Estimation of Treatment Effects from Combined Data: Identification versus Data Security

Tatiana Komarova (LSE),

Denis Nekipelov (UC Berkeley),

Evgeny Yakovlev (UC Berkeley)

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

THE UNIVERSITY OF CALIFORNIA BERKELEY LET THERE BE LIGHT 1868

Data combination

- Data from multiple sources are routinely combined for business purposes
    - Insurance
    - Bank loans
    - Online ad targeting
    - TV network programming

Data combination

- Addressing common data problems
  - Sample Selection
  - Omitted variables
  - Missing data
  - Measurement errors

Tradeoff

- How much do we need to know about the individuals in combined data to be able to
  - Successfully combine the data
  - Identify the parameters of interest

- Unfortunately, there is a tradeoff between model identification and identity disclosure

Data security threats

1. Sensitive data are “anonymized” but can be retrieved publicly
   – Threat: re-identification of the entire data entry for at least some individuals

2. Sensitive data are never released but the estimated model is publicly observable (research project, policy implementation)
   – Threat: re-identification of sensitive information for at least some individuals

Model of interest

- We focus on estimation of treatment effects
- Treatment status is sensitive information and stored in separate “anonymized” dataset
  - Results of HIV test, credit score, etc.
- The effect of sensitive treatment is policy-relevant
  - Side effects of medications for HIV-positive individuals, access to emergency loans for individuals with bad credit

Disclosure

- Our approach: disclosure occurs when
  - Data combination leads to a successful match between two datasets
  - Or, the treatment status can be established with high precision for at least one individual

Results in this paper

- Formalize the definition of identification from combined data
  - Different from traditional approach to identification; idea is based on the limits of experiments
- Use two notions of disclosure: statistical partial disclosure and identity disclosure
- Provide and empirically relevant example and describe the impact of guarantee for bound on the disclosure risk on identification

Identity disclosure and privacy

- Consumer privacy is a much more complicated subject

- In this paper we are only concerned with analyzing the restrictions on identity disclosure

- Even if the data is secure, consumers may still be negative regarding their data collected and used

- We do not aim at welfare evaluations: consumer responses to privacy may complex (see Acquisti, Friedman, and Telang (2006), Miller and Tucker (2009) , Goldfarb and Tucker (2010))

Threats to disclosure in split data

- Two datasets: (1) “anonymized” dataset with research-relevant information; (2) dataset with demographic information

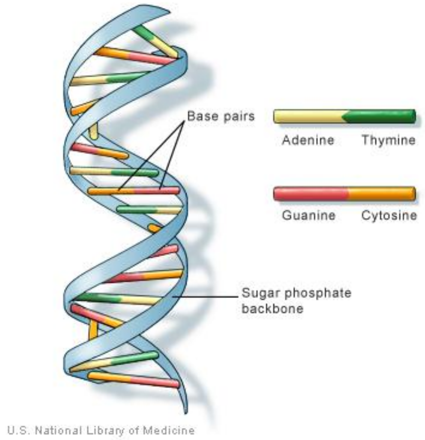
- Linkage privacy attack: data mining techniques may allow the “adversary” to associate entry from anonymized data file with demographic (identity) information

- Examples:
  - Sweeney (2002): identified the medical records of William Weld (MA governor) by linking voter registration records to “anonymized" Massachusetts Group Insurance Commission (GIC) medical encounter data
  - Narayanan and Shmatikov (2008) linked multiple consumers from Netflix prize dataset with imdb.com users

Disclosure risk in GWAS

- Genome-wide association studies (GWAS): relationship between genotype and factors of interest (health conditions, responsiveness to viral infections, etc.)

- Conclusions from minor allele frequencies



- Studies are based on analysis of ~100k minor allele frequencies

Disclosure risk in GWAS

- HapMap project sponsored by NIH reports the population-wide frequencies

- Then if we find out the allele sequence for a particular individual, we can compare this sequence with the one in a particular GWAS and the average in HapMap to determine whether that particular individual was in GWAS

- If GWAS was concentrating on a specific disease, we find very sensitive information regarding a particular individual.

- Demonstration of “optimal” adversary behavior in GWAS changed the standards of data release in the field

Disclosure and Measures of Disclosure Risk

- Types of disclosure:
  - Identity disclosure: entry *j* belongs to *Jane Doe*
  - Partial disclosure: One of the entries {*j,k,l*} belongs to *Jane Doe*, each entry {*j,k,l*} belongs to a cancer patient => *Jane Doe* is cancer patient

- Measuring disclosure risk:
  - We use pessimistic identity disclosure risk (Lambert (1993)): maximum probability of identity disclosure
  - Introduce the definition of partial disclosure as the p-value of the test for positive treatment status

Protection against linkage attacks

- *k*-anonymity (Samarati and Sweeney, 1998): there are at least *k* equally good links for each data entry
  - We use this notion in our paper
  - Can attain k-anonymity by suppressing data attributes
  - Downside: does not deliver protection from linkage with other databases

- Synthetic data approach
  - Extensive literature in statistics: Duncan and Lambert (1986), Duncan and Mukherjee (1991), Duncan and Pearson (1991), Fienberg (1994), Fienberg (2001), Abowd and Woodcock (2001)
  - Based on the use of dataset with variables (or entries)
  - In principle, can even release the entire dataset
  - Downside: harder to compute the measure of disclosure

Econometric model

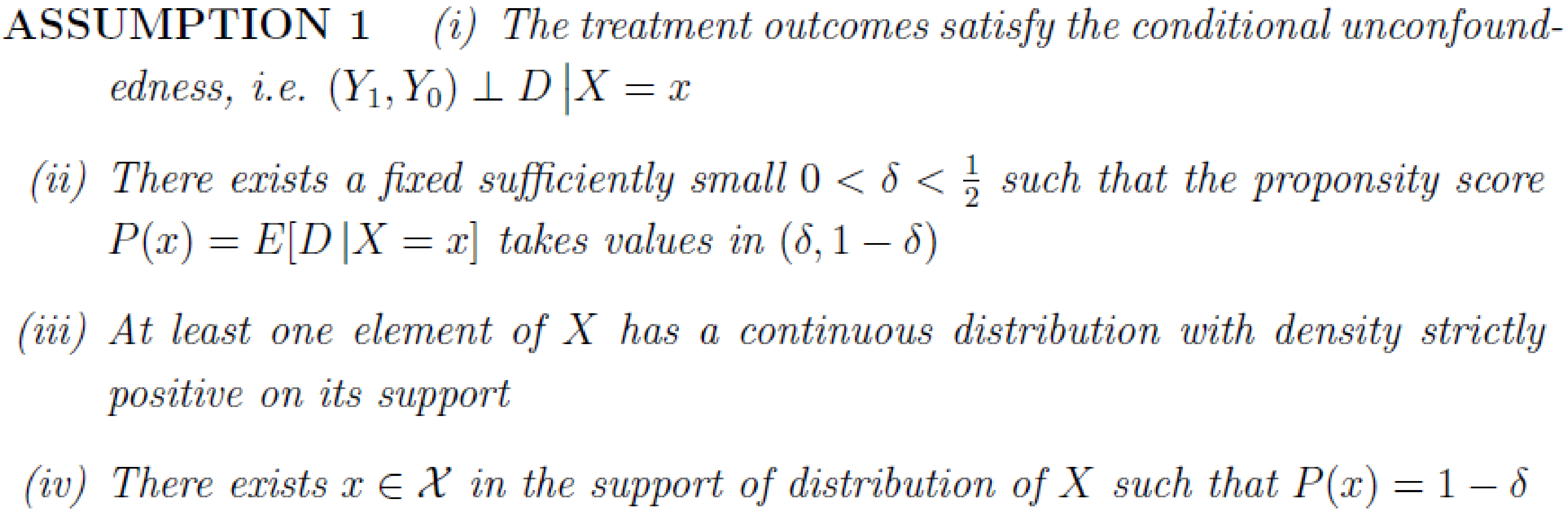
- Consider the standard treatment effect setup

$$Y = D\,Y_1 + (1 - D)\,Y_0$$

- Object of interest: ATE

$$t_{ATE} = E\left[Y_1 - Y_0\right]$$

Standard regularity conditions



ASSUMPTION 1 *(i) The treatment outcomes satisfy the conditional unconfoundedness, i.e.* $(Y_1, Y_0) \perp D \,|X = x$

(ii) There exists a fixed sufficiently small $0 < \delta < \frac{1}{2}$ *such that the proponsity score* $P(x) = E[D\,|X = x]$ *takes values in* $(\delta, 1 - \delta)$

(iii) At least one element of X *has a continuous distribution with density strictly positive on its support*

(iv) There exists $x \in \mathcal{X}$ *in the support of distribution of* X *such that* $P(x) = 1 - \delta$

Split samples

- Information regarding the treatment status, covariates and outcome is contained in different sources

- In our empirical application
  - Y are ratings that users on Yelp give businesses
  - X is the vector of consumer demographics
  - D is the indicator that an individual has visited a doctor
  - (Y, X) are never observed together with D

- This is very different from the situation with validation samples: there is not commonly observed variables

Data structure

- We assume that individuals in “treatment status” and “treatment outcome” samples overlap

- In addition to variables of interest, samples contain additional information that may be non-numeric (address, name, date of birth, etc.)
  - This information can be used for re-identification of individuals
  - We approach this by using the CS techniques used for combination of string data: consider string-valued random variables and define distance in the space of strings

Identification problem

- “Classical” approach to identification fails: on the population level there is only information on marginal distributions of Y and X
- If interested in the linear regression coefficient,

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

  can provide Frechet bounds

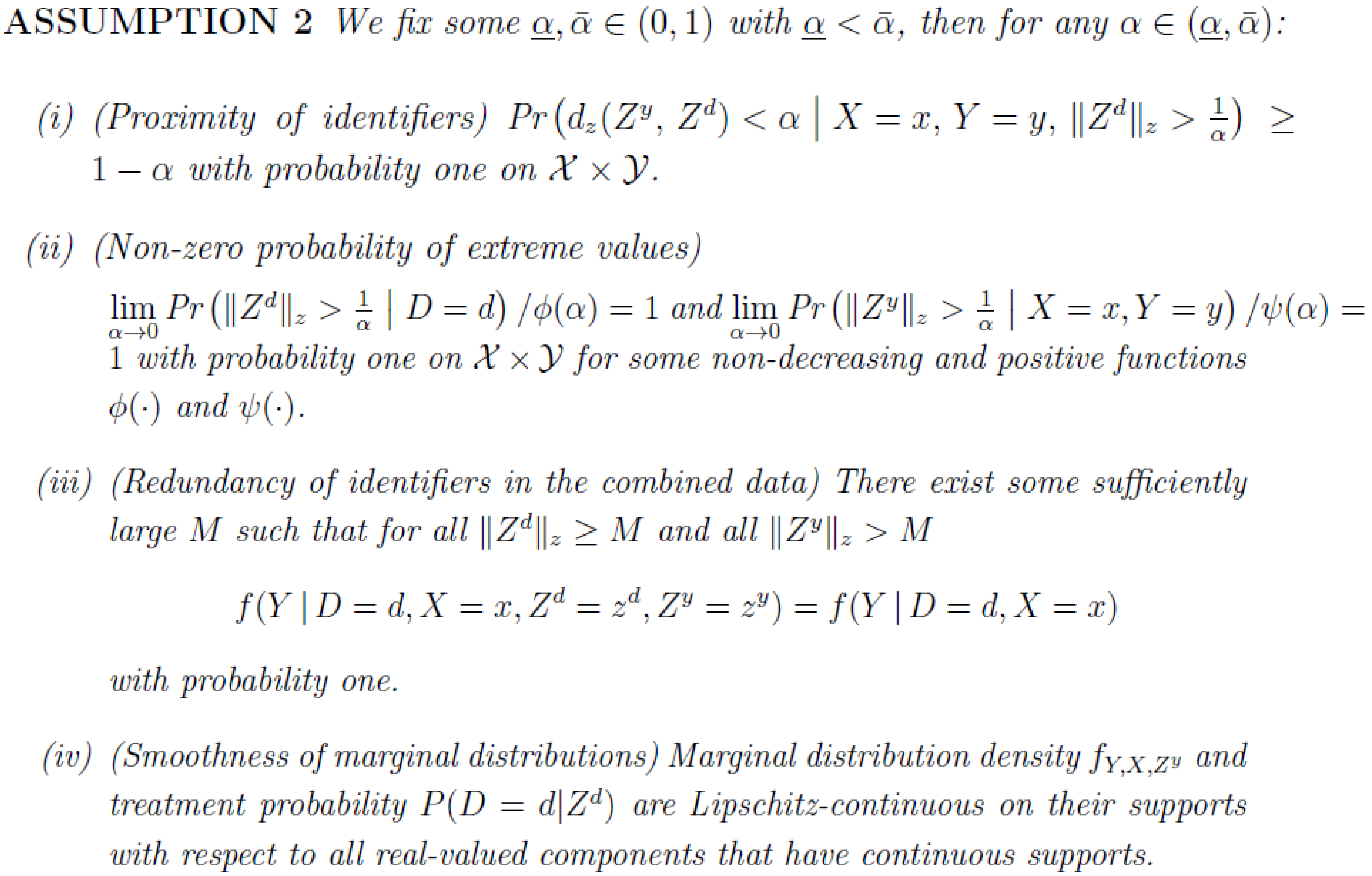
  $$-\sqrt{\frac{\mathrm{var}\,(Y)}{\mathrm{var}\,(X)}} \leq \beta \leq \sqrt{\frac{\mathrm{var}\,(Y)}{\mathrm{var}\,(X)}}$$

- This does not mean that parameter of interest cannot be recovered!

Combining finite samples of data

- Real datasets consist of numeric and string information
- For instance, we can have name, location, age and other variables
- Available numeric and string information can be combined to provide matches between entries in two databases
- Note: this is an intrinsically finite sample procedure

Data combination via constructed identifiers



ASSUMPTION 2 *We fix some* $\underline{\alpha}, \bar{\alpha} \in (0,1)$ *with* $\underline{\alpha} < \bar{\alpha}$*, then for any* $\alpha \in (\underline{\alpha}, \bar{\alpha})$*:*

(i) (Proximity of identifiers) $Pr\left(d_z(Z^y, Z^d) < \alpha \,\middle|\, X = x, Y = y, \|Z^d\|_z > \frac{1}{\alpha}\right) \geq 1 - \alpha$ *with probability one on* $\mathcal{X} \times \mathcal{Y}$.

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

$\lim_{\alpha \to 0} Pr\left(\|Z^d\|_z > \frac{1}{\alpha} \,\middle|\, D = d\right) / \phi(\alpha) = 1$ *and* $\lim_{\alpha \to 0} Pr\left(\|Z^y\|_z > \frac{1}{\alpha} \,\middle|\, X = x, Y = y\right) / \psi(\alpha) = 1$ *with probability one on* $\mathcal{X} \times \mathcal{Y}$ *for some non-decreasing and positive functions* $\phi(\cdot)$ *and* $\psi(\cdot)$.

(iii) (Redundancy of identifiers in the combined data) There exist some sufficiently large M *such that for all* $\|Z^d\|_z \geq M$ *and all* $\|Z^y\|_z > M$

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

with probability one.

(iv) (Smoothness of marginal distributions) Marginal distribution density f_{Y,X,Z^y} *and treatment probability* $P(D = d | Z^d)$ *are Lipschitz-continuous on their supports with respect to all real-valued components that have continuous supports.*

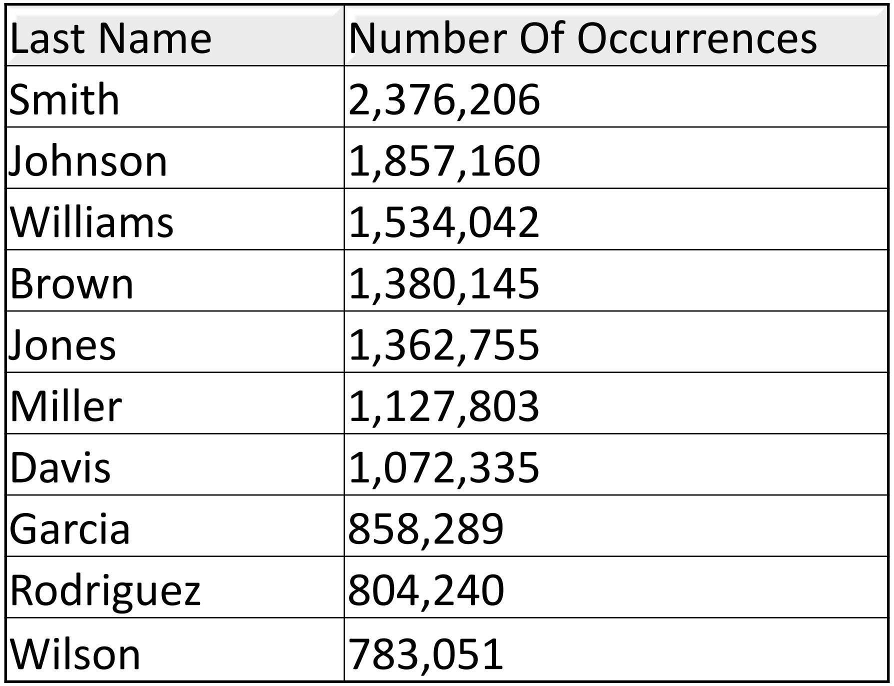
Structure of identifiers

- First, we require constructed identifiers to work better, the more rare value they take

- In our empirical example, we use user name and location in one database and actual name and location in the second database.

- If observe “Denis Nekipelov” in two databases (e.g. from Durham county, NC), the probability that entries belong to the same individual is higher than if we observe “John Smith” in two databases

Structure of identifiers



Last Name	Number Of Occurrences
Smith	2,376,206
Johnson	1,857,160
Williams	1,534,042
Brown	1,380,145
Jones	1,362,755
Miller	1,127,803
Davis	1,072,335
Garcia	858,289
Rodriguez	804,240
Wilson	783,051

* Source: US Census, 2000

Structure of identifiers

- Assumption 3 (iii): for correct matches, identity of an individual is just a label
- Once we match the individual in Yelp and property tax data, name becomes obsolete
- Z^x and Z^y do not add any more information to the model
- All “model-relevant” information absorbed by covariates. E.g. learn that Tatiana Komarova is female name with Russian origin, it only servea as label (for analysis can replace it by numeric label with no information loss)

Where do identifiers come from?

- There is an extensive literature on record linkage in the CS

- Record linkage tasks routinely arise in database management (large portion of business for)

- The idea is to define a “hybrid” similarity metric between the entries in two databases and then label each pair as
  - Match if the metric is sufficiently small
  - Non-match if the metric is sufficiently large
  - Uncertain otherwise (we ignore this)

Where do identifiers come from?

- Probabilistic record linkage:
  - Very similar to hypothesis tests: fixing the probabilities of incorrect match and incorrect non-match, minimize the proportion of “uncertain” observations
- Literature dates back to Kennedy, Axfold, James (1959) and Fellegi and Sunter (1969)
- Biggest problem: measuring similarity between string variables, e.g. Denis Nekipelov vs. Nekipelov Denis vs. Dennis Nekipelov

Where do identifiers come from?

- Commonly used distance measures for strings
  - Edit distance: number of edits (term replacements and deletions) required to turn one string into another
  - Jaro and Winkler’s measure: based on the length of two strings, number of common characters and number of required transpositions in common character blocks
  - TF/IDF: based on the term frequency (number of times the term is observed in the document) and inverse document frequency (number of documents containing the term)

Identification Strategy

- For each finite sample size, identify the set of matched observations

- The set of matched observations will corresponding approximate joint distribution

- Then we can get the parameter of interest from that distribution

- Identification will be the property of parameter sequence that we construct by increasing the size of split samples

Identification methodology

- We use decision rule

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

- Then our analysis will be based on distribution of matches identified with this decision rule
- This produces a sequence of densities indexed by *N*
- Identification results require verification of parameter vector that satisfies the analyzed conditional moment

Model identification: matched data

- Match indicator

$$m_{ij} = \begin{cases} 1, & \text{if } z_i^x \text{ and } z_j^y \text{ are characteristics of the same individual,} \\ 0, & \text{otherwise,} \end{cases}$$

- Probability of successful match

$$\pi_{ij}^N(x) = Pr\left(m_{ij} = 1 \mid X_i = x, |Z_i^x| > \frac{1}{\alpha_N}, |Z_j^y - Z_i^x| < \alpha_N \right)$$

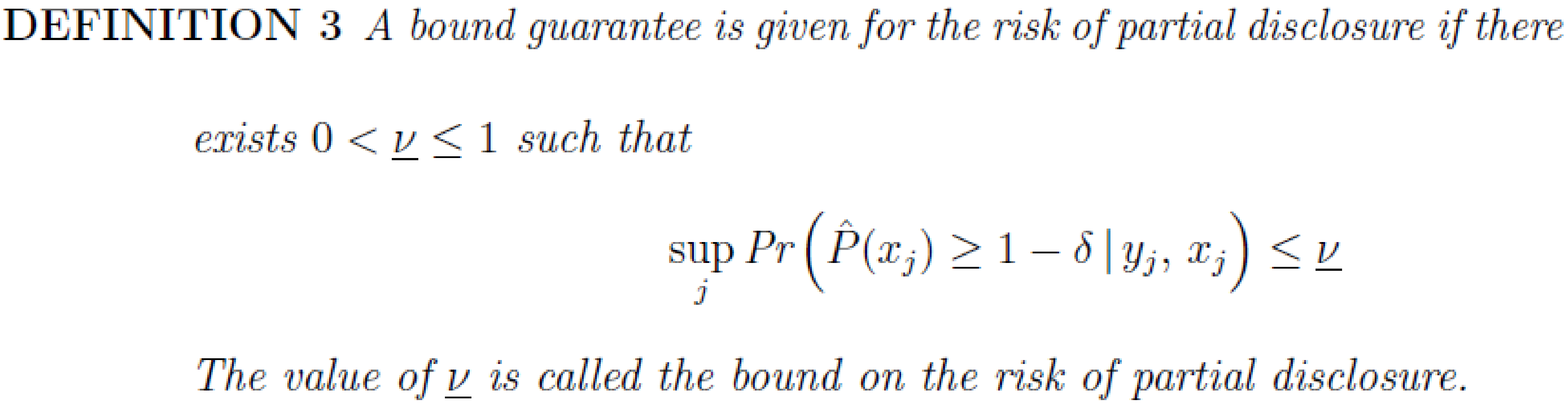
Identification results

1. The ATE and the propensity score are identified if we only consider “tail” observations, if they are matched correctly
2. There exist threshold decision rules for which the proportion of incorrect matches approaches zero in large samples
3. The ATE and the propensity score are point identified

Risk of partial disclosure

- For some individual we can guess the (positive) treatment status with very high confidence

- This is very undesirable, we introduce the notion of partial disclosure: *Can we test for positive treatment status?*

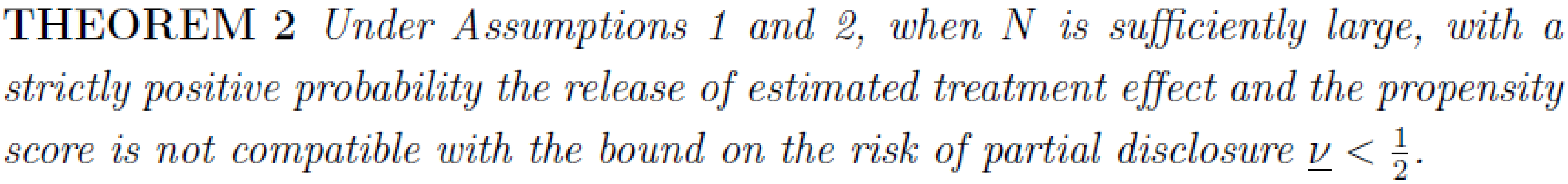


DEFINITION 3 *A bound guarantee is given for the risk of partial disclosure if there exists $0 < \underline{\nu} \leq 1$ such that*

$$\sup_{j} Pr\left(\hat{P}(x_j) \geq 1 - \delta \,|\, y_j,\, x_j\right) \leq \underline{\nu}$$

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

Threat of partial disclosure

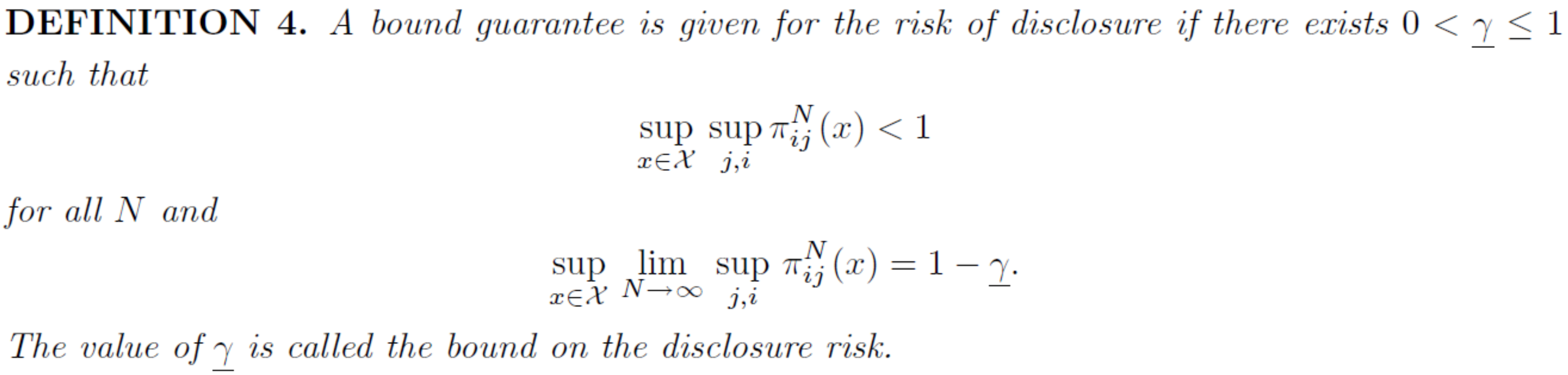


THEOREM 2 *Under Assumptions 1 and 2, when N is sufficiently large, with a strictly positive probability the release of estimated treatment effect and the propensity score is not compatible with the bound on the risk of partial disclosure $\underline{\nu} < \frac{1}{2}$.*

Identity disclosure risk

- If the probability of successful match is very high – disclosure has occurred (NB: datasets are finite). This needs to hold for every value of conditioning variable (vs. almost every)

- Need to deliver guarantee of no disclosure with high probability



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*

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

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

Identity disclosure and identification

1. Model identification is incompatible with any non-trivial bound on the risk of identity disclosure (at least one individual can be re-identified from the data)

2. When the bound on the disclosure risk is imposed the model is not identified

3. Consider partial identification: identified set is compatible with an interval for the ATE and the set of propensity scores

Disclosure without intent

- In general, there is no need to release the data to generate the risk of disclosure

- Consumers will be the most vulnerable if the model is used for some “optimal” allocation

  - Re-assignment of some student to a different school may imply that this is minority student

  - Showing the online ad to a specific individual means that the model predicts that the ad is relevant

Online micro-targeted ads

- Korolova (2010) provides example of a privacy attack using Facebook targeting
- Advertisers do not have direct access to user IDs (cannot directly ask for ad being shown to Bill Gates)
- Ads can be targeted for some very large set of user attributes
- Some users can be “singled out”
- This can be used to infer the values of “private” use attributes

Disclosure from click prediction models

- The ad is shown if the system estimates that it is relevant (the probability of click is high)

- In combination with ad targeting, can see how estimated “clickability” varies over some very small group of consumers: this may lead to partial disclosure or even identity disclosure

- Ad display allows one to isolate the users with high potential clickability. Knowing who is most likely to click on
  - Treatment for mesothelioma
  - Cannabis
  - IRS audit

  reveals a lot of information!

Disclosure from observed policies

- Disclosure can occur whenever the policy is driven by micro-data:
  - Retail network decides to distribute particular coupons in a particular location
  - Hospital network makes decision to transfer a specialty doctor to a particular location
  - TSA strengthens security in a particular airport on a specific day

Empirical application

- Illustrates that many online datasets are vulnerable to linkage attacks

- Demonstrates how data combination can be used to account for selection bias

- Show how data can be protected from linkage attacks via honoring *k*-anonymity

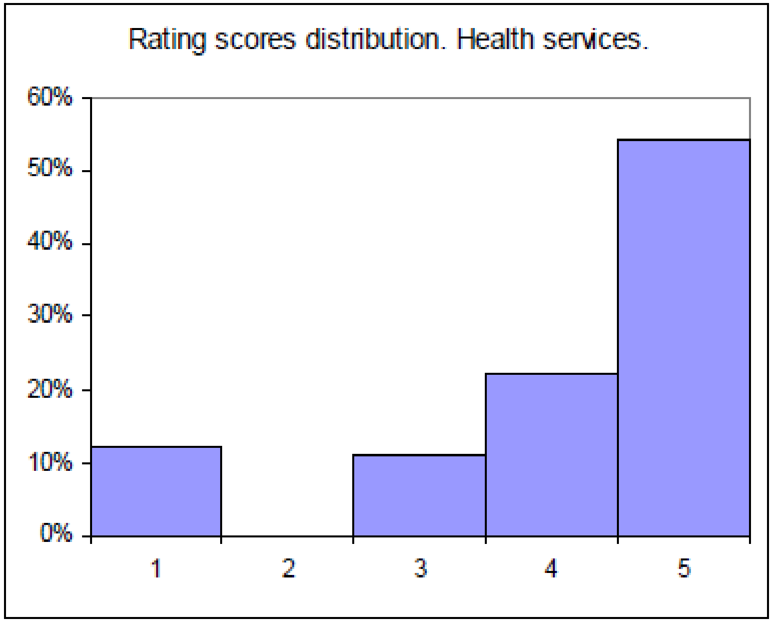
- Explicitly demonstrate the tradeoff between data protection and identification

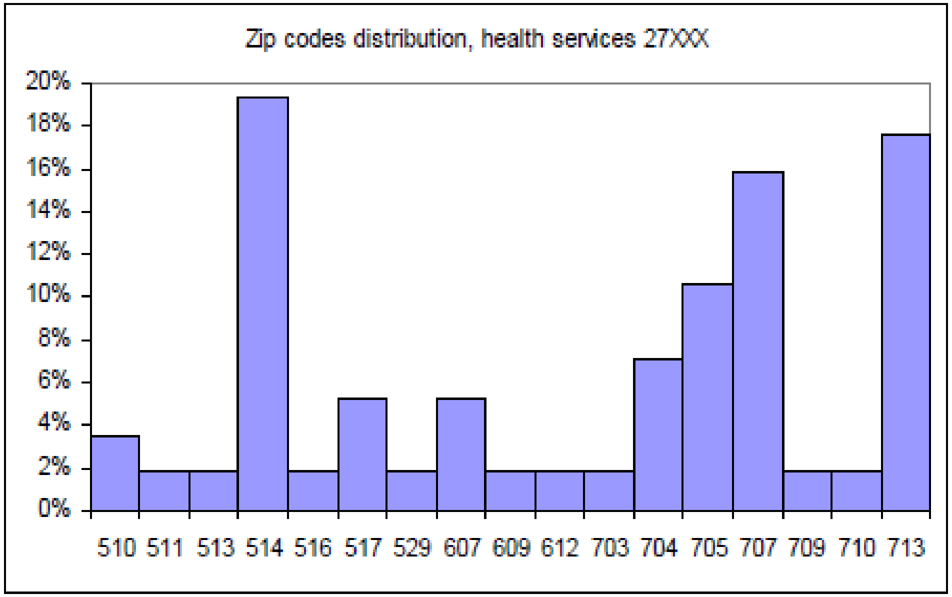
Do doctors make people happier?

- We want to study the effect of a visit to a doctor on individual ratings of businesses on Yelp.com
- How does a visit to medical facility change the ratings (comparing before and after)?
- Use the fact that an individual rated a medical facility as an indicator of a visit to a doctor

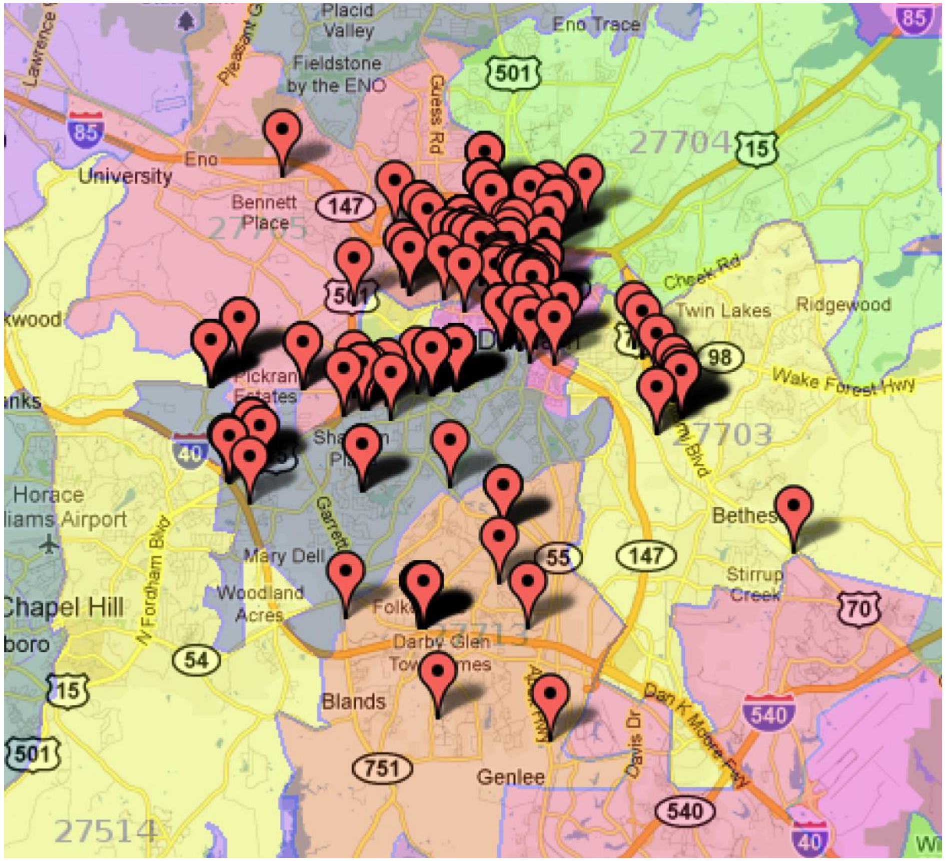
Data collection from Yelp.com

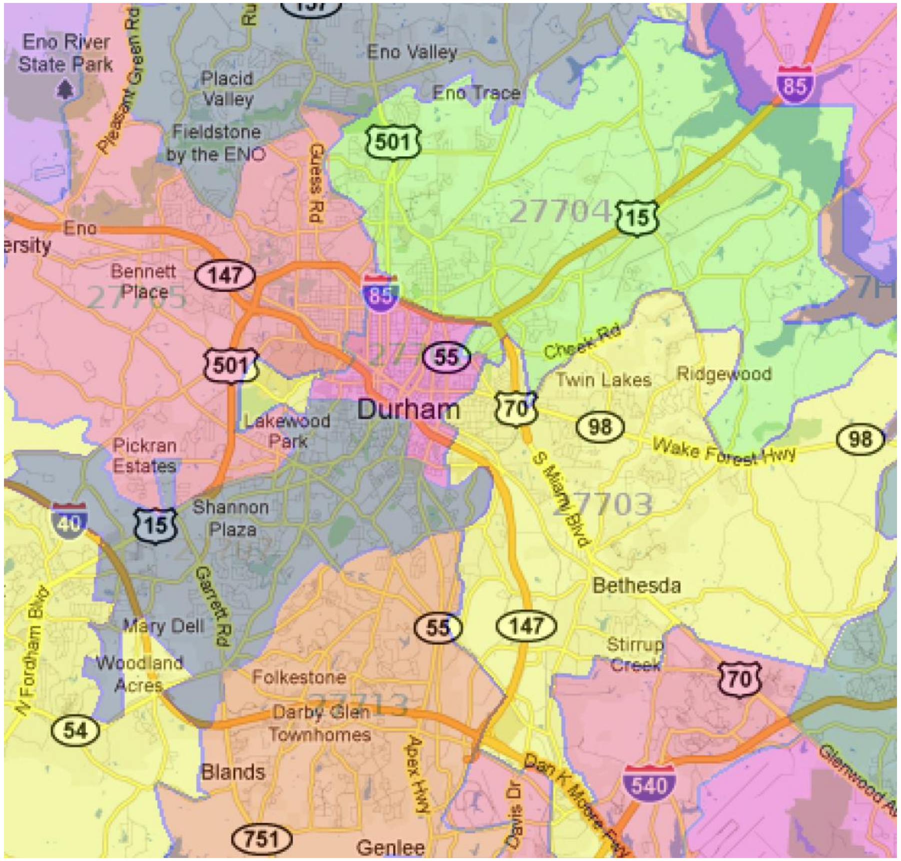
- Uploaded individual reviews and business information



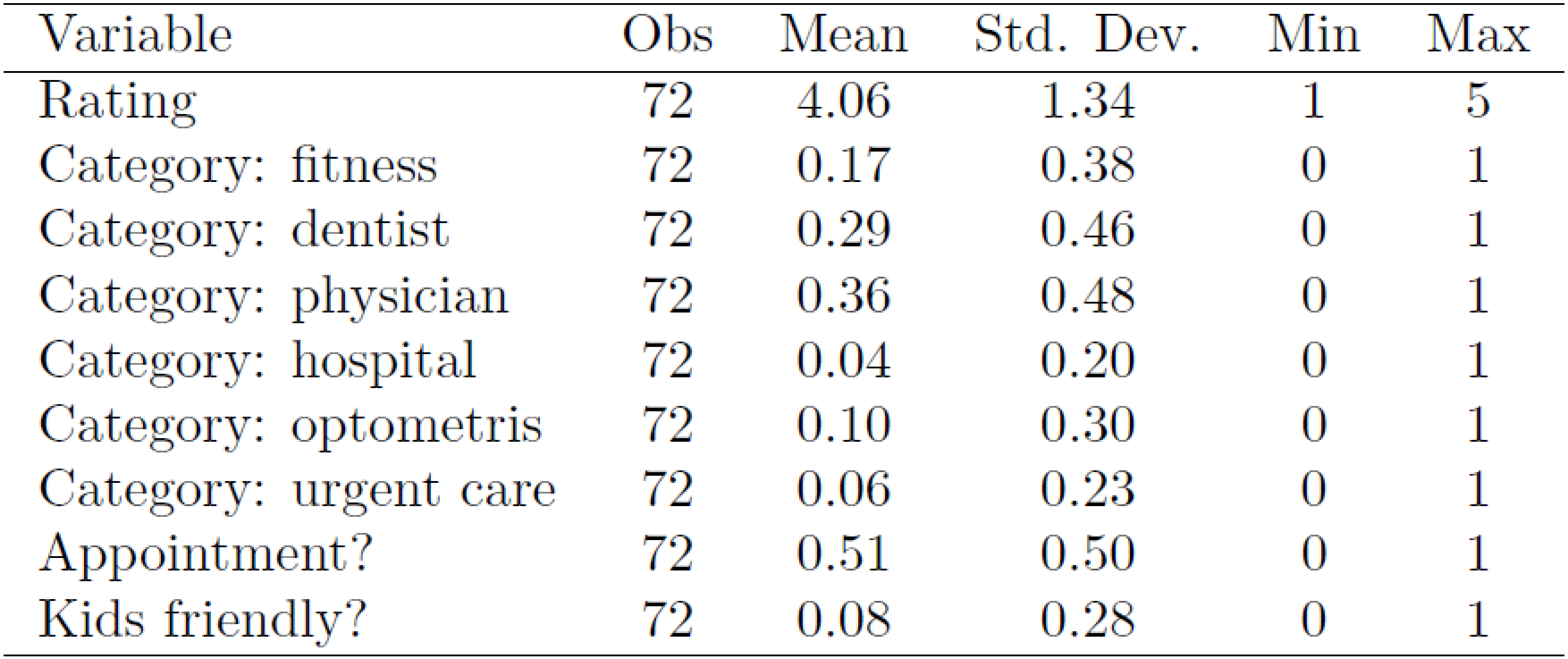


Coverage of Durham Healthcare Businesses





Basic statistics



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

What is problematic in these data?

- Yelp.com data cannot be used to predict the score that a “representative” user would assign

- Yelp data will over-sample:
  - People who use a particular business more frequently
  - People who had an unusual experience (e.g. amazing vs awful)

- To correct for “activity bias” we collect more data

Additional dataset: Durham county property tax bills

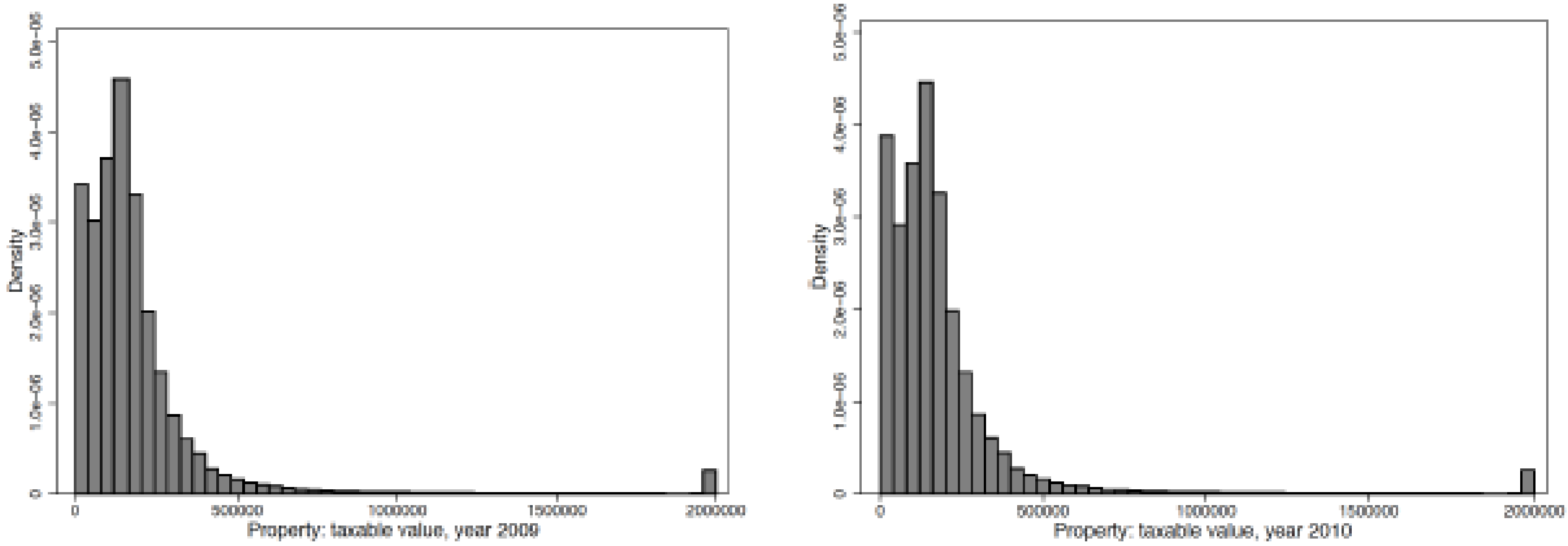
- The “activity bias” is mainly associated with individual income and distance to the restaurant
- We use the additional dataset of property tax bills:
  - The value of the house will be a proxi for income
  - The zip code will be a proxi for mutual location
  - In principle, can be even more precise in terms of distance

Durham county tax data collection

- Use the indexing of bills by parcel numbers
- Loop over all parcel numbers and upload the data for each property tax bill
- In the data can see residential and commercial properties
- Collect information on the property owner and precise address

Durham county property values

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

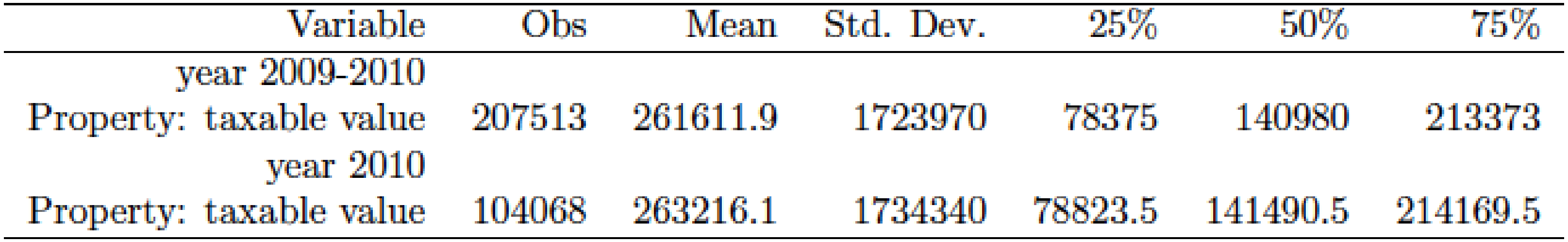


(a) (b)

Largest Property Owners

COUNTY OF DURHAM	750 STADIUM DR	128460	$156,300,000
SOUTHPOINT MALL LLC	7901 FAYETTEVILLE RD	149664	$169,000,000
DUKE UNIVERSITY	2100 DUKE UNIVERSITY	108792	$278,200,000

Durham county property tax values

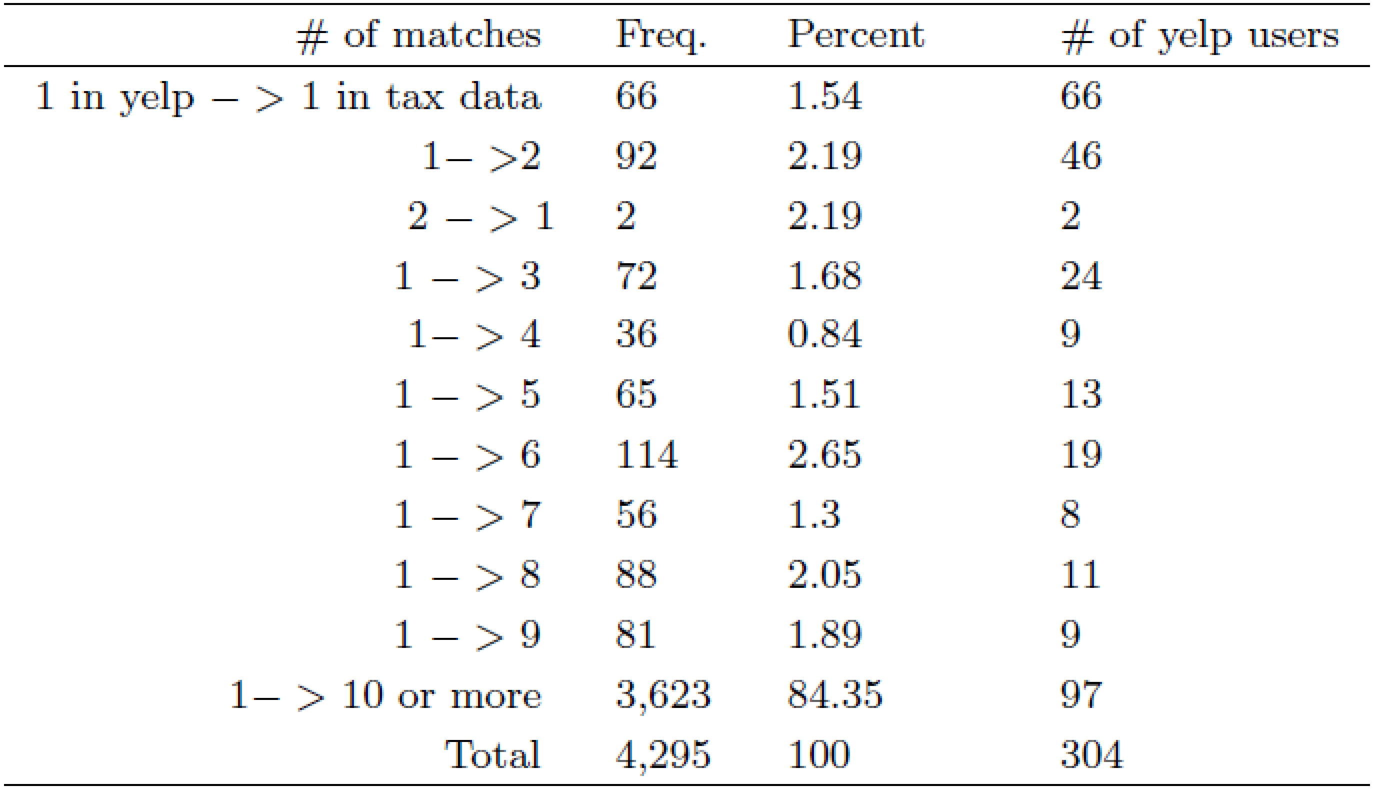


Variable	Obs	Mean	Std. Dev.	25%	50%	75%
year 2009-2010						
Property: taxable value	207513	261611.9	1723970	78375	140980	213373
year 2010						
Property: taxable value	104068	263216.1	1734340	78823.5	141490.5	214169.5

How the data are matched

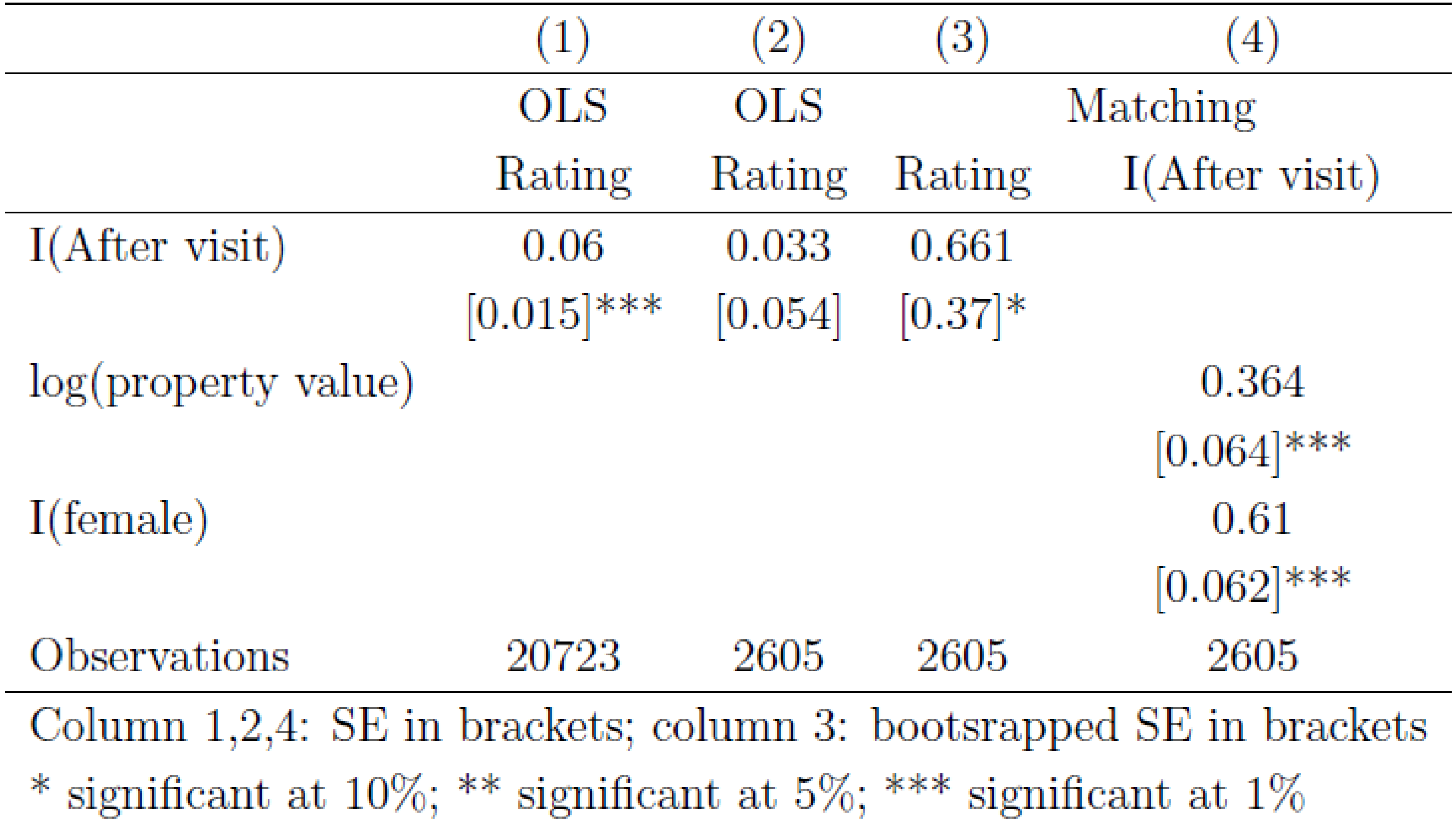
- Use the names of taxpayers and user names of Yelp.com reviewers
- Use edit distance between the string identifiers
- Additionally use the frequency ranks:
  - More likely to visit businesses in the same zip code
  - Used the square footage of the property to find whether has kids, thus would prefer a kid-friendly business
- Edit distance turns out to be most important

Combined dataset: number of observations within distance threshold



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

Treatment effects

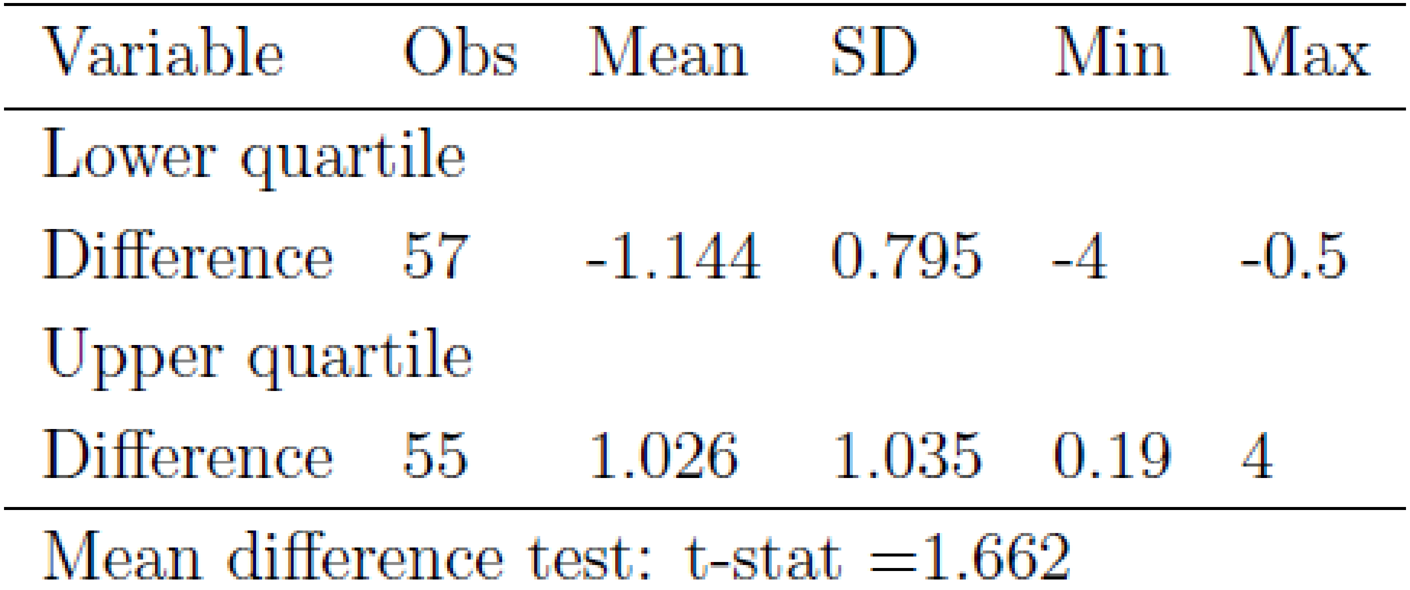


	(1)	(2)	(3)	(4)
	OLS	OLS	Matching	
	Rating	Rating	Rating	I(After visit)
I(After visit)	0.06	0.033	0.661	
	[0.015]***	[0.054]	[0.37]*	
log(property value)				0.364
				[0.064]***
I(female)				0.61
				[0.062]***
Observations	20723	2605	2605	2605

Column 1,2,4: SE in brackets; column 3: bootsrapped SE in brackets

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

Treatment quantiles



Variable	Obs	Mean	SD	Min	Max
Lower quartile					
Difference	57	-1.144	0.795	-4	-0.5
Upper quartile					
Difference	55	1.026	1.035	0.19	4

Mean difference test: t-stat =1.662

Disclosure risk guarantees

- *k*-anonymity assures that each entry will have at least *k* matches

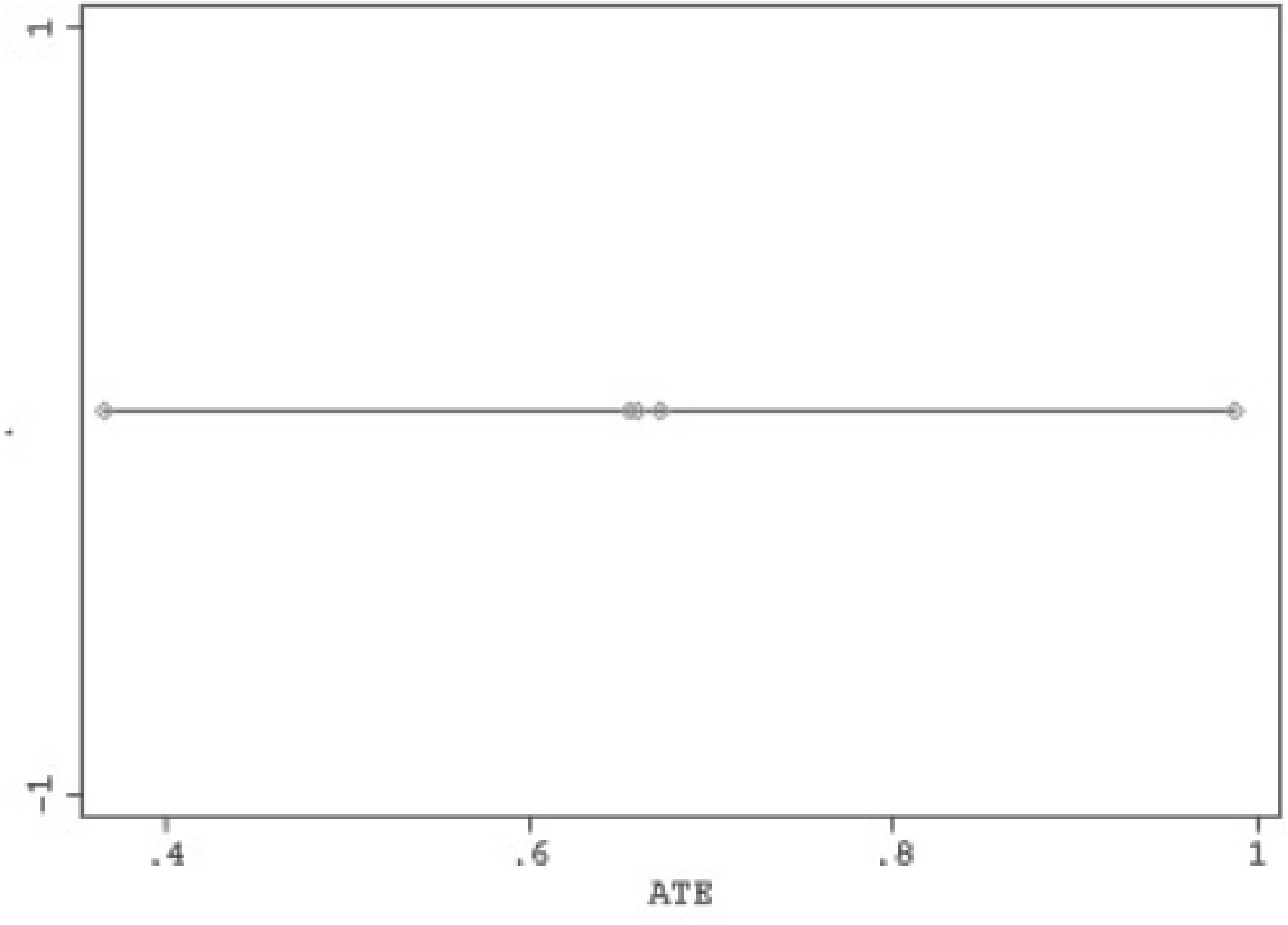
- Main source of matches: edit distance between strings

- We suppress symbols until each observation will have at least k counterparts

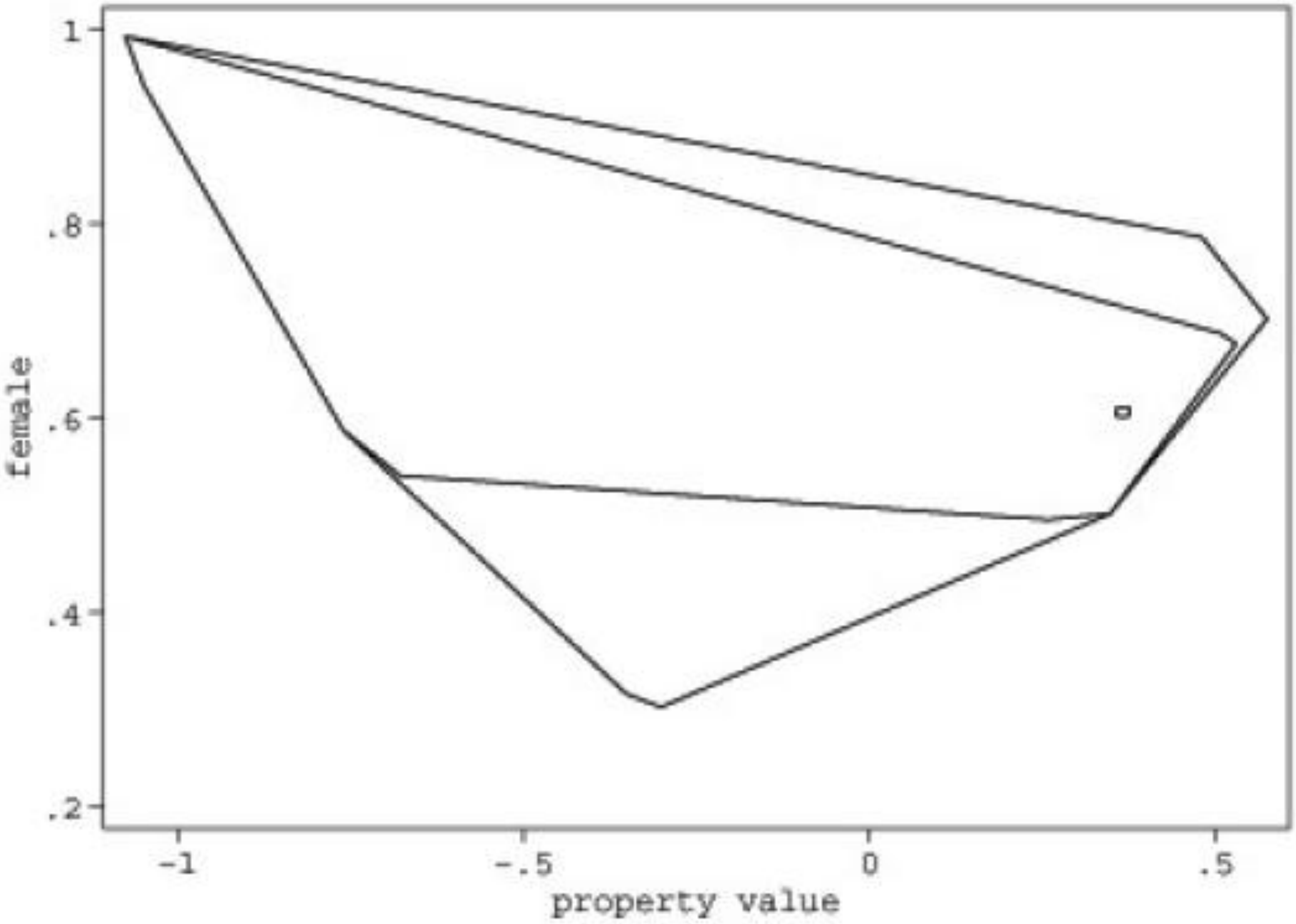
- d(Dennis,Denis)=1>d(Denis,Denis)

- Transform: Denis =>Den*
  d(Denis,Den*)=d(Dennis,Den*)

Loss of point identification: 2,3-anonymity



Loss of point identification: 2,3-anonymity and projections of propensity score



Conclusion

- Data mining techniques and econometric analysis can be used identify models from combined data under mild distributional assumptions

- Point identification from combined data is incompatible with restrictions on the risk of disclosure

- Without restrictions on the disclosure risk, confidential information regarding some individuals can be learnt from the estimated model even if the dataset is not released