta_\infty = \left\{\theta_\pi, \pi \in [\underline{\gamma}, 1]\ :\ \theta_\pi = (1 - \pi)\theta_0 + \pi\theta_1\right\}. \quad \square$$