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

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

This version: April 20, 2015

ABSTRACT

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

JEL Classification: C35, C14, C25, C13.

Keywords: Data protection, model identification, data combination.

[1] *First version: December 2011.* Support from the NSF and STICERD is gratefully acknowledged. We appreciate helpful comments from P. Haile, M. Jansson, C. Manski, A. de Paula, M. Pesendorfer, J. Powell, C. Tucker and E.Tamer. We appreciate feedback from seminar participants at various universities.

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

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

[4] Higher School of Economics, Moscow, Russia, e-mail: evgeny.yakovlev@gmail.com.

1 Introduction

Often, 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. These algorithms step from the long-existing literature in Econometrics and Statistics on validation samples. Such procedures are on the list of routine daily tasks for database management companies and are applied in a variety of settings, from medical to tax and employment databases.[1]

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

In this paper we introduce a novel framework for inference from combined data when individual datasets used for combination do not contain unique individual identifiers. Our framework is only requires *partial* information regarding the quality of the matches between the observations of combined datasets (e.g. upper and lower bounds on these probabilities) and allows to avoid parametric assumptions regarding the joint distribution of combined variables. This contrasts many existing approaches that are either based on the assumption of the known parametric form of the joint distribution of individual observations in the combined data, or a known distribution of the data combination errors. Our framework embeds both these settings as special cases. We develop an approach to identification of the parameters of an econometric model specified on the combined

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

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

dataset and demonstrate when these parameters can and cannot be identified.

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

Our approach to identification is novel as we notice that the data combination procedure in non-parametric settings can only be defined and implemented in the finite sample and not in the population. As a result, the identification is characterized as the property of limits of sequences of data combination rules (as opposed to the property of the population distribution as in the standard literature on identification). This is a crucial aspect in our identification method as we provide a new approach to model identification from combined datasets as a limiting property in the sequence of statistical experiments. Namely, we introduce the notion of identification from combined data through a limit of the set of parameters inferred from the combined data as the sizes of both datasets approach infinity. These sets and their limiting behavior depend, first, on the properties of the data combination procedure and, second, on what kind of information about this procedure is provided to the researcher by the data curator.

Our framework naturally applies to the analysis of situations where the identifying information is intentionally removed from the data by the data curators to reduce the "sensitivity" of the data. In this case, an instance of a successful combination of two observations from two disjoined datasets means that the variables contain enough information to attribute these two observations to the same individual. This implies that the corresponding individual information can be de-anonymized, i.e. the *individual disclosure* can occur. Our novel econometric framework allows us to study estimators that use combined data in the settings where the data curators explicitly limit such cases of individual disclosure. We also study the tradeoff between disclosure limitation (defined by the probability that an individual disclosure can occur) and the quality of identification of the parameters of interest. To our knowledge, our paper is the first one to study such a tradeoff.

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. If one of the data curator's objectives is to provide some privacy guarantees and prevent disclosure when conducting the task of combing the data, then we argue that the issues of model identification/estimation and the risk of disclosure should be analyzed jointly. In particu-

lar, we investigate how the limitations imposed on the risk of disclosure of confidential personal data affect the amount of information that researchers and policy makers can obtain about the empirical model of interest.

Among our findings is that there is a trade-off 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. Further, we analyze the partial identification issue and what estimates our consumer behavior model can deliver under the constraints on the identity disclosure. We note that the goal of our work is not to demonstrate the vulnerability of online personal data but to provide a real example of the tradeoff between privacy and identification.

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

In our empirical application, we illustrate a scenario where the researcher has access to both the sensitive and public datasets, and thus, the researcher essentially takes the role of the data curator. Provided that in this case the researcher can control the properties of the data combination procedure, it becomes her responsibility to ensure that a required bound on the risk of disclosure is imposed. 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.

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

The setup of this paper can be applied to situations when there are several independent data curators having access to separate datasets. Private firms and large government agencies collect large socio-economic datasets. The Internal Revenue Service, Social Security Administration and the US Census Bureau collect large comprehensive datasets that have large or complete overlaps over individuals whose data has been collected. Each of these agencies operate as independent data curators meaning that each of them has full control over their data, full exclusion rights over access to these data. Most existing data curators operate based on the vault storage model where the data is stored locally in a secure location and raw disaggregated data cannot be taken outside of the vault. Within their data management programs, each such a data owner allows researchers to access the data vault upon passing some clearance procedure. With this data analysis model there could be many researchers who can access many of such data vaults. However, provided that the raw data cannot be removed from the vault, neither of these researchers can combine individual data from two or more such vaults. Thus, this is the situation where each of the researchers knows the marginal distribution of the data in each of the vaults. However, none of the researchers knows the joint distribution of the data across the vaults and thus cannot estimate the model that contains the variables from multiple sources. Recently, several empirical researchers have been able to obtain permissions to merge separate administrative data sources. We note that while each data curator controls their own dataset, they also control the "sensitivity" of the variables contained in the dataset. For instance, some variables can be removed from the researcher's access based on the disclosure risk considerations. Such a risk cannot be controlled if the data from one source controlled by one data curator are combined with the data controlled by another data curator. Provided that the marginal data distributions from different sources are already known to the researchers the disclosure threat in this case comes precisely from the data combination.

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 pseudo-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 pseudo-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.

Related literature.

Our paper is related to several strands in the computer science literature. One of them is on the optimal structures of linkage attacks as well as the requirements in relation to data releases. The structure of linkage attacks is based on the optimal record linkage results that have been long used in the analysis of databases and data mining. To some extent, these results have been used used in econometrics for combination of datasets as described in Ridder and Moffitt [2007]. In record linkage, one provides a (possibly) probabilistic rule that can match the records from one dataset with the records from the other dataset in an effort to link the data entries corresponding to the same individual.[3] In several striking examples, computer scientists have shown that a simple removal of personal information such as names and social security numbers does not protect data from individual disclosure. For instance, Sweeney [2002b] identified the medical records of William Weld, then governor of Massachusetts, by linking voter registration records to "anonymized" Massachusetts Group Insurance Commission (GIC) medical encounter data, which retained the birthdate, sex, and zip code of the patient.

In relation to the security of individual data, the computer science literature, e.g. Samarati and Sweeney [1998], Sweeney [2002a], Sweeney [2002b], LeFevre et al. [2005], Aggarwal et al. [2005], LeFevre et al. [2006], Ciriani et al. [2007], has developed and implemented the so-called k-anonymity approach. A database instance is said to provide k-anonymity, for some number k, if every way of singling an individual out of the database returns records for at least k individuals. In other words, anyone whose information is stored in the database can be "confused" with k others. Under k-anonymity, a data combination procedure will respect the required bound on the disclosure risk. We describe it in Section 2.3 and use it in the empirical part. An alternative solution is in the use of synthetic data and a related notion of differential privacy, e.g. Dwork and Nissim [2004], Dwork [2006], Abowd and Vilhuber [2008], as well as Duncan and Lambert [1986], Duncan and Mukherjee [1991], Duncan and Pearson [1991], Fienberg [1994], and Fienberg [2001] Duncan et al. [2001], Abowd and Woodcock [2001].

We note that while the computer science literature has alluded to the point that data protection may lead to certain trade-offs in data analysis, data protection has never been considered in the context of model identification. For instance, a notion of "data utility" has been introduced that

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

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.

Our paper is also related to the literature on partial identification of models with contaminated or corrupted data, even though our identification approach is new. Manski [2003], Manski [2007] and Horowitz and Manski [1995] note that data errors or data modifications pose identification problems and generally result in only set identification of the parameter of interest. Manski and Tamer [2002] and Magnac and Maurin [2008] give examples where – for confidentiality or anonymity reasons – the data may be transformed into interval data or some attributes may be suppressed, leading to the loss of point identification of the parameters of interest. Consideration of the general setup in Molinari [2008] allows one to assess the impact of some data "anonymization" as a general misclassification problem. Cross and Manski [2002] and King [1997] study the ecological inference problem where a researcher needs to use the data from several distinct datasets to conduct inference on a population of interest. In ecological inference, several datasets usually of aggregate data are available. Making inferences about micro-units or individual behavior in this case is extremely difficult because variables that allow identification of units are not available. Cross and Manski [2002] show that the parameters of interest are only partially identified. We note that in our case the data contain individual observation on micro-units and there is a limited overlap between two datasets, making the inference problem dramatically different from ecological inference.

Though less directly related to our analysis, there is also a literature within economics that considers privacy as something that may have a subjective value for consumers (see Acquisti [2004]) rather than a formal guarantee against intruders' attacks. Considering personal information as a "good" valued by consumers leads to important insights in the economics of privacy. As seen in Varian [2009], this approach allows researchers to analyze the release of private data in the context of the tradeoff between the network effects created by the data release and the utility loss associated with this release. The network effect can be associated with the loss of competitive advantage of the owner of personal data, as discussed in Taylor [2004], Acquisti and Varian [2005], Calzolari and Pavan [2006]. Consider the setting where firms obtain a comparative advantage due to the possibility of offering prices that are based on the past consumer behavior. Here, a subjective individual perception of privacy is important. This is clearly shown in both the lab experiments in Gross and Acquisti [2005], Acquisti and Grossklags [2008], as well as in the real-world environment in Acquisti et al. [2006], Miller and Tucker [2009] and Goldfarb and Tucker [2010]. Given all these findings, we believe that disclosure protection is a central theme in the privacy discourse, as privacy protection is impossible without the data protection.

2 Econometric model

2.1 Model and data structure

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

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

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

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 $x \in \mathcal{X}$,

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

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

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

ASSUMPTION 1. *Parameter θ_0 is uniquely determined from the moment equation (2.1) and the 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 deal with the problem of 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 healthcare business and healthcare 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 is 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 $\mathcal{D}^{xv} = \{x_i, v_i\}_{i=1}^{N^x}$.[4]

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

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

EXAMPLE 1. *For linear models, without any additional information identification with split sample data comes down to computing Fréchet bounds. For example, in a bivariate linear regression of random variable Y on random variable X with $Var[X] > 0$, the slope coefficient can be expressed as*

$$b_0 = \frac{cov(Y, X)}{Var[X]}.$$

Because the joint distribution of Y and X is unknown, $cov(Y, X)$ cannot be calculated even if the marginal distributions of Y and X are available.

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

[5] This assumption is imposed for technical simplicity and can be relaxed.

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

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

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

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

2.2 Identifiers and decisions rules for data combination

Our data linkage procedure is based on comparing the value of an identifier Z^y constructed for each observation in the main dataset with the value of an identifier Z^x constructed for each observation in the auxiliary dataset. These identifiers are random vectors that can consist of both numerical and string variables. $Z^y = Z^y(Y,W)$ is the multivariate function of Y and an auxiliary random vector W observed together with Y, while $Z^x = Z^x(X,V)$ is the multivariate function of X and some an auxiliary random vector V observed together with X. We suppose that these identifiers Z^y and

Z^x are constructed in such a way that they have the same dimension and the same support. Our combination rule is based on comparing the values of z_i^y and z_i^x for each $j = 1, \ldots, N^y$ and each $i = 1, \ldots, N^x$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.3 Data combination from observations with infrequent values

Let us define the norm of z_i^x as

$$\|z_i^x\| = \omega_n \|z_i^{x,n}\|_E + \omega_s \|z_i^{x,s}\|_S.$$

Analogously, the norm of z_j^y is

$$\|z_j^y\| = \omega_n \|z_j^{y,n}\|_E + \omega_s \|z_j^{y,s}\|_S.$$

By infrequent attributes we mean the values of identifiers in the tails.

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

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

(i) (Proximity of identifiers with extreme values)

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

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

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

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

(iii) (Redundancy of identifiers in the full data)

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

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

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

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

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

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

Assumption 3 implies that the ordering of the values of $\|Z^y\|$ and $\|Z^x\|$ is meaningful and that the tails of the distributions of $\|Z^x\|$ and $\|Z^y\|$ contains extreme values. If we considered a situation when all the variables in Z^y and Z^x were discrete, this would mean that at least one of these variables has an infinite support – for instance, it takes integer values 1, 2, 3, ... with positive probabilities. Ridder and Moffitt [2007] overview cases where *a priori* available numeric identifiers Z^y and Z^x are jointly normally distributed random variables, but we avoid making such specific distributional assumptions.

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

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

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

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

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

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

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

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

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

We recognize that Assumption 3 puts restrictions on the behavior of infrequent (tail) realizations of identifiers Z^x and Z^y. Specifically, we expect that conditional on $\|Z^x\|$ taking a high value, the values of identifiers constructed from two datasets must be close. We illustrate this assumption with our empirical application, where 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.

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

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

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

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

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

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

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

either

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

or

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

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

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

Clealry, it always holds that

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

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

3 Implementation of data combination and implications for identity disclosure

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

3.1 Implementation of data combination

In our model, the realizations of random variables Y and X are contained in disjoint datasets. After constructing identifiers Z^y and Z^x, we directly observe the empirical distributions of (Y, Z^y) and (X, Z^x). Even though these two distributions provide some information about the joint distribution of (Y, X), such as Fréchet bounds, they do not fully characterize it if no data combination whatsoever is conducted, and thus, there are many joint distributions of (Y, X) (or, more generally, joint distributions of (Y, Z^y, X, Z^x)) consistent with the observed distributions of (Y, Z^y) and (X, Z^x). This means that we would have to consider all such compatible joint distributions of (Y, X) when trying to determine the parameter of interest using (2.1). Intuitively, any compatible joint distribution of (Y, X) would give us a different value of the parameter of interest, which means that the

parameter of interest can only be determined up to a set. Thus, the econometric model of interest is not identified from the available information about the distributions of (Y, Z^y) and (X, Z^x).

The identification of the econometric model is potentially possible if the two datasets are combined for at least some observations and thus, more information becomes available about the dependence structure between vectors (Y, Z^y) and (X, Z^x), from which we can consequently obtain more information about the dependence structure between Y and X. The best case scenario from the identification point of view occurs if our data combination procedure allows us to learn the copula describing the true joint distribution of (Y, Z^y, X, Z^x) as a function of two separate distributions of (Y, Z^y) and (X, Z^x). This would automatically give us the copula describing the true joint distribution of (Y, X) as a function of the marginal distributions of Y and X, and then we would be able to point identify θ_0 using (2.1). Whether this scenario will occur clearly depends on the quality of the data combination procedure.

An important feature of data combination that has to be taken into account is that it is inherently a finite-sample procedure. Therefore, in Section 4 we define identification from combined data as a property of the limit of statistical experiments (as the finite-sample increases). To the best of our knowledge, this is a new approach to analyzing parameter identification from combined data.

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

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

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

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

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

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

Consider an observation i from $\mathcal{D}^x$ such that $\|z_i^x\| \geq 1/\alpha_N$. If we find a data entry j from the dataset $\mathcal{D}^y$ such that $d(z_j^y, z_i^x) < \alpha_N$, then we consider i and j as a potential match. In other words, if identifiers z_i^x and z_j^y are both large and are close, then we consider (x_i, z_i^x) and (y_j, z_j^y) as observations possibly corresponding to the same individual. This seems to be a good strategy when α_N is small because, according to Assumption 3, when the pair (Z^x, Z^y) is drawn from their

true joint distribution, the conditional probability of Z^x and Z^y taking proximate values when Z^x is large in the absolute value is close to 1. Even though the decision rule is independent of the values of x_i and y_j, the probability $Pr\left(m_{ij} = 1 \mid \mathcal{D}_N(y_j, x_i, z_j^y, z_i^x), x_i = x, y_j = y, \mathcal{D}^x, \mathcal{D}^y\right)$ for a finite $N = (N^x, N^y)$ can depend on these values (and also depend on the sizes of datasets $\mathcal{D}^x$ and $\mathcal{D}^y$) and therefore can differ across pairs of i and j.

Using the combination rule $\mathcal{D}_N(\cdot)$, for each $j \in \{1, \ldots, N^y\}$ from the database $\mathcal{D}^y$ we try to find an observation i from the database $\mathcal{D}^x$ that satisfies our matching criteria and thus presents a potential match for j. We can then add the "long" vector (y_j, z_j^y, x_i, z_i^x) to our combined dataset if neither (y_j, z_j^y) for this specific j nor (x_i, z_i^x) for this specific i enter the combined dataset as subvectors of other "long" observations. In other words, if there are several possible matches i from $\mathcal{D}^x$ for some j in $\mathcal{D}^y$(or several possible matches j from $\mathcal{D}^y$ for some i in $\mathcal{D}^x$), we can put only one of them in our combined dataset. Mathematically, each combined dataset $\mathcal{G}_N$ can be described by an $N^y \times N^x$ matrix $\{d_{ji}, j = 1, \ldots, N^y; i = 1, \ldots, N^x\}$ of zeros and ones, which satisfies the following conditions:

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

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

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

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

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

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

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

split datasets, the probability of the error of this kind is

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

or

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

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

(2) Data combination errors of the second kind occur when observations j and i do belong to the same individual but the procedure does not identify these two observations as a potential match (we still consider i such that $\|z_i^x\| \geq 1/\alpha_N$). For the two given split datasets, the probability of the error of this kind is

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

or

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

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

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

3.2 Risk of disclosure

What we notice so far is that given that there is no readily available completely reliable similarity metric between the two databases we rely on the probabilistic properties of the data. As a result, in estimation we have to resort to only using the pairs of combined observations. If correct matches are made with a sufficiently high probability, this may pose a potential problem if one of the two datasets contains sensitive individual-level information. In fact, if the main dataset contains de-personalized but highly sensitive individual data and the auxiliary dataset, which is being combined with the main dataset, contains publicly available individual-level information (such as demographic data, names and addresses, etc.), 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 such an information leakage is to control the accuracy of utilized data combination procedures. In particular, we consider controlling the error of the first kind.

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

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

Propositions 1 and 2, which appear later in this section, give conditions on the sequence of α_N, $\alpha_N \to 0$, that are sufficient to guarantee that the probability of the error of the first kind vanishes as $N^y \to \infty$. Proposition 5 give conditions on α_N, $\alpha_N \to 0$, under which the probability of the error of the first kind is separated away from 0 as $N^y \to \infty$.

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

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

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

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

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

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

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

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

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

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

Our definition of the disclosure guarantee requires, first of all, that for any two finite datasets $\mathcal{D}^y$ and $\mathcal{D}^x$ and any matched pair, the value of $p_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y)$ is strictly less than one. In other words, there is always a positive probability of making a linkage mistake. However, even if probabilities $p_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y)$ are strictly less than 1, they may turn out to be very high when N^y is sufficiently large and α_N is sufficiently small. 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, 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. The value of $\bar{\gamma}$ is the extent of the non-disclosure risk guarantee.

We emphasize that the risk of disclosure needs to be controlled in split datasets of any sizes with any realizations 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 specified bound. This requirement is very different from the guarantee with probability one as here we need to ensure that even for datasets observed with an extremely low probability the probability of a correct match honors the stipulated bound.

An important practical question is whether there exist (the classes of the) decision rules that guarantee a specified bound on the disclosure risk. Below we present results that indicate, first, that for a given bound on the disclosure risk we can find sequences of thresholds such that the corresponding decision rules honor this bound, and second, that the rates of convergence for these sequences depend on the tail behavior of identifiers used in the data combination procedure. Propositions 1 and 5 give general results. Propositions 3, 4 and 6, 7 consider two important cases where the tails of the distributions of identifiers are geometric and exponential.

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

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

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

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

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

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

Then non-disclosure is not guaranteed.

PROPOSITION 3. *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*

$$\alpha_N = o\left(\frac{1}{(N^x)^{\frac{1}{c_2+2}}} \right) \tag{3.11}$$

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

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

and, thus, non-disclosure is not guaranteed.

PROPOSITION 4. ***(Absence of 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* $\alpha_N \to 0$ *(as* $N^y \to \infty$*) be chosen in such a way that*

$$\lim_{N^y \to \infty} N^x e^{-\frac{c_2}{\alpha_N}} \alpha_N = 0. \tag{3.12}$$

Then

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

and, thus, non-disclosure is not guaranteed.

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

The proofs of propositions 1 - 4 are in Appendix B.

The next three propositions describe instances in which non-disclosure can be guaranteed.

PROPOSITION 5. ***(Non-disclosure risk guarantee).*** *Suppose that for given non-decreasing and positive for* $\alpha \in (0, \bar{\alpha})$ *functions* $\phi(\cdot)$ *and* $\psi(\cdot)$ *the sequence of* $\alpha_N \to 0$ *(as* $N^y \to \infty$*) is chosen in such a way that*

$$\liminf_{N^y \to \infty} \frac{N^x}{\phi(\alpha_N)} \int_{\frac{1}{\alpha_N}}^{\infty} \left(\psi\left(\frac{1}{z - \alpha_N} \right) - \psi\left(\frac{1}{z + \alpha_N} \right) \right) \frac{\phi'\left(\frac{1}{z}\right)}{z^2} \, dz > 0. \tag{3.13}$$

Then non-disclosure is guaranteed.

PROPOSITION 6. ***(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 the sequence of* $\alpha_N \to 0$ *(as* $N^y \to \infty$*) be chosen in such a way that*

$$\liminf_{N^y \to \infty} \alpha_N \left(N^x\right)^{\frac{1}{c_2+2}} > 0. \tag{3.14}$$

Then non-disclosure is guaranteed.

PROPOSITION 7. ***(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 the sequence of* $\alpha_N \to 0$ *(as* $N^y \to \infty$*) be chosen in such a way that*

$$\liminf_{N^y \to \infty} N^x e^{-\frac{c_2}{\alpha_N}} \alpha_N > 0. \tag{3.15}$$

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

The proofs of propositions 5 - 7 can be found in Appendix B.

Propositions 2 and 5 demonstrate that the compliance of the decision rule generated by a particular threshold sequence with a given bound guarantee for the disclosure risk depends on the rate at which the threshold sequence converges towards zero as the sizes of $\mathcal{D}^y$ and $\mathcal{D}^x$ increase. Informally, consider two threshold sequences α_N and α_N^* where the former converges to zero much faster than the latter so that $\frac{\alpha_N^*}{\alpha_N} \to \infty$. Clearly, for large enough sizes of the datasets $\mathcal{D}^y$ and $\mathcal{D}^x$, the sequence α_N^* not only allows more observations to be included in the combined dataset but also gives a greater number of possible combined datasets. In fact, all observations with the values of the constructed identifiers z_i^x between $\frac{1}{\alpha_N^*}$ and $\frac{1}{\alpha_N}$ are rejected by the decision rule implied by the sequence α_N but could be approved by the decision rule implied by the sequence α_N^*. In addition, the sequence α_N^* is much more liberal in its definition of the proximity between the identifiers z_j^y and z_i^x. As a result, the decision rule implied by the sequence α_N^* 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 linkages for observations with larger distances between the identifiers are decreasingly reliable, the sequence α_N^* results in a larger proportion of incorrect matches. The effect can be 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 faster than those in Proposition 5, where non-disclosure is guaranteed.

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

It can be seen in propositions 3, 4 and 6, 7 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 $\mathcal{D}^x$. This is quite intuitive because in Assumption 2 that database we assumed that $\mathcal{D}^y$ 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.

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 can be used for data combination and that guarantee the avoidance of identity disclosure with a given probability. The rate of these sequences depends on the tail 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 constructed using the proposed data combination procedure. We also show the implications of the bound on the disclosure risk for identification.

We emphasize that the structure of our identification argument is non-standard. In fact, the most common identification argument in the econometrics literature is based on finding a mapping between the population distribution of the data and parameters of interest. If the data distribution leads to a single parameter value, this parameter is called point identified. However, as we explained in the previous section, the population distribution of the immediately available data in our case is not informative, because it consists of two unrelated marginal distributions corresponding to population distributions generating split samples $\mathcal{D}^y$ and $\mathcal{D}^x$. Combination of these two samples and construction of a combined subsample is only possible when these samples are finite. In other words, knowing the probability that a given 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 point identify and estimate the model of interest.

PROPOSITION 8. *For any $\theta \in \Theta$ and any $\alpha \in (0, \bar{\alpha})$,*

$$E\left[\rho(Y, X; \theta) \mid X = x, |Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\right] = E\left[\rho(Y, X; \theta) \mid X = x\right]. \tag{4.16}$$

The proof of this proposition is in Appendix B.

The result in Proposition 8 is quite intuitive. Record linkage is based on Z^x and Z^y, which are by Assumption 3 are unrelated to Y and hence to $\rho(Y, X, \theta)$ given X. This immediately makes $E[\rho(Y, X, \theta) \,|\, X] = E[\rho(Y, X, \theta) \,|\, X, G(Z^x, Z^y)]$ for any function G, so we can in particular define G to indicate a high probability of correctly matched data. In short, we can identify the parameters in the model just using a subpopulation with relatively infrequent characteristics because are the observations that are very likely to be correctly matched, because information used for matching is by assumption conditionally independent of the model.

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

$$E\left[\rho(Y, X; \theta_0) \mid X = x, |Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\right] = 0 \tag{4.17}$$

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 8 is that

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

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

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

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

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

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

When using infrequent observations only, we can apply Proposition 8 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 pseudo-identified set based on the combined data. This notion incorporates several features. First, it takes into account the result of Proposition 8, which tells us that the information obtained from the correctly linked data is enough to point identify the model. Second, it takes into consideration the fact that it is possible to make some incorrect matches, and the extent to which the data are mismatched determines how much we can learn about the model. Third, it takes into account the fact that the data combination procedure is a finite-sample technique and identification must therefore be treated as a limiting property as the sizes of both datasets increase. We start with a discussion of the second feature and then conclude this section with a discussion of the third feature.

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

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

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

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

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

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

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

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

This value is not defined otherwise (that is, if (y_j, z_j^y) and (x_i, z_i^x) with $y_j = y$, $x_i = x$ are never combined).

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

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

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

In light of the result in (4.17), we want to consider $E^N\left[\rho(y,x;\theta)\,\middle|\,X=x\right]$ and analyze how close this conditional mean is to 0, and how close it gets to 0 as $\alpha_N \to 0$. If, for instance, $\pi^N(y,x)$ approaches 0 almost everywhere, then in the limit we expect this conditional mean to coincide with the left-hand side in (4.17), and thus, take the value of 0 if and only if $\theta = \theta_0$. Intuitively, the situation is going to be completely different if even for arbitrarily small thresholds the values of $\pi^N(y,x)$ will be separated away from 0 for a positive measure of (y,x) .

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

$$r\left(E\left[\rho(Y,X;\theta)\,\middle|\,X=x\right]\right)=0 \quad \Longrightarrow \quad \theta=\theta_0 \tag{4.18}$$

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

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

where

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

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

$$E\left[\sup_{\theta\in\Theta}\|h(X)\rho(Y,X;\theta)\|\right]<\infty, \quad E^*\left[\sup_{\theta\in\Theta}\|h(X)\rho(\widetilde{Y},X;\theta)\|\right]<\infty \tag{4.19}$$

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

Condition (4.18) is satisfied if and only if for $\pi^N(y,x)=0$ a.e.,

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

In rare situations this condition can be violated for some choices of instruments $h(\cdot)$[6], so $h(\cdot)$ has to be chosen in a way to guarantee that it holds. Here and thereafter we suppose that (4.18) is satisfied.

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

Let Π^N denote the information available to the researcher about the proportions $\pi^N(\cdot,\cdot)$. We can interpret Π^N as the set of all functions $\pi^N(\cdot,\cdot)$ that are possible under the available to the researcher

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

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

$$\Theta_N = \bigcup_{\pi^N \in \Pi^N} \underset{\theta \in \Theta}{\text{Argmin}} \; r\left(E^N\left[\rho(y_j, x_i; \theta) \,\middle|\, x_i = x\right]\right). \tag{4.20}$$

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

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

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

Proposition 9 below shows that in this situation the following set Θ_∞ is a limit of the sequence of N-identified sets Θ_N:

$$\Theta_\infty = \bigcup_{\pi \in \Pi^\infty} \underset{\theta \in \Theta}{\text{Argmin}} \; r\left((1-\pi)E\left[\rho(Y, X; \theta) \,\middle|\, X = x\right] + \pi E^*\left[\rho(\widetilde{Y},\, X; \theta) | X = x\right]\right), \tag{4.21}$$

where

$$r\left((1-\pi)E\left[\rho(Y, X; \theta) \,\middle|\, X = x\right] + \pi E^*\left[\rho(\widetilde{Y},\, X; \theta) | X = x\right]\right) = g_\pi(\theta)' W_0 g_\pi(\theta)$$

with

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

PROPOSITION 9. *Suppose that Π^∞ consists of constant values and for any $\pi \in \Pi^\infty$ there exists $\pi^N(\cdot,\cdot) \in \Pi^N$ such that*

$$\sup_{y \in Y, x \in X} |\pi^N(y,x) - \pi| \to 0 \text{ as } N^y \to \infty. \tag{4.22}$$

Also suppose that for any $\pi \in \Pi^\infty$ the function $g_\pi(\theta)' W_0 g_\pi(\theta)$ has a unique minimizer. Consider Θ_N defined as in (4.20) and Θ_∞ defined as in (4.21). 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 $g_\pi(\theta)' W_0 g_\pi(\theta)$ for each $\pi \in \Pi^\infty$ gives that $d(\Theta_\infty, \Theta_N) \to 0$ as $N^y \to \infty$.

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

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

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

DEFINITION 3. *We say that parameter θ_0 is point identified (partially pseudo-identified) from infrequent attribute values if $\Theta_\infty = \{\theta_0\}$ ($\Theta_\infty \neq \{\theta_0\}$).*

Whether the model is point identified depends on the properties of the model, the distribution of the data, and the matching procedure. Definition 3 implies that if θ_0 is point identified, then at infinity we can construct only one combined data subset using a chosen matching decision rule and that all the matches are correct ($\Pi^\infty = \{0\}$). If for a chosen $h(\cdot)$ in the definition of the distance $r(\cdot)$ parameter θ_0 is point identified in the sense of Definition 3, then θ_0 is point identified under any other choice of function $h(\cdot)$ that satisfies (4.18), and (4.19).

If the parameter of interest is only partially pseudo-identified from infrequent attribute values, then Θ_∞ is the best approximation to θ_0 in the limit in terms of the distance $r(\cdot)$ under a chosen $h(\cdot)$. In this case, Θ_∞ is sensitive to the choice of $h(\cdot)$ and W_0 and in general will be different for different $r(\cdot)$ satisfying (4.18) and (4.19). In the case of partial pseudo-identification, $0 \in \Pi^\infty$ implies that $\theta_0 \in \Theta_0$, but otherwise θ_0 does not necessarily belong to Θ_0.

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

contains y as one of the values y_j, we have that if $\inf_{i,j} \pi^N\left(y_j = y, x_i = x, \{y_j, z_j^y\}_{j=1}^{N^y}, \{x_i, z_i^x\}_{i=1}^{N^x}\right) > 0$ if $\pi^N\left(y_j = y, x_i = x, \{y_j, z_j^y\}_{j=1}^{N^y}, \{x_i, z_i^x\}_{i=1}^{N^x}\right)$ is defined.

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

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

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

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

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

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

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

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

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

Proof. Condition

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

can equivalently be written as

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

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

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

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

As we can see, Theorem 1 provides the identification result when there is no bound imposed on disclosure risk. The rates of the sequences of thresholds for which the condition of this theorem is satisfied are established in propositions 1, 3 and 4 in Section 3.

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

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

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

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

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 generally Θ_{∞} is not a singleton. □

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

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

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

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

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

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

Te proof of Corollary 1 is in Appendix B.

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

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

$$\theta_1 = E_X[XX']^{-1} E_X[X] E_Y[Y], \tag{4.23}$$

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

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

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

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

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

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

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

The pseudo-identified set for the intercept is

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

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

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

In our application, we want to illustrate a scenario when the researcher herself has access to both the sensitive and public datasets (possibly with some variables that can be used for data linkage purposefully removed) and thus she essentially takes the role of the data curator. Provided that in

such a case the researcher can control the properties of the data combination procedure, it becomes her responsibility to ensure that a required bound on the risk of disclosure is imposed. One of the main reasons for this to occur is that it is guaranteed that one of the constructed combined datasets (when a correct data combination rule is used) will contain all correct matches between the two split datasets. Given that the researcher in this case will necessarily recover the correct dataset, the corresponding set of estimated parameters that will be constructed for each of combined datasets will contain the true parameter. This set of estimates (with the bound on the disclosure risk in place) will be the identified set for the parameter of interest.

We study the 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 commercial 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 individuals 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 friendliness, 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 suppression are not disclosed by Yelp). Though we collected information on suppressed 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 healthcare 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 the OLS estimates is the 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 which is described in Section 2.3. 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, respectively, 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^y observations in the Yelp database each having exactly k counterparts in the property tax database, then the number of combined datasets we can construct is of the rate k^{N^y}.

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 treatment 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*, Springer, 239–246.

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

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

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

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

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

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

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

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

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

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*, ACM, 313–324.

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

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

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

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

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

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

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

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

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

Fellegi, I. and A. Sunter (1969): "A theory for record linkage," *Journal of the American Statistical Association*, 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, 1347–1356.

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

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

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

Horowitz, J. and C. Manski (1995): "Identification and robustness with contaminated and corrupted data," *Econometrica*, 63, 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*, 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 (2015): "Estimation of Treatment Effects from Combined Data: Identification versus Data Security," in *Economic Analysis of the Digital Economy*, ed. by A. Goldfarb, S. Greenstein, and C. Tucker, Chicago: The University of Chicago Press.

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

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

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

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

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

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

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

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

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

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

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

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," Tech. rep., Citeseer.

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

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

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

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

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 et al. [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 et al. [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 $p_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y)$ in (3.8) is equal to

$$\frac{\overline{p}_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y) Pr(m_{ij} = 1 \,|\, x_i = x, y_j = y, \mathcal{D}^x, \mathcal{D}^y)}{\overline{p}_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y) Pr(m_{ij} = 1 \,|\, x_i = x, y_j = y, \mathcal{D}^x, \mathcal{D}^y) + \underline{p}_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y) Pr(m_{ij} = 0 \,|\, x_i = x, y_j = y, \mathcal{D}^x, \mathcal{D}^y)}, \tag{B.24}$$

where

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

Note that $Pr(m_{ij} = 1 \,|\, x_i = x, y_j = y, \mathcal{D}^x, \mathcal{D}^y) = \frac{1}{N^x}$.

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

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

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

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

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

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

converges to 0.

Note that

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

From Assumption 3, for small α_N,

$$\underline{p}_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y) = \int_{|z_i^x| > \frac{1}{\alpha_N}} \left(\psi\left(\frac{1}{|z_i^x| - \alpha_N} \right) - \psi\left(\frac{1}{|z_i^x| + \alpha_N} \right) \right) (1 + o_y(1)) g_1(|z_i^x|)(1 + o_{xz^x}(1)) \, dz_i^x, \tag{B.25}$$

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

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

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

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

□

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

□

Proof of Proposition 3. Let us check that if a sequence α_N is chosen as in (3.11), then it satisfies (3.10). In other words, let us check that

$$\frac{N^x}{\alpha_N^{c_1}} \int\limits_{\frac{1}{\alpha_N}}^{\infty} \left(\left(\frac{1}{z-\alpha_N} \right)^{c_2} - \left(\frac{1}{z+\alpha_N} \right)^{c_2} \right) \frac{1}{z^{c_1+1}}\, dz \to 0 \quad \text{as } N^y \to \infty.$$

Indeed,

$$\frac{N^x}{\alpha_N^{c_1}} \int\limits_{\frac{1}{\alpha_N}}^{\infty} \left(\left(\frac{1}{z-\alpha_N} \right)^{c_2} - \left(\frac{1}{z+\alpha_N} \right)^{c_2} \right) \frac{1}{z^{c_1+1}}\, dz = \frac{N^x}{\alpha_N^{c_1}} \int\limits_{\frac{1}{\alpha_N}}^{\infty} \left(1 - \left(\frac{z-\alpha_N}{z+\alpha_N} \right)^{c_2} \right) \left(\frac{1}{z-\alpha_N} \right)^{c_2} \frac{1}{z^{c_1+1}}\, dz \tag{B.26}$$

$$= \frac{N^x}{\alpha_N^{c_1}} \int\limits_{\frac{1}{\alpha_N}}^{\infty} \left(1 - \left(1 - \frac{2\alpha_N}{z+\alpha_N} \right)^{c_2} \right) \left(\frac{1}{z-\alpha_N} \right)^{c_2} \frac{1}{z^{c_1+1}}\, dz. \tag{B.27}$$

If α_N is small enough, then for all $z \geq \frac{1}{\alpha_N}$,

$$1 - \left(1 - \frac{2\alpha_N}{z+\alpha_N} \right)^{c_2} \leq q_1 \frac{\alpha_N}{z+\alpha_N}$$

for some constant $q_1 > 0$. Therefore, if α_N is small enough, then for all $z \geq \frac{1}{\alpha_N}$ we have

$$\left(1 - \left(1 - \frac{2\alpha_N}{z+\alpha_N} \right)^{c_2} \right) \left(\frac{1}{z-\alpha_N} \right)^{c_2} \frac{1}{z^{c_1+1}} \leq q_2 \frac{\alpha_N}{z^{c_1+c_2+2}}$$

for some constant $q_2 > 0$. Finally, note that

$$\frac{q_2 N^x}{\alpha_N^{c_1-1}} \int\limits_{\frac{1}{\alpha_N}}^{\infty} \frac{1}{z^{c_1+c_2+2}}\, dz = \frac{q_2 N^x}{1+c_1+c_2} \alpha_N^{c_2+2} \to 0 \quad \text{as } N^y \to \infty$$

if α_N is chosen as in (3.11).

□

Proof of Proposition 4. Let us check that if a sequence α_N is chosen as in (3.12), then it satisfies (3.10). In other words, let us check that

$$N^x e^{\frac{c_1}{\alpha_N}} \int\limits_{\frac{1}{\alpha_N}}^{\infty} \left(e^{-c_2(z-\alpha_N)} - e^{-c_2(z+\alpha_N)} \right) e^{-c_1 z}\, dz \to 0 \quad \text{as } N^y \to \infty.$$

Indeed,

$$\begin{aligned} N^x e^{\frac{c_1}{\alpha_N}} \int\limits_{\frac{1}{\alpha_N}}^{\infty} \left(e^{-c_2(z-\alpha_N)} - e^{-c_2(z+\alpha_N)} \right) e^{-c_1 z}\, dz &= N^x e^{\frac{c_1}{\alpha_N}} \left(e^{c_2\alpha_N} - e^{-c_2\alpha_N} \right) \int\limits_{\frac{1}{\alpha_N}}^{\infty} e^{-(c_1+c_2)z}\, dz \\ &= N^x e^{-\frac{c_2}{\alpha_N}} \ \frac{e^{c_2\alpha_N} - e^{-c_2\alpha_N}}{c_1+c_2}. \end{aligned}$$

Note that for some constant $r > 0$

$$e^{c_2\alpha_N} - e^{-c_2\alpha_N} \le r\alpha_N.$$

Now it is obvious that if α_N is chosen as in (3.12), then (3.10) holds.

□

Proof of Proposition 5. From (B.24), using Assumption 3 obtain that for $\alpha_N \in (0, \bar{\alpha})$

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

and, thus,

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

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

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

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

$$\liminf_{N^y \to \infty} \frac{N^x}{\phi(\alpha_N)} \inf_{x,y} \inf_{\mathcal{D}^x, \mathcal{D}^y} \inf_{i,j} \underline{p}_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y) > 0. \tag{B.28}$$

Using (B.25), obtain that for small α_N,

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

Clearly then, $\frac{N^x}{\phi(\alpha_N)} \inf\limits_{x,y} \inf\limits_{\mathcal{D}^x, \mathcal{D}^y} \inf\limits_{i,j} \underline{p}_{ij}^N(x, y, \mathcal{D}^x, \mathcal{D}^y)$ is bounded from below by

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

Taking into account the relationship between $g_1(z)$ and $\phi\left(\frac{1}{z}\right)$, and the fact that $\sup_{|z_i^x|>\frac{1}{\alpha_N}} \sup_{x_i\in\mathcal{X}} |o_{xz^x}(1)| \to 0$ and $\sup_{y_i\in\mathcal{Y}} |o_y(1)| \to 0$ as $\alpha_N \to 0$, we obtain that the condition (3.13) guarantees then that (B.28) holds.

□

Proof of Proposition 6. Let us check that if a sequence α_N is chosen as in (3.14), then it satisfies (3.13). In other words, let us check that

$$\liminf_{N^y\to\infty} b_2c_1\frac{N^x}{\alpha_N^{c_1}}\int\limits_{\frac{1}{\alpha_N}}^{\infty}\left(\left(\frac{1}{z-\alpha_N}\right)^{c_2}-\left(\frac{1}{z+\alpha_N}\right)^{c_2}\right)\frac{1}{z^{c_1+1}}\,dz>0.$$

Use (B.26) and note that if α_N is small enough, then for all $z\geq\frac{1}{\alpha_N}$,

$$1-\left(1-\frac{2\alpha_N}{z+\alpha_N}\right)^{c_2}\geq\tilde{q}_1\frac{\alpha_N}{z+\alpha_N}$$

for some constant $\tilde{q}_1>0$. Therefore, if α_N is small enough, then for all $z\geq\frac{1}{\alpha_N}$ we have

$$\left(1-\left(1-\frac{2\alpha_N}{z+\alpha_N}\right)^{c_2}\right)\left(\frac{1}{z-\alpha_N}\right)^{c_2}\frac{1}{z^{c_1+1}}\geq\tilde{q}_2\frac{\alpha_N}{z^{c_1+c_2+2}}$$

for some constant $\tilde{q}_2>0$. Finally, note that

$$\liminf_{N^y\to\infty}\tilde{q}_2b_2c_1\frac{N^x}{\alpha_N^{c_1-1}}\int\limits_{\frac{1}{\alpha_N}}^{\infty}\frac{1}{z^{c_1+c_2+2}}\,dz=\liminf_{N^y\to\infty}\tilde{q}_2b_2c_1\frac{N^x}{1+c_1+c_2}\alpha_N^{c_2+2}>0$$

if α_N is chosen as in (3.14).

□

Proof of Proposition 7. Let us check that if a sequence α_N is chosen as in (3.15), then it satisfies (3.13). In other words, we want to check that

$$\liminf_{N^y\to\infty} c_1N^xe^{\frac{c_1}{\alpha_N}}\int\limits_{\frac{1}{\alpha_N}}^{\infty}\left(e^{-c_2(z-\alpha_N)}-e^{-c_2(z+\alpha_N)}\right)e^{-c_1z}\,dz>0.$$

Note that

$$N^xe^{\frac{c_1}{\alpha_N}}\int\limits_{\frac{1}{\alpha_N}}^{\infty}\left(e^{-c_2(z-\alpha_N)}-e^{-c_2(z+\alpha_N)}\right)e^{-c_1z}\,dz=N^xe^{-\frac{c_2}{\alpha_N}}\,\frac{e^{c_2\alpha_N}-e^{-c_2\alpha_N}}{c_1+c_2}$$

and for some constant $\tilde{r}>0$

$$e^{c_2\alpha_N}-e^{-c_2\alpha_N}\geq\tilde{r}\alpha_N.$$

Thus, if α_N is chosen as in (3.15), then (3.13) holds. □

Proof of Proposition 8. Using Assumption 3 (iii) and the law of iterated expectations,

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

By Assumption 3 (i) and (iii),

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

This implies

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

which is equivalent to (4.16). □

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

$$Q(\theta,\pi)\equiv g_\pi(\theta)'W_0g_\pi(\theta).$$

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

$$Q_N(\theta,\pi^N)\equiv g^N(\theta)'W_0g^N(\theta)$$

for the chosen $\pi^N(\cdot,\cdot)$. 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

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

Therefore,

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

Conditions (4.19) imply that $\sup_{\theta\in\Theta}\|g_\pi(\theta)\| < \infty$. Thus, we only need to establish that $\sup_{\theta\in\Theta}\|g^N(\theta) - g_\pi(\theta)\| \to 0$. Using condition (4.22), we can show that $g^N(\theta)$ can be represented as the sum of four terms –

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

where

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

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

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

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

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

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

and thus,

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

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

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

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

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:

$$\pi E^*\left[h(X)\rho(\widetilde{Y},X;\theta)\right]\cdot\left(\frac{C_{N1}}{C_{N1}+\int\int D_{N1}(y_j,x_i)f_Y(y_j)f_X(x_i)\,dy_j\,dx_i}-1\right) \tag{B.29}$$

and

$$\pi\cdot\frac{\int\int h(x_i)\rho(y_j,x_i;\theta)D_{N1}(y_j,x_i)f_Y(y_j)f_X(x_i)\,dy_j\,dx_i}{C_{N1}+\int\int D_{N1}(y_j,x_i)f_Y(y_j)f_X(x_i)\,dy_j\,dx_i}, \tag{B.30}$$

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.29) 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.29) converges to 0 uniformly over $\theta\in\Theta$.

The $\sup_{\theta\in\Theta}$ of the norm of the term in (B.30) 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

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

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

$$\sup_{\theta\in\Theta}|Q_N(\theta,\pi^N)-Q(\theta,\pi)|\to 0. \tag{B.31}$$

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

As established in Theorem 2, the identified set is

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

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

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

where

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

Note that

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

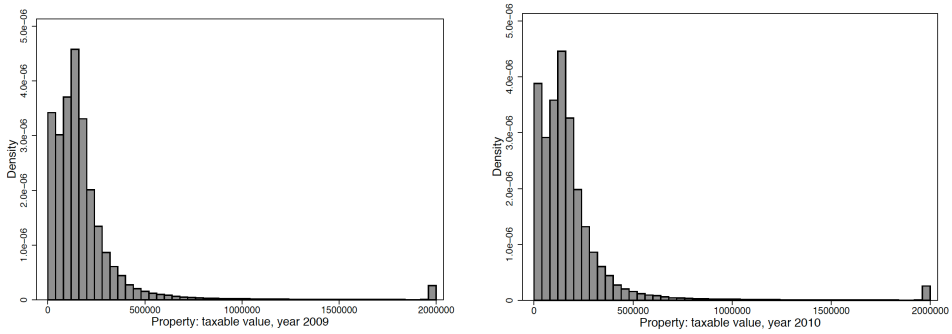
Clearly, $g_\pi(\theta)' g_\pi(\theta)$ takes the value of 0 if and only if $g_\pi(\theta)$ takes the value of 0, which happens if and only if $\theta = (1-\pi)\theta_0 + \pi\theta_1$. Thus for each $\pi \in [\underline{\gamma}, 1]$,

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

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

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

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

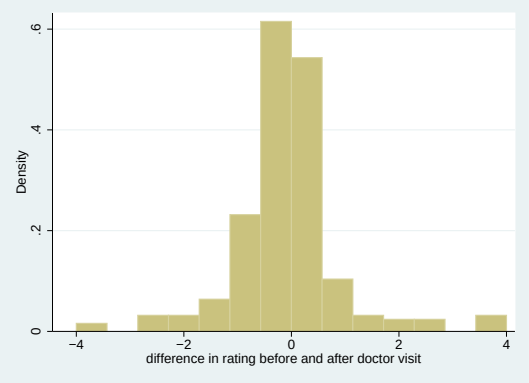


Figure 3: Average treatment effect

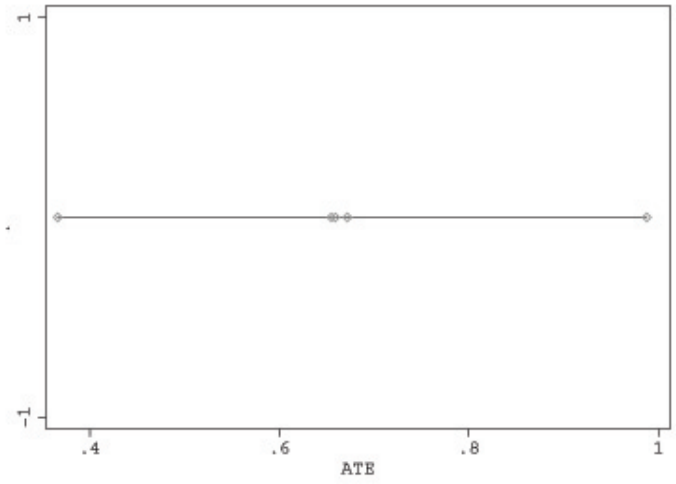


Figure 4: Identified sets for propensity score coefficients

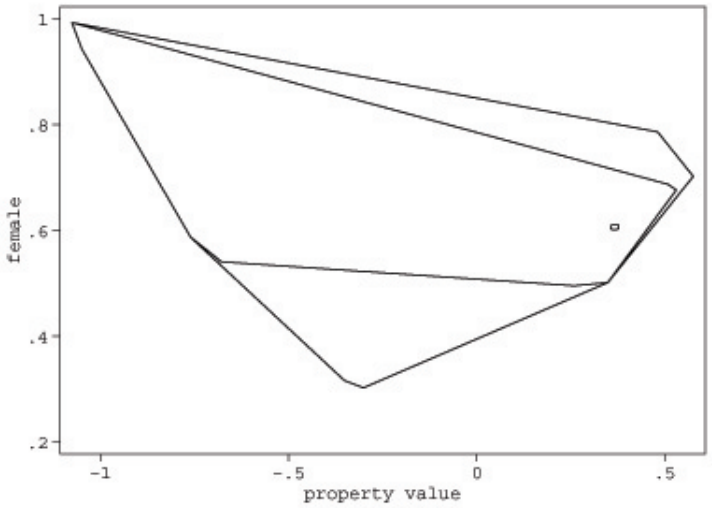


Table 1: Summary statistics from property tax bills in Durham County, NC.

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: bootstrapped SE in brackets
* significance at 10%; ** significance at 5%; *** significance 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					