UNIVERSITY OF Southampton

Southampton Statistical Sciences Research Institute

BLUE ETS

Assessing the Disclosure Risk of Perturbed Enterprise Data

Natalie Shlomo
Southampton Statistical Sciences Research Institute
University of Southampton
N.Shlomo@soton.ac.uk

EUROPEAN COMMISSION
European Research Area
SEVENTH FRAMEWORK PROGRAMME
Funded under Socio-economic Sciences & Humanities

Topics Covered

- Introduction and motivation
- Theory of disclosure risk assessment for identity disclosure
  - Probabilistic modelling extended for misclassification
  - Probabilistic record linkage – linking the frameworks
- Disclosure risk assessment for attribute disclosure of enterprise data
- Discussion

Introduction

- EU 7th Framework funded Blue-ETS project deals with the access and release of enterprise microdata
- Enterprise microdata rarely released as PUF but some agencies release highly perturbed (synthetic) datasets
- How to assess disclosure risk for perturbed enterprise microdata?

Introduction

Types of disclosure risks:

- Identity Disclosure – relevant for microdata from social surveys with small sample fractions
  - Disclosure risk scenario: ‘intruder’ attack on microdata through linking to available public data sources
  - Linkage via identifying key variables common to both sources, eg. gender, age, region, ethnicity
  - Need to take into account protection afforded by the sampling
  - Disclosure risk measured through the notion of population uniqueness

Introduction

Types of disclosure risks (cont):

- Attribute Disclosure – relevant for microdata from business surveys and whole population counts
  - Disclosure risk scenario: ‘intruder’ attack on microdata via the sensitive variables which may be publically available
  - Microdata treated as a census

Introduction

- For identity disclosure, need to quantify the risk of identification
- Probabilistic models based on population uniqueness on set of identifying key variables
- Population counts in contingency table spanned by key variables unknown
- Distribution assumptions to draw inference from the sample for estimating population parameters
- Take into account misclassification/perturbation

Introduction

- Risk assessment for perturbative methods typically based on probabilistic record linkage
  - Conservative assessment of risk of identification
  - Assumes that intruder has access to original dataset and does not take into account protection afforded by sampling
- Fit probabilistic record linkage into the probabilistic modelling framework for categorical matching variables
- Show that probabilistic record linkage can be used to assess attribute disclosure

Disclosure Risk Assessment

Probabilistic Modelling

- Let $\mathrm{f} = \{f_k\}$ denote a q-way frequency table $k = (k_1, \ldots, k_q)$ which is a sample from a population table $\mathrm{F} = \{F_k\}$ where F_k indicates a cell population count and f_k sample count in cell k

- Disclosure risk measure:

$$\tau_1 = \sum_k I(f_k = 1, F_k = 1) \qquad \tau_2 = \sum_k I(f_k = 1)\frac{1}{F_k}$$

- For unknown population counts, estimate from the conditional distribution of $F_k \mid f_k$

$$\hat{\tau}_1 = \sum_k I(f_k = 1)\hat{P}(F_k = 1 \mid f_k = 1) \qquad \hat{\tau}_2 = \sum_k I(f_k = 1)\hat{E}(\frac{1}{F_k} \mid f_k = 1)$$

Disclosure Risk Assessment

- Natural assumption: $F_k \sim Poisson(\lambda_k)$

  Bernoulli sampling: $f_k \mid F_k \sim Bin(F_k, \pi_k)$

  π_k is the sampling fraction in cell *k*

  It follows that: $f_k \sim Poisson(\pi_k \lambda_k)$ and

$$F_k \mid f_k \sim Poisson(\lambda_k (1 - \pi_k))$$

  where $F_k \mid f_k$ are conditionally independent

Disclosure Risk Assessment

- Skinner and Holmes, 1998, Elamir and Skinner, 2006 use log linear models to estimate parameters $\{\lambda_k\}$
- Sample frequencies f_k are independent Poisson distributed with a mean of $\mu_k = \pi_k \lambda_k$
- Log-linear model for estimating $\{\mu_k\}$ expressed as:

$$\log(\mu_k) = \mathbf{x}'_k \beta$$

where $\mathbf{x}$ design matrix of key variables and their interactions

- MLE's calculated by solving score function:

$$\sum_k [f_k - \exp(\mathrm{x}'_k \beta)]\mathrm{x}_k = 0$$

Disclosure Risk Assessment

- Fitted values calculated by: $\hat{u}_k = \exp(\mathbf{x}'_k \hat{\beta})$ and $\hat{\lambda}_k = \dfrac{\hat{u}_k}{\pi_k}$
- Individual risk measures estimated by:

$$\hat{P}(F_k = 1 | f_k = 1) = \exp(-\hat{\lambda}_k (1 - \pi_k))$$

$$\hat{E}(\frac{1}{F_k} | f_k = 1) = [1 - \exp(-\hat{\lambda}_k (1 - \pi_k))]/[\hat{\lambda}_k (1 - \pi_k)]$$

- Skinner and Shlomo (2009) develop goodness of fit criteria which minimize the bias of disclosure risk estimates, for example, for τ_1

$$\hat{B}_1 = \sum_k \hat{\lambda}_k \exp(-\hat{\lambda}_k)(1 - \pi_k)\{(f_k - \hat{\mu}_k) + (1 - \pi_k)[(f_k - \hat{\mu}_k)^2 - f_k]/(2\pi_k)\}$$

Disclosure Risk Assessment

- Criteria related to tests for over and under-dispersion:
  - over-fitting - sample marginal counts produce too many random zeros, leading to expected cell counts too high for non-zero cells and under-estimation of risk
  - under-fitting - sample marginal counts don't take into account structural zeros, leading to expected cell counts too low for non-zero cells and over-estimation of risk

- Criteria selects the model using a forward search algorithm which minimizes $\hat{B}_i / \sqrt{\hat{v}_i}$ for $\hat{\tau}_i$, $i = 1,2$ where $\hat{\nu}_i$ is the variance of $\hat{B}_i$

Disclosure Risk Assessment

Example: Population of 944,793 from UK 2001 Census

SRS sample size 9,448

Key: Area (2), Sex (2), Age (101), Marital Status (6), Ethnicity (17), Economic Activity (10) - 412,080 cells

Model Selection:

Starting solution: main-effects log-linear model which indicates under-fitting (minimum error statistics too large)
Add in higher interaction terms until minimum error statistics indicate fit

Model Search Example (SRS n=9,448)

True values $\tilde{\tau}_1 = 159$ $\tilde{\tau}_2 = 355.9$

Area–ar, Sex-s, Age–a, Marital Status–m, Ethnicity–et, and Economic Activity-ec

	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{B}_1 / \sqrt{v_1}$	$\hat{B}_2 / \sqrt{v_2}$
Independence - I	386.6	701.2	48.54	114.19
All 2 way - II	104.9	280.1	-1.57	-2.65
1: I + {a*ec}	243.4	494.3	54.75	59.22
2: 1 + {a*et}	180.1	411.6	3.07	9.82
3: 2 + {a*m}	152.3	343.3	0.88	1.73
4: 3 + {s*ec}	149.2	337.5	0.26	0.92
5a: 4 + {ar*a}	148.5	337.1	-0.01	0.84
5b: 4 + {s*m}	147.7	335.3	0.02	0.66
6b: 5b + {ar*a}	147.0	335.0	-0.24	0.56
6c: 5b + {ar*m}	148.9	337.1	-0.04	0.72
6d: 5b + {m*ec}	146.3	331.4	-0.24	0.03
7c: 6c + {m*ec}	147.5	333.2	-0.34	0.06
7d: 6d + {ar*a}	145.6	331.0	-0.44	-0.03

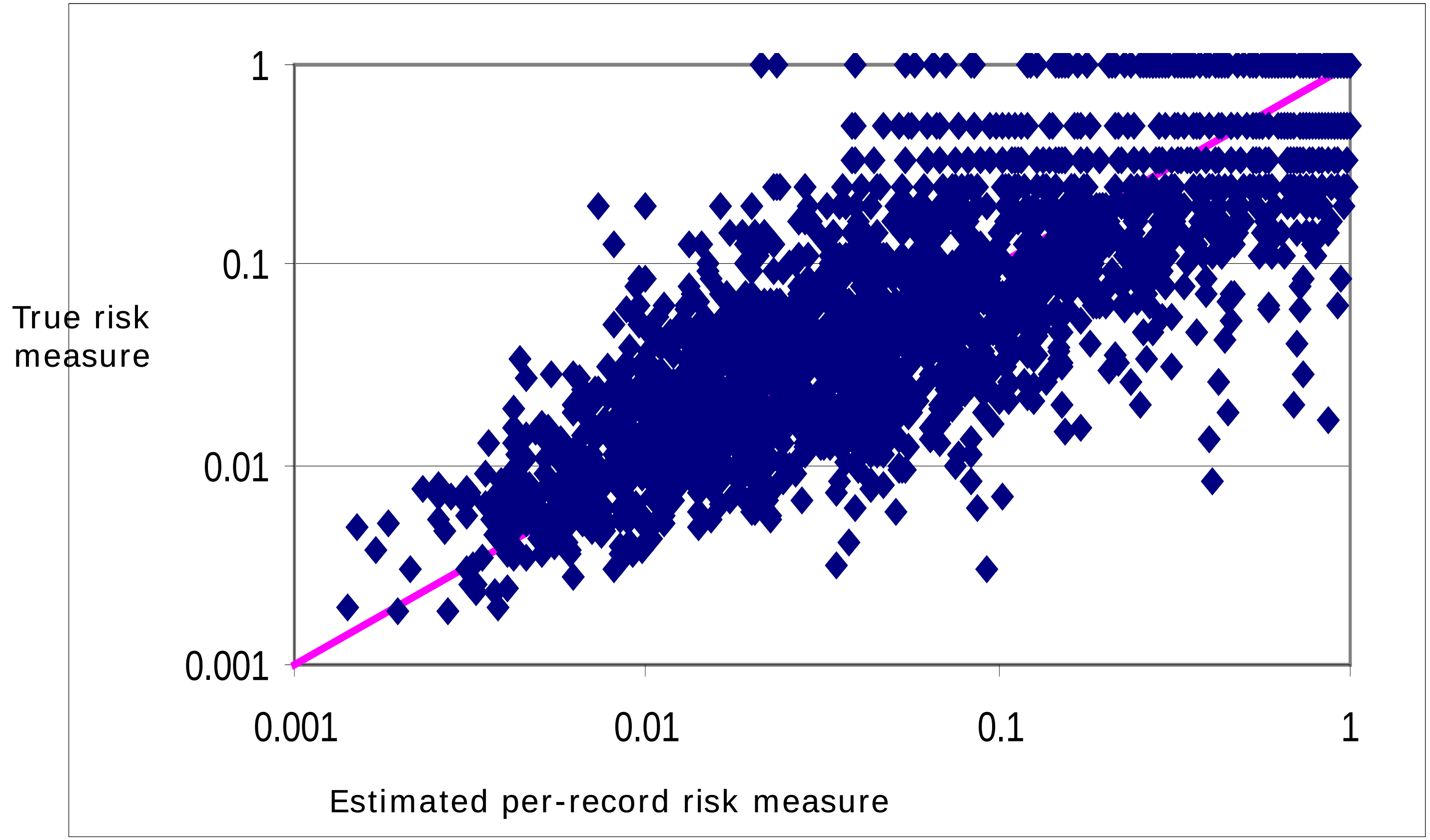
Model Search Example

Preferred Model: {a*ec}{a*et}{a*m}(s*ec}{ar*a}

True Global Risk: $\tilde{\tau}_1 = 159$ $\tilde{\tau}_2 = 355.9$

Estimated Global Risk $\hat{\tau}_1 = 148.5$ $\hat{\tau}_2 = 337.1$

Log-scale



Disclosure Risk Assessment Under Misclassification

- Model assumes no misclassification errors either arising from data processes or purposely introduced for SDL

- Shlomo and Skinner, 2010 address misclassification errors

Let: $M_{kj} = P(\tilde{X} = k \mid X = j)$

where X cross-classified key variables:

X in population fixed

$\tilde{X}$ in microdata subject to misclassification

Disclosure Risk Assessment Under Misclassification

- The per-record disclosure risk measure of a match of external unit B to a unique record in microdata A that has undergone misclassification:

$$P(A=B \mid \tilde{f}_k = 1) = \frac{M_{kk} / (1-\pi M_{kk})}{\sum_j F_j M_{kj} / (1-\pi M_{kj})} \leq \frac{1}{F_k} \qquad (1)$$

- For small misclassification and small sampling fractions:

$$\frac{M_{kk}}{\sum_j F_j M_{kj}} \quad \text{or} \quad \frac{M_{kk}}{\tilde{F}_k} \qquad (2)$$

- Global measure: $\tau_2 = \sum_k I(f_k = 1) \frac{M_{kk}}{\tilde{F}_k}$ estimated by:

$$\hat{\tau}_2 = \sum_k I(\tilde{f}_k = 1) M_{kk} \hat{E}\left(\frac{1}{\tilde{F}_k} \mid \tilde{f}_k \right) \qquad (3)$$

where per-record risk: $M_{kk} \hat{E}\left(\frac{1}{\tilde{F}_k} \mid \tilde{f}_k = 1 \right)$

Misclassification Example

- Population of individuals from 2001 United Kingdom (UK) Census N=1,468,255

- 1% srs sample n=14,683

- Six key variables: Local Authority (LAD) (11), sex (2), age groups (24), marital status (6), ethnicity (17), economic activity (10) K=538,560.

Misclassification Example

- Record Swapping: LAD swapped randomly, eg. for a 20% swap:

  Diagonal: $M_{kk}^{c} = 0.8$

  Off diagonal: $M_{kj}^{c} = 0.2 \times n_k / \left(\sum_{l \neq k} n_l \right)$ where n_k is the number of records in the sample from LAD k

- Pram: LAD misclassified, eg. for a 20% misclassification

  Diagonal: $M_{kk}^{c} = 0.8$

  Off diagonal: $M_{kj}^{c} = 0.02 \quad (0.2/10)$

  Parameter: $\alpha = 0.55$

Misclassification Example

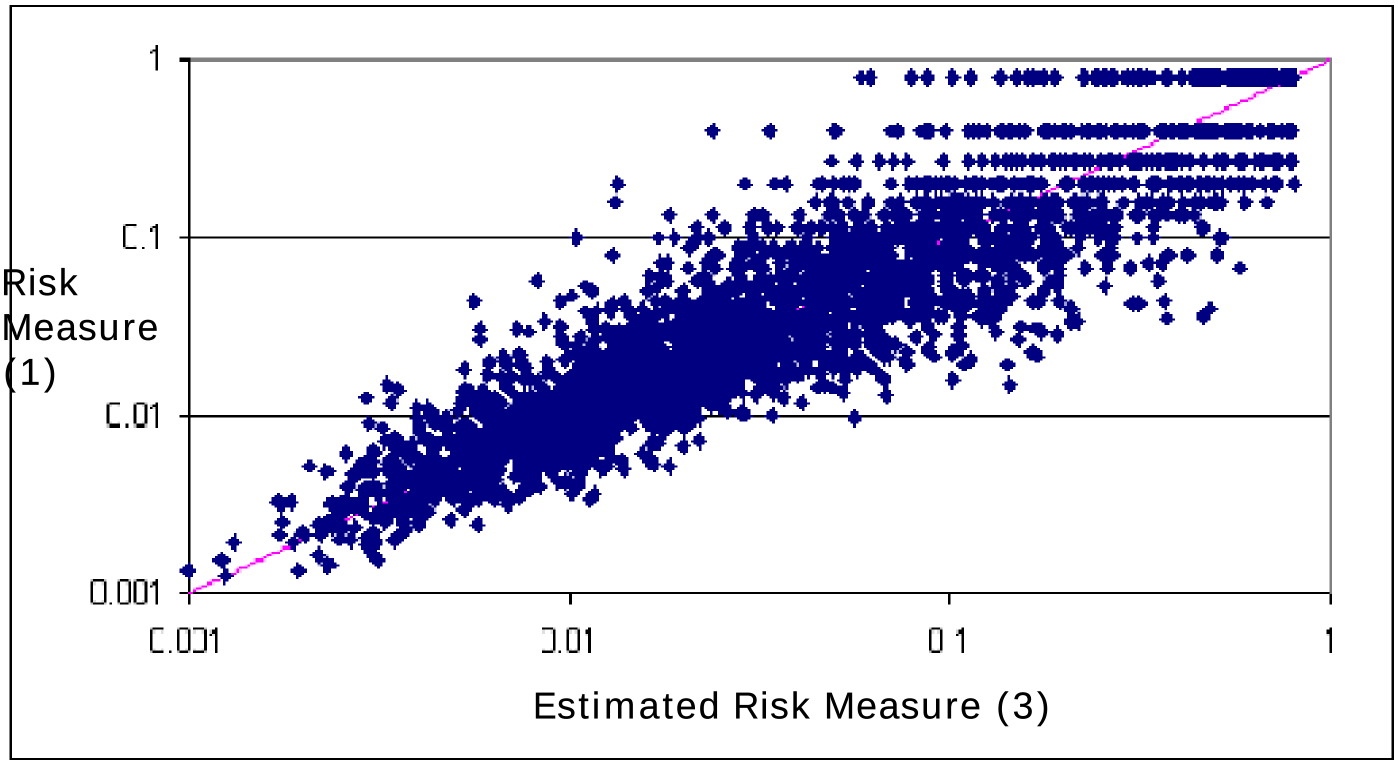
- Random 20% perturbation on LAD
- Global risk measures: Expected correct matches from SU's

Global Risk Measure	PRAM	Swapping
True risk measure in original sample	358.1	362.4
Estimated naïve risk measure ignoring misclassification	349.5	358.6
Risk measure on non-perturbed records	292.2	292.8
Risk measure under misclassification (1) Sample uniques	299.7 2,779	298.9 2,831
Approximation based on diagonals M_{kk}^{c} (2)	299.8	298.9
Estimated risk measure under misclassification (3)	283.1	286.8

Expected correct match per sample unique:
Pram: 10.8% Record swapping: 10.6%

Misclassification Example

- Estimating individual per-record risk measures for 20% random swap based on log linear modelling (log scale):



- From perspective of intruder, difficult to identify high risk (population unique) records

Disclosure Risk Assessment for Identity Disclosure

Probabilistic Record Linkage

- $\tilde{X}_a$ value of vector of cross-classified identifying key variables for unit a in the microdata ($a \in s_1$)
- X_b corresponding value for unit b in the external database ($b \in s_2$) ($s_2 \subseteq P$)
- Misclassification mechanism via probability matrix:

$$P(\tilde{X}_a = k \mid X_a = j) = M_{kj}$$

- Comparison vector $\gamma(\tilde{X}_a, X_b)$ for pairs of units $(a,b) \in s_1 \times s_2$
- For subset $\tilde{s} \subset s_1 \times s_2$ partition set of pairs in $\tilde{s}$

Matches (M) Non-matches (U)

through likelihood ratio: $m(\gamma)/u(\gamma)$ where

$$m(\gamma) = P(\gamma(\tilde{X}_a, X_b) = \gamma \mid (a,b) \in M)$$

$$u(\gamma) = P(\gamma(\tilde{X}_a, X_b) = \gamma \mid (a,b) \in U)$$

Probabilistic Record Linkage

- $p = P((a,b) \in M)$ probability that pair is in M
- Probability of a correct match:

$$p_{M|\gamma} = P((a,b) \in M \mid \gamma(\tilde{X}_a, X_b)) = m(\gamma)p / [m(\gamma)p + u(\gamma)(1-p)]$$

- Estimate parameters using previous test data or EM algorithm and assuming conditional independence

$$m(\gamma) = P(\gamma(\tilde{X}_a, X_b) \mid (a,b) \in M)$$

$$= P(\gamma_1(\tilde{X}_a, X_b) \mid (a,b) \in M) \times P(\gamma_2(\tilde{X}_a, X_b) \mid (a,b) \in M) \ldots P(\gamma_K(\tilde{X}_a, X_b) \mid (a,b) \in M)$$

Probabilistic Record Linkage

- Estimate parameters using EM algorithm and assuming conditional independence:
  - Let γ_q^a 1,0 agreement for $a`th$ pair on $q`th$ key variable
  - Complete data: $\{\gamma^a, g\}$ where $\gamma^a = (\gamma_1^a, \gamma_2^a, ..., \gamma_Q^a)$ and g unknown indicator variable: $\{g_{am}, g_{au}\}$ where $g_{am} = 1$ if pair a is in M and $g_{au} = 1$ if pair a is in U
  - Estimates g_{am} and g_{au} are conditional probabilities of being in M or U given observed data for pair a

Probabilistic Record Linkage

- EM algorithm (cont.)
  - Let $\hat{p}$ estimated proportion of correct matches
  - From Bayes theorem, E-step

$$\hat{g}_{am} = \frac{\hat{p}\prod_{q=`}^{Q} m_q^{\gamma_q^a}(1-m_q)^{1-\lambda_q^a}}{\hat{p}\prod_{q=1}^{Q} m_q^{\gamma_q^a}(1-m_q)^{1-\gamma_q^a} + (1-\hat{p})\prod_{q=1}^{Q} u_q^{\gamma_q^a}(1-u_q)^{1-\gamma_q^a}}$$

$$\hat{g}_{au} = \frac{(1-\hat{p})\prod_{q=1}^{Q} u_q^{\gamma_q^a}(1-u_q)^{1-\gamma_q^a}}{\hat{p}\prod_{q=1}^{Q} m_q^{\gamma_q^a}(1-m_q)^{1-\gamma_q^a} + (1-\hat{p})\prod_{q=1}^{Q} u_q^{\gamma_q^a}(1-u_q)^{1-\gamma_q^a}}$$

  - M-step

$$\hat{m}(\gamma_q) = \sum_{i=1}^{R}\hat{g}_{am}\gamma_{qi}^{a} \Big/ \sum_{i=1}^{R}\hat{g}_{am} \qquad \hat{u}(\gamma_q) = \sum_{i=1}^{R}\hat{g}_{au}\gamma_{qi}^{a} \Big/ \sum_{i=1}^{R}\hat{g}_{au}$$

$$\hat{p} = \sum_{i=1}^{R}\hat{g}_{am} / R$$

Linking the Frameworks

- No misclassification

	Non-match	Match	Total
Disagree	$n(N-1)-f_k(F_k-1)$	$n-f_k$	$Nn-f_kF_k$
Agree	$f_k(F_k-1)$	f_k	f_kF_k
Total	$n(N-1)$	n	Nn

$m(\gamma)=f_k/n \qquad u(\gamma)=f_k(F_k-1)/n(N-1) \qquad p=1/N$

$$p_{M|\gamma}=\frac{1/N\times f_k/n}{1/N\times f_k/n+(1-1/N)f_k(F_k-1)/n(N-1)}=\frac{1}{F_k}$$

Linking the Two Frameworks

- Misclassification observed misclassified sample count $\tilde{f}_k$ with $\tilde{X}_a = k$ derived by: $\tilde{f}_k = M_{kk} f_k + \sum_{k \neq j} M_{kj} f_j$

	Non-match	Match	Total
Disagree	$Nn - n - \tilde{f}_k F_k + M_{kk} f_k$	$n - M_{kk} f_k$	$Nn - \tilde{f}_k F_k$
Agree	$\tilde{f}_k F_k - M_{kk} f_k$	$M_{kk} f_k$	$\tilde{f}_k F_k$
Total	$Nn - n$	n	Nn

$m(\gamma) = M_{kk} f_k / n$ $\quad u(\gamma) = (\tilde{f}_k F_k - M_{kk} f_k) / n(N-1)$ $\quad p = 1/N$

$$p_{M|\gamma} = \frac{1/N \times M_{kk} f_k / n}{1/N \times M_{kk} f_k / n + (1 - 1/N)(\tilde{f}_k F_k - M_{kk} f_k)/n(N-1)} \approx \frac{M_{kk}}{\pi \tilde{f}_k} \approx \frac{M_{kk}}{\tilde{F}_k}$$

Empirical Study

- Matching 2,853 sample uniques to the population and blocking on all key variables except LAD result in 1,534,293 possible pairs

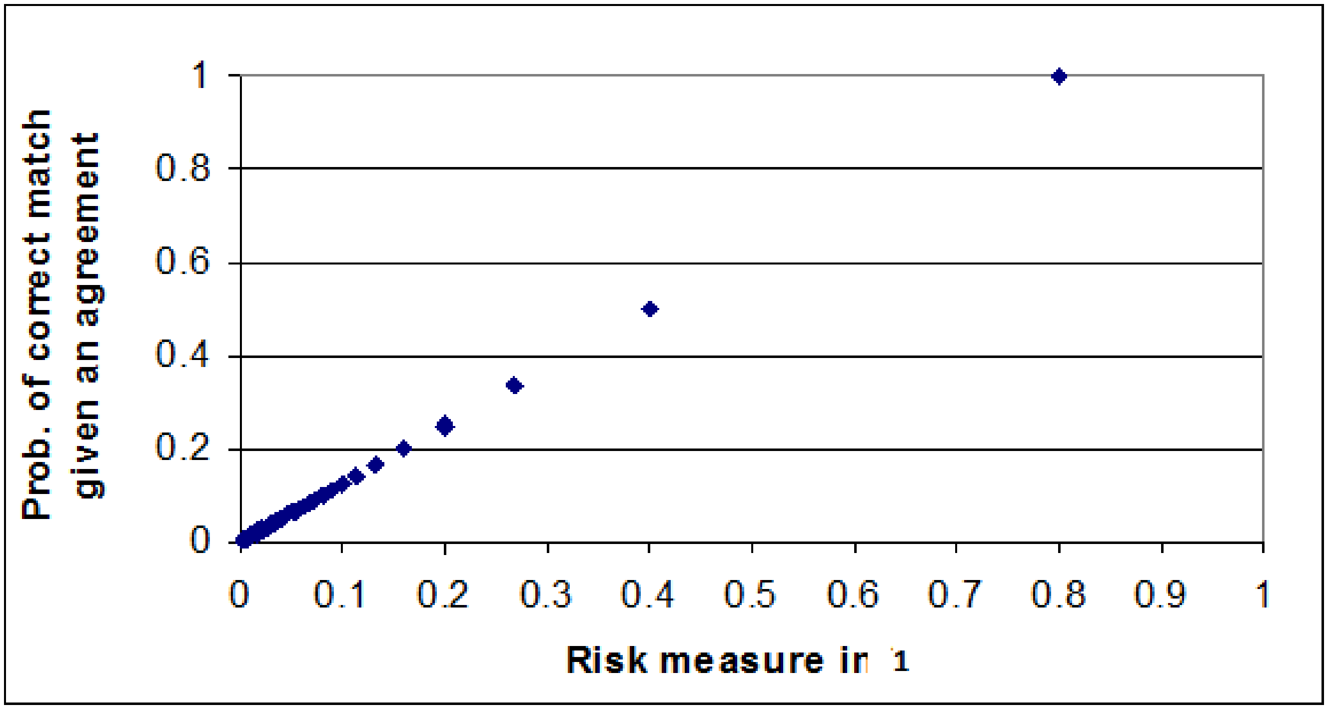
	Non-match	Match	Total
Disagree LAD	1,388,069	619	1,388,688
Agree LAD	143,321	2,234	145,555
Total	1,531,390	2,853	1,534,293

$m(\gamma) = 0.78 \quad u(\gamma) = 0.09 \quad p = 0.002$

- On average, probability of a correct match given an agreement on LAD $p_{M|\gamma} = 0.015$

Empirical Study

- Probability of a correct match given on agreement $p_{M|\gamma}$ for each $\gamma(\tilde{X}_a, X_b) = k$
- Compare to risk measure $M_{kk} / \tilde{F}_k$



- Summing over $p_{M|\gamma}$ the global disclosure risk measure of 289.5.

Empirical Study

- Estimation via EM algorithm for one $\gamma(\tilde{X}_a, X_b) = k$

	Non-match	Match	Total
Disagree LAD	2,283	1	2,284
Agree LAD	48	2	50
Total	2,331	3	2,334

- True parameters: $m(\gamma) = 0.667$ $u(\gamma) = 0.021$ $p = 0.0013$

  $\hat{p}_{M|\gamma} = 2/50 = 0.040$

- Estimation: $\hat{m}(\gamma) = 0.726$ $\hat{u}(\gamma) = 0.020$ $\hat{p} = 0.0015$

$$\hat{p}_{M|\gamma} = \frac{0.0015(0.726)}{0.0015(0.726) + (1 - 0.0015)(0.020)} = 0.052$$

- Difficult to estimate parameters
- Accuracy of EM algorithm depends on a large number of pairs and a relatively large number of correct matches (approximately over 5%)

Disclosure Risk Assessment for Attribute Disclosure

- Use record linkage techniques to assess disclosure risk for attribute disclosure in enterprise microdata
- Assumes that the data is taken from a Census and $f_k = F_k$ so that the probability of a correct match depends on M_{kk} the probability of not being perturbed
- Use a string comparator to measure the distance between original and perturbed values of a variable $p,\quad p = 1, \dots, P$ (Yancy et al. 2002)
- String comparator takes a value between 0 and 1 for each variable
- Assuming conditional independence assumption of F & S, combine individual string comparators to estimate M_{kk}

Disclosure Risk Assessment for Attribute Disclosure

- String comparator for variable p :

  Calculate the noise: $\varepsilon_i = Y_i - \tilde{Y}_i$ where $\tilde{Y}_i$ is the perturbed value for record i

  - $Z_i = (\varepsilon_i - E(\varepsilon_i))/Var(\varepsilon_i)$ and $STR^p{}_i = 1 - |1 - 2\Phi(Z_i)|$
  - $STR^p{}_i = \exp\{-|\varepsilon_i|/med(|\varepsilon_i|)\}$

- Calculate a weighted average of string comparators where the weights W_p are the normalized odds of a correct match given an agreement (similar to u-probability of F&S record linkage)

- Calculate odds via a logistic regression model where the response variable is the true match indicator and the explanatory variables the string comparators

Disclosure Risk Assessment for Attribute Disclosure

- Probability of a correct match for record i :

  $p_i = \sum W_p \, STR^p{}_i$ and $\sum W_p = 1$

- Decide on a type I error (probability of declaring a match when the null is no match) and determine threshold to declare the pairs that are matches

- Disclosure risk measures:
  - Proportion of correct matches out of declared links
  - Odds of a correct match given an agreement: declared links that are true matches / declared links that are false matches
  - $\sum_{i \in M} p_{i|}$ expected number of correct matches

Example

- Sugar Farms Data from a 1982 survey of sugar cane industry in Queensland, Australia: Region (4 categories) and 5 continuous variables: Area, Harvest, Receipts, Costs, Profits (=Receipts-Costs)

- Data Protection:
  - 5 outliers removed resulting in 333 farms
  - Region not perturbed
  - Area (identifying variable) coarsened 9 categories
  - Remaining continuous variables perturbed with multivariate random Gaussian noise within quintiles of receipts (index for quintiles dropped):

$$(\varepsilon_H, \varepsilon_R, \varepsilon_C, \varepsilon_P)^T \sim N(\mu', \Sigma)$$

where $\boldsymbol{\mu}'^{\mathrm{T}} = (\mu'_H, \mu'_R, \mu'_C, \mu'_P,) = (\frac{1-d_1}{d_2}\mu_H, \frac{1-d_1}{d_2}\mu_R, \frac{1-d_1}{d_2}\mu_C, \frac{1-d_1}{d_2}\mu_P)$

and $\boldsymbol{\Sigma}$ is the original covariance matrix

Example

- The vector $\boldsymbol{\mu}'$ contains the corrected means of each of the four variables in the quintile with $d_1 = \sqrt{(1-\delta^2)}$ and $d_2 = \sqrt{\delta^2}$ and δ is the perturbation parameter

- For each variable on record i, calculate a linear combination, for example, for receipts:

$$\tilde{R}_i = d_1 R_i + d_2 \varepsilon_{Ri}$$

- Mean vector and covariance matrix remain the same as the original data and the edit constraint: Profits=Receipts-Costs is exactly preserved

- Assume one dataset released

- Create all possible pairs: 333^2=110,889 however Region not perturbed so use as blocking variable: 31,367 possible pairs

Results

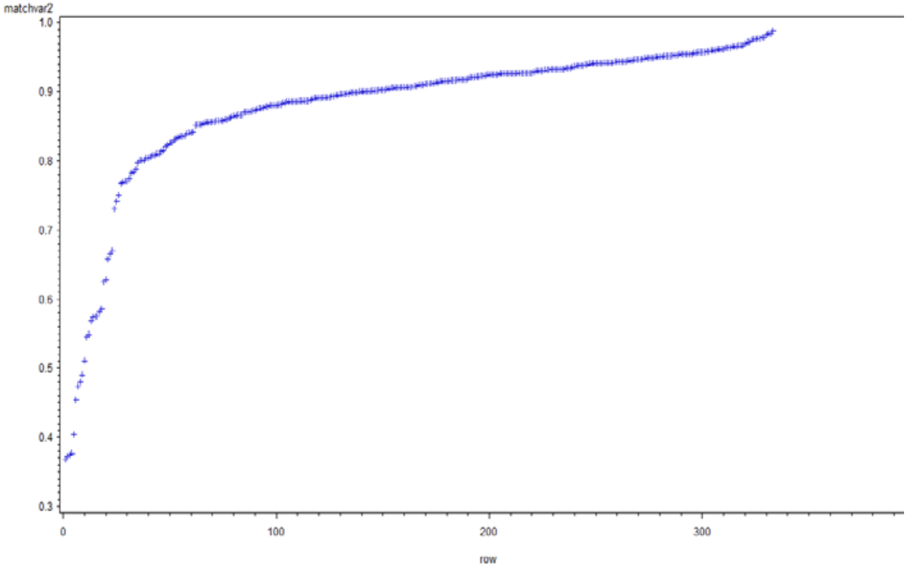
- Threshold: Type I error 1.4%

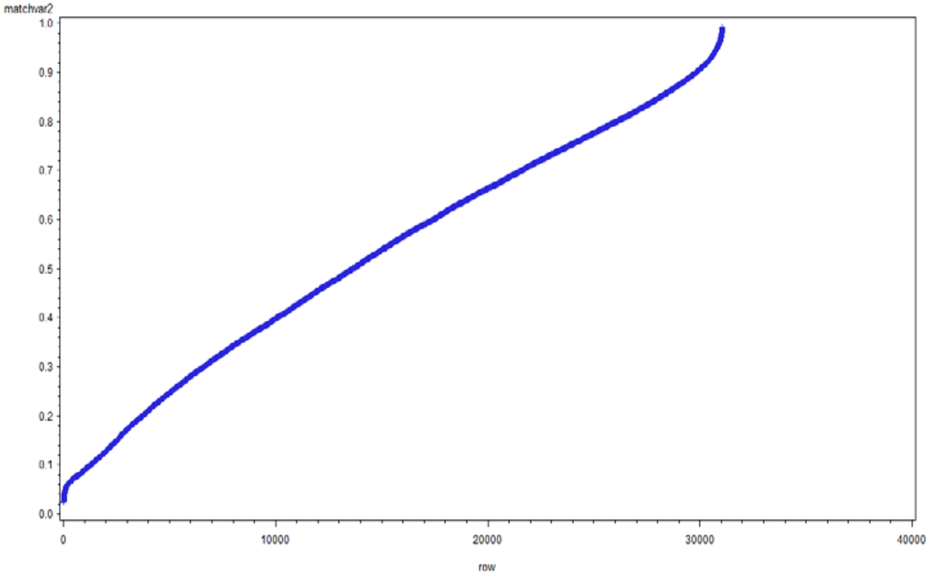
		Delta=0.4		Delta=0.7	
		Distribution	Exponential	Distribution	Exponential
Equal Weights	Matches/Links	0.297	0.290	0.160	0.151
	Matches/False Matches	0.423	0.409	0.191	0.178
	Sum of p_i	307.5	290.0	289.8	263.9
Weights Odds	Matches/Links	0.307	0.313	0.168	0.175
	Matches/False Matches	0.443	0.455	0.201	0.213
	Sum of p_i	309.0	295.6	299.9	292.7

Results

Probability of a Match

Delta=0.7 String Comparator=exponential function





Matches Non-matches

Σ

Discussion

- Empirical evidence of connection between F&S record linkage and the probabilistic modelling for estimating identification risk

- Statistical agencies can accurately estimate global disclosure risk measures for a risk-utility assessment assuming known non-misclassification probability
  - Estimation is carried out through log linear modelling for the probabilistic modelling or the EM algorithm for the F&S record linkage

- Based on the connection between F&S record linkage and probabilistic modelling for identity disclosure, use record linkage techniques to assess attribute disclosure of enterprise microdata

Thank you for your attention