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

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

This version: February 10, 2014

ABSTRACT

Data combination is routinely used by researchers and businesses to improve the quality of data to deliver more precise inference in data-driven decisions and policies. The original raw data may be enriched with additional information to help deal with the omitted variables problem, control for measurement errors, or account for sample selection. 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. When one of the datasets used in the combination procedure contains sensitive individual information, in the combined dataset the identity information gets supplemented with previously anonymized information. We define the sensitivity of the combined data by the *quality* of the link between the two datasets for a particular individual and consider the notion of individual disclosure as the existence of a link with a quality exceeding a certain threshold (meaning that one can conclude with a high confidence that the constructed link is correct). The main question of our paper is *whether parametric inference from combined data is possible when it is guaranteed that the data combination procedure does not lead to statistical disclosure*. 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 P. Haile, M. Jansson, C. Manski, A. de Paula, J. Powell, C. Tucker and E.Tamer.

[2] Department of Economics, London School of Economics and Political Science

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

[4] New Economic School, Moscow, Russia

1 Introduction

Data combination is a vital step in the comprehensive analysis of industrial and government data and resulting policy decisions. Typical industrial data are contained in large, well-indexed databases and combining 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 matched 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 record errors. Given that most industrial databases have a homogenous structure, prediction algorithms can be "trained" using a dataset of manually resolved matching errors and those algorithms can be further used for error control. 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, 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 disclosure and privacy considerations. As a result, data combination tasks rarely reduce to a simple merge 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 may contain individual-level variables that either the individuals in the data or data curators may find inappropriate to disclose to the general public.

In this paper we consider the case where the "main" dataset used for inference contains potentially sensitive but anonymized individual information. However, for the purpose of dealing with the omitted variable problem, this dataset is combined with an auxiliary dataset that does not contain sensitive personal data but contains personal identifiers (for instance, names and addresses). In this case, data combination leads to combination of sensitive anonymized individual information with individual identifiers. We associate an event when we are able to match an individual identifier to the sensitive data for a particular individual with a very high probability with *individual disclosure*. If we measure the quality of the match as an appropriately defined distance between the observations in two separate datasets, the most valuable observations are those for which this distance is small. This means that once we use the combined data for inference, these are the observations that will make the highest contribution to parameter inference. Our main research question is whether

[1] See, e.g. Wright (2010) and Bradley, Penberthy, Devers, and Holden (2010) among others.

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

identification of a parametric model from combined data is compatible with restrictions imposed on individual disclosure. Our main finding is that there is a tradeoff between the identification of the model and limitations on individual disclosure. Whenever a non-zero disclosure restriction is imposed, the model of interest that is based on the dataset combined from two separate datasets is not point identified.

The contribution of our paper is an analysis of the implications of data security and privacy protection considerations on identification of an empirical model from the data. The risk of potential disclosure is related to the probability of the recovery of private information for each individual in the dataset once the estimated model and implemented policy decisions become publicly observable. 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. In this paper, the issues of model estimation and the risk of disclosure are analyzed jointly. In particular, we investigate how the limits imposed on the risk of disclosure of confidential consumer data affect the amount of information a policy maker can obtain about the empirical model of interest.

Our important finding is that inference from the datasets that need to be combined is only well-defined in a finite sample. Data are combined individual-by-individual based on a measure of distance between the observations in the two datasets. The restrictions imposed on individual disclosure can be expressed as restrictions on the choice of the data combination procedure (such that one can provide a probabilistic guarantee that a high-quality match between any two observations will not occur). We introduce the notion of identification from combined data through a limit of the set of parameters inferred from the combined data as the size of both datasets approaches infinity. In this context, we study the identification of a semiparametric model provided that only data combination procedures that provide the non-disclosure guarantee can be used.

In our empirical application we illustrate both the data combination procedure itself and the impact of the choice of this procedure on the identification of a semiparametric model. We use review data from the Healthcare section and general business sections on Yelp.com, where Yelp users rank health care facilities based on their experiences. The data pertain to facilities located in Durham county, North Carolina. The empirical question that we address in our work is whether a Yelp.com user's visit to a doctor has an impact on the user's reviewing behavior for other businesses. However, a user profile on Yelp.com does not contain any demographic or location information about the user. Without controlling for this information, inference based solely on the review data would be prone to a selection bias because consumers who use the healthcare facilities more frequently may be more prone to writing a review. On the other hand, active Yelp users may be more likely

to review a healthcare business among other businesses. To control for sample selection using the individual-level demographic variables, we collected a database of individual property tax records in Durham county. Applying a record linkage technique from the data mining literature, we merge the health service review data with the data on individual locations and property values, which we use to control for sample selection bias. To be more precise, when combining data with the aim of bias correction we rely on observing data entries with infrequent attribute values (extracted from usernames, and individual names and locations) in the two datasets. Accurate links between these entries may disclose the identities of Yelp.com users.

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. We find that *limitations on the risk of identity disclosure lead to the loss of point identification of the model of interest*. Further, we analyze the partial identification issue and what estimates the consumer behavior model can deliver under the constraints on the identity disclosure. We provide a new approach to model identification from combined datasets as a limiting property in the sequence of statistical experiments.

An important aspect of the combination of data from the dataset that contains sensitive individual data with public information from an auxiliary, external dataset is the possibility of so-called linkage attacks. A linkage attack can be defined as decision rule that generates a link between at least one data entry in the sensitive anonymized dataset and public information that contains individual identifiers which is correct with probability exceeding a selected confidence threshold. The optimal structure of such attacks as well as the requirements in relation to data releases have been studied in the computer science literature. The structure of linkage attacks is based on the optimal record linkage results that have been long used in the analysis of databases and data mining. To some extent, these results have been used used in econometrics for combination of datasets as described in Ridder and Moffitt (2007). In several striking examples, computer scientists have shown that 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.

Overall, modern medical databases pose a large threat to individual disclosure. A dramatic example of a large individual-level database is the data from genome-wide association studies (GWAS), which provide an in-depth analysis of genetic origins of human health conditions and susceptibility to diseases. A common practice of such studies until recently has been to publish data on minor allele frequencies. The analysis of such data allows researchers to demonstrate the evidence of a genetic origin of the condition that is being studied. Homer et al. (2008) showed that by using the reported averages of the minor allele frequencies together with the publicly available single nucleotide polymorphism (SNP) dataset from the NIH HapMap they could infer the presence of an individual

with a known genotype in a mix of DNA samples. As a result, if a particular study is devoted to the analysis of a particular health condition or a disease, the discovery that a particular individual belongs to the studied subsample means that this individual has that condition or disease.

Our definition of the disclosure risk evaluates the probability of a successful linkage attack. We use what Lambert (1993) calls a *pessimistic* measure of the risk of disclosure: the maximum upper bound on the probability of linking a record in the released, anonymized sample with individual information from public data. Methods for controlling the risk of identity disclosure have been developed, so the bounds on disclosure risk are practical and enforceable.[3]

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 beem introduced that characterizes the accuracy of a statistical function that can be evaluated from the released data (e.g. see Lindell and Pinkas (2000), Brickell and Shmatikov (2008)), 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.

Although our identification approach is new, to understand the impact of the bounds on the individual disclosure risk we exploit ideas from the literature on partial identification of models with contaminated or corrupted data. 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 in-

[3]Computer science literature, e.g. Samarati and Sweeney (1998), Sweeney (2002b), Sweeney (2002a), LeFevre, DeWitt, and Ramakrishnan (2005), Aggarwal, Feder, Kenthapadi, Motwani, Panigrahy, Thomas, and Zhu (2005), LeFevre, DeWitt, and Ramakrishnan (2006), Ciriani, di Vimercati, Foresti, and Samarati (2007), 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 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. 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, Fienberg, Krishnan, Padman, and Roehrig (2001), Abowd and Woodcock (2001).

terest 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.

In this paper we find the suggested approach of constructing identified sets for the parameter of interest to be highly informative. As we show, the size of the identified set is directly linked to the pessimistic measure of the disclosure risk, and in the case of a linear model the size is directly proportional to this measure. This is a powerful result which shows that there is a direct conflict between the informativeness of the data used in the model of interest and the security of individual data. As a result, combination of the sensitive anonymized data with publicly available individual data is not compatible with the non-disclosure of individual identities.

In this paper we connect the risk of individual disclosure to the possibility of recovering the true identity of individuals in the anonymized dataset with sensitive individual information. 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, Nekipelov, and Yakovlev (2012) where we introduce the notion of a partial disclosure.

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

The rest of the paper is organized as follows. In Section 2 we describe the problem of econometric inference and characterize the structure of the data generating process. In Section 3 we describe the class of data combination rules used in this paper and demonstrate the implications of these 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 introduce the notion of identification from combined data and characterize the structure of the identified set of model parameters when one uses the data combination rules that we propose. We also analyze the relationship between the structure of the identified set and the bound on disclosure risk. In Section 5 using an empirical example we demonstrate the implications of the tradeoff between identification and disclosure protection. In Section 6 we provide final remarks and conclude.

2 Econometric model

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 $\rho(\cdot)$ 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 population joint distribution of (Y, X).*

As an empirical illustration, in Section 5 we estimate a model of consumer ratings on the online rating website Yelp.com for Yelp users located in Durham, NC, where ratings are expressed as rank scores from 1 to 5 (5 is the highest and 1 is the lowest score). Our goal is to explore the impact of a visit of a particular Yelp.com user to a local doctor on this user's subsequent rating behavior. In this context, we are concerned with potential selection induced by the correlation of rating behavior, frequency

of visits to entertainment and food businesses (disproportionately represented on Yelp.com), and patronage of health care businesses with consumer-level demographics. However, the individual demographic information on Yelp.com is limited to the self-reported user location and self-reported first name, in addition to all reviews by the user.

To obtain reliable additional demographic variables that can be used to control for sample selection, we collected an additional dataset that contains the property tax information for local taxpayers in Durham county. The data reflect the property tax paid for residential real estate along with characteristics of the property owner such as name, location, and the appraised value of the property. If we had data from Yelp.com merged individual-by-individual with the property tax records, then for each consumer review we would know both the score assigned by the consumer to the health care business and health care business and consumer characteristics. In reality, however, there is no unique identifier that labels observations in both data sources.

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 $\{y_i, w_i, x_i, v_i\}_{i=1}^n$ is a random sample from the population distribution of the data.

(iii) The observable data are formed by two independently created random data subsamples from the sample of size n such that the first data subsample is $\mathcal{D}^{yw} = \{y_j, w_j\}_{j=1}^{N^y}$ and the second subsample is $\mathcal{D}^{xv} = \{x_i, v_i\}_{i=1}^{N^x}$.

(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.

Assumption 2 characterizes the observable variables as independently drawn subsamples of the infeasible "master" dataset. Without any additional information, one can only re-construct marginal distributions $f_X(\cdot)$ and $f_Y(\cdot)$ but not the joint distribution $f_{Y,X}(\cdot)$.

EXAMPLE 1. *Without any additional information, identification in linear models 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, 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 us to draw conclusions about the joint moments of the regressor and the outcome can be summarized by the Cauchy-Schwartz inequality $|cov(Y,X)| \leq \sqrt{Var[Y]}\sqrt{Var[X]}$. *We can, therefore, determine the slope coefficient only up to a set:*

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

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 variables (or vectors) V and W is not immediately useful for the econometric model that is being estimated. However, this information helps us to construct measures of similarity between observations y_j in dataset $\mathcal{D}^{yw}$ and observations x_i in dataset $\mathcal{D}^{xv}$. W and V are the variables (or vectors) that are very likely to be highly correlated for a given individual but are uncorrelated across individuals. In our empirical example, the Yelp.com dataset contains the username of each Yelp reviewer while the property tax bills dataset has the full name of each individual taxpayer. As the first component of V we use the Yelp.com username and as the first component of W we consider the first name of the taxpayer. Other elements of V constructed from the Yelp data are the modal zip code of the businesses rated by the user, the modal cuisine (by ethnicity) of the restaurants rated by the user, and the presence of "mostly female" businesses in the ratings (such as day spa's, pilates and yoga studios and nail salons). The corresponding elements of W include the zip code of the taxable property and the following three binary variables: a) whether the first name of the taxpayer is in the list of 500 most popular white, black, and hispanic names as per 2010 US Census; b) whether the last name of taxpayer is in the list of 500 most popular white, black, and hispanic last names; c) whether the name of the taxpayer is in the list of 500 most popular female names in the US Census. For instance, we can expect that consumers tend to rate businesses that are located closer to where they live. It is also likely that the self-reported name in the user review on Yelp.com is highly correlated with her real name (the default option offered by Yelp.com generates the user name as the first name and the last name initial). We can thus consider a notion of similarity between the observations in the "anonymous" rating data on Yelp.com and the demographic variables in the property tax data. This measure of similarity will be used to combine observations in the two datasets.

We formalize the detection of "similarity characteristics" in split sample data as the construction of vector-valued functions of the data Z^x and Z^y that we expect to take similar values if observations in two datasets correspond to the same individual and to take more distant values otherwise. We can construct variables in Z^x and Z^y by "standardizing" the components of V and W in such a way that the corresponding constructed components have the same support and structure (such that, e.g., a binary component in Z^x correspond to a binary component in Z^y and the supports of the remaining components coincide in the two vectors). The construction of such classifiers is widely discussed in the modern computer science literature especially in relation to record linkage and data

recovery. In this paper, we take the procedure for the construction of such classifiers as given and illustrate a practical implementation of such a procedure in our empirical example.

In our theoretical framework, for simplicity we consider univariate Z^x and Z^y.

The following assumption describes the requirements on the data similarity characteristics.

ASSUMPTION 3. *There exist observable random variables $Z^x = Z^x(X, V)$ and $Z^y = Z^y(Y, W)$, which we call classifiers (or identifiers), and $\bar{\alpha} > 0$ such that for any $0 < \alpha < \bar{\alpha}$ the following hold:*

(i) (Proximity of identifiers with extreme values) For $x \in \mathcal{X}, y \in \mathcal{Y}$,

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

(ii) (Non-zero probability of extreme values) For $x \in \mathcal{X}, y \in \mathcal{Y}$,

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

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

(iii) (Redundancy of identifiers in the full data) For $x \in \mathcal{X}$,

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

where $f_{Y|X,Z^x,Z^y}$ denotes the conditional density of Y conditional on X, Z^x and Z^y, and $f_{Y|X}$ denotes the conditional density 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[4]

$$\lim_{|z| \to \infty} \sup_{x \in \mathcal{X}} \left| \frac{f_{Z^x|X}(z|X = x)}{g_1(z)} - 1 \right| = 0,$$
$$\lim_{|z| \to \infty} \sup_{y \in \mathcal{Y}} \left| \frac{f_{Z^y|Y}(z|Y = y)}{g_2(z)} - 1 \right| = 0,$$

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

The distributions of identifiers Z^y and Z^x are fully determined by the distributions of (Y, W) and (X, V), respectively. Z^y can be constructed for each observation in the first dataset $\mathcal{D}^{yw}$ and Z^x can be constructed for each observation in the second dataset $\mathcal{D}^{xv}$.

[4] If only one tail of distributions $f_{Z^x|X}$ and $f_{Z^y|Y}$ contains extreme values, then conditions (iv) must be satisfied only in that tail.

In our theoretical analysis we assume that the joint distribution of Y, Z^y, X and Z^x is absolutely continuous with respect to the Lebesgue measure. This is mainly done for algebraic convenience. Our analysis extends to cases when there are discrete variables among Y, Z^y, X and Z^x. We will denote the density of the random vector (Y, Z^y, X, Z^x) as f_{Y,Z^Y,X,Z^x} and denote the marginal densities of the subvectors (Y, Z^y) and (X, Z^x) as f_{Y,Z^Y} and f_{X,Z^x}, respectively.

Assumption 3 implies that the ordering of the values of variables Z^y and Z^x is meaningful. It also implies that at least one tail of the distributions of Z^x and Z^y contains extreme values. In the situation of discrete Z^y and Z^x this would mean that the distributions Z^y and Z^x have infinite supports – for instance, this is the case when Z^x and Z^y take each of integer values 1, 2, 3, ... with a positive probability. Ridder and Moffitt (2007) overview cases where *a priori* available numeric identifiers Z^y and Z^x are jointly normally distributed, but we avoid making such specific distributional assumptions.

Assumption 3 (i) is formulated in terms of the joint distribution of variables Z^y and Z^x and a distance between any two their realizations. We choose $|Z^y - Z^x|$, the absolute value of the difference between two values, as our distance measure. The assumption states that for infrequent observations – those for which the values of Z^x are in the tails of the distribution f_{Z^x} – 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. In practice, numeric identifiers are typically unavailable and the data entries that may potentially be useful for matching are often just strings of characters, such as names or zip codes. We argue that in this case a data combination procedure can be used if one chooses a suitable distance measure between non-numeric identifiers.

When the datasets can be combined based on both numeric and string-valued identifiers, we can use an appropriate distance measure between data entries by conjoining the distance measure used for strings and the Euclidean distance which can be used for the numeric data. For instance, in our empirical application we make use of such variables in individual entries as the number of reviews given by a particular user to businesses located in a specific zip code in the Yelp.com data and the zip code of an individual in the property tax data (this is our numerical identifier), as well as the name of the user (as our string-valued variables used for combination). There are multiple procedures for computationally optimal construction of individual identifiers which are based on clustering the data with some priors on the relationship between the variables in two datasets. For instance, Narayanan and Shmatikov (2008) use the collection of individual movie choices in the Netflix dataset and on imdb.com.

The econometric model is then estimated using the trimmed subset of combined pairs of observations for which the distance between the entries is below a selected threshold. Identification is achieved if our chosen identifiers, distance measure, and a threshold lead to correct matches. In Appendix A we provide a brief overview of distance measures for string data commonly used in data mining.

Functions $\phi(\cdot)$ and $\psi(\cdot)$ in Assumption 3 (ii) characterize the decay of the marginal distributions of

the constructed identifiers 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.

As illustrated in Example 2 for linear models, if Y and X are not observed jointly then variables Z^x and Z^y add useful information about the joint distribution of Y and X, allowing us to go beyond the worst case result of obtaining only Fréchet bounds for parameter θ_0.

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.

Function g_1 (g_2) in Assumption 3 (iv) describes the uniform over $x \in \mathcal{X}$ ($y \in \mathcal{Y}$) rate of the conditional density of Z^x conditional on X (Z^y conditional on Y) for extreme values of Z^x (Z^y).

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 the identifier taking a high value, the values of identifiers constructed from two datasets must be close. We illustrate this assumption with our empirical application, where we construct a categorical variable from the first names of individuals which we observe in two datasets. We can rank the names by their general frequencies in the population. Those frequencies tend to decline exponentially with the frequency rank of the name. As a result, conditioning on rare names in both datasets, we will be able to identify a specific person with a high probability. In other words, the entries with same rare name in the combined datasets are likely to correspond to the same individual.

3 Data combination: implementation and implications for identity disclosure

In this section we characterize 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.

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 joint empirical distributions of

(Y, Z^y) and (X, Z^x). Even though these two distributions provide some information about the joint distribution of Y and X, they do not fully characterize it. As a result, the econometric model of interest is not identified from the available information because θ_0 cannot be recovered from (2.1) uniquely and can only be determined up to a set by computing, for example, Fréchet bounds. The identification of the econometric model is only possible if the two datasets are combined for at least some observations. Because data combination is inherently a finite-sample procedure, we define identification from combined data as a property of the limit of statistical experiments. To the best of our knowledge, this is a new approach to parameter identification from combined data.

The decision rule that we construct allows us to take two disjoint datasets that need to be combined and construct a new dataset using a subset of observations from each of the disjoint datasets. We now describe a class of data combination procedures.

The original disjoint datasets $\mathcal{D}^y$ and $\mathcal{D}^x$ are the samples $\{y_j, z_j^y\}_{j=1}^{N^y}$ and $\{x_i, z_i^x\}_{i=1}^{N^x}$. Provided that the indices 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.

Notation. *Let* $N = (N^y, N^x)$.

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 $N^x \to \infty$.

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

Let $\mathcal{D}_N(\cdot)$ denote the decision rule used for data combination. This decision rule is a mapping from the entries in two disjoint databases into the binary decision outcomes. If $\mathcal{D}_N(y_j, z_j^y, x_i, z_i^x) = 1$, we consider entries i and j as possibly corresponding to the same individual. If $\mathcal{D}_N(y_j, z_j^y, x_i, z_i^x) = 0$, then we either consider them as corresponding to different individuals or we remain uncertain regarding these two entries belonging to the same individual.

The decision rule constructed in this way is deterministic. We base it on the postulated properties of the distribution of observation identifiers in Assumption 3:

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

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 $N^y \to \infty$. Provided that the focus of this paper is identification rather than estimation in the context of data combination, in the remainder of the paper, discussion about a data combination rule referes the whole class of data combination rules characterized by the threshold sequences with a given rate.

This decision rule uses identifiers Z^y and Z^x to gain some knowledge about the joint distribution of (Y, X) in the following way. If for a data entry j from the dataset $\mathcal{D}^y$ we find a data entry from the dataset $\mathcal{D}^x$ such that $|z_i^x| > 1/\alpha_N$ and $|z_j^y - z_i^x| < \alpha_N$, 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, for Z^x and Z^y with the joint distribution f_{Z^x,Z^y}, the conditional probability of them 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 $Pr(m_{ij} = 1 \mid \mathcal{D}_N(y_j, x_i, z_j^y, z_i^x))$ depends on these values and can therefore differ across pairs of i and j. For simplicity, the notation $Pr(m_{ij} = 1 \mid \mathcal{D}_N(y_j, x_i, z_j^y, z_i^x))$ suppresses such dependence on x_i and y_j.

Using the combination rule $\mathcal{D}_N(\cdot)$, for each $j \in \{1, \ldots, N^y\}$ from the database $\mathcal{D}^y$ we find observation i from the database $\mathcal{D}^x$ that satisfies our matching criteria and thus presents a potential match for j. We then add (y_j, z_j^y, x_i, z_i^x) to our combined dataset. If there are several possible matches i for some j, we put only one of them (arbitrarily chosen) in our combined dataset. Analogously, each i can be added to our combined dataset with at most one j. Mathematically, each combined dataset $\mathcal{D}^N$ 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 $j = 1, \ldots, N^y$, $\sum_{i=1}^{N^x} d_{ji} \leq 1$.

(c) For each $i = 1, \ldots, N^x$, $\sum_{j=1}^{N^y} d_{ji} \leq 1$.

Because some j in $\mathcal{D}^y$ or some i in $\mathcal{D}^x$ can have several possible matches, several different combined datasets can be constructed. In our identification approach we will take into account all of them.

Consider an observation i such that $|z_i^x| \geq 1/\alpha_N$. We may make two kinds of errors when finding entry i's counterpart in the dataset $\mathcal{D}^y$. Data combination errors of the first kind occur when we make an incorrect match – that is, when we believe that observations i and j correspond to the same individual, but in fact they do not. The probability of the error of this kind can be expressed as

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

where (X, Z^x) and $(\widetilde{Y}, \widetilde{Z^y})$ are independent random vectors with the densities f_{X,Z^x} and f_{Y,Z^y}, respectively. The errors of data combination of the second kind occur when observations i and j belong to the same individual but our procedure does not identify these two observations as a match (we still consider i such that $|z_i^x| \geq 1/\alpha_N$). The probability of error of this kind is

$$Pr\left(|Z^x - Z^y| \geq \alpha_N \;\middle|\; |Z^x| > 1/\alpha_N, X = x_i, Y = y_j\right),$$

where (Y, X, Z^x, Z^y) is distributed with the density f_{Y,X,Z^x,Z^y}. By Assumption 3, the size of this kind of error vanishes as $\alpha_N \to 0$.

What we notice so far is that, given that there is no readily available reliable similarity metric between the two databases, we rely on the probabilistic properties of the data. As a result, we have to resort

to only using the pairs of combined observations for which the correct match is identified with a sufficiently high probability. This poses a potential problem, especially if one of the two datasets contains sensitive individual-level information. In fact, if the main dataset contains de-personalized but highly sensitive individual data and the auxiliary dataset that needs to be combined with the main dataset contains publicly available individual-level information (such as demographic data, names and addresses, rtc.), then the combined dataset contains highly sensitive personal information together with publicly available demographic identifiers at least for some individuals. The only way of avoid information leakage is to control the accuracy of utilized data combination procedures. In particular, we consider controlling the error of the first kind in data combinations. Propositions 2 and 5, 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 $N^y \to \infty$. Propositions 3 and 6 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 $N^y \to \infty$.

As discussed previously, under our matching rule the conditional probability

$$p_{ij}^N(x) = Pr\left(m_{ij} = 1 \mid x_i = x, |z_i^x| > \frac{1}{\alpha_N}, |z_j^y - z_i^x| < \alpha_N\right) \tag{3.2}$$

of a successful match of (y_j, z_j^y) from $\mathcal{D}^y$ with (x_i, z_i^x) from $\mathcal{D}^x$ in general depends on the values of x_i and y_j. For the same pair (y_j, z_j^y) and (x_i, z_i^x) this probability may be different if other finite datasets that contain (y_j, z_j^y) and (x_i, z_i^x) are considered. According to our discussion, potential privacy threats occur when one establishes that a particular combined data pair is correct with high probability. This is the idea that we use to define the notion of the risk of 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 as the maximum probability of a successful linkage attack over all individuals in a database.

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

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

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

$$\sup_{x \in \mathcal{X}} \lim_{N^y \to \infty} \sup_{j,i} p_{ij}^N(x) \leq 1 - \underline{\gamma}. \tag{3.3}$$

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

Our definition of the disclosure guarantee requires that for any two finite datasets $\mathcal{D}^y$ and $\mathcal{D}^x$ and any matched pair, the value of $p_{ij}^N(x)$ is strictly less than one. In other words, there is always a positive probability of making a matching mistake. Even if probabilities $p_{ij}^N(x)$ are strictly less than 1, they can be very high for sufficiently small α_N. If this happens, it means that a pair of entries in two databases correspond to the same individual with a very high level of confidence and that

the linkage attack on a database is likely to be successful. Moreover, Assumption 3 implies that these probabilities may approach 1 arbitrarily closely if $\alpha_N \to 0$ at a certain rate as $N^y \to \infty$. Our definition of the disclosure guarantee requires that such situations do not arise.

We emphasize that the risk of disclosure needs to be controlled in any size dataset with any realization of the values of the covariates. In other words, one needs to provide an *ad omnia* guarantee that the probability of a successful match will not exceed the provided bound. This requirement is very different from the guarantee with probability one, as here we need to ensure that even for the datasets that may be observed with an extremely low probability, the match probability honors the imposed bound. For example, if the limit of $p_{ij}^N(x)$ is equal to $1-\bar{\gamma}$, then for any dataset incorrect matches occur with probability at least $\bar{\gamma}$, and the value of $\bar{\gamma}$ is thus the extent of non-disclosure risk guarantee. This means that in any dataset of size N there must be at least $O(N\bar{\gamma})$ matches per observation.[5] This clearly describes the relation between the data combination procedure, which determines $p_{ij}^N(x)$, and the risk of disclosure.

An important practical question to address is whether the (classes of the) decision rules that assure a particular bound for the disclosure risk exist. Below we present the results that indicate, first, that for a given bound on the disclosure risk we can find the sequence of thresholds such that the corresponding decision rule will honor a particular bound for the disclosure risk, and second, that the rates of convergence for these sequences depend on the tail behavior of identifiers used for data combination. In the main text we consider two important cases where the tails of the distributions of identifiers are geometric and exponential. In the Appendix we also discuss generalizations for arbitrary functions $\phi(\cdot)$ and $\psi(\cdot)$.

PROPOSITION 1. *Suppose for $\alpha \in (0, \bar{\alpha})$, $\phi(\alpha) = b_1\alpha^{c_1}$, $b_1, c_1 > 0$ and $\psi(\alpha) = b_2\alpha^{c_2}$, $b_2, c_2 > 0$. Let $\alpha_N > 0$ be chosen in such a way that*

$$\alpha_N = o\left(\frac{1}{(N^x)^{1/c_2}}\right) \tag{3.4}$$

as $N^y \to \infty$.

Then for any $x \in \mathcal{X}$,

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

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.

[5] As a result, for some very small datasets the bound will be attained trivially. For instance, if $\bar{\gamma} = 0.1$ and each matched dataset has 2 elements, then to provide the disclosure risk guarantee, each element must have 2 elements in the other datasets as matches. This means that the actual probability of an incorrect match is 1/2.

PROPOSITION 3. ***(Non-disclosure risk guarantee).*** *Suppose for* $\alpha \in (0, \bar{\alpha})$, $\phi(\alpha) = b_1 \alpha^{c_1}$, $b_1, c_1 > 0$, *and* $\psi(\alpha) = b_2 \alpha^{c_2}$, $b_2, c_2 > 0$. *Let* $\alpha_N > 0$ *be chosen in such a way that*

$$\begin{aligned}
&\lim_{N^y \to \infty} \alpha_N \left(N^x\right)^{\frac{1}{c_2+4}} > 0 \quad \text{if } c_1 \geq 1, c_2 \geq 1 \\
&\lim_{N^y \to \infty} \alpha_N \left(N^x\right)^{\frac{1}{c_2+6}} > 0 \quad \text{if } c_1 \geq 1, 0 < c_2 < 1, \text{ or } 0 < c_1 < 1, c_2 \geq 1 \\
&\lim_{N^y \to \infty} \alpha_N \left(N^x\right)^{\frac{1}{c_2+8}} > 0 \quad \text{if } 0 < c_1 < 1, 0 < c_2 < 1
\end{aligned} \tag{3.5}$$

Then non-disclosure is guaranteed.

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 shrinks towards zero as the sample size increases. Consider two threshold sequences α_N and α_N^* where the first sequence converges to zero much faster than the second one such that $\frac{\alpha_N^*}{\alpha_N} \to \infty$. Clearly, for any sizes of the datasets $\mathcal{D}^y$ and $\mathcal{D}^x$, the second sequence allows more observations to be included in the combined dataset. In fact, all observations with the values of the constructed identifiers z_i^x between $\frac{1}{\alpha_N}$ and $\frac{1}{\alpha_N^*}$ will be rejected by the decision rule implied by the first sequence and may not be rejected by the decision rule implied by the second sequence. In addition, the second sequence is much more liberal in its definition of the proximity between the identifiers z_j^y and z_i^x constructed in $\mathcal{D}^y$ and $\mathcal{D}^x$, respectively. As a result, the decision rule implied by the second sequence generates larger combined datasets. Because the matching information in $(-\frac{1}{\alpha_N}, -\frac{1}{\alpha_N^*}) \cup (\frac{1}{\alpha_N^*}, \frac{1}{\alpha_N})$ is less reliable than that in $(-\infty, -\frac{1}{\alpha_N}) \cup (\frac{1}{\alpha_N}, \infty)$ and matches for observations with larger distances between the identifiers are decreasingly reliable, the second sequence results in a larger proportion of incorrect matches. The effect is so significant that even for arbitrarily large datasets the probability of making a data combination error does not approach 0. In Proposition 2, where non-disclosure is not guaranteed, and the probability of making a data combination error of the first kind approaches 0 as N^y and N^x increase, thresholds used for the decision rule shrink to zero more slowly than $(N^x)^{-1/c_2}$. In Proposition 3, where non-disclosure is guaranteed, thresholds used for the decision rule converge to zero faster than $(N^x)^{-1/c_2}$.

The result in Proposition 1 is stronger than the result in 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.

It is intuitive that the rates of the threshold sequences used for the decision rule can be described in terms of the size of the dataset $\mathcal{D}^x$ alone rather than both $\mathcal{D}^y$ and $\mathcal{D}^x$: we assume (in Assumption 2) that database $\mathcal{D}^y$ contains the subset of individuals from the database $\mathcal{D}^x$, and hence $\mathcal{D}^x$ is larger. The size of the larger dataset is the only factor determining how many potential matches from this dataset we are able to find for any observation in the smaller dataset without using any additional information from the identifiers.

The next three propositions consider the case of exponential tails.

PROPOSITION 4. *Suppose for $\alpha \in (0, \bar{\alpha})$, $\phi(\alpha) = b_1 e^{-c_1/\alpha}$, $b_1, c_1 > 0$, and $\psi(\alpha) = b_2 e^{-c_2/\alpha}$, $b_2, c_2 > 0$. Let $\alpha_N > 0$ be chosen in such a way that*

$$\lim_{N^y \to \infty} N^x e^{-\frac{c_2}{\alpha_N}} \alpha_N = 0.$$

Then for any $x \in \mathcal{X}$,

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

For instance, sequences $\alpha_N = \frac{a}{(N^x)^d}$ when $a, d > 0$, satisfy this condition.

The result of Proposition 4 implies the following result in Proposition 5.

PROPOSITION 5. ***(Absence of non-disclosure risk guarantee in the case of exponential tails)***
Suppose the conditions in Proposition 1 hold.

Then non-disclosure is not guaranteed.

PROPOSITION 6. ***(Non-disclosure risk guarantee in the case of exponential tails)*** *Suppose for $\alpha \in (0, \bar{\alpha})$, $\phi(\alpha) = b_1 e^{-c_1/\alpha}$, $b_1, c_1 > 0$ and $\psi(\alpha) = b_2 e^{-c_2/\alpha}$, $b_2, c_2 > 0$. Let $\alpha_N > 0$ be chosen in such a way that*

$$\lim_{N^y \to \infty} N^x e^{-\frac{c_2}{\alpha_N}} \alpha_N > 0.$$

Then non-disclosure is guaranteed.

For instance, sequences $\alpha_N = \frac{a}{\log N^x}$ when $a > c_2$, satisfy this condition (in this case, $\lim_{N^y \to \infty} N^x e^{-\frac{c_2}{\alpha_N}} \alpha_N = \infty$).

As we can see again, non-disclosure is guaranteed when the sequences of thresholds converge to ∞ more slowly than in the absence of disclosure guarantees.

With this discussion we find that the decision rules that we constructed are well-defined and there exists a non-empty class of sequences of thresholds that may 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 behavior of the identifiers' distributions.

4 Identification 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 that is 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 each 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 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 individual might be named "Tatiana" 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 Tatiana and not just two individuals with the same name. 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 estimate the model of interest.

PROPOSITION 7. *For any $\theta \in \Theta$ and any $\alpha \in (0, \bar{\alpha})$,*

$$E\left[\rho(Y, X; \theta) \mid X = x, |Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\right] = E\left[\rho(Y, X; \theta) \mid X = x\right]. \tag{4.6}$$

The proof of this proposition is in the Appendix.

Proposition 7 is an important part of our argument because it allows us to use a subpopulation with relatively infrequent characteristics to identify the parameter in the moment equation that applies to the entire population.

For example, if in the data from Durham, NC we find that two datasets both contain last names "Komarova", "Nekipelov" and "Yakovlev", we can use that subsample to identify the model for the rest of the population in North Carolina. Another important feature of this moment equation is that it does not require the distance between two identifiers to be equal to zero. In other words, if we see last name "Nekipelov" in one dataset and "Nikipelov" in the other dataset, we can still associate both entries with the same individual.

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 θ_0 can be estimated from the moment equation

$$E\left[\rho(Y, X; \theta) \mid X = x, |Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\right] = 0 \tag{4.7}$$

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.

A useful implication of Proposition 7 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 2. *Here we illustrate identification 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 7 and identify θ_0 from the "trimmed" moments. The solution can be expressed as

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

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 identified set based on the combined data. This notion incorporates several features. First, it takes into account the result of Proposition 7, which tells us that the information obtained from the correctly matched 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 thereforebe treated as a limiting property as the size of both datasets increases. We start with a discussion of the second feature and then conclude this section with a discussion of the third feature.

Let D^N, where $N = (N^y, N^x)$, denote a combined dataset constructed from the dataset $\mathcal{D}^x$ of size N^x and the dataset $\mathcal{D}^y$ of size N^y. D^N consists of observations (y_j, z_j^y, x_i, z_i^x) where j and i were matched by our data combination procedure. The density of (y_j, z_j^y, x_i, z_i^x) in D^N can be expressed in terms of the 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 $\pi^N(y, x, D^N)$ denote the proportion of incorrect matches in the set

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

If this set is empty, then $\pi^N(y, x, D^N)$ is not defined. By $\pi^N\left(y, x, \{y_j, z_j^y\}_{j=1}^{N^y}, \{x_i, z_i^x\}_{i=1}^{N^x}\right)$ we denote the average proportion of incorrect matches across all possible combined datasets D^N that can be obtained from $\mathcal{D}^y$ and $\mathcal{D}^x$. Then we find that

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

This value is not defined otherwise.

Now let $\pi^N(y, x)$ denote 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 any possible dataset of N^y observations of (y_j, z_j^y) and any possible dataset of N^x observations of (x_i, z_i^x) that satisfy the following conditions:

(a) Any individual in the first dataset is present in the second dataset. In other words, for each (y_j, z_j^y) in the first dataset there exists (x_i, z_i^x) in the second dataset such that (y_j, z_j^y) and (x_i, z_i^x) correspond to the same individual.

(b) If (y_j, z_j^y) and (x_i, z_i^x) 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}.

(c) Any (x_i, z_i^x) that does not correspond to any individual in the first dataset is a drawing from the marginal distribution f_{X,Z^x}.

These conditions imply that (y_j, z_j^y) has the density f_{Y,Z^y} and (x_i, z_i^x) has the density f_{X,Z^x}.

Next, we define the distribution density for an observation in a "generic" combined dataset of size N:

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

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 (4.7), we want to consider $E^N\left[\rho(y, x; \theta) \mid X = x\right]$ 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, $\pi^N(y, x) = 0$ almost everywhere, then this conditional mean coincides with the left-hand side in (4.7), and thus, takes the value of 0 if and only if $\theta = \theta_0$. In general, however, $\pi^N(y, x)$ is nonzero.

The next proposition describes the limiting behavior of the moment function $E^N[\rho(y_j, x_i; \theta)|x_i = x]$. To obtain its result, we impose the following condition on the moment of function $\|\rho(\cdot, \cdot; \cdot)\|$ in the case when X and Y are independent: for $x \in \mathcal{X}$,

$$E^*\left[\|\rho(\widetilde{Y}, X; \theta)\| \mid X = x\right] < \infty \quad \text{for all } \theta \in \Theta,$$

where E^* denotes the expectation taken over the distribution $f_Y(\widetilde{y}) f_X(x)$.

PROPOSITION 8. *For $x \in \mathcal{X}$, let $\pi^N(y, x) \to \pi(x)$ as $N^y \to \infty$ uniformly over all $y \in \mathcal{Y}$:*

$$\sup_{y \in \mathcal{Y}} |\pi^N(y, x) - \pi(x)| \to 0 \text{ as } N^y \to \infty. \tag{4.8}$$

Then, for $x \in \mathcal{X}$,

$$E^N[\rho(y_j, x_i; \theta)|x_i = x] \to (1 - \pi)E[\rho(Y, X; \theta)|X = x] + \pi E^*[\rho(\widetilde{Y},\, X; \theta)|X = x] \tag{4.9}$$

as $N^y \to \infty$.

The proof of this proposition is in the Appendix.

Next, 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 two conditions in the special case when $\pi_N(y, x)$ is equal to 0 almost everywhere:

$$r\left(E\left[\rho(Y, X; \theta) \mid X = x\right]\right) = 0 \quad \Longleftarrow \quad \theta = \theta_0 \tag{4.10}$$

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

$$r\left(E^N\left[\rho(y_j, x_i; \theta) \mid x_i = x\right]\right) = g^N(h, \theta)' W_0 g^N(h, \theta),$$

where

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

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

$$E\left[\sup_{\theta \in \Theta} \|h(X)\rho(Y, X; \theta)\|\right] < \infty, \quad E^*\left[\sup_{\theta \in \Theta} \|h(X)\rho(\widetilde{Y}, X; \theta)\|\right] < \infty \tag{4.11}$$

where $E_X[\cdot]$ denotes the expectation over the distribution of X.

Condition (4.10) is satisfied if and only if for $\pi^N(y,x) = 0$ almost everywhere,

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

In rare situations this condition may be violated for some instruments $h(\cdot)$, so $h(\cdot)$ has to be chosen in a way to guarantee that it holds. Here and thereafter we suppose that (4.10) is satisfied.

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 for given N under the chosen $r(\cdot)$.

Let Π^N denote the set of all $\pi^N(\cdot,\cdot)$ that are possible under the available prior information about the matching process. For instance, if the available information about π^N is that it takes values between π_1 and π_2 only, then any measurable function taking values between π_1 and π_2 has to be considered. The empirical evidence thus generates a set of values for θ approximating θ_0. Let Θ_N denote this set:

$$\Theta_N = \bigcup_{\pi^N \in \Pi^N} \underset{\theta\in\Theta}{\text{Argmin}}\ r\left(E^N\left[\rho(y_j,x_i;\theta) \mid x_i = x\right]\right). \tag{4.12}$$

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 $\sup\limits_{y\in\mathcal{Y},x\in\mathcal{X}} |\pi^N(y,x) - \pi(y,x)| \to 0$.

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. The next proposition shows that in this situation, the following set is a limit of sthe equence of Θ_N:

$$\Theta_\infty = \bigcup_{\pi\in\Pi^\infty} \underset{\theta\in\Theta}{\text{Argmin}}\ r\left((1-\pi)E\left[\rho(Y,X;\theta) \mid X = x\right] + \pi E^*\left[\rho(\widetilde{Y},\, X;\theta)|X = x\right]\right), \tag{4.13}$$

where

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

with

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

PROPOSITION 9. *Suppose that Π^∞ consists of constant values, and that for any $\pi \in \Pi^\infty$ function $g_\pi(h,\theta)' W_0 g_\pi(h,\theta)$ has a unique minimizer. Consider Θ_N defined as in (4.12) and Θ_∞ defined as in (4.13). 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 Appendix.

Proposition 9 can be rewritten in terms of the distances between sets Π^∞ and Π^N and sets Θ_∞ and Θ_N:

$$d(\Pi^\infty, \Pi^N) = \sup_{\pi \in \Pi^\infty} \inf_{\pi^N \in \Pi^N} \sup_{y \in Y, x \in X} |\pi^N(y,x) - \pi|$$
$$d(\Theta_\infty, \Theta_N) = \sup_{\theta \in \Theta_\infty} \inf_{\theta_N \in \Theta_N} \|\theta_N - \theta\|.$$

Indeed, the definition of Π^∞ gives that $d(\Pi^\infty, \Pi^N) \to 0$ as $N^y \to \infty$. Proposition 9 establishes that this condition together with the condition on the uniqueness of the minimizer of $g_\pi(h,\theta)' W_0 g_\pi(h,\theta)$ for each $\pi \in \Pi^\infty$ gives that $d(\Theta_\infty, \Theta_N) \to 0$ as $N^y \to \infty$.

Set Θ_∞ is what we call the set of parameter values identifiable from infrequent attribute values. The definition below provides notions of partial and point identification.

DEFINITION 2. *We say that parameter θ_0 is point identified (partially identified) from infrequent attribute values if $\Theta_\infty = \{\theta_0\}$ ($\Theta_\infty \neq \{\theta_0\}$).*

Whether the model is point identified is determined by the properties of the model, the distribution of the data, and the matching procedure. This definition 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 2, then θ_0 is point identified under any other choice of function $h(\cdot)$ that satisfies (4.10), and (4.11).

If the parameter of interest is only partially 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 (4.10) and (4.11). In the case of partial identification, $0 \in \Pi^\infty$ implies that $\theta_0 \in \Theta_0$, but otherwise θ_0 does not necessarily belong to Θ_0.

Thus far, we have not used the properties of the specific data combination rule to provide a concrete characterization of the set of parameters identified from infrequent attribute values. Now we consider identification from combined data samples obtained using a decision rule that honors a particular bound on the risk of individual disclosure. The bound on the risk of individual disclosure does not mean that making a correct match in a particular dataset is impossible. The bound on the individual disclosure risk implies that there will be multiple versions of a combined dataset. One of these versions can correspond to the "true" dataset for which $d_{ji} = m_{i,j}$ (using the notation from Section 3). However, in addition to this dataset we can also construct combined datasets with varying fractions of incorrect matches. This implies that for any $x \in \mathcal{X}$, $\inf_{j,i} \pi^N\left(y_j, x_i = x, \{y_j, z_j^y\}_{j=1}^{N^y}, \{x_i, z_i^x\}_{i=1}^{N^x}\right) > 0$. As a result, we can establish that $\inf_{y \in \mathcal{Y}} \pi^N(y,x) > 0$.

Condition (3.3) implies that $\inf_{x\in\mathcal{X}} \lim_{N^y\to\infty} \inf_{y\in\mathcal{Y}} \pi^N(y,x) \geq \underline{\gamma}$, which gives

$$\lim_{N^y\to\infty} \inf_{y\in\mathcal{Y}} \pi^N(y,x) \geq \underline{\gamma} \quad \text{for any } x \in \mathcal{X}. \tag{4.14}$$

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

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

If the only available information about the disclosure risk is a bound $\underline{\gamma}$, then for each $x \in \mathcal{X}$ we can only infer that $\pi(x) \in [\underline{\gamma}, 1]$. The results of Proposition 8 then give that for each $x \in \mathcal{X}$, any value among

$$(1-\pi(x))E[\rho(Y,X;\theta)|X=x] + \pi(x)E^*[\rho(\widetilde{Y},X;\theta)|X=x], \quad \pi(x) \in [\underline{\gamma},1]$$

may be the limit of of $E^N[\rho(y_j, x_i;\theta)|x_i = x]$.

To determine the set of parameters identifiable from infrequent attribute values, we need to find Π^∞ – the set of all possible uniform over both $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ limits of $\pi^N(y,x)$. Again, using the properties in Assumptions 3 (i)-(ii) for $\alpha_N \to 0$ and the property of our matching method, we can conclude that the limit of $\pi^N(y,x)$ does not depend on the values of y or x. Let us denote this limit as π. Uniformity over $y \in \mathcal{Y}$ and $x \in \mathcal{X}$ in Assumptions 3 (i)-(ii) implies that

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

From the information about the bound on disclosure risk we can conclude that π can take any value in $[\underline{\gamma}, 1]$. In other words, $\Pi^\infty = [\underline{\gamma}, 1]$. This fact will allow us to establish results on point (partial) identification of θ_0 in Theorem 1 (Theorem 2).

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

THEOREM 1. ***(Point identification of*** θ_0***).*** *Let* $\alpha_N \to 0$ *as* $N^y \to \infty$ *in such a way that* $\inf_{i,j} p_{ij}^N(x) \to 1$ *as* $N^y \to \infty$ *for any* $x \in \mathcal{X}$*. Then* θ_0 *is point identified from matches of infrequent values of the attributes.*

Proof. Condition

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

can equivalently be written as

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

which means that for any $\varepsilon > 0$, any $x \in \mathcal{X}$ and any sequence of datasets $\{y_j, z_j^y\}_{j=1}^{N^y}$ and $\{x_i, z_i^x\}_{i=1}^{N^x}$, when N^x and N^y are large enough,

$$\sup_{j,i} \pi^N \left(y_j, x_i = x, \{y_j, z_j^y\}_{j=1}^{N^y}, \{x_i, z_i^x\}_{i=1}^{N^x} \right) < \varepsilon.$$

This implies that when N^x and N^y are large enough,

$$\sup_{y \in \mathcal{Y}} \pi^N(y, x) < \varepsilon.$$

Since $\varepsilon > 0$ can be chosen arbitrarily small, we obtain that

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

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 Propositions 1 and 4 in Section 3.

Theorem 2 gives a partial 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. ***(Partial identification of*** θ_0***).*** *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 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,

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

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 Θ_∞ is not necessarily 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 there is a bound* $\underline{\gamma} > 0$ *on the disclosure risk. Then* θ_0 *is only partially identified from matches on infrequent attribute values, and, under the distance* $r(\cdot)$ *chosen*

in the spirit of least squares, the 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.

Note that $\theta_0 = E[XX']^{-1}E[XY]$. $E[XX']$ can be found from the marginal distribution of X and is thus 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

$$\theta_1 = E_X[XX']^{-1}E_X[X]E_Y[Y], \tag{4.15}$$

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 complete proof of Theorem 1 can be found in the Appendix.

EXAMPLE 3. *As a special case, consider a bivariate linear regression model*

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

Using our previous calculations, we obtain that the 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 identified set for the intercept is

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

Thus far, we have shown that using a data combination rule that selects observations with infrequent values of the attributes used for matching allows us to 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 information on those individuals. For instance, suppose that outcome variable Y is a discrete variable taking values 0 and 1 and the consumer attribute X is a continuous variable with support on $[0, 1]$, and suppose Y is contained in the anonymized sensitive data and X is a public information on individuals. Suppose that as a result of data combination we construct

a dataset with two observations $(0,0)$ and $(1,1)$. We fit a linear model $y = x$ to the data. If this estimate is revealed, one would be able to correctly predict the response of an individual with the attribute value $X = 1$ using the model. The implication is that estimates themselves may generate a threat of disclosure concretely the response Y might correspond to the purchase of a specific product by an individual, or a visit to a specific webpage, or the answer to an online questionnaire.

5 The Impact of Health Care on Consumer Satisfaction (Measured by Individual Ratings on Yelp.com)

To illustrate our theoretical analysis, we study identification of a simple linear model using data from Yelp.com and the public property tax records. The "main" dataset that we use contains ratings of local businesses in Durham, NC by Yelp users. The question we seek to answer can be informally formulated as: Does a visit to a doctor change the general attitudes of individuals in rating businesses on Yelp.com?

We can answer this question if for each given individual we are able to provide a prediction of whether and by how much the rating scores this individual gives to Yelp businesses change on average after this individual visits a doctor. The Yelp dataset corresponds to the dataset $\mathcal{D}^y$ in our theoretical analysis, and our outcome of interest Y has two elements: one is the mean of individual Yelp ratings before visiting a health-related business and the other is the mean of individual Yelp ratings after visiting a health-related business. It is clear, however, that producing such a prediction using data from Yelp.com alone will be problematic due to a familiar sample selection problem. In fact, the data sources solely from Yelp.com will over-sample the most active Yelp users who give reviews most frequently because (i) they have relatively higher incomes and thus they can "sample" more businesses; (ii) live more active lifestyles and reside closer to business locations; (iii) have more time at their disposal and can regularly write Yelp reviews. Sample selection that arises for these reasons can be controlled by including individual-level demographic characteristics into the model (such as income, age, location, etc.). However, for individual privacy and other reasons such information is not immediately available for Yelp users.

To control for sample selection, we reconstruct individual-level demographic information by combining the ratings from Yelp with information contained in individual property tax records that are publicly available for taxpayers in Durham county, NC. Combination of two datasets leads to the reconstruction of proxy variables for individual demographics for a subset of records from Yelp.com. Given that the property tax records contain the full name and address of the taxpayer, such a procedure will lead to the discovery of the exact name and residence for at least some Yelp users with high confidence, i.e. lead to individual disclosure. Below we show how our obtained point estimates behave with and without limits on the risk of individual disclosure.

The property tax data were extracted from a public access website via tax administration record search (see http://www.ustaxdata.com/nc/durham/). Property tax records are identified by parcel

numbers. We collected data from property tax records for years 2009/2010, in total collecting 104,068 tax bills for 2010 and 103,445 tax bills for 2009. Each tax record contains information on the taxable property value, first and last names of the taxpayer and the location of the property (house number, street, and zip code). We then merged the data across years by parcel number and property owner, removing properties that changed owner between 2009 and 2010. Property tax data allow us to assemble information on the names and locations of individuals as well as the taxable value of their properties, which we use as a proxy for individual wealth. Table 1 summarizes the distribution of taxable property values in the dataset constructed from property tax records.

[Table 1 about here.]

Histograms of the distribution of property tax values are presented in Figure 1. The outliers seen in the histograms are caused by commerical properties. We manually remove all commercial properties by removing properties in commercial zones as well as all properties valued above $ 3M.

[Figure 1 about here.]

The property tax dataset corresponds to the dataset $\mathcal{D}^x$ in our theoretical analysis. We used this dataset to construct the vector of identifiers for each taxpayer Z^x which contains the zip code of the residence, as well as binary variables that correspond to our "guesses" of gender and ethnicity of the taxpayer based on comparing the taxpayer's first and last names to the lists of 500 most popular male, female, white, hispanic and black first and last names in the 2010 US Census.

We collected the dataset from Yelp.com with the following considerations. First, we collected information for all individuals who ever rated health-related businesses. This focus was to find the subset of Yelp users for whom we can identify the effect of a visit to a doctor. Second, for each such individual, we collected all information contained in this individual's Yelp profile as well as all ratings the individual has ever made on Yelp.com. Third, we collected all available information on all businesses that were ever rated by the inviduals in our dataset. This includes the location and nature of the business. For businesses like restaurants we collected additional details, such as the restaurant's cuisine, price level, child frendliness, and hours of operation. We further use this information to construct a vector of identifiers Z^y as in our theoretical analysis. Vector Z^y contains location variables (e.g. the modal zip code of the rated business) as well as binary variables corresponding to "guesses" of individual demographics such as gender and ethnicity as well as a guess for the user's name constructed from the Yelp username.

The indicator for a visit to a health care business was constructed from the ratings of health care businesses. We treat an individual's rating of a health care business as evidence that this individual actually visited that business. We were able to extract reliable information from 59 Yelp.com users who rated health care services in Durham. We focused on only publicly released ratings: Yelp.com has a practice of filtering particular ratings that are believed to be unreliable (the reasons for rating

supression are not disclosed by Yelp). Though we collected information on supressed ratings, we chose not to use them in our empirical analysis. The final dataset contains a total of 72 reviews for Durham health care businesses. We show the summary statistics for the constructed variables in Table 2.

[Table 2 about here.]

As mentioned above, for data combination purposes we used the entire set of Yelp ratings in Durham, NC. We use our threshold-based record linkage technique to combine the Yelp and property tax record datasets. We construct a distance measure for the individual identifiers by combining the edit distance using the first and last names in the property tax dataset and the username on Yelp.com, and the sum of ranks corresponding to the same values of binary elements in Z^y and Z^x, for instance corresponding to the modal zip code of the rated business and the zip code of the residence of the taxpayer or the guessed gender and ethnicity of the Yelp user and guessed gender and ethnicity of the property taxpayer. Using this simple matching rule, we identified 397 individuals in the tax record data as positive matches. Fourteen people are uniquely identifiable in both databases. Table 3 shows the distribution of obtained matches.

[Table 3 about here.]

The matched observations characterize the constructed combined dataset of Yelp reviews and the property tax bills. We were able to find Yelp users and property owners for whom the the combined edit distance and the sum of ranks for discrepancies between the numeric indicators (zip code and location of most frequent reviews) are equal to zero. We call this dataset the set of "one-to-one" matches. Based on matched first names we evaluate the sex of each Yelp reviewer and construct dummy variable indicating that the name is on the list of 500 most common female names in the US from the Census data, as a proxy that the corresponding taxpayer is a female. We also constructed measures of for other demographic indicators, but they did not improve the fit of our estimated ratings model and we exclude them from our analysis.

To answer our empirical question and measure the effect of a visit to a doctor on individual ratings of businesses on Yelp, we associate the measured outcome with the average treatment effect. The treatment in this framework is the visit to a doctor and the outcome is the average of Yelp ratings of other businesses. Yelp ratings are on the scale from 1 to 5 where 5 is the highest score and 1 is the lowest. We find first that on average, after attending a health related business, Yelp users tend to have slightly (0.05 SD) higher average rating than before (see column 1 of Table 4).

[Table 4 about here.]

To visualize the heterogeneity of observed effects across individuals, we evaluate the difference in their average Yelp ratings before and after visiting a health care business. The histogram in Figure

2 illustrates differences in the average rating changes after a visit to a healthcare business across Yelp users.

[Figure 2 about here.]

We find a significant difference between the average ratings before and after the visit to a helthcare business: means in lower and upper quartile of average ratings significantly different from zero (at 10% significant level). Those in the upper quartile report an average increase in ratings of 1.02 points whereas those in the lower quartile reported an average decrease in ratings by 1.14 points (see Table 5).

[Table 5 about here.]

One of the caveats of OLS estimates is selection bias due to selection of those who are treated. For example, people with higher incomes may use the services of Yelp-listed businesses more frequently or, for instance, males might visit doctors less frequently than females. This selection may result in bias of the estimated ATE. The omission of demographic characteristics such as income or sex may result in inability to control for this selection and inability to get consistent estimates of treatment effects. Columns 2, 3 and 4 of Table 4 illustrate this point. To control for possible selection bias, we use a matching estimator for the subsample of data for which we have data on the housing value. Column 4 of Table 4 shows evidence of selection, with a positive correlation between the housing value and participation in the sample. Column 2 and column 3 exhibit the OLS and the matching estimates of treatment effect. After controlling for selection, the effect of visiting a doctor is much higher. According to the matching estimates, visiting a doctor increases rating by 0.66 points (0.5 SD) compared to the estimate of 0.03 points obtained from OLS procedure.

We can now analyze how the parameters will be affected if we want to enforce a bound on the disclosure risk. To do that we use the notion of k-anonymity. k-anonymity requires that for each observation in the main database there are at least k equally good matches in the auxiliary database corresponding to the upper bound on the disclosure risk of $1/k$ (which is the maximal probability of constructing a correct match for a given observation). In our data the main attribute that was essential for construction of correct matches was the first and the last name of the individual in the property tax data. To break the link between the first and last name information in the property tax data and the username in Yelp data, we suppress letters from individual names. For instance, we transform the name "Denis" to "Deni*" and then to "Den*". If in the Yelp data we observe users with names "Dennis" and "Denis" and in the property tax data we observe the name "Denis", then the edit distance between "Denis" and "Denis" is zero, whereas the edit distance between "Dennis" and "Denis" is 1. However, if in the property tax data we suppressed the last two letters leading to transformation "Den*", the distance between "Dennis", "Denis" and "Den*" is the same.

Using character suppression we managed to attain k-anonymity with $k = 2$ and $k = 3$ by erasing, respectivelu 3 and 4 letters from the name recorded in the property tax database. The fact that there

is no perfect matches for a selected value of the distance threshold, leads to the set of minimizers of the distance function. To construct the identified set, we use the idea from our identification argument by representing the identified set as a convex hull of the point estimates obtained for different combinations of the two datasets. We select the edit distance equal to k in each of the cases of k-anonymity as the match threshold. For each entry in the Yelp database that has at least one counterpart in the property tax data with the edit distance less than or equal to k, we then construct the dataset of potential matches in the Yelp and the property tax datasets. We then construct matched databases using each potentially matched pair. As a result, if we have N observations in the Yelp database each having exactly k counterparts in the property tax database, we construct k^N matched datasets. For each such matched dataset we can construct the point estimates. Figure 3 shows the identified set for the average treatment effect and Figure 4 demonstrates the identified set for the linear projection of the propensity score.

[Figure 3 about here.]

[Figure 4 about here.]

Even with a tight restriction on the individual disclosure, the identified set of the ATE lies strictly above zero. This means that even with limits on the disclosure risk, the sign of the average treament effect is identified. The identified set for the linear projection of the propensity score the for the effect of property value contains the origin, but does not contain the origin for effect of gender and thus the sign of the gender coefficient in the propensity score remains identified.

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. Using the example of demand for health care services, we show that the identity disclosure may occur even when the data is not publicly shared.

References

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

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

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

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

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

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

Aggarwal, G., T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu (2005): "Approximation algorithms for k-anonymity," *Journal of Privacy Technology*, 2005112001.

Bradley, C., L. Penberthy, K. Devers, and D. Holden (2010): "Health services research and data linkages: issues, methods, and directions for the future," *Health services research*, 45(5(2)), 1468–1488.

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*, pp. 70–78. ACM.

Calzolari, G., and A. Pavan (2006): "On the optimality of privacy in sequential contracting," *Journal of Economic Theory*, 130(1), 168–204.

Chaudhuri, S., K. Ganjam, V. Ganti, and R. Motwani (2003): "Robust and efficient fuzzy match for online data cleaning," in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pp. 313–324. ACM.

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): "Inference on regressions with interval data on a regressor or outcome," *Econometrica*, 70(1), 357–368.

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

Duncan, G., and D. Lambert (1986): "Disclosure-limited data dissemination," *Journal of the American statistical association*, 81(393), 10–18.

Duncan, G., and S. Mukherjee (1991): "Microdata Disclosure Limitation in Statistical Databases: Query Sizeand Random Sample Query Control," .

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

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

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

Fellegi, I., and A. Sunter (1969): "A theory for record linkage," *Journal of the American Statistical Association*, pp. 1183–1210.

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

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

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

Gross, R., and A. Acquisti (2005): "Information revelation and privacy in online social networks," in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pp. 71–80. ACM.

Gusfield, D. (1997): *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press.

Homer, N., S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. Pearson, D. Stephan, S. Nelson, and D. Craig (2008): "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays," *PLoS Genetics*, 4(8), e1000167.

Horowitz, J., and C. Manski (1995): "Identification and robustness with contaminated and corrupted data," *Econometrica*, 63(2), 281–302.

Jaro, M. (1989): "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida," *Journal of the American Statistical Association*, pp. 414–420.

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

Komarova, T., D. Nekipelov, and E. Yakovlev (2012): "Estimation of Treatment Effects from Combined Data with Sensitive Treatment," *LSE and UC Berkeley Working Paper*.

Lahiri, P., and M. Larsen (2005): "Regression analysis with linked data," *Journal of the American statistical association*, 100(469), 222–230.

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

LeFevre, K., D. DeWitt, and R. Ramakrishnan (2005): "Incognito: Efficient full-domain k-anonymity," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 49–60. ACM.

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

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

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

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

——— (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(2), 519–546.

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

Molinari, F. (2008): "Partial identification of probability distributions with misclassified data," *Journal of Econometrics*, 144(1), 81–117.

Narayanan, A., and V. Shmatikov (2008): "Robust de-anonymization of large sparse datasets," in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pp. 111–125. IEEE.

Newcombe, H., J. Kennedy, S. Axford, and A. James (1959): "Automatic linkage of vital and health records," *Science*, 130, 954–959.

Ridder, G., and R. Moffitt (2007): "The econometrics of data combination," *Handbook of Econometrics*, 6, 5469–5547.

Salton, G., and D. Harman (2003): *Information retrieval*. John Wiley and Sons Ltd.

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

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

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

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

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

Winkler, W. (1999): "The state of record linkage and current research problems," in *Statistical Research Division, US Census Bureau*. Citeseer.

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

Appendix

A Construction of individual identifiers

The key element of our identification argument is based on the construction of the identifying variables Z^y and Z^x such that we can merge some or all observations in the disjoint databases to be able to estimate the econometric model of interest. While we took the existence of these variables as given, their construction in itself is an important issue and there is a vast literature in applied statistics and computer science that is devoted to the analysis of the broken record linkage. For completeness of the analysis in our paper we present some highlights from that literature.

In general the task of merging disjoint databases is a routine necessity in may practical applications. In many cases there do exist perfect cross-database identifiers of individual entries. There could be multiple reasons why that is the case. For instance, there could be errors in data entry and processing, wrong variable formatting, and duplicate data entry. The idea that has arisen in Newcombe, Kennedy, Axford, and James (1959) and was later formalized in Fellegi and Sunter (1969) was to treat the record linkage problem as a problem of classification of record subsets into matches, non-matches and uncertain cases. This classification is based on defining the similarity metric between each two records. Then given the similarity metric one can compute the probability of particular pair

of records being a match or non-match. The classification of pairs is then performed by fixing the probability of erroneous identification of a non-matched pair of records as a match and a matched pair of records as a non-match by minimizing the total proportion of pairs that are uncertain. This matching technique is based on the underlying assumption of randomness of records being broken. As a result, using the sample of perfectly matched records one can recover the distribution of the similarity metric for the matched and unmatched pairs of records. Moreover, as in hypothesis testing, one needs to fix the probability of record mis-identification. Finally, the origin of the similarity metric remains arbitrary.

A large fraction of the further literature was devoted to, on one hand, development of classes of similarity metrics that accommodate non-numeric data and, on the other hand, development of fast and scalable record classification algorithms. For obvious reasons, measuring the similarity of string data turns out to be the most challenging. Edit distance (see, Gusfield (1997) for instance) is a metric that can be used to measure the string similarity. The distance between the two strings is determined as the minimum number of insert, delete and replace operations required to transform one string into another. Another measure developed in Jaro (1989) and elaborated in Winkler (1999) is based on the length of matched strings, the number of common characters and their position within the string. In its modification it also allows for the prefixes in the names and is mainly intended to linking relatively short strings such as individual names. Alternative metrics are based on splitting strings into individual "tokens" that are substrings of a particular length and then analyzing the power of sets of overlapping and non-overlapping tokens. For instance, Jaccard coefficient is based on the relative number of overlapping and overall tokens in two strings. More advanced metrics include the "TF/IDF" metric that is based on the term frequency, or the number of times the term (or token) appears in the document (or string) and the inverse document frequency, or the number of documents containing the given term. The structure of the TF/IDF-based metric construction is outlined in Salton and Harman (2003). The distance measures may include combination of the edit distance and the TF/IDF distance such as a fuzzy match similarity metric described in Chaudhuri, Ganjam, Ganti, and Motwani (2003).

Given a specific definition of the distance, the practical aspects of matching observations will entail calibration and application of a particular technique for matching observations. The structure of those techniques is based on, first, the assumption regarding the data structure and the nature of the record errors. Second, it depends on the availability of known matches, and, thus, allows empirical validation of a particular matching technique. When such a validation sample is available, one can estimate the distribution of the similarity measures for matched and non-matched pairs for the validation sample. Then, using the estimated distribution one can assign the matches for the pairs outside the validation sample. When one can use numeric information in addition to the string information, one can use hybrid metrics that combine the known properties of numeric data entries and the properties of string entries.

Ridder and Moffitt (2007) overviews some techniques for purely numeric data combination. In the

absence of validation subsamples that may incorporate distributional assumptions on the "similar" numeric variables. For instance, joint normality assumption with a known sign of correlation can allow one to invoke likelihood-based techniques for record linkage.

B Proofs

Proof of Proposition 1. Probability $p_{ij}^N(x)$ in (3.2) is equal to

$$\frac{Pr_x\left(|z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N} \,|\, m_{ij} = 1\right) Pr_x(m_{ij} = 1)}{Pr_x\left(|z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N} \,|\, m_{ij} = 1\right) Pr_x(m_{ij} = 1) + Pr_x\left(|z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N} \,|\, m_{ij} = 0\right) Pr_x(m_{ij} = 0)}, \tag{B.16}$$

where Pr_x is the notation for the probability being conditioned on $x_i = x$. Note that $Pr_x(m_{ij} = 1) = \frac{1}{N^x}$.

By Assumption 3, for $\alpha_N \in (0, \bar{\alpha})$ and any j and i,

$$Pr_x\left(|z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N} \mid m_{ij} = 1\right) \geq (1 - \alpha_N)(\phi(\alpha_N) + o(\phi(\alpha_N))).$$

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

$$\inf_{j,i} \frac{(1 - \alpha_N)(\phi(\alpha_N) + o(\phi(\alpha_N)))Pr_x(m_{ij} = 1)}{(1 - \alpha_N)(\phi(\alpha_N) + o(\phi(\alpha_N)))Pr_x(m_{ij} = 1) + Pr_x\left(|z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N} \,|\, m_{ij} = 0\right)}.$$

The last expression will converge to 1 as $N^y \to \infty$ if

$$\sup_{j,i} \frac{Pr_x\left(|z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N} \,|\, m_{ij} = 0\right)}{\phi(\alpha_N) Pr_x(m_{ij} = 1)} = \sup_{j,i} \frac{N^x Pr_x\left(|z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N} \,|\, m_{ij} = 0\right)}{\phi(\alpha_N)}$$

converges to 0.

For the sake of notational simplicity, assume that Z^x takes only positive values. Now obtain that

for any j and i, the probability $Pr_x\left(|z_j^y - z_i^x| < \alpha, |z_i^x| > \frac{1}{\alpha} \,|\, m_{ij} = 0\right)$ is bounded from above by

$$
\begin{aligned}
&Pr_x\left(\bigcup_{k=0}^{\infty}\left(\frac{1}{\alpha}+k\alpha < z_i^x \le \frac{1}{\alpha}+(k+1)\alpha\right), |z_j^y - z_i^x| < \alpha \,\middle|\, m_{ij}=0\right)\\
&\le Pr_x\left(\bigcup_{k=0}^{\infty}\left(\left(\frac{1}{\alpha}+k\alpha < z_i^x \le \frac{1}{\alpha}+(k+1)\alpha\right)\bigcap\left(\frac{1}{\alpha}+(k-1)\alpha < z_j^y \le \frac{1}{\alpha}+(k+2)\alpha\right)\right) \,\middle|\, m_{ij}=0\right)\\
&= \sum_{k=0}^{\infty} Pr_x\left(\left(\frac{1}{\alpha}+k\alpha < z_i^x \le \frac{1}{\alpha}+(k+1)\alpha\right)\bigcap\left(\frac{1}{\alpha}+(k-1)\alpha < z_j^y \le \frac{1}{\alpha}+(k+2)\alpha\right) \,\middle|\, m_{ij}=0\right)\\
&\le \sum_{k=0}^{\infty} Pr_x\left(\frac{1}{\alpha}+k\alpha < Z^x \le \frac{1}{\alpha}+(k+1)\alpha \middle|\, X=x\right) P\left(\frac{1}{\alpha}+(k-1)\alpha < Z^y \le \frac{1}{\alpha}+(k+2)\alpha\right)\\
&= \sum_{k=0}^{\infty}\left(\phi\left(\frac{\alpha}{k\alpha^2+1}\right)+o\left(\phi\left(\frac{\alpha}{k\alpha^2+1}\right)\right)-\phi\left(\frac{\alpha}{(k+1)\alpha^2+1}\right)-o\left(\phi\left(\frac{\alpha}{(k+1)\alpha^2+1}\right)\right)\right)\cdot
\end{aligned}
\tag{B.17}
$$

$$
\cdot\left(\psi\left(\frac{\alpha}{(k-1)\alpha^2+1}\right)+o\left(\psi\left(\frac{\alpha}{(k-1)\alpha^2+1}\right)\right)-\psi\left(\frac{\alpha}{(k+2)\alpha^2+1}\right)-o\left(\psi\left(\frac{\alpha}{(k+2)\alpha^2+1}\right)\right)\right)
$$

The same final expression in the inequality is obtained if Z^x can take negative as well as positive values.

Note that

$$
\phi\left(\frac{\alpha}{k\alpha^2+1}\right)-\phi\left(\frac{\alpha}{(k+1)\alpha^2+1}\right) = b_1\alpha^{c_1}\left(\frac{1}{(k\alpha^2+1)^{c_1}}-\frac{1}{((k+1)\alpha^2+1)^{c_1}}\right).
$$

Let $k \ge 1$. Function $\frac{1}{(k\alpha^2+1)^{c_1}} - \frac{1}{((k+1)\alpha^2+1)^{c_1}}$ is positive and takes the maximum value at

$$
\alpha^* = \sqrt{\frac{(k+1)^{\frac{1}{c_1+1}} - k^{\frac{1}{c_1+1}}}{(k+1)k^{\frac{1}{c_1+1}} - k(k+1)^{\frac{1}{c_1+1}}}}.
$$

This maximum value is equal to

$$
\left(k+1-k^{\frac{c_1}{c_1+1}}(k+1)^{\frac{1}{c_1+1}}\right)^{c_1} - \left((k+1)^{\frac{c_1}{c_1+1}}k^{\frac{1}{c_1+1}}-k\right)^{c_1} = k^{c_1}\left(\left(1+\frac{1}{k}\right)^{\frac{c_1}{c_1+1}}-1\right)^{c_1+1}.
$$

The Taylor expansion of $\left(1+\frac{1}{k}\right)^{\frac{c_1}{c_1+1}}$ gives that $\left(1+\frac{1}{k}\right)^{\frac{c_1}{c_1+1}} - 1 \le q_1\frac{1}{k}$ for some constant $q_1 > 0$ that does not depend on k. Therefore,

$$
\phi\left(\frac{\alpha}{k\alpha^2+1}\right)-\phi\left(\frac{\alpha}{(k+1)\alpha^2+1}\right) \le r_1\alpha^{c_1}\frac{1}{k},
$$

where $r_1 = b_1 q_1^{c_1+1}$. In a similar way, we can show that for some constant $r_2 > 0$ that does not depend on k,

$$
\psi\left(\frac{\alpha}{(k-1)\alpha^2+1}\right)-\psi\left(\frac{\alpha}{(k+2)\alpha^2+1}\right) \le r_2\alpha^{c_2}\frac{1}{k}, \quad k \ge 1.
$$

For $k=0$, the term $1-\frac{1}{(\alpha^2+1)^{c_1}}$ is bounded from above by $1-\frac{1}{(\bar{\alpha}^2+1)^{c_1}}$, and the term $\frac{1}{(-\alpha^2+1)^{c_2}}-\frac{1}{(2\alpha^2+1)^{c_2}}$ is bounded from above by $\frac{1}{(-\bar{\alpha}^2+1)^{c_2}}-\frac{1}{(2\bar{\alpha}^2+1)^{c_2}}$. Since series $\sum_{k=1}^{\infty}\frac{1}{k^2}$ is convergent, it follows that series (B.17) is bounded from above by $\alpha^{c_1+c_2}$ multiplied by some positive constant that does not depend on α.

Taking into account this result, we conclude that if (3.4) holds, then

$$\lim_{N^y\to\infty}\sup_{j,i}\frac{N^x Pr_x\left(|z_j^y-z_i^x|<\alpha_N,|z_i^x|>\frac{1}{\alpha_N}\,|\,m_{ij}=0\right)}{b_1\alpha_N^{c_1}}=0,$$

and hence,

$$\lim_{N^y\to\infty}\inf_{j,i}P\left(m_{ij}=1\mid x_i=x,|z_j^y-z_i^x|<\alpha_N,|z_i^x|>\frac{1}{\alpha_N}\right)=1.$$

From our proof it is clear that for general functions $\phi(\cdot)$ and $\psi(\cdot)$, the following condition will be sufficient to establish the result of this proposition:

$$\lim_{N^y\to\infty}\frac{N^x}{\phi(\alpha_N)}\sum_{k=0}^{\infty}\left(\phi\left(\frac{\alpha_N}{k\alpha_N^2+1}\right)-\phi\left(\frac{\alpha_N}{(k+1)\alpha_N^2+1}\right)\right)$$
$$\left(\psi\left(\frac{\alpha_N}{(k-1)\alpha_N^2+1}\right)-\psi\left(\frac{\alpha_N}{(k+2)\alpha_N^2+1}\right)\right)=0.$$

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

Proof of Proposition 3. Probability $p_{ij}^N(x)$ in (B.16) is bounded from above by

$$\frac{1}{1+\frac{N^x}{\phi(\alpha_N)+o(\phi(\alpha_N))}Pr_x\left(|z_j^y-z_i^x|<\alpha_N,|z_i^x|>\frac{1}{\alpha_N}\,|\,m_{ij}=0\right)\left(1-\frac{1}{N^x}\right)}.$$

We can suppose that $N^x\geq 2$. Then $p_{ij}^N(x)$ is bounded from above by

$$\frac{1}{1+0.5\frac{N^x}{\phi(\alpha_N)+o(\phi(\alpha_N))}Pr_x\left(|z_j^y-z_i^x|<\alpha_N,|z_i^x|>\frac{1}{\alpha_N}\,|\,m_{ij}=0\right)}.$$

For the sake of notational simplicity, assume that Z^x takes only positive values. Now obtain that

$Pr_x\left(|z_j^y - z_i^x| < \alpha, |z_i^x| > \frac{1}{\alpha} \,|\, m_{ij} = 0\right)$ is bounded from below by

$$
\begin{aligned}
&Pr_x\left(\bigcup_{k=0}^{\infty}\left(\frac{1}{\alpha} + k\alpha < z_i^x \le \frac{1}{\alpha} + (k+1)\alpha\right), |z_i^x - z_j^y| < \alpha \,\middle|\, m_{ij} = 0\right)\\
&\ge Pr_x\left(\bigcup_{k=0}^{\infty}\left(\left(\frac{1}{\alpha} + k\alpha < z_i^x \le \frac{1}{\alpha} + (k+1)\alpha\right)\bigcap\left(\frac{1}{\alpha} + k\alpha < z_j^y \le \frac{1}{\alpha} + (k+1)\alpha\right)\right) \,\middle|\, m_{ij} = 0\right)\\
&= \sum_{k=0}^{\infty} Pr_x\left(\left(\frac{1}{\alpha} + k\alpha < z_i^x \le \frac{1}{\alpha} + (k+1)\alpha\right)\bigcap\left(\frac{1}{\alpha} + k\alpha < z_j^y \le \frac{1}{\alpha} + (k+1)\alpha\right) \,\middle|\, m_{ij} = 0\right)\\
&\ge \sum_{k=0}^{\infty} Pr_x\left(\frac{1}{\alpha} + k\alpha < Z^x \le \frac{1}{\alpha} + (k+1)\alpha\right) P\left(\frac{1}{\alpha} + k\alpha < Z^y \le \frac{1}{\alpha} + (k+1)\alpha\right)\\
&= \sum_{k=0}^{\infty}\left(\phi\left(\frac{\alpha}{k\alpha^2+1}\right) + o\left(\phi\left(\frac{\alpha}{k\alpha^2+1}\right)\right) - \phi\left(\frac{\alpha}{(k+1)\alpha^2+1}\right) - o\left(\phi\left(\frac{\alpha}{(k+1)\alpha^2+1}\right)\right)\right)\cdot\\
&\cdot\left(\psi\left(\frac{\alpha}{k\alpha^2+1}\right) + o\left(\psi\left(\frac{\alpha}{k\alpha^2+1}\right)\right) - \psi\left(\frac{\alpha}{(k+1)\alpha^2+1}\right) - o\left(\psi\left(\frac{\alpha}{(k+1)\alpha^2+1}\right)\right)\right)
\end{aligned}
$$

The same final expressions in the inequalities are obtained if Z^x can take negative as well as positive values.

Note that

$$\phi\left(\frac{\alpha}{k\alpha^2+1}\right) - \phi\left(\frac{\alpha}{(k+1)\alpha^2+1}\right) = b_1\frac{\alpha^{c_1}}{((k+1)\alpha^2+1)^{c_1}}\left(\left(1 + \frac{\alpha^2}{k\alpha^2+1}\right)^{c_1} - 1\right).$$

From the Taylor expansion of function $(1+x)^{c_1}$ it follows that for $x \in [0,1)$,

$$
\begin{aligned}
&(1+x)^{c_1} \ge 1 + c_1 x, \quad \text{if } \ c_1 \ge 1;\\
&(1+x)^{c_1} \ge 1 + c_1 x + \frac{c_1(c_1-1)}{2}x^2, \quad \text{if } \ 0 < c_1 < 1.
\end{aligned}
$$

Hence, when $c_1 \ge 1$,

$$\phi\left(\frac{\alpha}{k\alpha^2+1}\right) - \phi\left(\frac{\alpha}{(k+1)\alpha^2+1}\right) \ge b_1 c_1\frac{\alpha^{c_1+2}}{((k+1)\alpha^2+1)^{c_1}(k\alpha^2+1)} \ge b_1 c_1\frac{\alpha^{c_1+2}}{((k+1)\bar{\alpha}^2+1)^{c_1}(k\bar{\alpha}^2+1)},$$

and when $0 < c_1 < 1$,

$$
\begin{aligned}
\phi\left(\frac{\alpha}{k\alpha^2+1}\right) - \phi\left(\frac{\alpha}{(k+1)\alpha^2+1}\right) &\ge b_1 c_1\frac{\alpha^{c_1+2}}{((k+1)\alpha^2+1)^{c_1}(k\alpha^2+1)}\left(1 + \frac{c_1-1}{2}\frac{\alpha^2}{k\alpha^2+1}\right)\\
&\ge b_1 c_1^2\frac{\alpha^{c_1+4}}{2((k+1)\alpha^2+1)^{c_1}(k\alpha^2+1)^2} \ge b_1 c_1^2\frac{\alpha^{c_1+4}}{2((k+1)\bar{\alpha}^2+1)^{c_1}(k\bar{\alpha}^2+1)^2}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&\psi\left(\frac{\alpha}{k\alpha^2+1}\right) - \psi\left(\frac{\alpha}{(k+1)\alpha^2+1}\right) \ge b_2 c_2\frac{\alpha^{c_2+2}}{((k+1)\bar{\alpha}^2+1)^{c_2}(k\bar{\alpha}^2+1)}, \quad \text{if } \ c_2 \ge 1\\
&\psi\left(\frac{\alpha}{k\alpha^2+1}\right) - \psi\left(\frac{\alpha}{(k+1)\alpha^2+1}\right) \ge b_2 c_2^2\frac{\alpha^{c_2+4}}{2((k+1)\bar{\alpha}^2+1)^{c_2}(k\bar{\alpha}^2+1)^2}, \quad \text{if } \ 0 < c_2 < 1
\end{aligned}
$$

Note that series $\sum_{k=0}^{\infty} \frac{1}{((k+1)\bar{\alpha}^2+1)^{c_i}(k\bar{\alpha}^2+1)}$ and series $\sum_{k=0}^{\infty} \frac{1}{((k+1)\bar{\alpha}^2+1)^{c_i}(k\bar{\alpha}^2+1)^2}$, $i = 1, 2$, are convergent.

Taking into account the bounds derived above, condition (3.5) and the fact that $\alpha_N > 0$, we conclude that for all N^x and N^y,

$$\frac{N^x}{\phi(\alpha_N) + o(\phi(\alpha_N))} Pr_x \left(|z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N} \mid m_{ij} = 0 \right) \geq \Delta$$

for some $\Delta > 0$. Then $p_{ij}^N(x) \leq \frac{1}{1+0.5\Delta}$, and thus,

$$\sup_{j,i} p_{ij}^N(x) \leq \frac{1}{1+0.5\Delta} < 1.$$

Clearly, $\lim_{N^y \to \infty} \sup_{i,j} p_{ij}^N(x) \leq \frac{1}{1+0.5\Delta}$, and therefore,

$$\sup_{x \in \mathcal{X}} \lim_{N^y \to \infty} \inf_{i,j} p_{ij}^N(x) \leq \frac{1}{1+0.5\Delta} < 1.$$

To summarize, non-disclosure is guaranteed.

From the proof of this proposition it is clear that for general functions $\phi(\cdot)$ and $\psi(\cdot)$, the following condition will be sufficient to guarantee non-disclosure:

$$\lim_{N^y \to \infty} \frac{N^x}{\phi(\alpha_N)} \sum_{k=0}^{\infty} \left(\phi\left(\frac{\alpha_N}{k\alpha_N^2 + 1} \right) - \phi\left(\frac{\alpha_N}{(k+1)\alpha_N^2 + 1} \right) \right) \left(\psi\left(\frac{\alpha_N}{k\alpha_N^2 + 1} \right) - \psi\left(\frac{\alpha_N}{(k+1)\alpha_N^2 + 1} \right) \right) > 0.$$

Proof of Proposition 4. Here we use the condition indicated in the end of the proof of Proposition 1. For our functions $\phi(\cdot)$ and $\psi(\cdot)$ the series given in that condition is equal to

$$b_2 N^x e^{-\frac{c_2}{\alpha_N} + c_2\alpha_N} (1 - e^{-c_1\alpha_N})(1 - e^{-3c_2\alpha_N}) \sum_{k=0}^{\infty} e^{-k(c_1+c_2)\alpha_N} = b_2 N^x e^{-\frac{c_2}{\alpha_N} + c_2\alpha_N} \frac{(1 - e^{-c_1\alpha_N})(1 - e^{-3c_2\alpha_N})}{1 - e^{-(c_1+c_2)\alpha_N}}$$

Since each of $1 - e^{-c_1\alpha_N}$, $1 - e^{-3c_2\alpha_N}$ and $1 - e^{-(c_1+c_2)\alpha_N}$ converges to 0 with the rate α_N, then the limit of the series is 0 if $N^x e^{-\frac{c_2}{\alpha_N}} \alpha_N \to 0$.

If $\alpha_N = \frac{a}{(N^x)^d}$, then $N^x e^{-\frac{c_2}{\alpha_N}} \alpha_N = a(N^x)^{1-d} e^{-\frac{c_2}{a}(N^x)^d}$. When $a, d > 0$ this sequence converges to 0.

Proof of Proposition 5. Obvious.

Proof of Proposition 6. Here we use the condition indicated in the end of the proof of Proposition 3. For our functions $\phi(\cdot)$ and $\psi(\cdot)$ the series given in that condition is equal to

$$b_2 N^x e^{-\frac{c_2}{\alpha_N}} (1 - e^{-c_1\alpha_N})(1 - e^{-c_2\alpha_N}) \sum_{k=0}^{\infty} e^{-k(c_1+c_2)\alpha_N} = b_2 N^x e^{-\frac{c_2}{\alpha_N}} \frac{(1 - e^{-c_1\alpha_N})(1 - e^{-c_2\alpha_N})}{1 - e^{-(c_1+c_2)\alpha_N}}$$

Since each of $1-e^{-c_1\alpha_N}$, $1-e^{-c_2\alpha_N}$ and $1-e^{-(c_1+c_2)\alpha_N}$ converges to 0 with the rate α_N, then the limit of the series is strictly positive if $\lim_{N^y\to\infty} N^x e^{-\frac{c_2}{\alpha_N}}\alpha_N > 0$.

If $\alpha_N = \frac{a}{\log N^x}$, then $N^x e^{-\frac{c_2}{\alpha_N}}\alpha_N = \frac{a(N^x)^{1-\frac{c_2}{a}}}{\log N^x}$. With $a > c_2$ this sequence converges to ∞.

Proof of Proposition 7. Using Assumption 3 (iii) and the law of iterated expectations,

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

By Assumption 3 (i) and (iii),

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

This implies

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

which is equivalent to (4.6).

Proof of Proposition 8. Note that $E^N[\rho(y_j,x_i;\theta)|x_i=x] = A_N(x)+B_N(x)$, where $A_N(x) =$

$$
\frac{\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_j^y-z_i^x|<\alpha_N}(1-\pi^N(y_j,x))\rho(y_j,x,\theta)f_{Y|X,Z^y,Z^x}(y_j|x_i=x,z_j^y,z_i^x)f_{Z^y,Z^x|X}(z_j^y,z_i^x|x_i=x)\,dz_j^y\,dz_i^x\,dy_j}{\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_j^y-z_i^x|<\alpha_N}f_{Y|X,Z^y,Z^x}(y_j|x_i=x,z_j^y,z_i^x)f_{Z^y,Z^x|X}(z_j^y,z_i^x|x_i=x)\,dz_j^y\,dz_i^x\,dy_j}
$$

$$
B_N(x) = \frac{\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}\pi^N(y_j,x)\rho(y_j,x;\theta)f_{Y,Z^y}(y_j,z_j^y)f_{Z^x|X}(z_i^x|x_i=x)\,dz_j^y\,dz_i^x\,dy_j}{\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}f_{Y,Z^y}(y_j,z_j^y)f_{Z^x|X}(z_i^x|x_i=x)\,dz_j^y\,dz_i^x\,dy_j}
$$

Using condition (4.8), we obtain that $A_N(x)$ can be represented as the sum of $A_{N1}(x) =$

$$
(1-\pi(x))\frac{\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_j^y-z_i^x|<\alpha_N}\rho(y_j,x,\theta)f_{Y|X,Z^y,Z^x}(y_j|x_i=x,z_j^y,z_i^x)f_{Z^y,Z^x|X}(z_j^y,z_i^x|x_i=x)\,dz_j^y\,dz_i^x\,dy_j}{\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_j^y-z_i^x|<\alpha_N}f_{Y|X,Z^y,Z^x}(y_j|x_i=x,z_j^y,z_i^x)f_{Z^y,Z^x|X}(z_j^y,z_i^x|x_i=x)\,dz_j^y\,dz_i^x\,dy_j}
$$

and $A_{N2}(x) =$

$$\frac{\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_j^y-z_i^x|<\alpha_N} o_{yx}(1)\rho(y_j,x,\theta)f_{Y|X,Z^y,Z^x}(y_j|x_i=x,z_j^y,z_i^x)f_{Z^y,Z^x|X}(z_j^y,z_i^x|x_i=x)\,dz_j^y\,dz_i^x\,dy_j}{\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_j^y-z_i^x|<\alpha_N} f_{Y|X,Z^y,Z^x}(y_j|x_i=x,z_j^y,z_i^x)f_{Z^y,Z^x|X}(z_j^y,z_i^x|x_i=x)\,dz_j^y\,dz_i^x\,dy_j},$$

where $\sup_{y_j\in\mathcal{Y}}|o_{yx}(1)|\to 0$ as $\alpha_N\to 0$. Proposition 7 gives that $A_{N1}(x)$ coincides with $(1-\pi(x))E[\rho(Y,X;\theta)\mid X=x]$.

$A_{N2}(x)$ converges to 0 because of uniform in y convergence property of $o_{yx}(1)$, the inequality

$$\begin{aligned}\|A_{N2}(x)\| &\le \sup_{y_j\in\mathcal{Y}}|o_{yx}(1)|\cdot E\left[\|\rho(Y,X;\theta)\|\mid X=x,|Z^x-Z^y|<\alpha_N,|Z^x|>\frac{1}{\alpha_N}\right]\\ &= \sup_{y_j\in\mathcal{Y}}|o_{yx}(1)|\cdot E\left[\|\rho(Y,X;\theta)\|\mid X=x\right]\end{aligned}$$

and condition $E\left[\|\rho(Y,X;\theta)\|\mid X=x\right]<\infty$. The fact that

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

can be established by using techniques in the proof of Proposition 7.

To summarize,

$$A_N(x)\to(1-\pi(x))E[\rho(Y,X;\theta)\mid X=x].$$

Now let us analyze $B_N(x)$, which can be represented as the sum of

$$B_{N1}(x)=\pi(x)\frac{\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}\rho(y_j,x;\theta)f_{Y,Z^y}(y_j,z_j^y)f_{Z^x|X}(z_i^x|x_i=x)\,dz_j^y\,dz_i^x\,dy_j}{\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}f_{Y,Z^y}(y_j,z_j^y)f_{Z^x|X}(z_i^x|x_i=x)\,dz_j^y\,dz_i^x\,dy_j}$$

and

$$B_{N2}(x)=\frac{\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}o_{yx}(1)\rho(y_j,x;\theta)f_{Y,Z^y}(y_j,z_j^y)f_{Z^x|X}(z_i^x|x_i=x)\,dz_j^y\,dz_i^x\,dy_j}{\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}f_{Y,Z^y}(y_j,z_j^y)f_{Z^x|X}(z_i^x|x_i=x)\,dz_j^y\,dz_i^x\,dy_j}.$$

The denominator in $B_{N1}(x)$ can be rewritten as

$$\int\left(\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}f_{Z^x|X}(z_i^x|x_i=x)f_{Z^y|Y}(z_j^y|y_j)\,dz_j^y\,dz_i^x\right)f_Y(y_j)\,dy_j.$$

From Assumption 3 (iv), for small α_N it is equal to

$$\int\left(\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}f_{Z^x|X}(z_i^x|x_i=x)g_2(z_j^y)(1+o_{yz^y}(1))\,dz_j^y\,dz_i^x\right)f_Y(y_j)\,dy_j,$$

where $\sup_{|z_j^y|>\frac{1}{\alpha_N}-\alpha_N} \sup_{y_j\in\mathcal{Y}} |o_{yz^y}(1)| \to 0$ as $\alpha_N \to 0$. Then we can write it as

$$\int_{|z_i^x|>\frac{1}{\alpha_N}} \int_{|z_i^x - z_j^y|<\alpha_N} f_{Z^x|X}(z_i^x|x_i = x) g_2(z_j^y)\, dz_j^y\, dz_i^x +$$

$$\int \left(\int_{|z_i^x|>\frac{1}{\alpha_N}} \int_{|z_i^x - z_j^y|<\alpha_N} o_{yz^y}(1) f_{Z^x|X}(z_i^x|x_i = x) g_2(z_j^y)\, dz_j^y\, dz_i^x \right) f_Y(y_j)\, dy_j.$$

Analogously, the numerator in $B_{N1}(x)$ can be written as

$$\int \rho(y_j, x; \theta) \left(\int_{|z_i^x|>\frac{1}{\alpha_N}} \int_{|z_i^x - z_j^y|<\alpha_N} f_{Z^x|X}(z_i^x|x_i = x) f_{Z^y|Y}(z_j^y|y_j)\, dz_j^y\, dz_i^x \right) f_Y(y_j)\, dy_j,$$

and for small α_N it is equal to

$$\int \rho(y_j, x; \theta) \left(\int_{|z_i^x|>\frac{1}{\alpha_N}} \int_{|z_i^x - z_j^y|<\alpha_N} f_{Z^x|X}(z_i^x|x_i = x) g_2(z_j^y)(1 + o_{yz^y}(1))\, dz_j^y\, dz_i^x \right) f_Y(y_j)\, dy_j =$$

$$\int \rho(y_j, x; \theta) f_Y(y_j)\, dy_j \cdot \int_{|z_i^x|>\frac{1}{\alpha_N}} \int_{|z_i^x - z_j^y|<\alpha_N} f_{Z^x|X}(z_i^x|x_i = x) g_2(z_j^y)\, dz_j^y\, dz_i^x +$$

$$\int \rho(y_j, x; \theta) \left(\int_{|z_i^x|>\frac{1}{\alpha_N}} \int_{|z_i^x - z_j^y|<\alpha_N} o_{yz^y}(1) f_{Z^x|X}(z_i^x|x_i = x) g_2(z_j^y)\, dz_j^y\, dz_i^x \right) f_Y(y_j)\, dy_j.$$

Thus, $B_{N1}(x)$ can be written as the sum of the following two terms:

$$\pi(x) \int \rho(y_j, x; \theta) f_Y(y_j)\, dy_j \cdot \frac{C_{N1}(x)}{C_{N1}(x) + \int D_{N1}(y_j, x) f_Y(y_j)\, dy_j}$$

and

$$\pi(x) \frac{\int \rho(y_j, x; \theta) D_{N1}(y_j, x) f_Y(y_j)\, dy_j}{C_{N1}(x)} \cdot \frac{C_{N1}(x)}{C_{N1}(x) + \int D_{N1}(y_j, x) f_Y(y_j)\, dy_j},$$

where

$$C_{N1}(x) = \int_{|z_i^x|>\frac{1}{\alpha_N}} \int_{|z_i^x - z_j^y|<\alpha_N} f_{Z^x|X}(z_i^x|x_i = x) g_2(z_j^y)\, dz_j^y\, dz_i^x,$$

$$D_{N1}(y_j, x) = \int_{|z_i^x|>\frac{1}{\alpha_N}} \int_{|z_i^x - z_j^y|<\alpha_N} o_{yz^y}(1) f_{Z^x|X}(z_i^x|x_i = x) g_2(z_j^y)\, dz_j^y\, dz_i^x.$$

Since

$$-\sup_{|z_j^y|>\frac{1}{\alpha_N}-\alpha_N} \sup_{y_j\in\mathcal{Y}} |o_{yz^y}(1)| \cdot C_{N1}(x) \le D_{N1}(y_j, x) \le \sup_{|z_j^y|>\frac{1}{\alpha_N}-\alpha_N} \sup_{y_j\in\mathcal{Y}} |o_{yz^y}(1)| \cdot C_{N1}(x)$$

and $\sup_{|z_j^y|>\frac{1}{\alpha_N}-\alpha_N} \sup_{y_j\in\mathcal{Y}} |o_{yz^y}(1)| \to 0$, then

$$\frac{C_{N1}(x)}{C_{N1}(x) + \int D_{N1}(y_j, x) f_Y(y_j)\, dy_j} \to 1$$

as $\alpha_N \to 0$.

Also, $\frac{\int \rho(y_j,x;\theta) D_{N1}(y_j,x) f_Y(y_j)\, dy_j}{C_{N1}(x)} \to 0$ because

$$\left\| \int \rho(y_j,x;\theta) D_{N1}(y_j,x) f_Y(y_j)\, dy_j \right\| \leq \sup_{|z_j^y| > \frac{1}{\alpha_N} - \alpha_N} \sup_{y_j \in \mathcal{Y}} |o_{yz^y}(1)| \cdot C_{N1}(x) \cdot \int \|\rho(y_j,x;\theta)\| f_Y(y_j)\, dy_j$$

and $\int \|\rho(y_j,x;\theta)\| f_Y(y_j)\, dy_j < \infty$.

Thus, we obtain that $B_{N1}(x)$ converges to

$$\pi(x) \int \rho(y_j,x;\theta) f_Y(y_j)\, dy_j.$$

As for $B_{N2}(x)$, the norm $\|B_{N2}(x)\|$ is bounded from above by the sum of

$$\sup_{y_j \in \mathcal{Y}} |o_{yx}(1)| \cdot \int \|\rho(y_j,x;\theta)\| f_Y(y_j)\, dy_j \cdot \left| \frac{C_{N1}(x)}{C_{N1}(x) + \int D_{N1}(y_j,x) f_Y(y_j)\, dy_j} \right|$$

and

$$\sup_{y_j \in \mathcal{Y}} |o_{yx}(1)| \cdot \sup_{|z_j^y| > \frac{1}{\alpha_N} - \alpha_N} \sup_{y_j \in \mathcal{Y}} |o_{yz^y}(1)| \cdot \int \|\rho(y_j,x;\theta)\| f_Y(y_j)\, dy_j \cdot \left| \frac{C_{N1}(x)}{C_{N1}(x) + \int D_{N1}(y_j,x) f_Y(y_j)\, dy_j} \right|,$$

and hence, $B_{N2}(x) \to 0$ as $\alpha_N \to 0$.

To summarize,

$$B_N(x) \to \pi(x) \int \rho(y_j,x;\theta) f_Y(y_j)\, dy_j = \pi(x) E^*[\rho(\widetilde{Y},\, X;\theta) | X = x].$$

Proof of Proposition 9. Fix $\widetilde{\theta} \in \Theta_\infty$. Let $\pi \in \Pi^\infty$ be such that $\widetilde{\theta}$ minimizes

$$Q(\theta,\pi) = g_\pi(h,\theta)' W_0 g_\pi(h,\theta).$$

We can find a sequence $\{\pi^N(\cdot,\cdot)\}$ that converges to π uniformly over all y and all x. Let θ_N be any value that minimizes

$$Q_N(\theta,\pi^N) = g^N(h,\theta)' W_0 g^N(h,\theta)$$

for $\pi^N(\cdot,\cdot)$. Clearly, $\theta_N \in \Theta_N$. Let us show that $\theta_N \to \widetilde{\theta}$.

First, we establish that $\sup_{\theta \in \Theta} |Q_N(\theta,\pi^N) - Q(\theta,\pi)| \to 0$. Note that

$$Q_N(\theta,\pi^N) - Q(\theta,\pi) = \left(g^N(h,\theta) - g_\pi(h,\theta)\right)' W_0 \left(g^N(h,\theta) - g_\pi(h,\theta)\right) + 2 g_\pi(h,\theta)' W_0 \left(g^N(h,\theta) - g_\pi(h,\theta)\right).$$

Therefore,

$$\sup_{\theta \in \Theta} |Q_N(\theta,\pi^N) - Q(\theta,\pi)| \leq \sup_{\theta \in \Theta} \|g^N(h,\theta) - g_\pi(h,\theta)\|^2 \|W\| + 2 \sup_{\theta \in \Theta} \|g_\pi(h,\theta)\| \sup_{\theta \in \Theta} \|g^N(h,\theta) - g_\pi(h,\theta)\| \|W_0\|.$$

Conditions (4.11) imply that $\sup_{\theta\in\Theta}\|g_\pi(h,\theta)\|<\infty$. Thus, we only need to establish that $\sup_{\theta\in\Theta}\|g^N(h,\theta)-g_\pi(h,\theta)\|\to 0$. Similarly to how it was done in Proposition 8, we can show that $g^N(h,\theta)$ can be represented as the sum of four terms –

$$g^N(h,\theta)=A_{N1}+A_{N2}+B_{N1}+B_{N2},$$

where

$$A_{N1}=(1-\pi)\frac{\int\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_j^y-z_i^x|<\alpha_N}h(x_i)\rho(y_j,x_i;\theta)f_{Y,X|Z^y,Z^x}(y_j|z_j^y,z_i^x)f_{Z^y,Z^x}(z_j^y,z_i^x)\,dz_j^y\,dz_i^x\,dy_j\,dx_i}{\int\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_j^y-z_i^x|<\alpha_N}f_{Y,X|Z^y,Z^x}(y_j|z_j^y,z_i^x)f_{Z^y,Z^x}(z_j^y,z_i^x)\,dz_j^y\,dz_i^x\,dy_j\,dx_i}$$

$$A_{N2}=\frac{\int\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_j^y-z_i^x|<\alpha_N}o_{yx}(1)h(x_i)\rho(y_j,x_i;\theta)f_{Y,X|Z^y,Z^x}(y_j|z_j^y,z_i^x)f_{Z^y,Z^x}(z_j^y,z_i^x)\,dz_j^y\,dz_i^x\,dy_j\,dx_i}{\int\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_j^y-z_i^x|<\alpha_N}f_{Y,X|Z^y,Z^x}(y_j|z_j^y,z_i^x)f_{Z^y,Z^x}(z_j^y,z_i^x)\,dz_j^y\,dz_i^x\,dy_j\,dx_i}$$

$$B_{N1}=\pi\frac{\int\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}h(x_i)\rho(y_j,x_i;\theta)f_{Y,Z^y}(y_j,z_j^y)f_{X,Z^x}(x_i,z_i^x)\,dz_j^y\,dz_i^x\,dy_j\,dx_i}{\int\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}f_{Y,Z^y}(y_j,z_j^y)f_{X,Z^x}(x_i,z_i^x)\,dz_j^y\,dz_i^x\,dy_j\,dx_i}$$

$$B_{N2}=\frac{\int\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}o_{yx}(1)h(x_i)\rho(y_j,x_i;\theta)f_{Y,Z^y}(y_j,z_j^y)f_{X,Z^x}(x_i,z_i^x)\,dz_j^y\,dz_i^x\,dy_j\,dx_i}{\int\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}f_{Y,Z^y}(y_j,z_j^y)f_{X,Z^x}(x_i,z_i^x)\,dz_j^y\,dz_i^x\,dy_j\,dx_i},$$

where terms $o_{yx}(1)$ do not depend on θ and are such that $\sup_{y_j\in\mathcal{Y},x_i\in\mathcal{X}}|o_{yx}(1)|\to 0$ as $\alpha_N\to 0$. Proposition 7 implies that $E\left[h(X)\rho(Y,X;\theta)||Z^x|>\frac{1}{\alpha},|Z^x-Z^y|<\alpha\right]=E[h(X)\rho(Y,X;\theta)]$. Therefore,

$$A_{N1}=(1-\pi)E[h(X)\rho(Y,X;\theta)],$$

and

$$g^N(h,\theta)-g_\pi(h,\theta)=A_{N2}+B_{N1}+B_{N2}-\pi E^*\left[h(X)\rho(\widetilde{Y},\,X;\theta)\right].$$

Note that

$$\begin{aligned}\sup_{\theta\in\Theta}\|A_{N2}\|&\le\sup_{y_j,x_i}|o_{yx}(1)|\cdot E\left[\sup_{\theta\in\Theta}\|h(X)\rho(Y,X;\theta)\|\;\middle|\;|Z^x|>\frac{1}{\alpha},|Z^x-Z^y|<\alpha\right]\\&=\sup_{y_j,x_i}|o_{yx}(1)|\cdot E\left[\sup_{\theta\in\Theta}\|h(X)\rho(Y,X;\theta)\|\right]\to 0\end{aligned}$$

as $\alpha_N\to 0$.

From Assumption 3 (iv), for small α_N the denominator in B_{N1} is the sum

$$\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}g_2(z_j^y)g_1(z_i^x)\,dz_j^y\,dz_i^x+$$

$$\int\int\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}\left(o_{xz^x}(1)+o_{z^yy}(1)+o_{z^yy}(1)o_{xz^x}(1)\right)g_2(z_j^y)g_1(z_i^x)f_Y(y_j)f_X(x_i)\,dz_j^y\,dz_i^x\,dy_j\,dx_i,$$

and, similarly, the numerator is the sum

$$\int\int h(x_i)\rho(y_j,x_i;\theta)\,dy_j\,dx_i\cdot\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N} g_2(z_j^y)g_1(z_i^x)\,dz_j^y\,dz_i^x+$$

$$\int\int h(x_i)\rho(y_j,x_i;\theta)\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}\left(o_{xz^x}(1)+o_{z^yy}(1)+o_{z^yy}(1)o_{xz^x}(1)\right)g_2(z_j^y)g_1(z_i^x)f_Y(y_j)f_X(x_i)\,dz_j^y\,dz_i^x\,dy_j\,dx_i,$$

where $o_{yz^y}(1)$ and $o_{xz^x}(1)$ do not depend on θ and are such that $\sup_{|z_j^y|>\frac{1}{\alpha_N}-\alpha_N}\sup_{y_j}|o_{yz^y}(1)|\to 0$ and $\sup_{|z_i^x|>\frac{1}{\alpha_N}}\sup_{x_i}|o_{xz^x}(1)|\to 0$ as $\alpha_N\to 0$. Then $B_{N1}-\pi E^*\left[h(X)\rho(\widetilde{Y},X;\theta)\right]$ is the sum of the following two terms:

$$\pi E^*\left[h(X)\rho(\widetilde{Y},X;\theta)\right]\cdot\left(\frac{C_{N1}}{C_{N1}+\int\int D_{N1}(y_j,x_i)f_Y(y_j)f_X(x_i)\,dy_j\,dx_i}-1\right) \tag{B.18}$$

and

$$\pi\cdot\frac{\int\int h(x_i)\rho(y_j,x_i;\theta)D_{N1}(y_j,x_i)f_Y(y_j)f_X(x_i)\,dy_j\,dx_i}{C_{N1}+\int\int D_{N1}(y_j,x_i)f_Y(y_j)f_X(x_i)\,dy_j\,dx_i}, \tag{B.19}$$

where

$$C_{N1}=\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N} g_2(z_j^y)g_1(z_i^x)\,dz_j^y\,dz_i^x$$

$$D_{N1}(y_j,x_i)=\int_{|z_i^x|>\frac{1}{\alpha_N}}\int_{|z_i^x-z_j^y|<\alpha_N}\left(o_{xz^x}(1)+o_{z^yy}(1)+o_{z^yy}(1)o_{xz^x}(1)\right)g_2(z_j^y)g_1(z_i^x)\,dz_j^y\,dz_i^x$$

The $\sup_{\theta\in\Theta}$ of the norm of the term in (B.18) is bounded from above by

$$\pi E^*\left[\sup_{\theta\in\Theta}\|h(X)\rho(\widetilde{Y},X;\theta)\|\right]\cdot\left|\frac{C_{N1}}{C_{N1}+\int\int D_{N1}(y_j,x_i)f_Y(y_j)f_X(x_i)\,dy_j\,dx_i}-1\right|.$$

Because

$$|D_{N1}(y_j,x_i)|\le\sup_{|z_i^x|>\frac{1}{\alpha_N}}\sup_{|z_j^y|>\frac{1}{\alpha_N}-\alpha_N}\sup_{y_j,x_i}|o_{yz^yxz^x}(1)|\cdot C_{N1}$$

with $\sup_{|z_i^x|>\frac{1}{\alpha_N}}\sup_{|z_j^y|>\frac{1}{\alpha_N}-\alpha_N}\sup_{y_j,x_i}|o_{yz^yxz^x}(1)|\to 0$, then $\frac{C_{N1}}{C_{N1}+\int\int D_{N1}(y_j,x_i)f_Y(y_j)f_X(x_i)\,dy_j\,dx_i}\to 1$ as $\alpha_N\to 0$. Hence, (B.18) converges to 0 uniformly over $\theta\in\Theta$.

The $\sup_{\theta\in\Theta}$ of the norm of the term in (B.19) is bounded from above by

$$\pi\cdot\frac{\int\int\sup_{\theta\in\Theta}\|h(x_i)\rho(y_j,x_i;\theta)\||D_{N1}(y_j,x_i)|f_Y(y_j)f_X(x_i)\,dy_j\,dx_i}{C_{N1}+\int\int D_{N1}(y_j,x_i)f_Y(y_j)f_X(x_i)\,dy_j\,dx_i}\le$$

$$\pi\cdot\sup_{|z_i^x|>\frac{1}{\alpha_N}}\sup_{|z_j^y|>\frac{1}{\alpha_N}-\alpha_N}sup_{y_j,x_i}|o_{yz^yxz^x}(1)|\cdot\frac{C_{N1}\cdot E^*\left[\sup_{\theta\in\Theta}\|h(X)\rho(\widetilde{Y},X;\theta)\|\right]}{C_{N1}+\int\int D_{N1}(y_j,x_i)f_Y(y_j)f_X(x_i)\,dy_j\,dx_i},$$

which converges to 0 as $\alpha_N\to 0$.

Thus, we obtain that

$$\sup_{\theta\in\Theta}\left\|B_{N1}-\pi E^*\left[h(X)\rho(\widetilde{Y},X;\theta)\right]\right\|\to 0.$$

Finally, consider $\sup_{\theta\in\Theta}\|B_{N2}\|$. This norm is bounded from above by the sum of

$$\sup_{y_j,x_i}|o_{yx}(1)|\cdot\int\sup_{\theta\in\Theta}\|h(x_i)\rho(y_j,x_i;\theta)\|f_Y(y_j)f_X(x_i)\,dy_j\,dx_i\cdot\frac{C_{N1}}{C_{N1}+\int\int D_{N1}(y_j,x_i)f_Y(y_j)f_X(x_i)\,dy_j\,dx_i}$$

and

$$\sup_{y_j,x_i}|o_{yx}(1)|\cdot\sup_{|z_i^x|>\frac{1}{\alpha_N}}\sup_{|z_j^y|>\frac{1}{\alpha_N}-\alpha_N}\sup_{y_j,x_i}|o_{yz^yxz^x}(1)|\cdot\frac{C_{N1}\int\sup_{\theta\in\Theta}\|h(x_i)\rho(y_j,x_i;\theta)\|f_Y(y_j)f_X(x_i)\,dy_j\,dx_i}{C_{N1}+\int\int D_{N1}(y_j,x_i)f_Y(y_j)f_X(x_i)\,dy_j\,dx_i},$$

and, hence, $\sup_{\theta\in\Theta}\|B_{N2}\|\to 0$ as $\alpha_N\to 0$.

To summarize our results so far, we showed that

$$\sup_{\theta\in\Theta}\|g^N(h,\theta)-g_\pi(h,\theta)\|\le\sup_{\theta\in\Theta}\|A_{N2}\|+\sup_{\theta\in\Theta}\left\|B_{N1}-\pi E^*\left[h(X)\rho(\widetilde{Y},\,X;\theta)\right]\right\|+\sup_{\theta\in\Theta}\|B_{N2}\|,$$

and, thus, $\sup_{\theta\in\Theta}\|g^N(h,\theta)-g_\pi(h,\theta)\|\to 0$ as $\alpha_N\to 0$. This implies that

$$\sup_{\theta\in\Theta}|Q_N(\theta,\pi^N)-Q(\theta,\pi)|\to 0. \tag{B.20}$$

Now, fix $\varepsilon>0$. Let us show that for large enough N^x,N^y, $Q(\theta^N,\pi)<Q(\widetilde{\theta},\pi)+\varepsilon$. Indeed, (B.20) implies that when N^x,N^y are large enough, $Q(\theta^N,\pi)<Q_N(\theta^N,\pi^N)+\varepsilon/3$. Also, $Q_N(\theta^N,\pi^N)<Q_N(\widetilde{\theta},\pi^N)+\varepsilon/3$ because θ^N is an argmin of $Q_N(\theta^N,\pi^N)$. Finally, (B.20) implies that when N^x,N^y are large enough, $Q_N(\widetilde{\theta},\pi^N)<Q(\widetilde{\theta},\pi)+\varepsilon/3$.

Let S be any open neighborhood of $\widetilde{\theta}$ and let S^c be its complement in $\mathbb{R}^l$. From the compactness of Θ and the continuity of $\rho(\cdot,\cdot,\cdot)$ in θ, we conclude that $\min_{S^c\cap\Theta}Q(\theta,\pi)$ is attained. The fact that $\widetilde{\theta}$ is the unique minimizer of $Q(\theta,\pi)$ gives that $\min_{S^c\cap\Theta}Q(\theta,\pi)>Q(\widetilde{\theta},\pi)$. Denote $\varepsilon=\min_{S^c\cap\Theta}Q(\theta,\pi)-Q(\widetilde{\theta},\pi)$. As we showed above, for this ε we have that when N^x,N^y are large enough,

$$Q(\theta^N,\pi)<Q(\widetilde{\theta},\pi)+\varepsilon=\min_{S^c\cap\Theta}Q(\theta,\pi),$$

which for large enough N^x,N^y gives $\theta^N\in S$. Since S can be chosen arbitrarily small, this means that $\theta^N\to\widetilde{\theta}$.

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, $\theta_0=E[XX']^{-1}E[XY]$. When $\widetilde{Y}$ is drawn from $f_Y(\cdot)$ independently of X, then $E^*\left[X(\widetilde{Y}-X'\theta_1)\right]=0$ gives $\theta_1=E[XX']^{-1}E_X[X]E_Y[\widetilde{Y}]=E[XX']^{-1}E_X[X]E_Y[\widetilde{Y}]$.

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

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

where

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

Note that

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

where

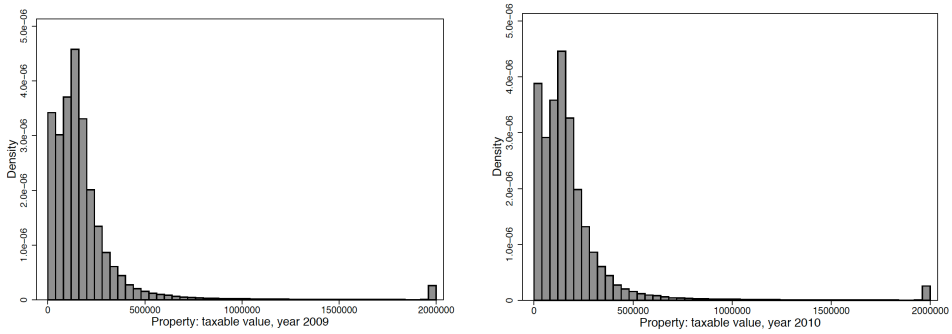
Clearly, $g_\pi(h, \theta)' g_\pi(h, \theta)$ takes the value of 0 if and only if $g_\pi(h, \theta)$ takes the value of 0, which happens if and only if $\theta = (1-\pi)\theta_0 + \pi\theta_1$. Thus for each $\pi \in [\underline{\gamma}, 1]$,

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

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,

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

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



(a) (b)

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

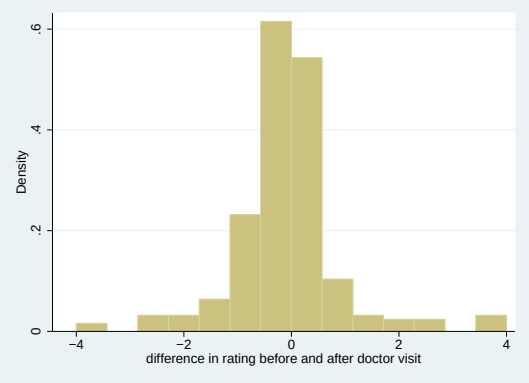


Figure 3: Average treatment effect

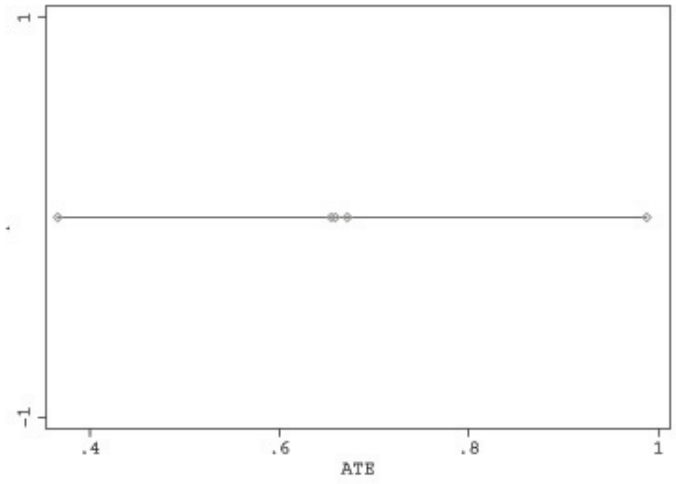


Figure 4: Identified sets for propensity score coefficients

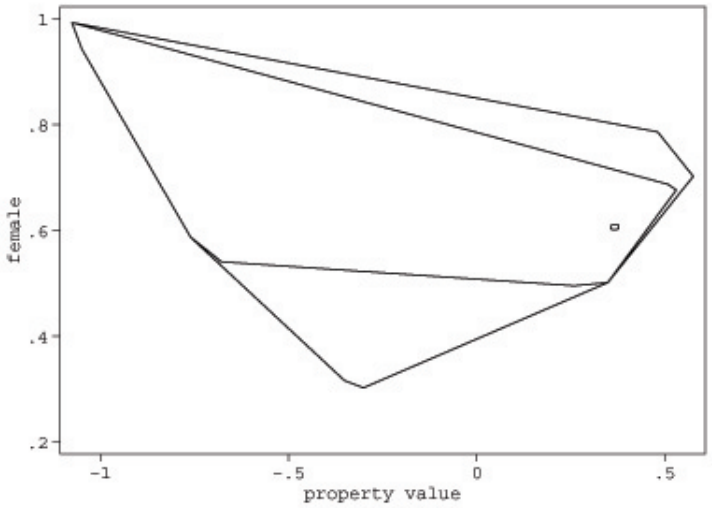


Table 1: Summary statistics from property tax bills in Durham County, NC.

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

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

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

Table 3: Features of edit distance-based matches

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

Table 4: Estimated treatment effects

	(1)	(2)	(3)	(4)
	OLS	OLS	Matching	
	Rating	Rating	Rating	I(After visit)
I(After visit)	0.06	0.033	0.661	
	[0.015]***	[0.054]	[0.37]*	
log(property value)				0.364
				[0.064]***
I(female)				0.61
				[0.062]***
Observations	20723	2605	2605	2605

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

Table 5: Quantile treatment effects

Variable	Obs	Mean	SD	Min	Max
Lower quartile					
Difference	57	-1.144	0.795	-4	-0.5
Upper quartile					
Difference	55	1.026	1.035	0.19	4
Mean difference test: t-stat =1.662					