

# Identification, data combination and the risk of disclosure

---

**Tatiana Komarova**  
**Denis Nekipelov**  
**Evgeny Yakovlev**

The Institute for Fiscal Studies  
Department of Economics, UCL

**cemmap** working paper CWP38/11

# Identification, Data Combination and the Risk of Disclosure<sup>1</sup>

Tatiana Komarova,<sup>2</sup> Denis Nekipelov,<sup>3</sup> Evgeny Yakovlev.<sup>4</sup>

This version: December 2011

## ABSTRACT

Businesses routinely rely on econometric models to analyze and predict consumer behavior. Estimation of such models may require combining a firm's internal data with external datasets to take into account sample selection, missing observations, omitted variables and errors in measurement within the existing data source. In this paper we point out that these data problems can be addressed when estimating econometric models from combined data using the data mining techniques under mild assumptions regarding the data distribution. However, data combination leads to serious threats to security of consumer data: we demonstrate that point identification of an econometric model from combined data is incompatible with restrictions on the risk of individual disclosure. Consequently, if a consumer model is point identified, the firm would (implicitly or explicitly) reveal the identity of at least some of consumers in its internal data. More importantly, we provide an argument that unless the firm places a restriction on the individual disclosure risk when combining data, even if the raw combined dataset is not shared with a third party, an adversary or a competitor can gather confidential information regarding some individuals from the estimated model.

**JEL Classification:** C35, C14, C25, C13.

**Keywords:** Data protection, model identification, data combination.

---

<sup>1</sup>Support from the NSF is gratefully acknowledged. We appreciate helpful comments from J. Powell, M. Jansson and P. Haile

<sup>2</sup>London School of Economics

<sup>3</sup>Corresponding author, Department of Economics, UC-Berkeley, e-mail: [nekipelov@econ.berkeley.edu](mailto:nekipelov@econ.berkeley.edu).

<sup>4</sup>Department of Economics, UC-Berkeley

## 1 Introduction

With the Internet now an established part of everyday life, issues and concerns regarding data security are now big news and generate front page headlines. Private businesses and government entities are collecting and storing increasing amounts of confidential personal data. This 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 others. Large businesses are routinely using multiple sources of information to study and predict the behavior of consumers. For instance, sponsored ads on Google are based on the consumer's information from their Gmail profile and their associated geographical or locational information.

Increasingly, businesses are declaring their commitment to the protection and security of personal data. Executives of several leading Internet companies have made multiple statements in the press stating the importance of customer privacy.<sup>1</sup> However, all those businesses rely on accurate and in-depth consumer behavior information and intelligence. In fact, advertising on the Internet is based on the estimation of the empirical models of consumer responses to ads (in terms of clicks, views, or purchases) from individual-level data. We can use Internet advertising as an example of the potential tradeoffs in using and collecting consumer-level data. To provide advertising targeted to a specific consumer, a company will be interested in the estimation of the model of the consumer's propensity to click on an ad (or purchase the advertised product or service) based on available data. First of all, available data that contains information regarding the actual page views and actual clicks or purchases will suffer from sample selection as most of the data will correspond to the customers who were already interested in a specific product and were searching for it.

As a result, the data on customers who *could* be interested in the product may not be recorded. Second, some individual-level variables may be missing. For instance, the age and gender information will only be available if the customer provides it themselves, for instance, whilst signing up for a free e-mail service. This will create a familiar omitted variable bias. Third, the data may be prone to errors of measurement. For instance, if the consumer uses a proxy server to connect to the internet, their location data will be obscured. To correct any bias in the estimates of behavioral consumer models, it is necessary to use auxiliary information.

In this paper we investigate how the potential need for data combination affects the probability of recovery of a customer's "true" identity (such as name and physical home address). In other words, we want to answer the following question: *'Is the scope for data combination compatible with the limitations imposed on the disclosure of consumer identity?'* In our empirical illustration we are interested in estimating the model of individual preferences for restaurants using the rating data from Yelp.com.

Yelp users rank restaurants based on their dining experience. However, this data will obviously be

---

<sup>1</sup>See, for instance, <http://bits.blogs.nytimes.com/2010/05/11/facebook-executive-answers-reader-questions/>

prone to selection bias: consumers who dine more frequently may be more likely to write reviews. Using the record linkage technique adopted from data mining literature, we merge the restaurant review data with the data on individual locations and property values which we use to control the sample selection. Our theoretical analysis shows that combining data with the aim of bias correction relies on observing data entries with infrequent attribute values in the two combined datasets. Accurate links for these entries can disclose consumer identities. Further, we analyse how the estimates of the consumer behavior model will be affected by the constraints on identity disclosure. As we find, any such limitation leads to a loss of point identification in the model of interest. In other words, we find that *there is a clear-cut tradeoff between the restrictions imposed on identity disclosure and the point identification of the consumer behavior model.*

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

We provide a new approach to model identification from combined datasets as a limit in the sequence of statistical experiments. Combination of data in the consumer dataset with individual information from auxiliary data may lead to the possibility of so-called linkage attacks. A linkage attack will be successful if one can provide a link between at least one data entry and auxiliary individual information with the probability exceeding the selected confidence threshold.

The optimal structure of such attacks, as well as the requirements in relation to the data release have been studied in computer science literature. The structure of linkage attacks is based on the optimal record linkage results that have been long used in the analysis of databases and data mining. To some extent, these results were used in econometrics for combination of datasets as described in Ridder and Moffitt (2007). In record linkage one provides a (possibly) probabilistic rule that can match the records from one dataset with the records from the other dataset in an effort to link the data entries corresponding to the same individual. In several striking examples, computer scientists have shown that the simple removal of personal information such as names and social security numbers does not protect the data from individual disclosure. Sweeney (2002b) identified the medical records of William Weld, then governor of Massachusetts, by linking voter registration records to “anonymised” Massachusetts Group Insurance Commission (GIC) medical encounter data, which retained the birthdate, sex, and zip code of the patient. Recent “de-personalised” data released for the Netflix prize challenge turned out to lead to a substantial privacy breach. As shown in Narayanan and Shmatikov (2008), using auxiliary information one can detect the identities of several Netflix users from the movie selection information and other data stored by Netflix.

As the identity disclosure threat is posed by the linkage attacks, we can define the restriction on disclosure risk in terms of probabilistic guarantees against the linkage attacks. We can use, what Lambert (1993) calls, a *pessimistic* measure of the risk of disclosure. It is the maximum upper bound on the probability of linking the individual information from the public data with the record in the

released anonymised data sample. The technology to control the risk of identity disclosure exists, so the bounds on disclosure risk are practically enforceable.

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

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

Although our identification approach is new, to understand the impact of the bounds on the individual disclosure risk we use ideas from literature on partial identification of models with contaminated or corrupted data. Manski (2003), Horowitz, Manski, Ponomareva, and Stoye (2003), Horowitz and Manski (2006), Magnac and Maurin (2008) have understood that many data modifications such as top-coding suppression of attributes and stratification lead to the loss of point identification of parameters of interest. Consideration of the general setup in Molinari (2008) allows one to assess the impact of some data “anonymisation” as a general misclassification problem. In this paper we find the approach of constructing identified sets for parameters of interest extremely informative. As we show in this paper, the size of the identified set for the parameter in the linear model is directly proportional to the pessimistic measure of disclosure risk. This is a powerful result that essentially states that there is a direct conflict between the informativeness of the data used in the consumer behavioral model and the security of individual data. As a result, combination of the company’s internal data with the auxiliary public individual data is not compatible with the non-disclosure

of individual identities. An increase in the complexity and nonlinearity of the model can further worsen the tradeoff.

In this paper we associate the ability of the company to recover the true identity of consumers from internal data with the risk of individual disclosure. This does not mean that we expect the company that constructs the consumer behavior model to misuse the data or intentionally compromise the identities of consumers. However, in some cases the consumer behavior model *may itself be disclosive*. For instance, Korolova (2010) shows examples of privacy breaches through micro ad targeting on Facebook.com. Facebook does not give advertisers direct access to user data. Instead, the advertiser interface allows them to create targeted advertising campaigns with a very granular set of targets. In other words, one can create a set of targets that will isolate a very small group of Facebook users (based on location, friends and likes). Korolova shows that certain users can be perfectly isolated from other users with a particularly detailed list of targets. Then, one can recover the “hidden” consumer attributes, such as age or sexual orientation, by constructing differential advertising campaigns such that a different version of the ad will be shown to the user depending on the value of the private attribute. Then the advertiser’s tools allow the advertiser to observe which version of the ad was shown to the Facebook user. When an online advertising company uses a consumer behavior model to show the ads to consumers who are likely to respond to them, this may lead to identity disclosure. Returning to the Facebook advertising example, if one targets the ad to isolate a very small group of consumers or a single consumer and the consumer behavior model suggests that the ad will be more effective for high-income individuals, then the fact that the ad was shown will indicate that the targeted consumer is likely a high-income individual.

Security of individual data is not synonymous to privacy, as privacy may have subjective value for consumers (see Acquisti (2004)). Privacy is a complicated concept that frequently cannot be expressed as a formal guarantee against intruders’ attacks. Considering personal information as a “good” valued by consumers leads to important insights in the economics of privacy. As seen in Varian (2009), this approach allowed the researchers to analyse the release of private data in the context the tradeoff between the network effects created by the data release and 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, subjective individual perception of privacy is important. This is clearly shown in both the lab experiments in Gross and Acquisti (2005), Acquisti and Grossklags (2008), as well as in the real-world environment in Acquisti, Friedman, and Telang (2006), Miller and Tucker (2009) and Goldfarb and Tucker (2010). Given all these findings, we believe that the disclosure protection plays a central role in the privacy discourse, as privacy protection is impossible without the data protection.

The rest of the paper is organised as follows. In Section 2 we describe the econometric problem and define the parameter of interest. In Section 3 we give sufficient condition for identification of

consumer behavior model from combined data. In Section 4 we give the definition of disclosure risk and study identification of the econometric model under restrictions on the disclosure risk. In Section 5 we provide the final remarks and conclude.

## 2 Econometric model

Suppose that consumer behavior model is based on the joint distribution of the vector-valued consumer response  $Y \in \mathcal{Y} \subset \mathbb{R}^m$  and consumer characteristics  $X \in \mathcal{X} \subset \mathbb{R}^k$ . The economic parameter of interest  $\theta_0$  (contained in the convex compact set  $\Theta$ ) defines the consumer response model

$$E[\rho(Y, X, \theta_0) | X = x] = 0. \quad (2.1)$$

We will focus on a linear separable model for  $\rho(\cdot)$  as our lead example, which can be directly extended to monotone nonlinear models. In a typical Internet environment consumer choices may include purchases in an online store, specific messages on a discussion board, comments on a rating website or a profile on a social networking or dating website. Consumer characteristics are the relevant socio-demographic characteristics such as location, demographic characteristics, and social links with other individuals. We assume that for the true joint distribution of  $Y$  and  $X$  there is only one  $\theta_0$  satisfying condition (2.1). Formally we write this as the following assumption.

**ASSUMPTION 1.** *Parameter  $\theta_0$  is uniquely determined from the moment equation (2.1) and the population conditional distribution  $Y | X$ .*

For empirical illustration we estimate the model of consumer’s rating of a restaurant (expressed as a rank score) as a function of the characteristics of the restaurant and demographic characteristics of consumers. We use the publicly available online data to estimate the model. We decided to focus on a specific geographic region, Durham, NC, to estimate the model. In order to obtain the restaurant information and the consumers opinions, we use the restaurant description and the user review from yelp.com, collecting information regarding all the restaurants located in Durham, NC. The data contains the user evaluation of the restaurant with the verbal description of the personal user experience as well as the restaurant details such as a price level category, cuisine type, hours of work and location. The information about users contains the self-reported user location, self-reported first name as well as all the reviews by each user. To obtain reliable personal information we collected the property tax data available for local taxpayers in Durham county. This data reflects the property tax paid for residential real estate along with some characteristics of the property owner (such as name), location and the appraisal value of the property. If we had the data from yelp.com merged individual-by-individual with the data on the property tax, then for each consumer review we would know both the score that the consumer assigned to the restaurant, as well as all restaurant and consumer characteristics. In reality, however, there is no unique identifier that labels the observations in both data sources.

This means that the variables of interest  $Y$  and  $X$  will not be observed simultaneously. One can separately observe the dataset containing values of  $Y$  and the dataset containing the values of  $X$

for the 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 a joint distribution of vector-valued random variables  $(Y, W, X, V)$  with values contained in  $\mathcal{Y} \times \mathcal{W} \times \mathcal{X} \times \mathcal{V} \subset \mathbb{R}^m \times \mathbb{R}^q \times \mathbb{R}^k \times \mathbb{R}^r$ .*

(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  $\mathcal{D}^y = \{y_j, w_j\}_{j=1}^{N^y}$  and the second subsample is  $\mathcal{D}^x = \{x_j, v_j\}_{j=1}^{N^x}$ .*

Assumption 2 characterizes the observable variables as independently drawn subsamples of the infeasible “master” data. This means that without any additional information, one can only reconstruct marginal distributions  $f_X(\cdot)$  and  $f_Y(\cdot)$ .

In case of a linear model identification with split sample data reduces to computing the familiar Fréchet bounds. Suppose that  $Y$  and  $X$  are continuous scalar random variables and the object of interest is the slope of linear regression of  $Y$  on  $X$ , which can be computed as

$$\beta = \frac{\text{cov}(Y, X)}{\text{var}(X)}.$$

Given that the information regarding the joint distribution of  $Y$  and  $X$  is not known, the covariance between  $Y$  cannot be directly estimated from the marginal distributions. As a result, only trivial information is available for the joint moments of the regressor and the outcome, which we may summarize as  $|\text{corr}(Y, X)| \leq 1$ . Therefore, we can find the identified set for the slope coefficient as

$$-\sqrt{\frac{\text{var}(Y)}{\text{var}(X)}} \leq \beta \leq \sqrt{\frac{\text{var}(Y)}{\text{var}(X)}}.$$

We can note that the constructed bounds are extremely wide especially when the regressor has small support. Moreover, we cannot even identify the direction of the relationship between the regressor and the outcome, which is extremely important in most economic applications.

Fréchet bounds for the estimated parameters are constructed on the premise that no additional information is available regarding the joint distribution of  $X$  and  $Y$ . Returning to our empirical example, we can note that consumer choice information and the consumer demographics are not completely unrelated. For instance, we may expect that consumers tend to go more frequently to the restaurants that are located closer to where they live. Also, it is likely that the self-reported name in the user review on yelp.com is highly correlated with the real name of the user. In general, we can formalize detection of the related variables in disjoint data as construction of vector-valued functions of the data which we expect to take proximate values if observations in two datasets correspond to the same individual and expect those values to be further for different individuals. Construction of

such classifiers is widely discussed in the modern computer science literature especially in relation to record linkage and data recovery. In this paper we take the procedure for construction of such classifiers as given and illustrate practical implementation of such a procedure in our empirical example. In the following assumption we express the requirements on the data classifiers.

**ASSUMPTION 3.** *We assume that there exist functions  $Z^x = Z^x(X, V)$  and  $Z^y = Z^y(Y, W)$  with the values in  $\mathbb{R}$  that are evaluated at the variables contained in the datasets  $\mathcal{D}^x$  and  $\mathcal{D}^y$ . For these functions there exists a distance  $d(\cdot, \cdot)$  and  $\bar{\alpha} > 0$  such that for any  $0 < \alpha < \bar{\alpha}$ :*

- (i)  *$\Pr(d(Z^y, Z^x) < \alpha \mid X = x, |Z^x| > \frac{1}{\alpha}) \geq 1 - \alpha$  for almost all  $x \in \mathcal{X}$ .*
- (ii) *For almost all  $x \in \mathcal{X}$  and almost all  $y \in \mathcal{Y}$ ,  $\Pr(|Z^x| > \frac{1}{\alpha} \mid X = x) = \phi(\alpha) + o(\phi(\alpha))$  and  $\Pr(|Z^y| > \frac{1}{\alpha} \mid Y = y) = \psi(\alpha) + o(\psi(\alpha))$  for some non-decreasing and positive at  $\alpha > 0$  functions  $\phi(\cdot)$  and  $\psi(\cdot)$ .*
- (iii) *For almost all  $x \in \mathcal{X}$ ,*

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

Functions  $Z^x$  and  $Z^y$  are adding more information regarding the joint distribution of  $Y$  and  $X$  allowing us to go beyond the Fréchet bounds for parameter  $\theta_0$ . Ridder and Moffitt (2007) overview the cases where numeric identifiers  $Z^x$  and  $Z^y$  are *a priori* available and their joint distribution is normal. Here, we argue that such numeric identifiers are typically unavailable and the data entries that may potentially be useful are typically strings of characters. However, we can still identify the consumer model of interest in the following way. First, we extract the parts of entries in the merged databases that can be used for matching. Second, we select an appropriate distance measure between the entries. And, third, we estimate the model for the trimmed subset of matches where the distance between the entries is below a selected threshold. In Appendix A we provide a brief overview of distance measures for string data, that are commonly used in data mining.

Assumption 3 (iii) states that for a pair of matched observations from 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 the regressor  $X$ . In other words, if the data is already matched, the constructed identifiers only label observations and do not improve any knowledge about the estimated economic model. As in the example of matching the observations by the names, once we extract all model-relevant information from the name (for instance, whether a specific individual is likely to be a male or a female, white, black or hispanic) and we already matched 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.

We recognize that Assumption 3 puts restrictions on the behavior of infrequent realizations of identifiers  $Z^x$  and  $Z^y$ . Specifically, we expect that conditional on the identifier taking a high value, the values of identifiers constructed from two datasets have to be close. We can illustrate this

assumption by our empirical application, where we construct a categorical variable from the first names of individuals that 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.

In general, the construction of identifiers allowing us to merge two datasets combines the information from the string entries of the data and numeric entries. In Appendix A we provide examples of possible distance measures that can be used for the string data. Then we can construct the distance between the data entries by combining the distance measure used for strings with the Euclidean distance which we can use for the numeric data. For instance, in the empirical application we use such variables in individual entries as the number of restaurant reviews given by a particular user in a specific zip code in the Yelp.com data, and the zip code of an individual in the property tax data. We also use indicators of the name of an individual in the property tax data belonging to the list of most common hispanic, black and asian names in 2009 US Census data and the cuisine of the restaurant reviewed, under the assumption that individuals of each specific ethnicity would prefer their ethnic cuisine. There are multiple examples of computationally optimal construction of individual identifiers which are based on clustering the data with some priors on the relationship between the variables in two datasets. For instance, Narayanan and Shmatikov (2008) uses the collection of individual movie choices in the Netflix dataset and on imdb.com.

Similar identifiers are constructed on a daily basis by online advertising companies trying to predict the probability of a consumer action (for instance, a click on the ad) based on the available characteristics of a consumer query. An advertising company usually considers a model of the probability of a consumer action as a function of consumer characteristics. Given that all consumer characteristics are not available for all the queries, they need to be inferred from the information contained in the query and the information that has already been collected by the advertising company.

### 3 Identification with Combined Data

In this section, we formalise the discussion in section 2 and introduce notions of point identification and partial identification of the econometric model from combined data. We suppose henceforth that Assumptions 1-3 hold.

In our model, variables  $Y$  and  $X$  are contained in separate datasets. Because from the dataset containing  $Y$  we can construct the identifier  $Z^y$  and from the dataset containing  $X$  we can construct the identifier  $Z^x$ , in the limit the available data will be represented by the joint distribution of  $(Y, Z^y)$  and the joint distribution of  $(X, Z^x)$ . These two joint distributions by themselves, however, will not be completely informative of the joint distribution of  $Y$  and  $X$ . The identification of the econometric model (which in our case reduces to identification of  $\theta_0$  from (2.1)) will only be possible

if two datasets are merged at least for some observations. On the other hand, data combination is an intrinsically finite sample procedure. This leads us to the idea of discussing identification in terms of the limit of statistical experiments. To our knowledge, this is a new approach to parameter identification from combined data.

The identification idea lies in considering each finite sample size and constructing a subsample, or a combined dataset, of what we believe to be matching entries in the two datasets. The parameter value of interest then can be constructed using the sample distribution. Considering databases of increasing sizes, we can build a sequence of estimated parameter values corresponding to the sequence of empirical distributions of observations in the combined dataset. We provide conditions under which this sequence of parameter values converges to the true value of the parameter of interest, leading to the (point) identification of  $\theta_0$ .

We start by creating samples  $\{y_j, z_j^y\}_{j=1}^{N^y}$  and  $\{x_i, z_i^x\}_{i=1}^{N^x}$  from datasets  $\{y_j, w_j\}_{j=1}^{N^y}$  and  $\{x_i, v_i\}_{i=1}^{N^x}$ . The joint distribution of the data in two combined datasets can be characterized as the distribution of random vectors  $(X_1, \dots, X_{N^x})'$ ,  $(Z_1^x, \dots, Z_{N^x}^x)'$  and  $N^y$ -dimensional random vectors  $(Y_1, \dots, Y_{N^y})'$ ,  $(Z_1^y, \dots, Z_{N^y}^y)'$ . Provided that the indices of matching entries are not known in advance, the same index entries do not necessarily belong to the same individual.

The largest combined dataset will contain  $N = \min\{N^x, N^y\}$  entries. We now characterize the joint distribution of an arbitrary pair of entries from two datasets as

$$f^N(y_j, x_i, z_j^y, z_i^x),$$

for each pair of elements  $i$  and  $j$ . Note that this density is equal to the product of marginal densities if  $i$  and  $j$  correspond to different individuals and it is equal to the joint density if it corresponds to the same individual.

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

We note that if two data entries do not belong to the same individual, provided that the data are independent across entries (recall the i.i.d. assumption) the distribution of interest can be expressed as

$$\begin{aligned} f^N(y_j, x_i, z_j^y, z_i^x) &= f_{Y, X, Z^y, Z^x}(y_j, x_i, z_j^y, z_i^x) Pr(m_{ij} = 1 \mid \mathcal{D}_N(y_j, x_i, z_j^y, z_i^x)) \\ &\quad + f_{Y, Z^y}(y_j, z_j^y) f_{X, Z^x}(x_i, z_i^x) Pr(m_{ij} = 0 \mid \mathcal{D}_N(y_j, x_i, z_j^y, z_i^x)), \end{aligned}$$

where  $\mathcal{D}_N(\cdot)$  is the decision rule used by the researcher to match observations to combine the data. If  $\mathcal{D}_N(y_j, x_i, z_j^y, z_i^x) = 1$ , the researcher assigns entries  $i$  and  $j$  to the same individual the same individual, otherwise, she considers them to belong to different individuals.

In this paper we focus on the deterministic combination rule

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

for a chosen  $\alpha_N$  such that  $0 < \alpha_N < \bar{\alpha}$ . This decision uses identifiers  $Z^y$  and  $Z^x$  to gain some knowledge about the joint distribution of  $(Y, X)$  in the following way. If, for a chosen  $\alpha_N$ , we find a data entry  $i$  with  $|z_i^x| > \frac{1}{\alpha_N}$  from one dataset and find a data entry  $j$  with the property  $d(z_j^y, z_i^x) < \alpha_N$  from the other dataset, then we believe  $i$  and  $j$  to be a match. In other words, if identifiers  $z_i^x$  and  $z_j^y$  are both large and are close in distance  $d(\cdot, \cdot)$ , then we consider  $(x_i, z_i^x)$  and  $(y_j, z_j^y)$  to be observations corresponding to the same individual. This seems to be a good strategy when  $\alpha_N$  is small because according to Assumption 3, in that case the conditional probability of  $Z^x$  and  $Z^y$  being close to each other when  $Z^x$  is large in the absolute value is close to 1. This probability however may still be strictly smaller than 1, which makes our matching rule imperfect. We can create incorrect matches, in which case

$$\mathcal{D}_N(y_j, x_i, z_j^y, z_i^x) \neq m_{ij},$$

and the probability of making incorrect matches is strictly positive for each individual in the given samples.

For the sake of notational simplicity, we use the absolute value as the distance measure  $d(\cdot)$ . This is appropriate when the data can be categorized. For instance, if the data contain names of individuals, we can assign numeric indices to the names according to their frequency rank in the Census. Our results will be valid for other definitions of the distances between the identifiers when such a numeric indexation will not be plausible.

Intuitively, if there is a “sufficient” number of data entries which we identify as matched observations, we have “enough” knowledge about the joint distribution of  $(Y, X)$  to estimate the model of interest.

The proposition below gives us an auxiliary result on the conditional moments of  $\rho(Y, X, \theta)$  for infrequent observations.

**Proposition 1.** *For any  $\theta \in \Theta$  and any  $\alpha \in (0, \bar{\alpha})$ ,*

$$E \left[ \rho(Y, X, \theta) \mid X = x, |Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha} \right] = E [\rho(Y, X, \theta) \mid X = x]. \quad (3.2)$$

The proof of this proposition is in the Appendix.

Equation (3.2) is an important part of our argument because it allows us to use the subpopulation with relatively infrequent characteristics to identify the parameter of the moment equation that is valid for the entire population. So 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 us to have 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.

A clear characterization of identification using infrequent data attributes can be given in the bivariate linear model. Let  $Y$  and  $X$  be two scalar random variables and let parameter  $\theta_0$  be  $\theta_0 = (a_0, b_0)$ .

The consumer model of interest is characterized by

$$E[Y - a_0 - b_0X | X = x] = 0.$$

This restriction implies the following two equations:

$$0 = E[Y - a_0 - b_0X] = E[X(Y - a_0 - b_0X)],$$

which have the unique solution if  $\text{Var}(X) > 0$ :

$$b_0 = \frac{\text{cov}(Y, X)}{\text{Var}(X)}, \quad a_0 = E[Y] - b_0E[X].$$

In order to characterize the identified parameters via conditioning only on the matching rule, we consider the equations for the coefficients as solutions to the following system of equations:

$$\begin{aligned} 0 &= E[(Y - a_0(\alpha) - b_0(\alpha)X)\mathbf{1}\{|Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\}] \\ 0 &= E[X(Y - a_0(\alpha) - b_0(\alpha)X)\mathbf{1}\{|Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\}]. \end{aligned}$$

Defining  $X^* = \frac{X\mathbf{1}\{|Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\}}{E[\mathbf{1}\{|Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\}]^{1/2}}$  and  $Y^* = \frac{Y\mathbf{1}\{|Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\}}{E[\mathbf{1}\{|Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\}]^{1/2}}$  we can represent the coefficients of interest via weighted least squares:

$$b_0(\alpha) = \frac{\text{cov}(X^*, Y^*)}{\text{Var}(X^*)}, \quad a_0(\alpha) = \frac{E[Y^*] - b_0(\alpha)E[X^*]}{E[\mathbf{1}\{|Z^x - Z^y| < \alpha, |Z^x| > \frac{1}{\alpha}\}]^{1/2}}.$$

Proposition 1 implies that  $b_0(\alpha) = b_0$  and  $a_0(\alpha) = a_0$  for any  $\alpha \in (0, \bar{\alpha})$ .

Thus, if the joint distribution of  $(Y, X)$  for infrequent observations with  $\{|Z^x| > \frac{1}{\alpha}, |Z^x - Z^y| < \alpha\}$  is known, then  $\theta_0$  can be estimated from moment equation

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

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. The observations with more common values of identifiers have a higher probability of creating false matches and, thus, are less valuable for the purpose of data combination.

Density  $f^N(\cdot)$  characterizes the joint distribution of matched pairs. We interpret the expectation  $E^N[\cdot]$  as the integral over density  $f^N(\cdot)$ . The distribution  $f^N(\cdot)$  does not coincide with the distribution of a sample we would have drawn from the joint distribution of  $(Y, X, Z^y, Z^x)$  because of a positive probability of making a matching error. As a result, even though (3.2) holds, expectation  $E^N[\cdot]$  does not coincide with  $E[\cdot]$ . We now evaluate the quality of approximation of the true expectation with  $E^N[\cdot]$  approaching the cutoff points  $\alpha_N$  to zero as the size of both matched datasets increases.

We denote by  $\mathcal{A}^N$  a collection of functions  $\Theta \times \mathcal{X} \mapsto \mathbb{R}^p$  (where  $p$  is the dimension of the moment function  $\rho(\cdot)$ ) which are pointwise partial limits of  $E^N[\rho(y_j, x_i, \theta) \mid x_i = x]$  for a given decision rule.

We introduce the distance  $r(\cdot)$  that measures the proximity of the conditional moment vector to the origin. An example of a distance we can consider is

$$r(a(\theta, x)) = g(a, \theta)' W g(a, \theta),$$

where

$$g(a, \theta) = E[h(X)a(\theta, X)],$$

with a (nonlinear)  $J \times p$ ,  $J \geq k$ , instrument  $h(X)$  and a  $J \times J$  positive definite matrix  $W$ .

We provide the following definition of the set of parameters of interest.

**DEFINITION 1.** *Set*

$$\Theta_N = \bigcup_{a(\cdot, \cdot) \in \mathcal{A}^N} \underset{\theta \in \Theta}{\text{Arg inf}} r(a(\theta, x))$$

*is called the set of parameter values identifiable from infrequent attribute values.*

The next definition gives a notion of point identification of  $\theta_0$ .

**DEFINITION 2.** *Let  $\Theta_\infty = \bigcap_{N=1}^\infty \Theta_N$ . We say that parameter  $\theta_0$  is point identified from infrequent attribute values if  $\Theta_\infty = \{\theta_0\}$ .*

Our notion of identification relies on our choice of distance  $r(\cdot)$  and on the selected decision rule for data combination, including the behavior of this rule as the dataset size increases. If the parameter of interest cannot be identified in the combined data, then  $\Theta_\infty$  is the best approximation to this parameter.

Next we can define partial identification of the parameter of interest.

**DEFINITION 3.** *We say that  $\theta_0$  is partially identified from infrequent attribute values if*

$$\theta_0 \in \Theta_\infty \text{ but } \Theta_\infty \neq \{\theta_0\}.$$

## 4 Data Combination and the Risk of Disclosure

Definition 2 characterizes our idea of point identification based on the limit of conditional means that are re-constructed from finite samples of merged observations. In this section, we explicitly show that data combination will be associated with the risk of disclosure. Using a specific definition of disclosure risk, we will show how identification of the model from infrequent attribute values will be affected with the required limits imposed on the risk of disclosure.

The limits on the risk of disclosure for each individual is the main reason why we define identification and non-disclosure guarantee through the limits of sequences of conditional means and probabilities

of correct matches and do not attempt to do this in terms of the population distributions of  $(Y, Z^y)$  and  $(X, Z^x)$ . Notions of identification and non-disclosure guarantee in terms of the population distributions would ignore what may happen to individuals from a subset of measure zero. In other words, it would require conditions for non-disclosure guarantee to hold for *almost every* individual in the population rather than for *every* individual. But the ability to guarantee non-disclosure for *every* individual is essential, and therefore such a population approach does not serve our purpose.

In light of Assumption 3, the conditional probability

$$\pi_{ij}^N(x) = Pr \left( m_{ij} = 1 \mid x_i = x, |z_i^x| > \frac{1}{\alpha_N}, |z_j^y - z_i^x| < \alpha_N \right) \quad (4.3)$$

of a successful match of  $z_i^x$  and  $z_j^y$  under our matching rule can be very high for sufficiently small  $\alpha_N$ . This means that the pair of entries in two databases will correspond to the same individual with a very high level of confidence, which means that the linkage attack on each database will be quite successful. To measure the risk of disclosure in the possible linkage attacks we use the definition of the pessimistic disclosure risk in Lambert (1993). In terms of our data model, we can formalize the pessimistic disclosure risk as the maximum probability of a linkage attack over all individuals in the database.

**DEFINITION 4.** *A bound guarantee is given for the risk of disclosure if there exists  $0 < \underline{\gamma} \leq 1$  such that*

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

for all  $N$  and

$$\sup_{x \in \mathcal{X}} \lim_{N \rightarrow \infty} \sup_{j,i} \pi_{ij}^N(x) = 1 - \underline{\gamma}.$$

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

It is important to note that the risk of disclosure needs to be controlled in any size dataset with any realization of the values of covariates. In other words one needs to provide an *ad omnia* guarantee that the probability of a successful match will not exceed the provided bound. This requirement is very different from the guarantee with probability one, as here we need to ensure that even for the datasets that may be observed with an extremely low probability, the match probability honors the imposed bound. For example, if the limit of  $\pi_{ij}^N(x)$  is equal to  $1 - \bar{\gamma}$ , this means that for any dataset one incorrect matches occur with probability at least  $\bar{\gamma}$ , and thus, the value of  $\bar{\gamma}$  is the extent of non-disclosure risk guarantee. This means that in any dataset of size  $N$  there will be at least  $O(N\bar{\gamma})$  matches per observation.<sup>2</sup>

The bound on the individual disclosure does not mean that making a correct match is impossible. Instead, in this case due to the “imperfect” matching rule along with correct matches the researcher

---

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

can find equally good incorrect matches. This means that there will be multiple versions of the combined dataset. One of these versions will correspond to the “true” combined dataset. However, along with it one can construct the datasets where there is a fraction of incorrect matches. The probability  $\bar{\gamma}$  indicates the highest proportion of the incorrect matches in the constructed version of the combined dataset. Then all possible versions will have the proportion of incorrect matches varying from 0 (in the “true” version) to  $\bar{\gamma}$  (in the most contaminated version). Next we consider how this idea of the dataset combination when we impose a bound on the disclosure risk translates into the properties of the estimator of interest.

The next proposition describes the limiting behavior of the moment function  $E^N[\rho(y_j, x_i; \theta)|x_i = x]$ .

**Proposition 2.** *Let  $\alpha_N \rightarrow 0$  as  $N \rightarrow \infty$ . Suppose that  $\Pr\left(m_{ij} = 0 \mid |z_i^x| > \frac{1}{\alpha_N}, |z_i^x - z_j^y| < \alpha_N\right) \rightarrow \gamma$  as  $N \rightarrow \infty$ . Then, for almost all  $x$ ,*

$$E^N[\rho(y_j, x_i; \theta)|x_i = x] \rightarrow (1 - \gamma)E[\rho(Y, X; \theta)|X = x] + \gamma E^*[\rho(\tilde{Y}, X; \theta)|X = x] \quad (4.4)$$

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

The proof of this proposition is in the Appendix.

Clearly, under the conditions of Proposition 2, the set of parameter values identifiable from infrequent attribute values is

$$\Theta_\infty(\gamma) = \text{Arg inf}_{\theta \in \Theta} r \left( (1 - \gamma)E[\rho(Y, X; \theta)|X = x] + \gamma E^*[\rho(\tilde{Y}, X; \theta)|X = x] \right).$$

In particular, if there is a value of  $\theta$  that gives for almost all  $x \in \mathcal{X}$  the value 0 to the limiting function in (4.4), then

$$\Theta_\infty(\gamma) = \left\{ \theta \in \Theta : (1 - \gamma)E[\rho(Y, X; \theta)|X = x] + \gamma E^*[\rho(\tilde{Y}, X; \theta)|X = x] = 0 \text{ for almost all } x \in \mathcal{X} \right\}.$$

The following result on point identification is an implication of Proposition 2. It establishes that  $\theta_0$  is point identified from observations with infrequent values of the attributes if non-disclosure is not guaranteed, that is, if in the limit all our matches are correct.

**THEOREM 1. (Point identification of  $\theta_0$ ).** *Let  $\alpha_N \rightarrow 0$  as  $N \rightarrow \infty$ . Suppose there is no non-disclosure guarantee. Then  $\theta_0$  is point identified from matches on infrequent values of the attributes.*

*Proof.* The absence of non-disclosure guarantee means that the only possible asymptotic behavior of the conditional probabilities  $\Pr\left(m_{ij} = 0 \mid |z_i^x| > \frac{1}{\alpha_N}, |z_i^x - z_j^y| < \alpha_N\right)$  is their convergence to 0. For  $\gamma = 0$ , the limiting function in (4.4) is  $E[\rho(Y, X; \theta)|X = x]$ , which according to Assumption 1 takes value 0 for almost all  $x \in \mathcal{X}$  only at the parameter value  $\theta_0$ . Hence,  $\Theta_\infty = \{\theta_0\}$ .  $\square$

The theorem below gives a partial identification result.

**THEOREM 2. (Partial identification of  $\theta_0$ ).** Let  $\alpha_N \rightarrow 0$  as  $N \rightarrow \infty$ . Suppose there is a bound  $\underline{\gamma} > 0$  imposed on the disclosure risk. Then in general  $\theta_0$  is identified from matches on infrequent values of the attributes only partially, and the identified set is

$$\Theta_\infty = \bigcup_{\gamma \in [0, \underline{\gamma}]} \Theta(\gamma),$$

where  $\Theta(\gamma)$  is as defined after Proposition 2.

Thus, the identified set is the collection of parameter values obtained in the limit under all possible extents of non-disclosure guarantee  $\gamma$  up to the disclosure risk bound  $\underline{\gamma}$ . Elements in  $\Theta(\gamma)$ ,  $\gamma \in (0, \underline{\gamma}]$ , are in general different from  $\theta_0$ , which means that  $\Theta_\infty$  is non-singleton.

*Proof.* First of all, note that  $\Theta(\gamma) = \{\theta_0\}$ , and therefore,  $\theta_0 \in \Theta_\infty$ . Second, having the bound  $\underline{\gamma}$  on the disclosure risk means that it is possible to have convergence  $Pr\left(m_{ij} = 0 \mid |z_i^x| > \frac{1}{\alpha_N}, |z_i^x - z_j^y| < \alpha_N\right) \rightarrow \gamma$  as  $N \rightarrow \infty$  for any  $\gamma \in [0, \underline{\gamma}]$ . These facts together with Proposition 2 and definitions 1 and 3 yield the result of the theorem.  $\square$

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

**THEOREM 3.** Consider a linear model with  $\theta_0$  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  $\theta_0$  is only partially identified from infrequent attribute values, and, under the distance  $r(\cdot)$  chosen in the spirit of least squares, the identified set is the following collection of convex combinations of parameters  $\theta_0$  and  $\theta_1$ :

$$\Theta_\infty = \{\theta_\gamma, \gamma \in [0, \underline{\gamma}] : \theta_\gamma = (1 - \gamma)\theta_0 + \gamma\theta_1\},$$

where  $\theta_1$  is the parameter value one would obtain using only incorrect matches. In terms of Proposition 2,  $\theta_1$  would be obtained if in the limit matches were incorrect with probability 1.

Note that  $\theta_0 = E[XX']^{-1}E[XY]$ .  $E[XX']$  can be found from the marginal distribution of  $X$  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. The key insight in Theorem 3 is that if the match is incorrect, then we are combining the values of  $X$  and  $Y$  that belong to different individuals and, therefore, these values are independent. When we consider independent  $X$  and  $Y$  with distributions  $f_X(\cdot)$  and  $f_Y(\cdot)$ , we have  $E^*[X(Y - X'\theta)] = 0$ . Solving the last equation we obtain

$$\theta_1 = E_X[XX']^{-1}E_X[X]E_Y[Y], \tag{4.5}$$

which can be found from split samples without using any matching methodology. When the combined data contains correct and incorrect matches, the resulting estimator will be a mixture of estimators are obtain for correct and incorrect matches ( $\theta_0$  and  $\theta_1$  correspondingly).

As a special case, consider a bivariate linear regression model

$$E[Y - a_0 - b_0X|X = x] = 0.$$

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

$$\{b_\gamma : b_\gamma = (1 - \gamma)b_0, \gamma \in [0, \underline{\gamma}]\}$$

because  $b_1 = 0$ , and for the intercept it is

$$\{a_\gamma : a_\gamma = (1 - \gamma)a_0 + \gamma E_Y[Y], \gamma \in [0, \underline{\gamma}]\} = \{a_\gamma : a_\gamma = E_Y[Y] - (1 - \gamma)b_0 E_X[X], \gamma \in [0, \underline{\gamma}]\}.$$

The complete proof of Theorem 3 can be found in the Appendix.

Next, we analyze what should be the restrictions on the marginal distributions of identifiers to allow for imposing bounds on the disclosure risk.

**Proposition 3. (Absence of non-disclosure risk guarantee).** *Let  $\alpha_N > 0$ ,  $\alpha_N \rightarrow 0$  be chosen in such a way that*

$$\lim_{N \rightarrow \infty} \frac{\max\{N^x, N^y\}}{\phi(\alpha_N)} \sum_{k=0}^{\infty} \left( \phi\left(\frac{\alpha_N}{k\alpha_N^2 + 1}\right) - \phi\left(\frac{\alpha_N}{(k+1)\alpha_N^2 + 1}\right) \right) \left( \psi\left(\frac{\alpha_N}{(k-1)\alpha_N^2 + 1}\right) - \psi\left(\frac{\alpha_N}{(k+2)\alpha_N^2 + 1}\right) \right) = 0, \quad (4.6)$$

where  $\phi(\cdot)$  and  $\psi(\cdot)$  are functions in Assumption 3. Then non-disclosure is not guaranteed.

As a consequence of this proposition, parameter  $\theta_0$  is point identified from matching observations.

**Proposition 4. (Non-disclosure risk guarantee).** *Let  $\alpha_N > 0$ ,  $\alpha_N \rightarrow 0$  be such that*

$$\lim_{N \rightarrow \infty} \frac{\max\{N^x, N^y\}}{\phi(\alpha_N)} \sum_{k=0}^{\infty} \left( \phi\left(\frac{\alpha_N}{k\alpha_N^2 + 1}\right) - \phi\left(\frac{\alpha_N}{(k+1)\alpha_N^2 + 1}\right) \right) \left( \psi\left(\frac{\alpha_N}{k\alpha_N^2 + 1}\right) - \psi\left(\frac{\alpha_N}{(k+1)\alpha_N^2 + 1}\right) \right) > 0. \quad (4.7)$$

Then non-disclosure is guaranteed.

Propositions 3 and 4 show that whether non-disclosure is guaranteed or not depends on whether the thresholds  $\frac{1}{\alpha_N}$ , above which the values of  $z_i^x$  or  $z_i^y$  are not revealed, grow slowly or fast as the sample size increases. The example below clarifies and sheds light on this issue.

**EXAMPLE 1.** *Suppose that  $\phi(\alpha) = \alpha$  and  $\psi(\alpha) = \alpha$  for  $0 < \alpha < \bar{\alpha} < 1$ . In the Appendix, we use propositions 3 and 4 and show in detail that if, for example,  $\alpha_N > 0$  are chosen in such a way that  $\alpha_N = o\left(\frac{1}{\max\{N^x, N^y\}}\right)$  as  $N \rightarrow \infty$ , then there is no non-disclosure guarantee, whereas if  $\alpha_N > 0$  are chosen in such a way that  $\max\{N^x, N^y\}^{1/5} \alpha_N \rightarrow c > 0$  as  $N \rightarrow \infty$ , then non-disclosure is guaranteed. Thus, in the former case, when non-disclosure is not guaranteed, thresholds  $\frac{1}{\alpha_N}$  grow faster than  $\max\{N^x, N^y\}$ . In the latter case, when non-disclosure is guaranteed, thresholds  $\frac{1}{\alpha_N}$  grow at the rate  $\max\{N^x, N^y\}^{1/5}$ .*

So far, we have shown that the presence of “thin sets” of consumers allows us to identify the parameters of the econometric model. However, the fact that we are using possibly small subset of consumers to estimate the behavior model, the obtained estimates may reveal information on those consumers. For instance suppose that consumer response  $Y$  is a discrete variable with values 0 and 1 and the consumer attribute  $X$  is a continuous variable with the support on  $[0, 1]$  and  $Y$  is contained in the internal firm data and  $X$  is the information from the public dataset. Suppose that as a result of the merge we constructed a dataset with two observations  $(0, 0)$  and  $(1, 1)$ . As a result, we fit a linear model  $y = x$  to the data. If the firm reveals this estimate, one would be able to correctly predict the response of an individual with the attribute value  $X = 1$  using the model. This means that the estimates themselves may cause a threat of disclosure, as in reality the response  $Y$  would correspond to the purchase of a specific product by an individual, or a visit to a specific webpage, or the answer in an online questionnaire.

Moreover, the example of micro-targeted online advertising in Korolova (2010) shows that the firm does not even need to release the model in order to create a disclosure threat. Korolova (2010) conducts a field experiment on Facebook.com where users are required to supply their information such as age, but they may also choose to make that information “private” which will not be observable by other Facebook users. Some users may also have attributes, such as their favorite band, the city of origin or their sexual orientation, that will only be visible to their “friends”. Facebook has a very advanced set of targeting tools available to its advertisers. Even though, the advertisers cannot explicitly request to have their ads shown to specific users, they can target ads to very narrow user groups. It turns out, one can “single out” some users that by setting the set of targets based on publicly observable user attributes (i.e. not those observable only by “friends”). Then, for instance, one can recover the unobservable age of the user by constructing different versions of the ad that will be shown if the targeted user is in a specific age group. Then knowing which ad was shown, the advertiser will recover the user’s age.

Returning to our previous example, suppose that the consumer model is used by the online advertising firm and evaluates the probability of consumer click. Suppose that the advertising firm does not reveal the estimates of the model but allows ad targeting. Then the advertiser can choose to target the consumer with the attribute value  $X = 1$  and ask the advertising company to show the ad only if the click probability is higher than 0.99... (so that the ad will be very “relevant” to this consumer). Then the fact that the ad was shown, allows the advertiser to correctly recover the

response of consumer with the attribute  $X = 1$ .

## 5 Empirical Application

### 5.1 Data collection

To illustrate our results on relationship between the bound on the disclosure risk and identification of economic model, we estimate the model of consumer restaurant choices determined by consumer demographics and mutual location of restaurants and consumer residences.

We collected the dataset from the public access websites on the Internet. To collect the information regarding the consumer ratings of restaurants we use the data from Yelp.com and to collect the data on the demographics we used the database of residential property taxes.

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

[Table 1 about here.]

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

[Figure 1 about here.]

Separately we collected the data on the individual restaurant reviews. For the source of that information we solely used the public data from Yelp.com corresponding to the restaurants located in Durham, NC (see <http://www.yelp.com/durham-nc>). First, we assembled the list of local restaurants in Durham that are represented on Yelp.com. Then for each restaurant we collected the information on that restaurant that is represented on its yelp.com page such as exact address, cuisine, price level (given by the brackets), family and children-friendly indicators. Then for each restaurant we collected the data on the personal reviews that were given by the Yelp.com users. Yelp.com has

reviewer user names that are in the format of the first name and the first letter of the last name. For each reviewer-restaurant pair we collected the data on the reviewer rating of the restaurant that can assign the grade from 1 to 5 to the restaurant with 5 being the highest grade. The dataset from Yelp.com for the Durham, NC produces the entries for 485 Yelp.com users who wrote 2326 reviews to 290 (out of 343 listed) Durham restaurants. We show the summary statistics for the constructed variables in Table 2. Figure 2 demonstrates the sample distributions of the attributes: distribution of restaurant ratings, distribution of the restaurant price levels, and the distribution of the restaurants by the zip codes.

[Table 2 about here.]

[Figure 2 about here.]

Next, we constructed the individual identifiers using the rank cutoff rule combining the edit distance using the first and last name in the property tax dataset and the user name on Yelp.com (see Appendix A), and the sum of ranks indicating that the taxpayer in the tax data is located in the same zip code as the restaurant. Given this simple matching rule, we identified 304 Yelp.com users as positive matches. Sixty six people are uniquely identifiable in both databases. Table 3 shows the distribution of obtained matches. One-to-one matches correspond to the edit distance zero, one-to-two matches correspond to the edit distance one, etc.

[Table 3 about here.]

The matched observations characterize the constructed merged dataset of Yelp reviews and the property tax bills. We were able to find the reviewers in Yelp and the property owners in the property tax bills for whom the the combined edit distance and the Euclidean distance between the numeric indicators (zip code and location of most frequent reviews) is equal to zero. We call this dataset the set of “one-to-one” matches. Based on reviewer first name we evaluate sex of reviewer and construct dummy variable indicating that the name of the individual in the taxpayer data has a name that belongs to the list of top 500 female names in the US from the Census data (as a proxy that the corresponding taxpayer is a female). We also constructed the indices for other demographic indicators, but their coefficients were insignificant in our structural model and we do not incorporate them into our analysis. The statistics in the “one-to-one” matches dataset is summarized in Table 4

[Table 4 about here.]

## 5.2 Individual restaurant rating model

In our empirical application we address an important problem of recovering individual preferences from split sample data. Such problems frequently arise in online ad targeting. Ad targeting requires

estimation of individual propensity to perform a certain action (such as a click or a purchase) conditional on individual attributes. If the advertising company possesses only observational data, estimation of such a model requires the merge between the purchase data and the data on the online activity of consumers. Without the merged data the consumer action model needs to be estimated only based on the observed recorded consumer activity. This leads to a familiar data selection problem. Our theoretical findings show a clear trade-off between the privacy restrictions and the ability to identify an econometric model. This means that the higher are the requirements imposed on the disclosure risk, the less information the researcher has regarding the size of the selection and, therefore, the less efficient targeting will be.

Table 5 provides evidence of the selection in our data. Columns 1 and 2 provide estimates for the probability of giving an review (probit model) for a certain restaurant considering the reviewers from the “one-to-one match” dataset with and without restaurant fixed effects. We can see that the reviews that we do observe are coming from the individuals who are more prone to give the restaurant reviews in the first place. We can model this selection using the individual characteristics that we can construct from the merged tax and yelp.com data. In other words, we will model the propensity to give review determined by, first, the propensity of an individual to go to a restaurant (which is function of income, location and other demographics). And then, conditional on the individual dining in a restaurant, the propensity to give a review will be determined by the individual’s (dis)satisfaction by the restaurant.

[Table 5 about here.]

We formalize this using the following individual decision model. An individual extracts the utility from dining in a restaurant that depends on the vector of restaurant-specific characteristics  $x_1$  and the vector of demographic characteristics of an individual (such as wealth, location, and ethnicity). The utility also depends on the individual-specific idiosyncratic component  $\eta$ , and on the restaurant-specific idiosyncratic component  $e$ , which are not observed by the econometrician. The full *ex post* utility of an individual is defined as

$$U = u(x_1, x) - \eta - e,$$

where we assume that it is separable in deterministic component  $u(\cdot, \cdot)$  and the stochastic component  $\eta + e$ . Then the individual decision problem is the following. First, the individual makes a decision to go to a restaurant based on his or her expectation of the restaurant quality:

$$d_0 = \mathbf{1} \{u(x_1, x) - \eta - E[e] \geq 0\}.$$

We assume that consumers can correctly evaluate the uncertainty regarding the restaurant quality. Second, after making the decision to dine at the restaurant, the individual decides to write a review highly rating the restaurant if the *ex post* utility from visiting the restaurant exceeds a certain threshold:

$$d_1 = \mathbf{1} \{u(x_1, x) - \eta - e \geq \underline{u}\}.$$

In other words, we expect the individual to write a review if he or she was either very happy or very unhappy with the dining experience. Finally, the restaurant rating will be positive if the individual was pleased with the dining experience:

$$d_2 = \mathbf{1} \{u(x_1, x) - \eta - e \geq \underline{u}\}.$$

In the data we observe the decision to write a favorable review along with the restaurant data  $y = (d_2, x_1)$  for all people who wrote a review and we can observe the individual characteristics  $x$ .

It is clear that without the additional demographic information we would not be able to correctly estimate the parameters of the decision problem only based on the restaurant rating data. In fact, *we only observe the data for individuals who indeed came to the restaurant and wrote a review*. This is the main source of the activity bias in this environment.

Now we map the structural elements of the model (individual's deterministic utility component) to the observable variables. Assume that utility shocks  $e$  and  $\eta$  are mutually independent and they are also independent from the observable characteristics of consumers and restaurants. We also normalize the distributions of unobserved shocks assuming that  $e \sim N(0, 1)$  and  $\eta \sim N(0, \sigma^2)$ . Then, the probability of decision to write a positive review, given that an individual writes a review and given the individual-specific unobserved shock can be written as

$$\begin{aligned} \Pr \left\{ d_2 = 1 \mid d_1 = d_0 = 1, x_1, x, \eta \right\} &= \frac{\Pr\{e \leq u(x_1, x) - \underline{u} - \eta \mid d_0 = 0, x, x_1, \eta\}}{\Pr\{|u(x_1, x) - e - \eta| \geq \underline{u} \mid d_0 = 0, x, x_1, \eta\}} \\ &= \frac{\Phi(u(x_1, x) - \underline{u} - \eta)}{\Phi(u(x_1, x) - \underline{u} - \eta) + \Phi(-u(x_1, x) - \underline{u} + \eta)}, \end{aligned}$$

where  $\Phi(\cdot)$  is the cdf of standard normal distribution. Finally, recalling that we normalized the restaurant-specific shock which leads to  $E[e] = 0$ . This means that we can determine the density of the distribution of individual-specific utility shocks for those people who choose to dine at the restaurant:

$$f(\eta \mid d_0 = 1, x_1, x) = \begin{cases} \frac{\varphi\left(\frac{\eta}{\sigma}\right)}{\sigma \Phi(u(x_1, x))}, & \text{if } \eta \leq u(x_1, x), \\ 0, & \text{otherwise,} \end{cases}$$

where  $\varphi(\cdot)$  is the standard normal density. As a result, we are able to express the observable probability of a favorable review by taking the expectation over the utility shocks for consumers who chose to dine in the restaurant:

$$\begin{aligned} \Pr \left\{ d_2 = 1 \mid d_1 = d_0 = 1, x_1, x \right\} \\ = (\sigma \Phi(u(x_1, x)))^{-1} \int_{-\infty}^{u(x_1, x)} \frac{\Phi(u(x_1, x) - \underline{u} - \eta) \varphi\left(\frac{\eta}{\sigma}\right)}{\Phi(u(x_1, x) - \underline{u} - \eta) + \Phi(-u(x_1, x) - \underline{u} + \eta)} d\eta. \end{aligned}$$

We can establish non-parametric identification of deterministic component of individual utility given the specified assumptions on unobservable variables and the individual decision.

**THEOREM 4.** *Suppose that there exist  $x_1^*$  and  $x^*$ , and  $x_1^{**}$  and  $x^{**}$  in the support of random variables  $X_1$  and  $X$  such that  $u(x_1^*, x^*) = 0$  and  $u(x_1^{**}, x^{**}) = 1$ . Then if there is a subset of the support of  $X_1$  and  $X$  where the observable probability  $Pr\left\{d_2 = 1 \mid d_1 = d_0 = 1, x_1, x\right\}$  has non-zero matrix of first derivatives (or first differences for discrete covariates) with respect to  $x_1$  and  $x$ , then structural parameters of the model  $\{u(\cdot, \cdot), \underline{u}, \sigma\}$  are identified.*

Provided that this identification result, we can use our data to estimate the structural parameters of the model. Taking into account the size of the merged dataset that were able to create as well as our desire to compare the results from standard linear models, we choose to further parametrise individual utility, assuming that it is linear.

Table 6 presents the estimated parameters of the structural model. As one can see, selection has a very large impact on the obtained estimates. The results indicate a large sizeable impact of the property value on the individual utility index. In other words an individual with a more expensive property is more likely to go to a restaurant. Also we find that Japanese and Mexican restaurants tend to have a high positive impact on the utility index as well.

[Table 6 about here.]

### 5.3 Data protection: $k$ -anonymity and the quality of point identification

As our analysis shows, using a simple notion of the edit distance for the string entries in combination with the Euclidean distance for numeric entries in the database of Yelp.com users and the property tax data from Durham county, allows us find 65 users for whom there exist counterparts in each database with the distance equal to zero. This means that we successfully performed the linkage attack on the Yelp reviews database. This in fact allowed us to construct the point estimates for our consumer behavior model.

Now we can analyze how the parameters will be affected if we want to enforce a bound on the disclosure risk. To do that we use the notion of  $k$ -anonymity.  $k$ -anonymity requires that for each observation in the main database there is at least  $k$  equally good matches in the auxiliary database. In our data the main attribute that was essential for correct matches was the name and the last name information. To break these links, we started erasing letters from individual names. For instance, we transform the name “Denis” to “Deni\*” then to “Den\*”. Then if in the Yelp data we observe the 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 which is definitely smaller than the edit distance between “Dennis” and “Denis” (equal to 1). Then in property tax data we suppressed the last two letters leading to transformation “Den\*”, the distance between both “Dennis” and “Denis” and “Den\*” is the same.

Using character suppression we managed to attain  $k$ -anonymity with  $k = 2$  and  $k = 3$  by erasing, correspondingly 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. Then for each entry in the Yelp database that has at least one counterpart in the property tax data with the edit distance less or equal to  $k$ , we construct the dataset of potential matches in Yelp and the dataset of possible matched observations in the property tax dataset. Then, we construct matched databases using each potentially matched pair. As a result, if we have, for instance,  $N$  observations in the Yelp database each having exactly  $k$  counterparts in the property tax database, then we construct  $k^N$  matched datasets. For each such matched dataset we can construct the point estimates. Figure 3 demonstrates the two-dimensional cuts of the obtained identified set of parameters under  $k$ -anonymity with the original point estimates.

[Figure 3 about here.]

As we can see, although some parameters maintain their sign when the identified set is constructed (such as price of the restaurant, property values, and gender), other parameters have the identified set including the origin. As a result, one is not even able to infer their correct signs if  $k$ -anonymity is enforced.

## 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 and 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 targeted online advertising, we show that the identity disclosure may occur even when the data is not publicly shared. We then investigate the possibility of imposing the bound on the disclosure risk. Such a bound can be enforced by using one of many available methods such as  $k$ -anonymity or synthetic data. However, we find that the presence of the bound on the disclosure risk will also lead to the loss of point identification of the model.

## References

- ABOWD, J., AND L. VILHUBER (2008): “How Protective Are Synthetic Data?,” in *Privacy in Statistical Databases*, pp. 239–246. Springer.
- ABOWD, J., AND S. WOODCOCK (2001): “Disclosure limitation in longitudinal linked data,” *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, pp. 215–277.
- ACQUISTI, A. (2004): “Privacy and security of personal information,” *Economics of Information Security*, pp. 179–186.
- ACQUISTI, A., A. FRIEDMAN, AND R. TELANG (2006): “Is there a cost to privacy breaches? An event study,” in *Fifth Workshop on the Economics of Information Security*. Citeseer.
- ACQUISTI, A., AND J. GROSSKLAGS (2008): “What can behavioral economics teach us about privacy,” *Digital Privacy: Theory, Technologies, and Practices*, pp. 363–377.
- ACQUISTI, A., AND H. VARIAN (2005): “Conditioning prices on purchase history,” *Marketing Science*, pp. 367–381.
- AGGARWAL, G., T. FEDER, K. KENTHAPADI, R. MOTWANI, R. PANIGRAHY, D. THOMAS, AND A. ZHU (2005): “Approximation algorithms for k-anonymity,” *Journal of Privacy Technology*, 2005112001.
- CALZOLARI, G., AND A. PAVAN (2006): “On the optimality of privacy in sequential contracting,” *Journal of Economic Theory*, 130(1), 168–204.
- CHAUDHURI, S., K. GANJAM, V. GANTI, AND R. MOTWANI (2003): “Robust and efficient fuzzy match for online data cleaning,” in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pp. 313–324. ACM.
- CIRIANI, V., S. DI VIMERCATI, S. FORESTI, AND P. SAMARATI (2007): “k-Anonymity,” *Secure Data Management in Decentralized Systems*. Springer-Verlag.
- DUNCAN, G., S. FIENBERG, R. KRISHNAN, R. PADMAN, AND S. ROEHRIG (2001): “Disclosure limitation methods and information loss for tabular data,” *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pp. 135–166.
- DUNCAN, G., AND D. LAMBERT (1986): “Disclosure-limited data dissemination,” *Journal of the American statistical association*, 81(393), 10–18.
- DUNCAN, G., AND S. MUKHERJEE (1991): “Microdata Disclosure Limitation in Statistical Databases: Query Size and Random Sample Query Control,” .
- DUNCAN, G., AND R. PEARSON (1991): “Enhancing access to microdata while protecting confidentiality: Prospects for the future,” *Statistical Science*, pp. 219–232.

- DWORK, C. (2006): “Differential privacy,” *Automata, languages and programming*, pp. 1–12.
- DWORK, C., AND K. NISSIM (2004): “Privacy-preserving datamining on vertically partitioned databases,” in *Advances in Cryptology–CRYPTO 2004*, pp. 134–138. Springer.
- FELLEGI, I., AND A. SUNTER (1969): “A theory for record linkage,” *Journal of the American Statistical Association*, pp. 1183–1210.
- FIENBERG, S. (1994): “Conflicts between the needs for access to statistical information and demands for confidentiality,” *Journal of Official Statistics*, 10, 115–115.
- (2001): “Statistical perspectives on confidentiality and data access in public health,” *Statistics in medicine*, 20(9-10), 1347–1356.
- GOLDFARB, A., AND C. TUCKER (2010): “Online display advertising: Targeting and obtrusiveness,” *Marketing Science*.
- GROSS, R., AND A. ACQUISTI (2005): “Information revelation and privacy in online social networks,” in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pp. 71–80. ACM.
- GUSFIELD, D. (1997): *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press.
- HOROWITZ, J., AND C. MANSKI (2006): “Identification and estimation of statistical functionals using incomplete data,” *Journal of Econometrics*, 132(2), 445–459.
- HOROWITZ, J., C. MANSKI, M. PONOMAREVA, AND J. STOYE (2003): “Computation of bounds on population parameters when the data are incomplete,” *Reliable computing*, 9(6), 419–440.
- JARO, M. (1989): “Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida,” *Journal of the American Statistical Association*, pp. 414–420.
- KOROLOVA, A. (2010): “Privacy violations using microtargeted ads: A case study,” in *IEEE International Workshop on Privacy Aspects of Data Mining (PADM’2010)*, pp. 474–482.
- LAMBERT, D. (1993): “Measures of disclosure risk and harm,” *Journal of Official Statistics*, 9, 313–313.
- LEFEVRE, K., D. DEWITT, AND R. RAMAKRISHNAN (2005): “Incognito: Efficient full-domain k-anonymity,” in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 49–60. ACM.
- (2006): “Mondrian multidimensional k-anonymity,” in *Data Engineering, 2006. ICDE’06. Proceedings of the 22nd International Conference*, pp. 25–25. IEEE.
- MAGNAC, T., AND E. MAURIN (2008): “Partial identification in monotone binary models: discrete regressors and interval data,” *Review of Economic Studies*, 75(3), 835–864.

- MANSKI, C. (2003): *Partial identification of probability distributions*. Springer Verlag.
- MILLER, A., AND C. TUCKER (2009): “Privacy protection and technology diffusion: The case of electronic medical records,” *Management Science*, 55(7), 1077–1093.
- MOLINARI, F. (2008): “Partial identification of probability distributions with misclassified data,” *Journal of Econometrics*, 144(1), 81–117.
- NARAYANAN, A., AND V. SHMATIKOV (2008): “Robust de-anonymization of large sparse datasets,” in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pp. 111–125. IEEE.
- NEWCOMBE, H., J. KENNEDY, S. AXFORD, AND A. JAMES (1959): “Automatic linkage of vital and health records,” *Science*, 130, 954–959.
- RIDDER, G., AND R. MOFFITT (2007): “The econometrics of data combination,” *Handbook of Econometrics*, 6, 5469–5547.
- SALTON, G., AND D. HARMAN (2003): *Information retrieval*. John Wiley and Sons Ltd.
- SAMARATI, P., AND L. SWEENEY (1998): “Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression,” Discussion paper, Cite-seer.
- SWEENEY, L. (2002a): “Achieving k-anonymity privacy protection using generalization and suppression,” *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, 10(5), 571–588.
- (2002b): “k-anonymity: A model for protecting privacy,” *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, 10(5), 557–570.
- TAYLOR, C. (2004): “Consumer privacy and the market for customer information,” *RAND Journal of Economics*, pp. 631–650.
- VARIAN, H. (2009): “Economic aspects of personal privacy,” *Internet Policy and Economics*, pp. 101–109.
- WINKLER, W. (1999): “The state of record linkage and current research problems,” in *Statistical Research Division, US Census Bureau*. Citeseer.

## 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 many practical applications. In many cases there do exist perfect cross-database identifiers of individual entries. There could be multiple reasons why that is the case. For instance, there could be errors in data entry and processing, wrong variable formatting, and duplicate data entry. The idea that has arisen in Newcombe, Kennedy, Axford, and James (1959) and was later formalized in Fellegi and Sunter (1969) was to treat the record linkage problem as a problem of classification of record subsets into matches, non-matches and uncertain cases. This classification is based on defining the similarity metric between each two records. Then given the similarity metric one can compute the probability of particular pair of records being a match or non-match. The classification of pairs is then performed by fixing the probability of erroneous identification of a non-matched pair of records as a match and a matched pair of records as a non-match by minimizing the total proportion of pairs that are uncertain. This matching technique is based on the underlying assumption of randomness of records being broken. As a result, using the sample of perfectly matched records one can recover the distribution of the similarity metric for the matched and unmatched pairs of records. Moreover, as in hypothesis testing, one needs to fix the probability of record mis-identification. Finally, the origin of the similarity metric remains arbitrary.

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

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

Ridder and Moffitt (2007) overviews some techniques for purely numeric data combination. In the absence of validation subsamples that may incorporate distributional assumptions on the “similar” numeric variables. For instance, joint normality assumption with a known sign of correlation can allow one to invoke likelihood-based techniques for record linkage.

## B Proofs

**Proof of Proposition 1.** Using Assumption 3 (iii) and the law of iterated expectations,

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

By Assumption 3 (i) and (iii),

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

This implies

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

which is equivalent to (3.2).

**Proof of Proposition 2.** Note that  $E^N[\rho(y_j, x_i; \theta)|x_i = x] =$

$$A_N(x)Pr\left(m_{ij} = 1 \mid x_i = x, |z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N}\right) + B_N(x)Pr\left(m_{ij} = 0 \mid x_i = x, |z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N}\right),$$

where

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

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

As shown in Proposition 1,

$$A_N(x) = E[\rho(Y, X, \theta) \mid X = x].$$

Now consider the numerator in  $B_N(x)$ .

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

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

Part (ii) of Assumption 3 in particular implies that for two independent random vectors  $(\tilde{Y}, \tilde{Z}^y)$  and  $(X, Z^x)$  with distributions  $f_{Y, Z^y}$  and  $f_{X, Z^x}$  respectively, for small  $\alpha > 0$ , and for almost all  $x \in \mathcal{X}$  and  $\tilde{y} \in \mathcal{Y}$ ,

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

for some non-decreasing positive function  $\xi(\cdot)$ . This means that the numerator in  $B_N(x)$  is

$$(\xi(\alpha_N) + o(\xi(\alpha_N))) \int \rho(y_j, x, \theta) f_Y(y_j) dy_j.$$

Analogously, for the denominator of  $B_N(x)$  we obtain that

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

Therefore,

$$\lim_{N \rightarrow \infty} B_N(x) = \lim_{N \rightarrow \infty} \frac{(\xi(\alpha_N) + o(\xi(\alpha_N))) \int \rho(y_j, x, \theta) f_Y(y_j) dy_j}{\xi(\alpha_N) + o(\xi(\alpha_N))} = \int \rho(y_j, x, \theta) f_Y(y_j) dy_j = E^*[\rho(\tilde{Y}, X; \theta) | X = x].$$

To complete the proof, we also take into account that the conditions of the proposition give

$$\lim_{N \rightarrow \infty} Pr(m_{ij} = 1 \mid x_i = x, |z_i^x| > \frac{1}{\alpha_N}, |z_i^x - z_j^y| < \alpha_N) = \lim_{N \rightarrow \infty} Pr(m_{ij} = 1 \mid |z_i^x| > \frac{1}{\alpha_N}, |z_i^x - z_j^y| < \alpha_N) = 1 - \gamma$$

$$\lim_{N \rightarrow \infty} Pr(m_{ij} = 0 \mid x_i = x, |z_i^x| > \frac{1}{\alpha_N}, |z_i^x - z_j^y| < \alpha_N) = \gamma$$

**Proof of Theorem 3.** Denote

$$a(\theta, x, \gamma) = (1 - \gamma)E[Y - X'\theta | X = x] + \gamma E^*[\tilde{Y} - X'\theta | X = x],$$

where  $\tilde{Y}$  is distributed according to  $f_Y(\cdot)$  and is independent of  $X$ . Introduce the distance  $r(a(\theta, x, \gamma))$  in the spirit of least squares in the following way:

$$r(a(\theta, x, \gamma)) = E_X[Xa(\theta, X, \gamma)]' E_X[Xa(\theta, X, \gamma)].$$

Note that

$$\begin{aligned} E_X[Xa(\theta, X, \gamma)] &= (1 - \gamma)E[X(Y - X'\theta)] + \gamma E^*[X(\tilde{Y} - X'\theta)] \\ &= (1 - \gamma)E[XY] - (1 - \gamma)E_X[XX']\theta + \gamma E_X[X]E_Y[\tilde{Y}] - \gamma E_X[XX']\theta \\ &= (1 - \gamma)E[XY] + \gamma E_X[X]E_Y[Y] - E_X[XX']\theta \\ &= E_X[XX'] \left( (1 - \gamma)E_X[XX']^{-1}E[XY] + \gamma E_X[XX']^{-1}E_X[X]E_Y[Y] - \theta \right) \\ &= E_X[XX'] \left( (1 - \gamma)\theta_0 + \gamma\theta_1 - \theta \right). \end{aligned}$$

Clearly,  $E_X[Xa(\theta, X, \gamma)]$  takes value 0 and, consequently,  $r(a(\theta, x, \gamma))$  takes its minimum value iff  $\theta = (1 - \gamma)\theta_0 + \gamma\theta_1$ . In other words,

$$\Theta_\gamma = \{(1 - \gamma)\theta_0 + \gamma\theta_1\}.$$

Theorem 2 the implies the result of this theorem.

**Proof of Proposition 3.** Probability  $\pi_{ij}^N(x)$  in (4.3) is equal to

$$\frac{P_x \left( |z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N} \mid m_{ij} = 1 \right) P_x(m_{ij} = 1)}{P_x \left( |z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N} \mid m_{ij} = 1 \right) P_x(m_{ij} = 1) + P_x \left( |z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N} \mid m_{ij} = 0 \right) P_x(m_{ij} = 0)}, \quad (\text{B.8})$$

where  $P_x$  is the notation for conditioning on  $x_i = x$ . Note that  $P_x(m_{ij} = 1) = \frac{1}{\max\{N_x, N_y\}}$ .

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

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

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

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

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

$$\inf_{j,i} \frac{P_x \left( |z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N} \mid m_{ij} = 0 \right)}{\phi(\alpha_N)P_x(m_{ij} = 1)} = \inf_{j,i} \frac{\max\{N_x, N_y\}P_x \left( |z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N} \mid m_{ij} = 0 \right)}{\phi(\alpha_N)}$$

converges to 0.

For the sake of notational simplicity, assume that  $Z^x$  takes only positive values. Now obtain that  $P_x(|z_j^y - z_i^x| < \alpha, |z_i^x| > \frac{1}{\alpha} | m_{ij} = 0)$  is bounded from above by

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

The same final expression in the inequality is obtained if  $Z^x$  can take negative as well as positive values. Taking into account this result, we conclude that if  $\alpha_N \rightarrow 0$  and (4.6) holds, then

$$\lim_{N \rightarrow \infty} \inf_{j,i} \frac{P_x \left( |z_j^y - z_i^x| < \alpha_N, |z_i^x| > \frac{1}{\alpha_N} \mid m_{ij} = 0 \right)}{\phi(\alpha_N) P_x(m_{ij} = 1)} = 0$$

and hence,

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

Therefore, there is no non-disclosure guarantee.

**Proof of Proposition 4.** Probability  $\pi_{ij}^N(x)$  in (B.8) is bounded from above by

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

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

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

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

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

$$\begin{aligned}
& P_x \left( \bigcup_{k=0}^{\infty} \left( \frac{1}{\alpha} + k\alpha < z_i^x \leq \frac{1}{\alpha} + (k+1)\alpha \right), |z_j^y - z_i^x| < \alpha \mid m_{ij} = 0 \right) \\
& \geq P_x \left( \bigcup_{k=0}^{\infty} \left( \left( \frac{1}{\alpha} + k\alpha < z_i^x \leq \frac{1}{\alpha} + (k+1)\alpha \right) \cap \left( \frac{1}{\alpha} + k\alpha < z_j^y \leq \frac{1}{\alpha} + (k+1)\alpha \right) \right) \mid m_{ij} = 0 \right) \\
& \geq \sum_{k=0}^{\infty} P_x \left( \left( \frac{1}{\alpha} + k\alpha < z_i^x \leq \frac{1}{\alpha} + (k+1)\alpha \right) \cap \left( \frac{1}{\alpha} + k\alpha < z_j^y \leq \frac{1}{\alpha} + (k+1)\alpha \right) \mid m_{ij} = 0 \right) \\
& \geq \sum_{k=0}^{\infty} P_x \left( \frac{1}{\alpha} + k\alpha < Z^x \leq \frac{1}{\alpha} + (k+1)\alpha \right) P \left( \frac{1}{\alpha} + k\alpha < Z^y \leq \frac{1}{\alpha} + (k+1)\alpha \right) \\
& \geq \sum_{k=0}^{\infty} \left( \phi \left( \frac{\alpha}{k\alpha^2 + 1} \right) - \phi \left( \frac{\alpha}{(k+1)\alpha^2 + 1} \right) \right) \left( \psi \left( \frac{\alpha}{k\alpha^2 + 1} \right) - \psi \left( \frac{\alpha}{(k+1)\alpha^2 + 1} \right) \right),
\end{aligned}$$

where we imposed that  $o(\phi(\alpha))$  and  $o(\psi(\alpha))$  in Assumption 3 are non-decreasing. This condition can be imposed without a loss of generality. The same final expression in the inequality is obtained if  $Z^x$  can take negative as well as positive values.

Taking into account this result, condition (4.7) and the fact that  $\alpha_N > 0$  for all  $N$ , we conclude that for all  $N^x$  and  $N^y$ ,

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

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

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

Clearly,  $\lim_{N \rightarrow \infty} \sup_{j,i} \pi_{ij}^N(x) \leq \frac{1}{1+0.5\Delta}$  if this limit exists, and therefore,

$$\sup_{x \in \mathcal{X}} \liminf_{N \rightarrow \infty} \pi_{ij}^N(x) \leq \frac{1}{1+0.5\Delta} < 1.$$

To summarize, non-disclosure is guaranteed.

**Proofs in Example 1.** First, consider  $\sum_{k=0}^{\infty} \left( \frac{\alpha}{k\alpha^2+1} - \frac{\alpha}{(k+1)\alpha^2+1} \right) \left( \frac{\alpha}{(k-1)\alpha^2+1} - \frac{\alpha}{(k+2)\alpha^2+1} \right)$ . It is equal to

$$\begin{aligned}
& \sum_{k=0}^{\infty} \frac{\alpha^3}{(k\alpha^2+1)((k+1)\alpha^2+1)} \frac{3\alpha^3}{((k-1)\alpha^2+1)((k+2)\alpha^2+1)} = \frac{3\alpha^6}{(\alpha^2+1)(-\alpha^2+1)(2\alpha^2+1)} + \\
& + \sum_{k=1}^{\infty} \frac{3\alpha^6}{(k\alpha^2+1)((k+1)\alpha^2+1)((k-1)\alpha^2+1)((k+2)\alpha^2+1)} = \frac{3\alpha^4}{(\alpha+\frac{1}{\alpha})(-\alpha^2+1)(2\alpha+\frac{1}{\alpha})} + \\
& + \sum_{k=1}^{\infty} \frac{3\alpha^2}{(k\alpha+\frac{1}{\alpha})((k+1)\alpha+\frac{1}{\alpha})((k-1)\alpha+\frac{1}{\alpha})((k+2)\alpha+\frac{1}{\alpha})}
\end{aligned}$$

Now use the fact that for  $m > 0$ , the function  $m\alpha + \frac{1}{\alpha}$  attains its minimum for  $\alpha > 0$  at the point  $\alpha = 1/\sqrt{m}$ . This minimum value is equal to  $2\sqrt{m}$ . Taking this account, we obtain that the infinite sum in the above formula is bounded from above by

$$\frac{3\alpha^4}{4\sqrt{2}(-\bar{\alpha}^2 + 1)} + 3\alpha^2 \sum_{k=1}^{\infty} \frac{1}{2\sqrt{k} \cdot 2\sqrt{k+1} \cdot 2\sqrt{k-1} \cdot 2\sqrt{k+2}}$$

Since

$$\sum_{k=1}^{\infty} \frac{1}{\sqrt{k} \cdot \sqrt{k+1} \cdot \sqrt{k-1} \cdot \sqrt{k+2}} < \infty,$$

and  $\phi(\alpha) = \alpha$ , we obtain that the convergence  $\max\{N^x, N^y\}_{\alpha_N} \rightarrow 0$  as  $N \rightarrow \infty$  implies that the condition (4.6) in Proposition 3 is satisfied and, thus, there is no uniform privacy guarantee.

Now consider  $\sum_{k=0}^{\infty} \left( \phi\left(\frac{\alpha}{k\alpha^2+1}\right) - \phi\left(\frac{\alpha}{(k+1)\alpha^2+1}\right) \right) \left( \psi\left(\frac{\alpha}{k\alpha^2+1}\right) - \psi\left(\frac{\alpha}{(k+1)\alpha^2+1}\right) \right)$ . It is equal to

$$\sum_{k=0}^{\infty} \frac{\alpha^6}{(k\alpha^2+1)^2((k+1)\alpha^2+1)^2} \geq \alpha^6 \sum_{k=0}^{\infty} \frac{1}{(k\bar{\alpha}^2+1)^2((k+1)\bar{\alpha}^2+1)^2}$$

Since

$$\sum_{k=0}^{\infty} \frac{1}{(k\bar{\alpha}^2+1)^2((k+1)\bar{\alpha}^2+1)^2} < \infty,$$

we conclude that the convergence  $\max\{N^x, N^y\}_{\alpha_N^5} \rightarrow c > 0$  implies that (4.7) in Proposition 4 is satisfied and, thus, privacy is guaranteed.

**Proof of Theorem 4.** Consider the observed positive rating probability at points  $(x_1^*, x^*)$  and  $(x_1^{**}, x^{**})$ . We note that

$$\begin{aligned} \Pr \left\{ d_2 = 1 \mid d_1 = d_0 = 1, x_1^*, x^* \right\} &= 2 \int_{-\infty}^0 \frac{\Phi(-\underline{u} - \sigma z) \varphi(z)}{\Phi(-\underline{u} - \sigma z) + \Phi(-\underline{u} + \sigma z)} dz, \\ \Pr \left\{ d_2 = 1 \mid d_1 = d_0 = 1, x_1^{**}, x^{**} \right\} &= \frac{1}{\Phi(1)} \int_{-\infty}^1 \frac{\Phi(1 - \underline{u} - \sigma z) \varphi(z)}{\Phi(1 - \underline{u} - \sigma z) + \Phi(-1 - \underline{u} + \sigma z)} dz. \end{aligned}$$

We note that for any  $\sigma > 0$  and  $\underline{u} > 0$  the gradients of the right-hand side of both equations are not equal to zero. Moreover, both right-hand sides are monotone increasing in  $\sigma$  and monotone decreasing in  $\underline{u}$  taking values from 0 to 1. By the intermediate value theorem for continuous functions the constructed system of equations has a solution. Moreover, due to strict monotonicity, this solution is unique.

Finally, given  $\sigma$  and  $\underline{u}$ , we can see that the right-hand side is depends on function  $u(x_1, x)$ . We can differentiate the right-hand side expression with respect to  $u(\cdot, \cdot)$  as an argument. Then we note that the gradient of the observed probability with respect to the unknown utility at point  $(x_1^*, x^*)$

and be expressed as

$$1 - \sqrt{\frac{2}{\pi}} \mathbf{P}^* + 2 \int_{-\infty}^0 \kappa(z) \frac{\Phi(-\underline{u} - \sigma z) \varphi(z)}{\Phi(-\underline{u} - \sigma z) + \Phi(-\underline{u} + \sigma z)} dz,$$

where  $\mathbf{P}^* = \Pr \left\{ d_2 = 1 \mid d_1 = d_0 = 1, x_1^*, x^* \right\}$  and  $\kappa(z) > 0$ . This expression is strictly positive. Therefore, integration from  $u(x_1^*, x^*) = 0$ , of the observed probability, allows us to identify the utility of consumers.

*Q.E.D.*

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

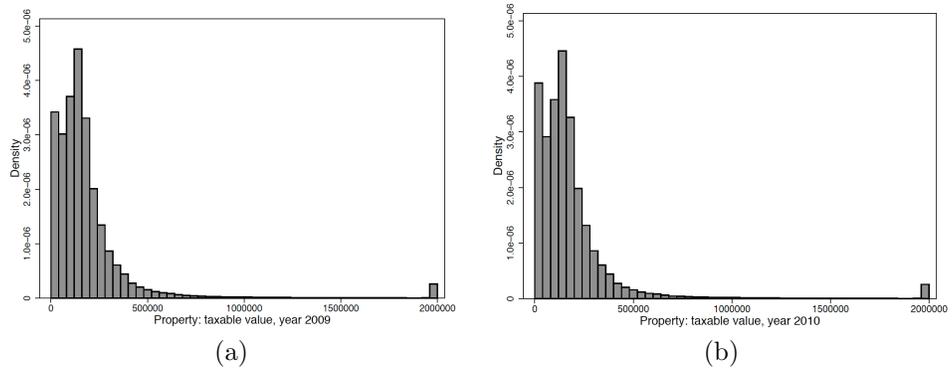
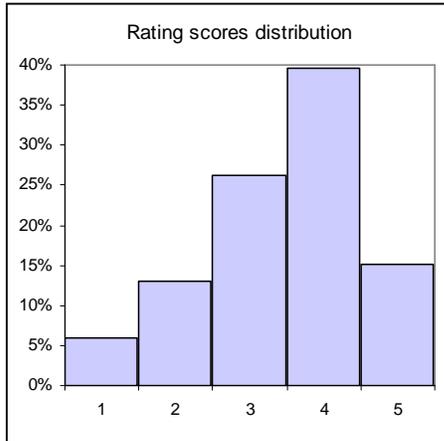
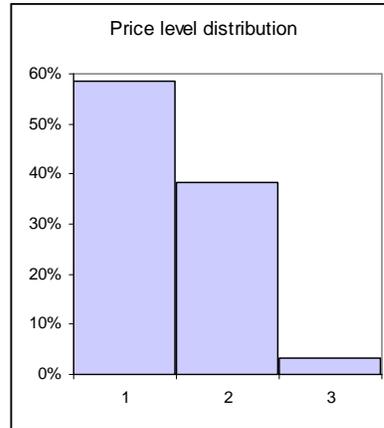


Figure 2: Distribution of features in the Yelp restaurant ratings data



(a): Restaurant ratings



(b): Restaurant price level



(c): Restaurant representation in the zip codes

Figure 3: Parameters of econometric model and identified sets under  $k$ -anonymity with  $k = 2$  and 3

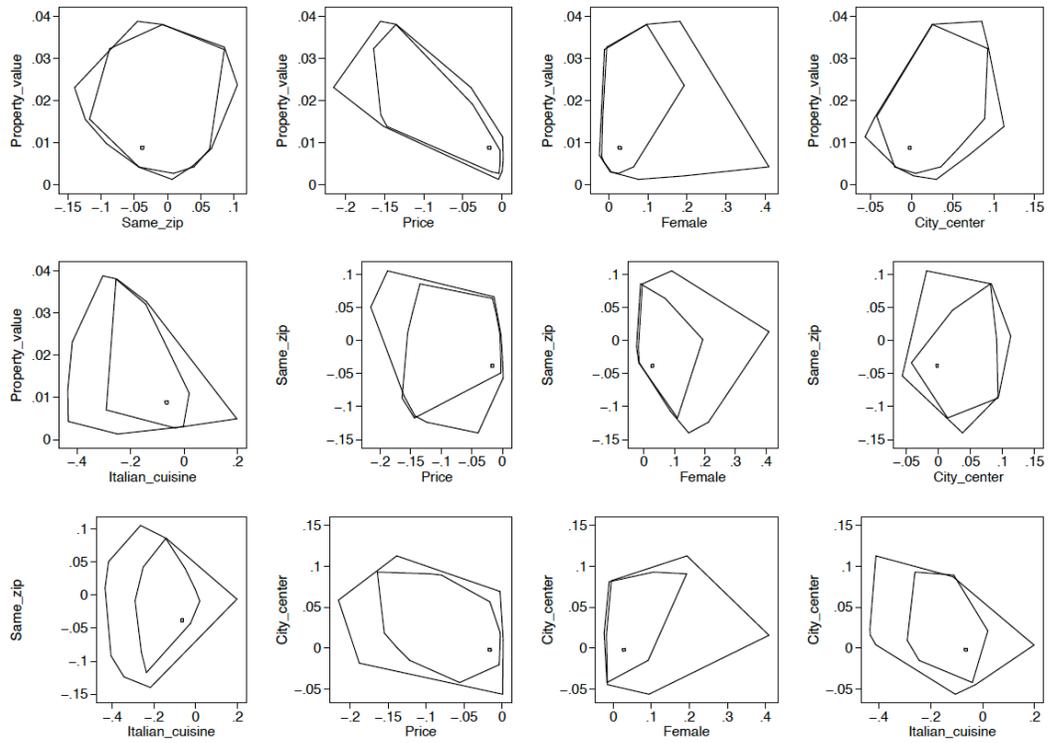


Table 1: Summary statistics from property tax bills in Durham County, NC.

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

Table 2: Summary statistics from Yelp.com for the restaurant information in Durham ,NC

Variable	Obs	Mean	Std.Dev.	Min	Max
Rating-level data					
Rating	2326	3.651	1.052	1	5
Price level	2265	1.631	0.573	1	3
cuisine: Mexican	2326	0.118	0.323	0	1
cuisine: Japanese	2326	0.062	0.242	0	1
cuisine: breakfast	2326	0.113	0.318	0	1
cuisine: Asian	2326	0.092	0.290	0	1
cuisine: American	2326	0.180	0.384	0	1
cuisine: Italian	2326	0.035	0.184	0	1
Restaurant-level data					
Average rating	290	3.479	0.796	1	5
Price level	251	1.446	0.558	1	3

Table 3: Features of edit distance-based matches

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

Table 4: Summary statistics in matched dataset of property tax bills and Yelp.com reviews

Variable	Obs	Mean	Std. Dev.	Min	Max
rating	429	3.492	1.001	1	5
price level	416	1.579	0.584	1	3
cuisine: Mexican	429	0.107	0.310	0	1
cuisine: Japanese	429	0.049	0.216	0	1
cuisine: breakfast	429	0.096	0.294	0	1
cuisine: Asian	429	0.084	0.278	0	1
cuisine: American	429	0.177	0.382	0	1
cuisine: Italian	429	0.051	0.221	0	1
I(city center)	429	0.233	0.423	0	1
I(same zip)	429	0.219	0.414	0	1
I(female)	429	0.214	0.411	0	1
log (property value)	429	12.26	0.634	10.34	13.14

Table 5: Probit estimates for review probability in matched dataset of property tax bills and Yelp.com reviews

	Pr(give review)	
log(property value)	0.129	0.144
	[0.038]***	[0.039]***
I(same zip)	0.252	0.289
	[0.059]***	[0.062]***
I(female)	-0.503	-0.539
	[0.051]***	[0.053]***
price level	0.095	
	[0.040]**	
I(city center)	0.103	
	[0.057]*	
cuisine: Mexican	0.077	
	[0.078]	
cuisine: Japanese	0.271	
	[0.115]**	
cuisine: breakfast	0.207	
	[0.083]**	
cuisine: Asian	0.077	
	[0.084]	
cuisine: American	0.092	
	[0.065]	
cuisine: Italian	0.07	
	[0.104]	
Restaurant FE	No	Yes
Constant	-3.474	-4.255
	[0.46]***	[1.03]***
Observations	11635	11635

Note: Robust standard errors in brackets

\* significant at 10%; \*\* significant at 5%; \*\*\* significant at 1%

Table 6: Estimates from the structural model

	Model with truncation ( $\bar{U}=1$ )	Model without truncation ( $\bar{U}=0$ )
cuisine: Mexican	0.254 <i>0.108</i>	0.541 <i>0.228</i>
cuisine: Japanese	0.546 <i>0.211</i>	1.064 <i>0.357</i>
cuisine: breakfast	0.088 <i>0.114</i>	0.210 <i>0.223</i>
cuisine: Asian	0.063 <i>0.112</i>	0.249 <i>0.234</i>
cuisine: American	0.024 <i>0.077</i>	0.145 <i>0.178</i>
cuisine: Italian	-0.153 <i>0.141</i>	-0.482 <i>0.302</i>
I(city center)	0.051 <i>0.074</i>	0.067 <i>0.163</i>
price	-0.060 <i>0.064</i>	-0.241 <i>0.114</i>
log(property value)	0.019 <i>0.009</i>	0.029 <i>0.017</i>
I(same zip)	-0.022 <i>0.036</i>	-0.155 <i>0.159</i>
I(female)	0.095 <i>0.067</i>	0.167 <i>0.156</i>
constant	0 (fixed)	0 (fixed)
$\bar{U}$	1 (fixed)	0 (fixed)
$\hat{\sigma}$	0.050 <i>0.005</i>	0.0001 <i>0.002</i>

Note: bootstrapped standard errors in italic