Прикладная эконометрика, 2017, т. 48, с. 44–62.
Applied Econometrics, 2017, v. 48, pp. 44–62.

T. Komarova, D. Nekipelov, A. Al Rafi, E. Yakovlev[1]

K-anonymity: A note on the trade-off between data utility and data security

Researchers often use data from multiple datasets to conduct credible econometric and statistical analysis. The most reliable way to link entries across such datasets is to exploit unique identifiers if those are available. Such linkage however may result in privacy violations revealing sensitive information about some individuals in a sample. Thus, a data curator with concerns for individual privacy may choose to remove certain individual information from the private dataset they plan on releasing to researchers. The extent of individual information the data curator keeps in the private dataset can still allow a researcher to link the datasets, most likely with some errors, and usually results in a researcher having several feasible combined datasets. One conceptual framework a data curator may rely on is k-anonymity, $k \geq 2$, which gained wide popularity in computer science and statistical community. To ensure k-anonymity, the data curator releases only the amount of identifying information in the private dataset that guarantees that every entry in it can be linked to at least k different entries in the publicly available datasets the researcher will use. In this paper, we look at the data combination task and the estimation task from both perspectives — from the perspective of the researcher estimating the model and from the perspective of a data curator who restricts identifying information in the private dataset to make sure that k-anonymity holds. We illustrate how to construct identifiers in practice and use them to combine some entries across two datasets. We also provide an empirical illustration on how a data curator can ensure k-anonymity and consequences it has on the estimation procedure. Naturally, the utility of the combined data gets smaller as k increases, which is also evident from our empirical illustration.

Keywords: data protection; data combination; *k*-anonymity.
JEL classification: C35; C14; C25; C13.

1. Introduction

Data needed to carry out empirical analyses are often stored in separate databases. Thus, accurately combining the observations in these disjoint databases often constitutes a first essential step in such analysis. For large and well-indexed databases, constructing a combined dataset often amounts to finding or constructing unique identifiers for the disjoint databases and selecting the unique matching pairs of observations based on these identifiers. Once a com-

[1] **Komarova Tatiana —** London School of Economics and Political Science, London, UK; t.komarova@lse.ac.uk.
Nekipelov Denis — University of Virginia, Charlottesville, US; denis@virginia.edu.
Al Rafi Ahnaf — London School of Economics and Political Science, London, UK; A.A.Rafi@lse.ac.uk.
Yakovlev Evgeny — New Economic School, Moscow, Russia; eyakovlev@nes.ru.

bined dataset of these unique matches has been constructed, these data can be used to estimate the requisite statistical model. However, as highlighted in the work by Komarova et al. (2015), the resulting estimated statistical model, and perhaps even policies based on the model, will reflect the characteristics of the individuals whose information was used to construct the combined dataset. Though the researcher may not release the combined dataset in its entirety, releasing the estimates along with the combination procedure used is common practice in the world of research

This enhanced information about the distribution of characteristics (as offered by published point estimates and data combination method) can present a threat to the security of the individuals' information, especially when one of the databases contains sensitive individual data. As shown by Komarova et al. (2015), among others, this threat can be posed even when each database itself poses no risk — because the sensitive data may be name-address anonymized and the other publicly available data may contain names and addresses along with other information about individual characteristics. For instance, the database containing sensitive information could be a name-address anonymized registry of individuals who have undergone or are currently undergoing psychiatric treatment at in-patient units, whilst another database could be a publicly available tax database containing names and addresses, along with information about other characteristics like taxable income, etc. A combined dataset will be able to assign names and addresses to the patients, and this will result in a security breach.

With such threats to security being made possible by the public release of estimates and combination method, those left in charge of handling the security of information present in the «sensitive databases» (who we shall call the data curators) may wish to impose further security measures before releasing the data to the researcher. For instance, they may require the researcher to sign an agreement which imposes an anonymity restriction on the researcher's available combination methods. It may also be the case that the researcher and the data curator are the same entity — as is the case with some experimental drug treatment RCT's in the medical field.

The goal of this paper is to consider the problem of estimating a linear regression model with selection problems (see Section 2) when the data is contained in separate databases, and the researcher (henceforth the econometrician) has a restriction placed on her research by the data curator regarding the anonymity of the individuals in the dataset. We present a simple procedure called «implementing *k*-anonymity» (Definition 2), where $k \in \mathbb{N}$, which the researcher can follow to ensure that she meets the anonymity requirements given to her by the curator. We also show that following this procedure comes at the cost of point estimation, and that empirical analysis following the implementation of *k*-anonymity for $k \geq 2$ can result in sets of parameter estimates instead of unique estimates. To do this, we consider an empirical example involving restaurants in the Durham, NC area and the ratings and reviews they receive on Yelp — a website which allows its users to rate and review listed businesses. We estimate a regression model of ratings (the utility individuals derived from dining at a given restaurant) on individual and restaurant characteristics. Yelp does not publish information on the users providing the ratings and we used the public personal and real estate property tax database to collect demographic data regarding potential Yelp users. Our data collection and summary statistics are presented in Section 3. The main objective of this paper is to illustrate how data combination can be conducted in practice when data come from different datasets and to illustrate the implementation of *k*-anonymity. With that purpose in mind, we treat the Yelp dataset as the sensitive or restricted dataset (we think of ranking data as sensitive data) and the property tax dataset as our publicly available, non-sensitive dataset. We first estimate the model presented in Section 2.4 based

on unique matches in the two datasets to get point estimates of the utility parameters of interest. These point estimates are presented in Table 5 in Section 4. Then we estimate the model again after implementing 2-anonymity and 3-anonymity on our combination procedure. These estimates are presented in Section 5. Section 2 can be broken down as follows.

Section 2.1 describes the assumed structure of the data we are using from a population perspective. We introduce the notion of an infeasible «master dataset» — a hypothetical combined dataset available from the two separate disjoint datasets consisting of all correct unique matches. Section 2.2 presents the problem we face in constructing identifiers for linking Yelp users with their corresponding entries in the property tax database. Section 2.3 outlines the choice problem faced by the econometrician selecting combination methods. We also discuss the anonymity restrictions placed on the econometrician's research in this subsection and the consequences these restrictions pose for combining data. Section 2.4 discusses the utility model we would like to estimate and its associated identification problem. We conclude our discussion in Section 6.

2. Model setup

We consider the problem of estimating a linearly parametrized utility function with selection problems (in the decision to engage in the utility-producing activity as well as in the decision to reveal the utility derived) when the data is contained in two separate datasets — one «private» and one «public». To do this, we follow a setup similar to (Komarova et al., 2015, 2017). However, we make a slight adjustment to the data generating model due to the fact that each observation in our hypothetical «master dataset» is not just a single individual, but an individual-firm combination.

2.1. Data structure: The «master» sample

A firm in our sample is described by a real-valued random vector Ξ and a string-valued random vector W^f. The realisations of Ξ belong to $\mathcal{X} \subseteq \mathbb{R}^{k_f}$ and the realisations of W^f belong to a finite set of strings we denote by $\mathcal{W}_f$.

Each individual's decision-utility responses for a given restaurant are described by a random vector $Y = (U, d_0, d_1)'$, which takes values in $\mathbb{Y} = \mathbb{U} \times \{0,1\}^2$, where $\mathbb{U} \subseteq \mathbb{R}$. Each individual is described by a random vector $X = (X^M, X^A)$ which takes values in $\mathbb{X} = \mathbb{X}_M \times \mathbb{X}_A$, where X^M takes values in $\mathbb{X}_M \subseteq \mathbb{R}^{k_M}$ and X^A takes values in $\mathbb{X}_A \subseteq \mathbb{R}^{k_A}$. The sample realisations of X^M come from our private dataset, henceforth referred to as the main dataset, and the sample realisations of X^A come from our public dataset, henceforth referred to as the auxiliary dataset.

Each individual is also labeled by vectors V and W containing combinations real and string valued variables that identify the individual, but do not interfere with the individual's utility-decision responses (e. g. names, date of birth, etc.). We assume that V and W serve simply as labels and do not contain any utility-relevant information. This is later incorporated more formally in Assumption 1. The realisations of V belong to the product space $\mathcal{W} = \mathcal{S}_A \times \mathbb{R}^{k_w}$ and the sample realisations of V are contained in our main dataset. Similarly, the realisations of W belong to the product space $\mathcal{W} = \mathcal{S}_A \times \mathbb{R}^{k_w}$ and the sample realisations of W are contained in the auxiliary dataset. We note that it is possible for some string valued variables to interfere with indi-

vidual responses. For instance one's address or the address of the restaurant could affect the decision to attend a certain restaurant due to the distance and associated cost of travel (though it may be unlikely to affect the rating given). However, we deal with this by noting that addresses can be well summarized by a numerical variable like zip-code (or if one wants to be extremely precise, some coordinate system). We assume that the addresses (and any other string variable that may affect responses) have corresponding numerical equivalents (procured in the same dataset) which account for these effects, and are the only channel through which responses are affected. As in (Komarova et al., 2015), we note that each string can be converted into the digital binary format, and that there are numerous examples of definitions of distances over strings (e.g. see (Wilson et al., 2006)). We can define the norm in $\mathcal{S}_M$ as the distance between a given point and a «generic» point corresponding to the most commonly observed set of attributes. Then, the norm in $\mathcal{V}$ is defined as a weighted combination of the norm in $\mathcal{S}_M$ and the Euclidean norm in $\mathbb{R}^{k_w}$. All of the above arguments can of course be replicated for $\mathcal{W}$.

We assume that the data generating process creates $N_f \cdot N_A$ i.i.d. draws from the joint distribution of the random vector $\left(Y, X, V, W, \Xi, W^f\right)$. N_f is the number of firms and N_A is the number of individuals. These draws form the infeasible «master» sample:

$$\left\{\left(y_{ip}, x_i, v_i, w_i, \xi_p, w_p^f\right) \middle| i = 1,\ldots,N_A,\ p = 1,\ldots,N_f\right\}.$$

2.2. Data structure: The linkage problem for users and taxpayers

We are however faced with the following problem: the observations for the covariates pertaining to individual characteristics are not contained in the same sample. Assume that no such restrictions are present for the firms. So we restrict our attention to dealing with the separated-data problem for individuals until Assumption 1, where we have to refer to the joint distribution of the responses, individual characteristics, individual identifiers and restaurant characteristics.

One sample containing N_A observations is the «public data» i.i.d. sample $\left\{\left(x_j^A, w_j\right)\right\}_{j=1}^{N_A}$, which as previously stated, we will refer to as the auxiliary dataset. To keep notation concise during conditioning, we will denote this dataset by $\mathcal{DS}_A$. The second dataset is a subset of $N \le N_A$ observations from the «master dataset» and contains information regarding the other individual-specific covariates $\left\{\left(x_i^M, v_i\right)\right\}_{i=1}^{N}$. As previously stated, we will refer to this dataset as the main dataset, and to keep notation concise during conditioning, we will denote this dataset by $\mathcal{DS}_M$. Following Komarova, Nekipelov, and Yakovlev (2015), we consider the case where there is no direct link between the main and auxiliary datasets, i.e. v_i and w_j do not provide immediate links between the two datasets. For instance, in our example, out of 4295 possible links between individuals in the main (Yelp) and auxiliary (Property Tax) datasets given by our linkage procedure, only 66 unique links were identifiable, without the imposition of any anonymizing restrictions.

Thus, before engaging in any identification or estimation exercises, the econometrician first needs to construct a linkage procedure that will correctly combine the two datasets with high probability. We will also assume that the econometrician has been instructed by those in charge of curating the data she is working with to impose restrictions on the linkage and estimation

procedures she uses to protect the anonymity of the individuals involved. The reasons for this are understandably numerous (as the discussion in Section 1 above should indicate) in the case of sensitive data in the «private» dataset. As stated, the anonymizing procedure we examine in this paper is called «*k*-anonymity» which is defined in Definition 2 given below.

We first consider a two-step procedure that uses the similarity of information contained in the identifiers and covariates to provide links between the two datasets. The parameter of interest is then estimated from the links established by our procedure. To establish said similarity, we assume that the econometrician constructs variables $Z^M = Z^M\left(X^M, V\right)$ and $Z^A = Z^A\left(X^A, W\right)$ which act as individual identifiers and take values in a common space $\mathcal{Z} = \mathcal{S} \times \mathbb{R}^{k_z}$. The space $\mathcal{S}$ is a finite set of non-numeric nature corresponding to the information contained in $\mathcal{S}_M$ and $\mathcal{S}_A$. Assume that $\mathcal{S}$ is a space of strings endowed with a metric $d_{\mathcal{S}}(\cdot,\cdot)$ from the set of commonly used string distances. The distance in $\mathcal{Z}$ is then defined as the weighted combination of $d_{\mathcal{S}}$ and the Euclidean metric in $\mathbb{R}^{k_z}$: $d_{\mathcal{Z}}\left(z^M, z^A\right) = \left(\omega_S d_{\mathcal{S}}\left(z_S^M, z_S^A\right)^2 + \omega_z \left\| z_z^M - z_z^A \right\|^2\right)^{1/2}$ where $Z^p = \left(Z_S^p, Z_z^p\right)$, for $p \in \{M, A\}$ and $\omega_S, \omega_z > 0$. We also define the «null» element of $\mathcal{S}$ as the observed set of attributes that has the most number of components with the other observed sets of attributes and denote this by $0_{\mathcal{S}}$. Then the norm in $\mathcal{Z}$ is defined as the distance from the null element: $\|z\|_{\mathcal{Z}} = \left(\omega_S d_{\mathcal{S}}\left(z_S, 0_S\right)^2 + \omega_z \|z_z\|^2\right)^{1/2}$.

When the set $\mathcal{Z}$ is sufficiently large or contains some potentially «difficult to replicate» information (e.g. full name, social security number, or a combination of both, etc.), then a match in the two datasets based on this infrequent information very likely singles out the data of one person. This is formalized by expecting that if the identifiers take infrequent values (modelled as the case identifiers having large norms), then the fact that the values of Z^M and Z^A are close implies that the two corresponding observations belong to the same individual with high probability. This probability is a decreasing function with respect to the frequency of the observed values of Z^M and Z^A. We maintain the following assumptions regarding the distribution of the random vector $\left(Y, X^M, Z^M, X^A, Z^A, \Xi\right)$:

Assumption 1. *There exists* $\bar{\alpha}$ *such that for any* $\alpha \in (0, \bar{\alpha})$:

1) *Proximity of identifiers*:

$$\mathbb{P}\left(d_{\mathcal{Z}}\left(Z^M, Z^A\right) < \alpha \middle| X^M = x^M,\ X^A = x^A,\ \left\|Z^A\right\|_{\mathcal{Z}} > 1/\alpha\right) \geq 1 - \alpha.$$

2) *Non-zero probability of extreme values in both datasets*:

$$\mathbb{P}\left(\left\|Z^M\right\|_{\mathcal{Z}} > 1/\alpha \middle| X^M = x^M\right) > 0,$$

$$\mathbb{P}\left(\left\|Z^A\right\|_{\mathcal{Z}} > 1/\alpha \middle| X^A = x^A\right) > 0.$$

3) *Redundancy of identifiers in the combined data*: *There exists a sufficiently large* $K > 0$ *such that for all* $\left\|z^M\right\|_{\mathcal{Z}} \geq K$ *and all* $\left\|z^A\right\|_{\mathcal{Z}} \geq K$

$$f\left(Y \middle| X^M = x^M,\ X^A = x^A,\ \Xi = \xi,\ Z^M = z^M,\ Z^A = z^A\right) = f\left(Y \middle| X^M = x^M,\ X^A = x^A,\ \Xi = \xi\right),$$

where f *is the joint density of* $\left(Y, X^M, Z^M, X^A, Z^A, \Xi\right)$ *and each* $f(\cdot|\cdot)$ *above pertains to the respective conditional densities.*

Assumption 1.1 embodies the idea that more reliable matches are provided by pairs of identifiers whose values are infrequent. For instance, in our example, if we found an observation in

the Yelp dataset for Durham, NC with the attribute «Aaditya C» and an observation in the Durham, NC Property Tax dataset with the attribute «Aaditya Chakraborty»[2], we might expect them to belong to the same individual with a higher probability than if we found attribute values «Jane D» in the Yelp dataset and «Jane Doe» in the Property Tax dataset. We stress here that infrequency of a particular identifier does not mean that the corresponding observation is an «outlier». If the two datasets contain very detailed individual information such as combinations of full name, address and social security number, most attribute values will be unique.

Assumption 1.2 requires that there are sufficiently many observations with infrequent attribute values. This can be established empirically in each of the observed datasets making this a testable assumption.

Assumption 1.3 is crucial for the purpose of identification. It implies that even for extreme values of the identifiers (i.e. for observations that fall below a prescribed level of frequency) and the associated observed covariates (for individuals as well as firms), the identifiers only serve the purpose of data labels as soon as the «master dataset» (or the feasible subsample of the «master dataset») is recovered.

We note that in our empirical example, prior to imposing our anonymity restrictions, we have to restrict our name-based identifier to the surname initial since this is what is available for each Yelp user — the entire surname of the individual is not revealed. Our identifiers also do not have a numerical component, since no numerical identifiers (e.g. zip codes, telephone numbers, social security numbers) are available in the Yelp dataset. Thus our identifier norm is based solely on the frequency of the strings observed as can be seen in the matching distribution given in Section 4, Table 2. The exposition above outlines more general constructions for the identifiers used in linking the two datasets.

2.3. Combining the data — decision rules and *k*-anonymity

Using the identifiers presented in Section 2.2, we describe the data combination procedure used by the econometrician in finite samples by means of a deterministic binary decision rule $\mathcal{D}_N : \mathcal{Z} \times \mathcal{Z} \to \{0,1\}$, where $N \in \mathbb{N}$ is the size of the main dataset $\mathcal{DS}_M$ and for each pair of observations i in the main dataset $\mathcal{DS}_M$ and j in auxiliary dataset $\mathcal{DS}_A$,

$$\mathcal{D}_N\left(z_i^M, z_j^A\right) = \begin{cases} 1, & \text{if } z_i^M \text{ and } z_j^A \text{ satisfy certain conditions,} \\ 0, & \text{otherwise.} \end{cases}$$

Thus, for each pair of observations i in $\mathcal{DS}_M$ and j in $\mathcal{DS}_A$, we have an indicator variable M_{ij} defined by $M_{ij} = \mathcal{D}_N\left(z_i^M, z_j^A\right)$ which labels the pair as a «match» $\left(M_{ij} = 1\right)$ if we think that the observations belong to the same individual or labels the pair as a «non-match» $\left(M_{ij} = 0\right)$ if we think that it is unlikely that the observations belong to the same individual or are simply uncertain about this.

We focus on the set of data combination rules that is generated by our Assumption 1.1. For example, for the prescribed $\overline{\alpha}$, we consider the data combination rule:

[2] This name does not belong to anyone known to the authors. It was generated by one co-author's knowledge of long Bengali first names and surnames that would be unlikely to observed exceedingly frequently in Durham, NC.

$$\mathcal{D}_N\left(z_i^M, z_j^A\right) = \mathbf{1}\left\{d_{\mathcal{Z}}\left(z_i^M, z_j^A\right) < \alpha_N, \left\|z_j^A\right\|_{\mathcal{Z}} > 1/\alpha_N\right\} \quad (1)$$

generated by a Cauchy sequence $(\alpha_N)_{N\in\mathbb{N}}$ such that $0 < \alpha_N < \bar{\alpha}$ for all $N \in \mathbb{N}$ and $\lim_{N\to\infty} \alpha_N = 0$. The sequence $(\alpha_N)_{N\in\mathbb{N}}$ gives a sequence of thresholds that impose more stringent «rare» identifier frequency and distance requirements on larger samples, and tries to isolate unique matches in the limit (identifier distance of zero). For a more thorough discussion of more general conditions that can be imposed on this sequence, the interested reader is encouraged to read (Komarova et al., 2015, 2017).

Once the decision rule for combining the two samples in question has been chosen, the econometrician also needs to consider the efficacy with which the chosen rule makes links between individuals and take into account the anonymity restrictions placed on her by the data curator. To do this, the following theoretical indicator is used:

Definition 1. *Let m_{ij} be the indicator of the event that i and j are actually the same individual, where i is an individual in the main dataset $\mathcal{DS}_M$ and j is an individual in the auxiliary dataset $\mathcal{DS}_A$. Thus m_{ij} is equal to 1 if and only if i and j correspond to the same individual in our infeasible «master dataset».*

Thus, m_{ij} is the «true match» indicator for any pair i, j. Since the «master dataset» is infeasible, it is impossible to actually observe m_{ij}. Given this infeasibility, we can expect the decision rule to indicate incorrect matches (under the true match m_{ij}) as being «possibly or likely correct» with positive probability. So, since we can make incorrect matches, the match indicator M_{ij} under the decision rule $\mathcal{D}_N$ is not necessarily equal to m_{ij}. The econometrician would however want M_{ij} and m_{ij} to be highly correlated given the data available and subject to any anonymity restrictions given to her by a data curator (including of course the case of no restrictions). We may assume that the econometrician has an objective function that depends on the choice of decision rule, the available data and m_{ij} which she wants to maximize subject to the restrictions. For instance, she may wish to maximize $\mathrm{Corr}\left[M_{ij}, m_{ij} \middle| \mathcal{D}_N'\left(z_i^M, z_j^A\right) = 1, \mathcal{DS}_M, \mathcal{DS}_A\right]$. This particular choice problem over decision rules is equivalent to solving[3]:

$$\mathcal{D}_N \in \underset{\mathcal{D}_N'}{\operatorname{argmax}}\, \mathbb{P}\left(m_{ij} = 1 \middle| \mathcal{D}_N'\left(z_i^M, z_j^A\right) = 1, \mathcal{DS}_M, \mathcal{DS}_A\right) \quad (2)$$

subject to $\mathcal{D}_N'$ satisfying the requisite anonymity restrictions.

The choice of a decision rule is a separate complex problem that deserves a detailed discussion in a different paper. We would not be able to do it justice in this paper. So here we just assume that the econometrician solves her particular problem and chooses an appropriate decision rule $\mathcal{D}_N$. She then constructs a combined dataset as follows:

1. If observation i in $\mathcal{DS}_M$ is included in the combined dataset, then $\sum_{j=1}^{N_A} \mathcal{D}_N\left(z_i^M, z_j^A\right) \geq 1$.

[3] For the interested reader who may be wondering how we derived the above optimisation problem involving a conditional probability from a statement about conditional correlation, notice that M_{ij}, m_{ij} and $M_{ij}m_{ij}$ are all indicator random variables. The conditional expectation and variance of each indicator are both equal to the conditional probability of the indicator taking the value 1. Also note that the conditional probability of m_{ij} taking the value 1 given the data available is fixed by definition of m_{ij} since it is independent to any identifiers used.

2. For each i satisfying $\sum_{j=1}^{N_A} \mathcal{D}_N\left(z_i^M, z_j^A\right) \geq 1$, we pick an observation j in $\mathcal{DS}_A$ such that $\mathcal{D}_N\left(z_i^M, z_j^A\right) = 1$ and add the combined vector $\left(x_i^M, z_i^M, x_j^A, z_j^A\right)$ to the combined dataset if neither $\left(x_i^M, z_i^M\right)$ for this specific i nor $\left(x_j^A, z_j^A\right)$ for this specific j enters the combined dataset as subvectors of other combined observation vectors in the combined dataset.

This process can of course result in several possible combined datasets due to the possibility of incorrect matches. We denote this collection of datasets resulting from the chosen decision rule $\mathcal{D}_N$ and the available data $\left(\mathcal{DS}_M, \mathcal{DS}_A\right)$ by $\mathcal{G}\left(\mathcal{D}_N, \mathcal{DS}_M, \mathcal{DS}_A\right)$. We also assume that this is a non-empty set for the chosen decision rule. Once the model to be estimated is identified, the econometrician runs the associated estimation procedure using a subset of the datasets in $\mathcal{G}\left(\mathcal{D}_N, \mathcal{DS}_M, \mathcal{DS}_A\right)$. This subset can be the entire set, or some randomly selected subset. Thus, once anonymity restrictions are satisfied, the estimates the econometrician can release become sets of point estimates instead of a single point estimate. Estimation from an arbitrarily chosen dataset in $\mathcal{G}\left(\mathcal{D}_N, \mathcal{DS}_M, \mathcal{DS}_A\right)$ can be subject to error stemming from mismatch and so, this particular course of action would be inadvisable.

We move on to describing a particular class (indexed by the half-open unit interval $(0,1]$) of anonymity restrictions that may be placed on the econometrician by the data curator based on disclosure risk. Suppose the econometrician is considering the decision rule choice problem and the data curator knows prior to the choice being made that the choice and its associated «combined dataset formation» process will be released to the public alongside the parameter estimates as part of the research documentation (as is standard or usual research practice). The sensitive data used however will not be released (as per the contract with the curator). The curator is concerned about the risk of individual disclosure. An individual disclosure occurs when an individual in the «master dataset» (available to the curator) is correctly matched in the two datasets by $\mathcal{D}_N$ — which is the event that $m_{ij} = 1$ given that $\mathcal{D}_N\left(z_i^M, z_j^A\right) = 1$ and given the available data $\mathcal{DS}_M$, $\mathcal{DS}_A$. For technical convenience, we assume that the curator is concerned with placing an upper bound on the conditional probability of the above event for any pair of observations in the respective datasets. That is, she is concerned with making sure that $\mathbb{P}\left(m_{ij} = 1 \middle| \mathcal{D}_N\left(z_i^M, z_j^A\right) = 1, \mathcal{DS}_M, \mathcal{DS}_A\right) \leq \delta$ for all observations i in $\mathcal{DS}_M$ and all observations j in $\mathcal{DS}_A$, where $\delta \in (0,1]$. For a full discussion of other measures of disclosure risk as well as measures of harm from disclosure, we refer the interested reader to (Komarova et al., 2017) and (Lambert, 1993). Notice that in the above set of anonymity restrictions, the case of the curator imposing $\delta = 1$ is equivalent to the curator not imposing any restrictions at all.

We now describe a set of anonymisation procedures (indexed by $k \in \mathbb{N}$) that a data curator may use. The econometrician then calibrates the estimator to the choice of the anonymisation procedure used by the data curator. This procedure is called «implementing k-anonymity». It became widely accepted in the computer science and data science community. Its description can be found e.g. in (Samarati, Sweeney, 1998; Sweeney, 2002a, 2002b). In our context it is defined as follows:

Definition 2 (*k*-anonymity). *Let* $k \in \mathbb{N}$ *be given. We say that the binary decision rule* $\mathcal{D}_N(\cdot,\cdot)$ *implements k-anonymity if for each observation i in the main dataset* $\mathcal{DS}_M$ $(i = 1,\ldots,N)$*, one of the following (mutually exclusive) conditions hold*:

1) $\mathcal{D}_N\left(z_i^M, z_j^A\right) = 0$ *for all* $j = 1,\ldots,N_A$; *that is,* i *cannot be combined with any observation* j *in the auxiliary dataset* $\mathcal{DS}_A$;

2) $\sum_{j=1}^{N_A} \mathcal{D}_N\left(z_i^M, z_j^A\right) \geq k$; *that is, for* i *there are at least* k *equally good matches in the auxiliary dataset.*

Remark 1. *Note that by implementing k-anonymity, we have that for any* i *from* $\mathcal{DS}_M$ *and any j from* $\mathcal{DS}_A$,

$$\mathbb{P}\left(m_{ij} = 1 \middle| \mathcal{D}_N\left(z_i^M, z_j^A\right) = 1, \mathcal{DS}_M, \mathcal{DS}_A\right) = \begin{cases} 0, & \text{if } \sum_{l=1}^{N_A} \mathcal{D}_N\left(z_i^M, z_l^A\right) = 0, \\ 1 \Big/ \sum_{j=1}^{N_A} \mathcal{D}_N\left(z_i^M, z_j^A\right), & \text{otherwise.} \end{cases}$$

So, it always holds that

$$\mathbb{P}\left(m_{ij} = 1 \middle| \mathcal{D}_N\left(z_i^M, z_j^A\right) = 1, \mathcal{DS}_M, \mathcal{DS}_A\right) \leq \frac{1}{k}.$$

The data curator first chooses a $k \in \mathbb{N}$ that would allow her to satisfy the upper bound on individual disclosure risk, δ, that yields the desired bound for the disclosure risk. For the given $\delta \in (0,1]$, the set of «correct» choices of k is non-empty, since $\left\{m \in \mathbb{N} \middle| m \geq 1/\delta\right\} \neq \varnothing$ (which follows since $\mathbb{R}$ has the Archimedean Property) and by Remark 1, we can choose any k in this set. Once this choice has been made, the econometrician restricts herself to the set of decision rules that account for the fact that the data curator used *k*-anonymity with a given k[4].

It is also worth noting that a choice rule $\mathcal{D}_N$ corresponding a higher *k* (as long as $k \leq N_A$) leads to a larger set of possible combined datasets $\mathcal{G}\left(\mathcal{D}_N, \mathcal{DS}_M, \mathcal{DS}_A\right)$. This follows from our observation that each k corresponds to an upper bound on disclosure risk of $1/k$. A more stringent restriction resulting from a choice of a higher k leads to each observation in the main dataset being matched under the associated rule $\mathcal{D}_N$ to more individuals in the auxiliary dataset. This results in more possible combined datasets if the combined dataset construction process above is followed.

We conclude this section by noting that the choice of decision rule can be controlled by restricting the identifiers used for the purpose of combination. This can be achieved for instance by adjusting the string norm by making the set of «frequently observed» characteristics larger. One could also achieve this by changing the set $\mathcal{Z}$ of values the identifier takes — making

[4] We note here that we have thus far left open the possibility of the curator imposing an anonymity restriction that is too stringent for the available data. For instance it is possible that $\delta^{-1} > N_A$, which would require each observation in $\mathcal{DS}_M$ to be linked by $\mathcal{D}_N$ to more observations in $\mathcal{DS}_A$ than it actually contains or none at all! We may therefore assume that the curator is reasonable, and imposes a δ which at least satisfies $\delta \geq 1/N_A$. One would of course usually prefer a δ that is much more lax to get more value from the available data, but still offers sufficient protection against individual disclosure to keep the curator satisfied.

it smaller would make it more difficult to find «rare» observations to construct unique matches with. We combine these approaches when implementing 2-anonymity and 3-anonymity in our empirical example in Section 5.

2.4. Individual restaurant rankings

2.4.1. Attendance and review decision process

We consider the model of individual restaurant rankings. The utility that an individual extracts from dining in at a restaurant depends on a vector of demographic characteristics of the individual (such as wealth, location, and ethnicity), x, and a vector of restaurant-specific characteristics, ξ. This utility also depends an individual-specific idiosyncratic component η on a restaurant-specific idiosyncratic component e, which is not observed by the econometrician. The full *ex post* utility of the individual is defined as:

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

where we assume that it is separable in a deterministic component $u(\cdot,\cdot)$ and the stochastic component $\eta + e$.

The individual decision problem is as follows. First, the individual makes a decision to go to a restaurant based on his or her expectation of the restaurant quality:

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

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 rating the restaurant if the *ex post* utility from visiting the restaurant exceeds a certain threshold:

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

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:

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

2.4.2. Identification

In the data we observe the decision to write a favourable review along with the restaurant data $y = (d_2, \xi)$ 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 came to the restaurant and wrote a review*. This is the main source of «activity bias» (similar to selection bias) in this environment.

Now we map the structural elements of the model (the individual's deterministic utility component) to the observable variables. Assume that utility shocks η and e are mutually indepen-

dent and that they are also independent from the observable characteristics of consumers and restaurants. We also normalize the distributions of unobserved shocks assuming that $e \sim \mathcal{N}(0,1)$ and $\eta \sim \mathcal{N}(0,\sigma^2)$. Then, the probability of the decision to write a positive review, given that the individual writes a review and given the individual-specific unobserved shock can be written as:

$$\mathbb{P}\left(d_2 = 1 \middle| d_1 = d_0 = 1, x, \xi, \eta\right) = \frac{\mathbb{P}\left(e \le u(x,\xi) - \underline{u} - \eta \middle| d_0 = 1, x, \xi, \eta\right)}{\mathbb{P}\left(\left|u(x,\xi) - e - \eta\right| \ge \underline{u} \middle| d_0 = 1, x, \xi, \eta\right)} =$$

$$= \frac{\Phi\left(u(x,\xi) - \underline{u} - \eta\right)}{\Phi\left(u(x,\xi) - \underline{u} - \eta\right) + \Phi\left(-u(x,\xi) - \underline{u} + \eta\right)},$$

where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution. Finally, recalling that we normalized the restaurant-specific shock, we necessarily have $\mathbb{E}[e] = 0$. This means that we can determine the density of the distribution of individual-specific utility shocks for an individual who chooses to dine at the restaurant:

$$f\left(\eta \middle| d_0 = 1, x, \xi\right) = \begin{cases} \dfrac{\varphi(\eta/\sigma)}{\sigma\Phi\left(u(x,\xi)\right)}, & \text{if } \eta \le u(x,\xi), \\ 0, & \text{otherwise,} \end{cases}$$

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

$$\mathbb{P}\left(d_2 = 1 \middle| d_1 = d_0 = 1, x, \xi\right) =$$

$$= \left(\sigma\Phi\left(u(x,\xi)\right)\right)^{-1} \int_{-\infty}^{u(x,\xi)} \frac{\Phi\left(u(x,\xi) - \underline{u} - \eta\right)\varphi(\eta/\sigma)}{\Phi\left(u(x,\xi) - \underline{u} - \eta\right) + \Phi\left(-u(x,\xi) - \underline{u} + \eta\right)} d\eta.$$

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

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

Proof. Consider the observed positive rating probability at points $\left(x^*, \xi^*\right)$, and $\left(x^{**}, \xi^{**}\right)$. We note that

$$\mathbb{P}\left(d_2 = 1 \middle| d_1 = d_0 = 1, x^*, \xi^*\right) = 2\int_{-\infty}^{0} \frac{\Phi\left(-\underline{u} - \sigma s\right)\varphi(s)}{\Phi\left(-\underline{u} - \sigma s\right) + \Phi\left(-\underline{u} + \sigma s\right)} ds,$$

$$\mathbb{P}\left(d_2 = 1 \middle| d_1 = d_0 = 1, x^{**}, \xi^{**}\right) = \frac{1}{\Phi(1)} \int_{-\infty}^{1} \frac{\Phi\left(1 - \underline{u} - \sigma s\right)\varphi(s)}{\Phi\left(1 - \underline{u} - \sigma s\right) + \Phi\left(-1 - \underline{u} + \sigma s\right)} ds.$$

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 σ 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 σ and $\underline{u}$, we can see that the right-hand side depends on the function $u(x,\xi)$. We can differentiate the right-hand side with respect to $u(\cdot,\cdot)$ as the argument. Then we note that the gradient of the observed probability with respect to the unknown utility at the point (x^*,ξ^*) can be expressed as

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

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

3. Data collection

Our data was collected from two sources using Perl scripts that gather information from webpage sources by taking advantage of page indexing.

Property tax data were extracted from the tax administration record search of the Durham county government web-site[5]. We looped over all unique individual record identifiers (parcel numbers) and collected data on property tax bills for the calendar years 2009 and 2010. As a result, we collected data on 103445 property tax bills for 2009 and 104068 property tax bills for 2010. Each bill contained information on the taxable value of the property, first and last names of the taxpayer and a description of the property (including property address). Data on the taxpayer website's source is stored in tabular format so we collected rows that correspond to the years 2009 and 2010, and then merged them into one dataset. This procedure took about three weeks. The resulting dataset is what we refer to in Section 2.2 as our «public» or auxiliary dataset, $\mathcal{DS}_A$.

Data on restaurants and reviewers were collected from the Yelp website[6]. First, we created data on the websites of Durham restaurants that are listed in Yelp. Then, we looped over the list of restaurants and collected data on restaurant street address, cuisine, price level (3-scale), reviewer's rating (5-scale), phone, zip and I (kid friendly). In addition we collected links to the webpages of reviewers of these restaurants. Finally, we went through reviewers' webpages, and collected data on each reviewer's first name, the initial of her/his surname (this is the only surname data available on Yelp pages), price level of restaurants she/he attended, and the average rating she/he gave to restaurants. To collect data on restaurant and reviewer characteristics we went through the corresponding webpage source and looked for key words that appear right before (or after) the needed information. For example, to collect data on restaurant phone numbers, we look for the key word 'bizPhone'. In the webpage source, this keyword appears right before phone number, so we extracted the 10 symbols that follow this keywords and got the desired restaurant phone number. Following the aforementioned process, we collected data on 485 users and 2326 reviews of 290

[5] See http://www.ustaxdata.com/nc/durham/.

[6] See http://www.yelp.com/durham-nc.

(out of 343 listed) Durham restaurants. This procedure took around one month. The resulting dataset is what we refer to in Section 2.2 as our «private» or main dataset, $\mathcal{DS}_M$.

Tables 1a and 1b present summary statistics of property tax, Yelp rating, cuisine and price data. Figures 1 and 2 present histograms of the taxable value of property in the years 2009 and 2010 respectively. Figures 3, 4 and 5 present histograms of restaurant rating, price levels and zip codes respectively.

Table 1a. Summary statistics from the property tax dataset

Variable	Observations	Mean	Standard deviation	Min	Max
Years 2009 and 2010					
Property: Taxable Value	207513	261611.9	1713482	3	2.78E+08
Year 2010					
Property: Taxable Value	104068	263216.1	1734340	3	2.78E+08

Table 1b. Summary statistics from the Yelp dataset

Variable	Observations	Mean	Standard deviation	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

Next, we merged the property tax and Yelp datasets using first name and the initial of the surname as identifiers. 304 Yelp users (out of 485) had at least one match with the same first name and same surname initial in the property tax data. Sixty-six of these users were uniquely identifiable in both databases. Table 2 presents the matching statistics.

Table 2. Matching distribution

# of matches	Frequency	Percentage	# of Yelp users
1 in Yelp → 1 in tax data	66	1.54	66
1 → 2	92	2.19	46
2 → 1	2	2.19	2
1 → 3	72	1.68	24
1 → 4	36	0.84	9
1 → 5	65	1.51	13
1 → 6	114	2.65	19
1 → 7	56	1.30	8
1 → 8	88	2.05	11
1 → 9	81	1.89	9
1 → 10 or more	3623	84.35	97
Total	4295	100	304

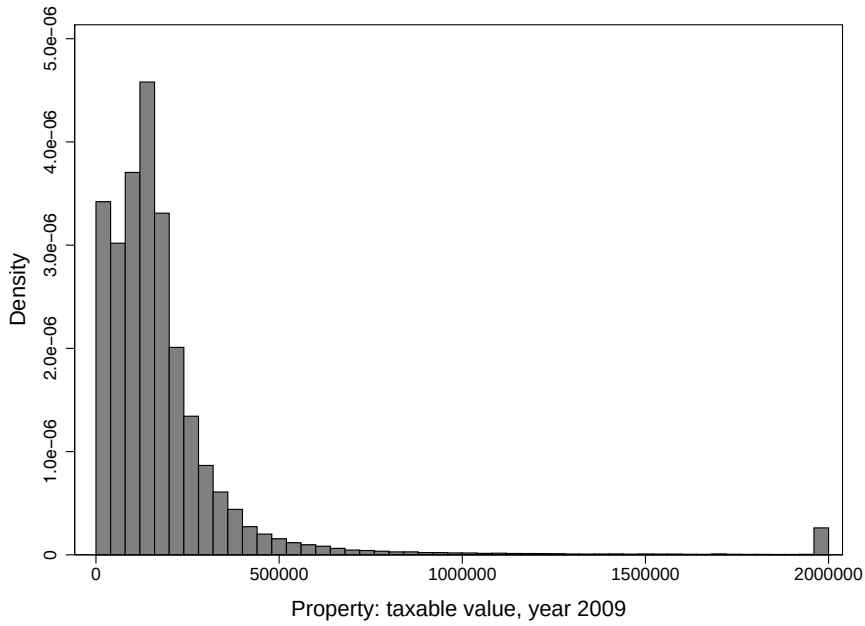

Fig. 1. Distribution of taxable values of property in the Durham, NC area in 2009

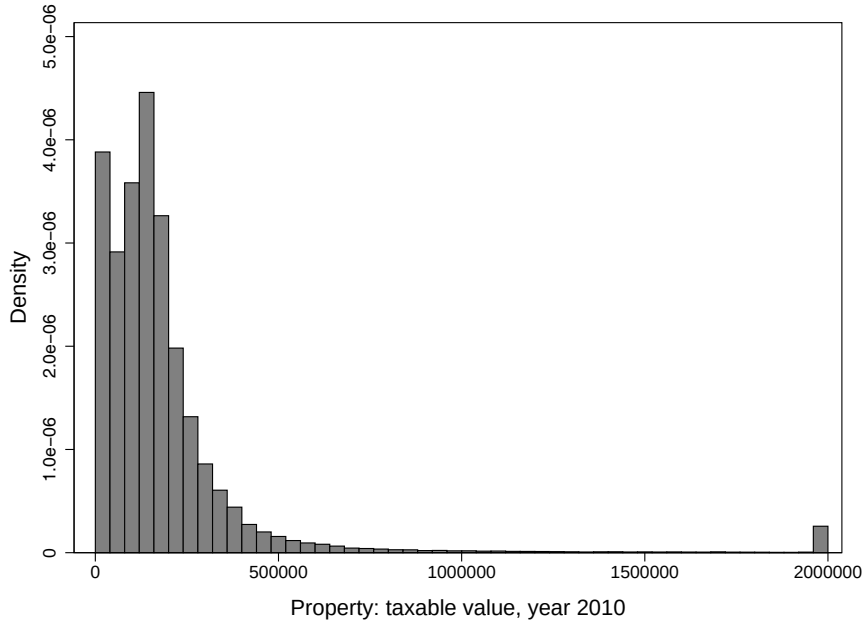

Fig. 2. Distribution of taxable values of property in the Durham, NC area in 2010

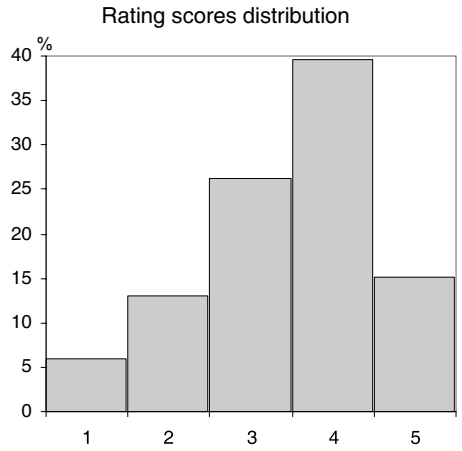

Fig. 3. Distribution of Yelp users' rating scores for restaurants

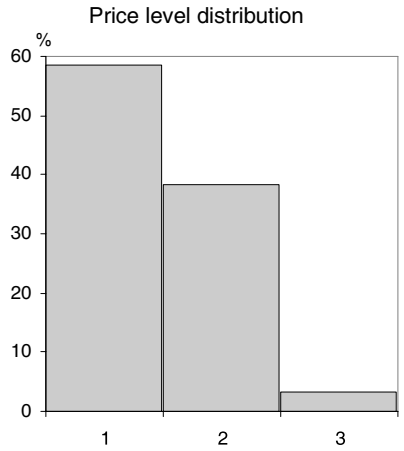

Fig. 4. Distribution of restaurants' price levels as reported on restaurants' Yelp pages

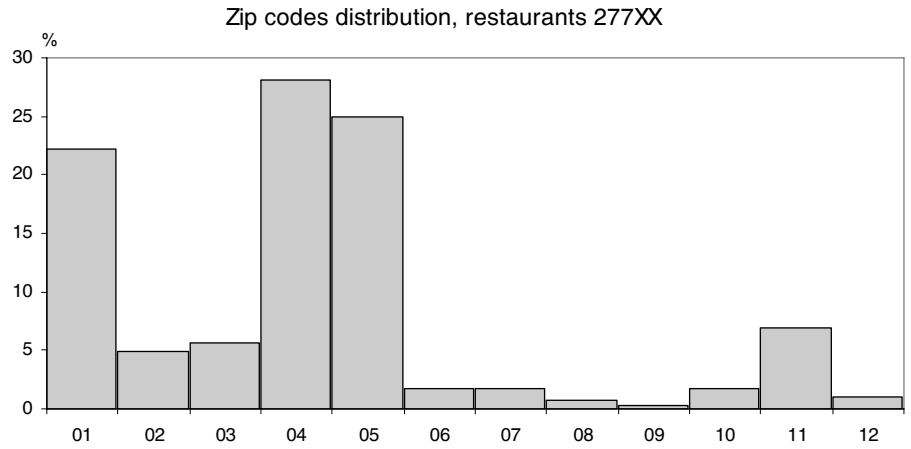

Fig. 5. Distribution of restaurants' zip codes

4. Empirical analysis based on «one-to-one matches» subsample

Table 3 presents summary statistics for those reviewers (and their reviews) who have unique matches in the property tax dataset, i.e. our «one-to-one match» subsample. In addition to the variables described above, we constructed a few more as follows. Based on the data for zip codes, we constructed a dummy variable that indicates whether or not the restaurant is in the city centre, as well as a dummy variable for whether or not the reviewer lives within the same zip code as the restaurant. Based on the reviewer's first name we evaluated the gender of the reviewer and constructed a dummy variable to indicate whether or not the reviewer is female (to be precise, has female name).

As we have already mentioned, estimation of utility parameters based only on «one-to-one match» data will likely suffer from «activity bias» due to oversampling of a selected group of individuals that submit reviews on Yelp (which corresponds to a familiar selection bias).

Table 3. Summary statistics for «one-to-one matches» subsample in combined dataset

Variable	Observations	Mean	Standard deviation	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 4 provides evidence of this selection in our data. Columns 1 and 2 provide estimates for probit regressions of the probability of giving a review for a certain restaurant for reviewers from the «one-to-one match» subsample without and with restaurant fixed effects. The results

Table 4. Probit regression estimates for probability of giving review, sample restricted to «one-to-one matches»

	Probability of giving a 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, 5 and 1%.

show that selection does take place on reviewer characteristics: people with more expensive properties (acting as a proxi for higher incomes) give more reviews, people give more reviews to restaurants within the same zip code area, and females gives fewer reviews. And again, some of these characteristics are not observable in the Yelp database, e.g. personal income and zip code of the area in which the reviewer lives.

Table 5 presents estimates of the utility parameters of the model presented above. As one can see, correction for «activity bias» mentioned above changes most of the estimated utility parameters.

Table 5. Estimates of utility parameters

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

Note. Bootstrapped standard errors are italicized.

5. Estimation under 2- and 3-anonymity

Finally, we check how estimates of our parameters would change if we impose 2-anonymity and 3-anonymity on individual identifiers in the Yelp dataset. In the case of 2-anonymity, for each of the 66 Yelp users from the «one-to-one match» subsample, we tried to find at least one alternative match in the property tax data. First, we tried to find at least one person with the same first name and same (or nearest if the same is not available) zip code with most of the rated restaurants, and different surname initial. Thus, we first try to «suppress» the surname initial. Out of these

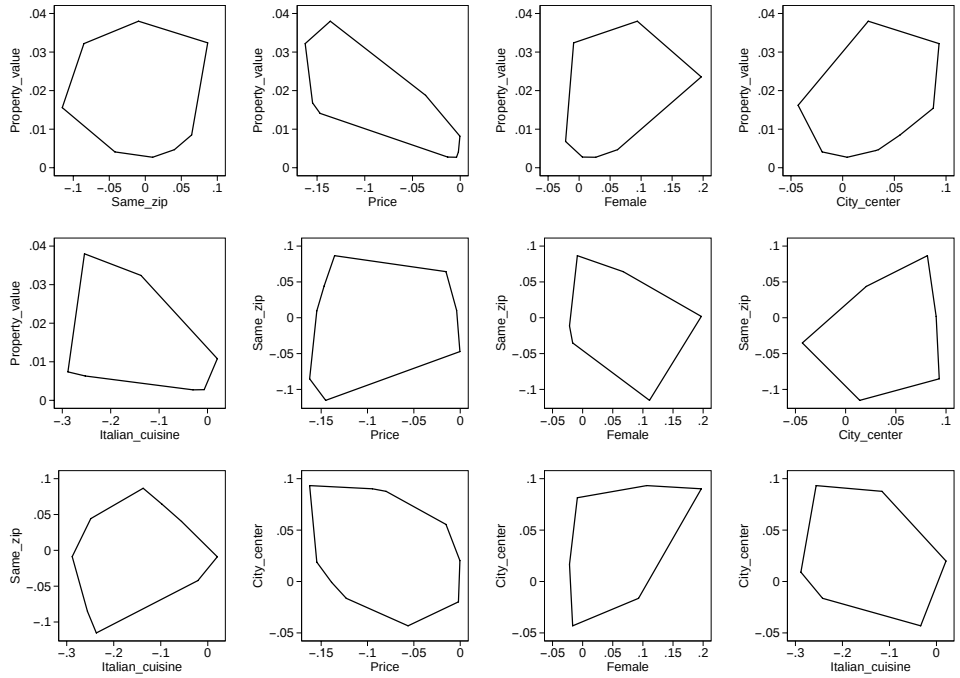

Fig. 6. Convex hulls of set of estimates of utility parameters in the case of 2-anonymity

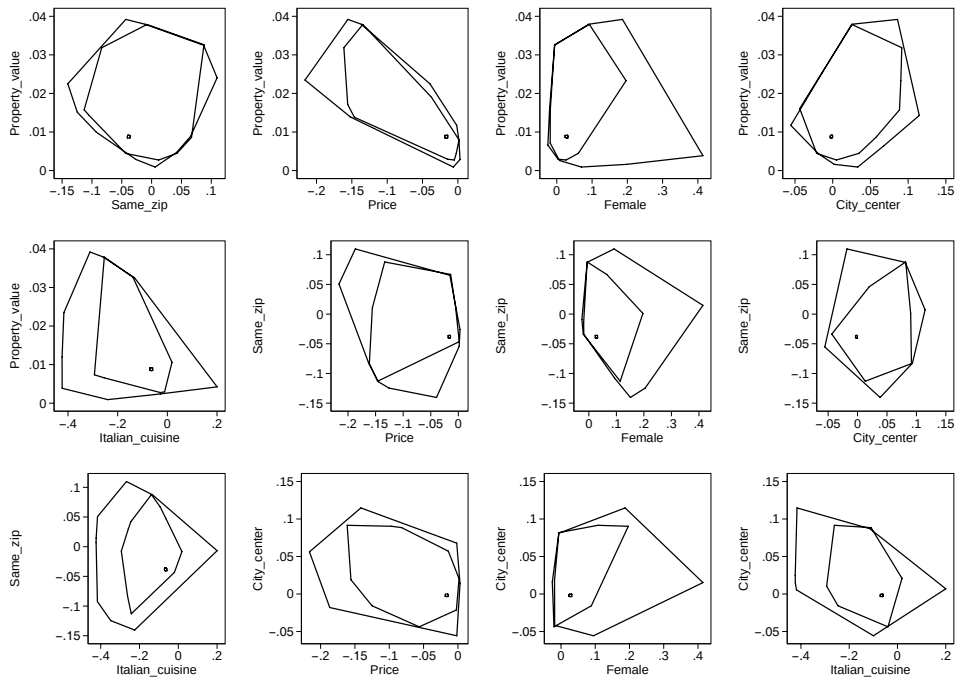

Fig. 7. Superimposed convex hulls of sets of estimates of utility parameters from the «one-to-one matches», and imposing 2-anonymity and 3-anonymity

66 users, 61 had this second match in tax data. For the remaining 5 users that do not have at least one other match in first name, we match them with people with the same surname initial. Then, for each individual in the Yelp «one-to-one match» subsample, we randomly pick a unique «suppressed match» in the property tax database for the respective individual and estimate the utility parameters for this «possibly mismatched» subsample. Figure 6 shows the convex hull of set of possible utility parameters on two-dimensional space of coefficients that arose from our particular random selection of the «possibly mismatched» subsamples. A similar process was followed for the case of 3-anonymity. Figure 7 shows the superimposed convex hulls of the sets of parameter estimates from the «one-to-one matches» subsamples, 2- and 3-anonymity imposed «possibly mismatched» subsamples. The largest convex hull in each case pertains to the imposition of 3-anonymity and the unique estimates of course pertain to the «one-to-one matches» subsample. As can be seen, more stringent anonymity restrictions result in larger sets of parameter estimates — thus leading to larger convex hulls. This illustrates the trade-off between the anonymity of data used in estimation and identification of parameters in our estimation problem.

6. Conclusion

This paper highlights the consequences for estimation using data from combined datasets that result from imposing stricter restrictions on data security. Given a restriction on anonymity from the data curator, this restriction can be satisfied by implementing *k*-anonymity for a «correct» choice of $k \in \mathbb{N}$. With stricter restrictions on anonymity however, larger *k*'s must be chosen. Our empirical example shows that as M_{ij} increases, the resulting set of parameter estimates get larger as the set of possible mismatched combined datasets get larger. Hence, a trade-off exists between the econometrician's goal of providing accurate estimates for the parameters of interest and the data curator's goal of preserving sensitive individual data from disclosure.

References

Komarova T., Nekipelov D., Yakovlev E. (2015). Estimation of treatment effects from combined data: Identification versus data security. In: *Economic Analysis of the Digital Economy*, eds. A. Goldfarb, S. M. Greenstein and C. E. Tucker, 279–308. University of Chicago Press.

Komarova T., Nekipelov D., Yakovlev E. (2017). Identification, data combination and the risk of disclosure. *Forthcoming in Quantitative Economics*.

Lambert D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics*, 9 (2), 313–331.

Samarati P., Sweeney L. (1998). Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression. Technical report, SRI International.

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

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

Wilson A. G., Graves T. L., Hamada M. S., Reese C. S. (2006). Advances in data combination, analysis and collection for system reliability assessment. *Statistical Science*, 21 (4), 514–531.

Received 31.08.2017; *accepted 22.09.2017.*