NCRM National Centre for Research Methods

A unified approach to the assessment of both identification and attribution risks

Duncan Smith and Mark Elliot

University of Manchester

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH

Introduction

In statistical disclosure control (SDC) we are generally concerned with identification risk or attribution risk

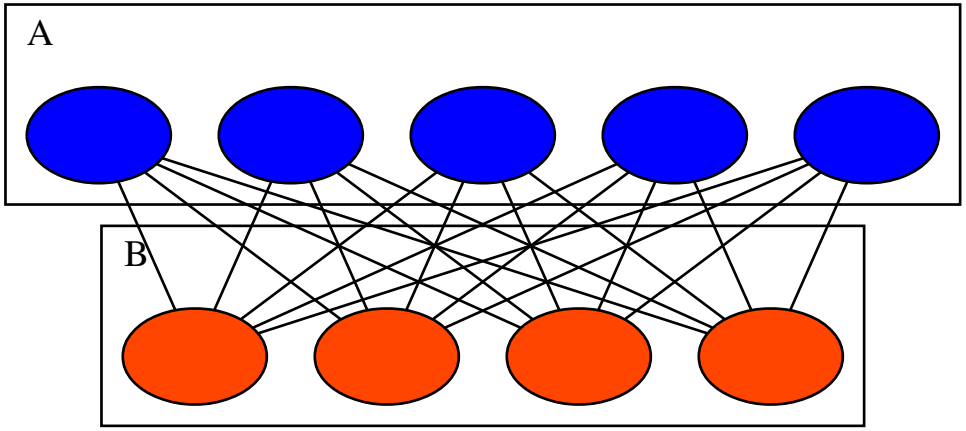
Identification

- Associating a target with a single record

Attribution

- Associating a target with a sufficiently detailed attribute value

Disclosure inevitably involves some form of linkage



Classical Record Linkage

We have a set of all possible matches

$$A \times B = \{(a, b); a \in A, b \in B\}$$

which can be partitioned into sets of matched and unmatched pairs

$$M = \{(a, b); a = b, a \in A, b \in B\}$$

$$U = \{(a, b); a \neq b, a \in A, b \in B\}$$

and the goal of record linkage is to allocate each possible match to either *M* or *U*

Fellegi-Sunter

- If the data are aligned so that each index corresponds to the same variable in *A* or *B*. Then, Fellegi-Sunter assumes that the posterior odds can be factorized as:

$$\frac{Pr((a,b)\in M|(a,b))}{Pr((a,b)\in U|(a,b))}=\left(\prod_{i=1}^{n}\frac{Pr((a_i,b_i)|(a,b)\in M)}{Pr((a_i,b_i)|(a,b)\in U)}\right)\frac{Pr((a,b)\in M)}{Pr((a,b)\in U)}$$

where comparisons on the key variables are for equality

Extended Latent Model

Consider an extended latent model

$$\frac{Pr((a,b)\in M|(a,b))}{Pr((a,b)\in U|(a,b))}=\frac{Pr((a,b)|(a,b)\in M)}{Pr((a,b)|(a,b)\in U)}\times \left(\prod_{i=1}^{n}\frac{Pr((a_i,b_i)|(a,b)\in M)}{Pr((a_i,b_i)|(a,b)\in U)}\right)\frac{Pr((a,b)\in M)}{Pr((a,b)\in U)}$$

where the additional Bayes factor concerns the values of variables rather than the binary comparison of value pairs

Given a full probability model $\mathcal{M}$,

We can directly generate

$$Pr((a,b)|(a,b)\in M)$$

We can also generate

$$Pr((a,b)|(a,b)\in U)$$

(although there are reasons why we might use a distinct model)

Expectation Maximization

The log likelihood function is,

$$\ln\left(f\left(x|\Phi,\Phi_m,\Phi_u\right)\right)=\sum_{j=1}^{N} g_j\cdot\left(\sum_{i=1}^{n}\ln\left(m_i^{\gamma_i^j}\left(1-m_i\right)^{1-\gamma_i^j}\right),\sum_{i=1}^{n}\ln\left(u_i^{\gamma_i^j}\left(1-u_i\right)^{1-\gamma_i^j}\right)\right)^T+$$
$$\sum_{j=1}^{N} g_j\cdot\left(\ln\left(p\right),\ln\left(1-p\right)\right)^T+$$
$$\sum_{j=1}^{N} g_j\cdot\left(\ln\left(Pr\left(\delta^j|\Phi_m\right)\right),\ln\left(Pr\left(\delta^j|\Phi_u\right)\right)\right)^T$$

where $\Phi=\left(m,u,p\right)$ and Φ_m and Φ_u are the parameter vectors associated with the models for the numerator and denominator terms of the additional Bayes factor

In principle we can find the parameter values that maximize this function via Expectation Maximization (Jaro, 1989)

Pragmatic approach

1. Perform classical linkage (using EM)
2. Estimate full probability model $\mathcal{M}$ from linked data
3. Repeat EM linkage with fixed $\Phi_m = \Phi_u$

The model $\mathcal{M}$ can be iteratively re-fitted

- It can be difficult to avoid over-fitting
- Maximum a Posteriori (MAP) estimation can help

Additional training data can be used to improve estimation of $\mathcal{M}$

Disclosure control

Identification risk → match probabilities

Attribution risk → match probabilities and / or $\mathcal{M}$

Use precision recall curves

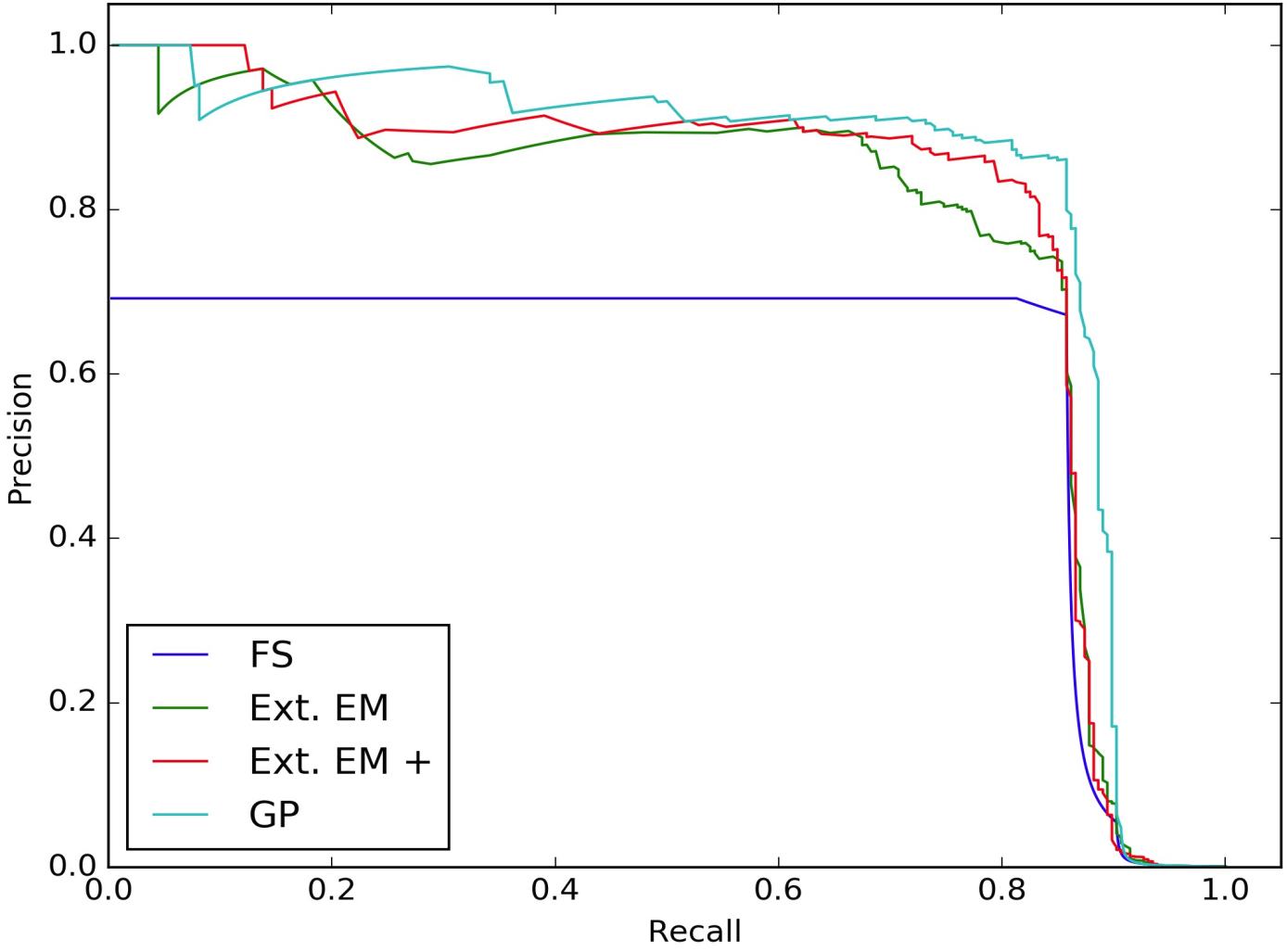
$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

Experiment using data sampled from the insurance network (http://www.bnlearn.com/bnrepository/#insurance)

Data partitioned and sub-sampled to generate datasets *A* and *B* (with known match statuses)

Perturbations applied to data in *B* to simulate typographical and coding errors


1.0
0.8
0.6
0.4
0.2
0.0
Precision
0.0
0.2
0.4
0.6
0.8
1.0
Recall
FS
Ext. EM
Ext. EM +
GP

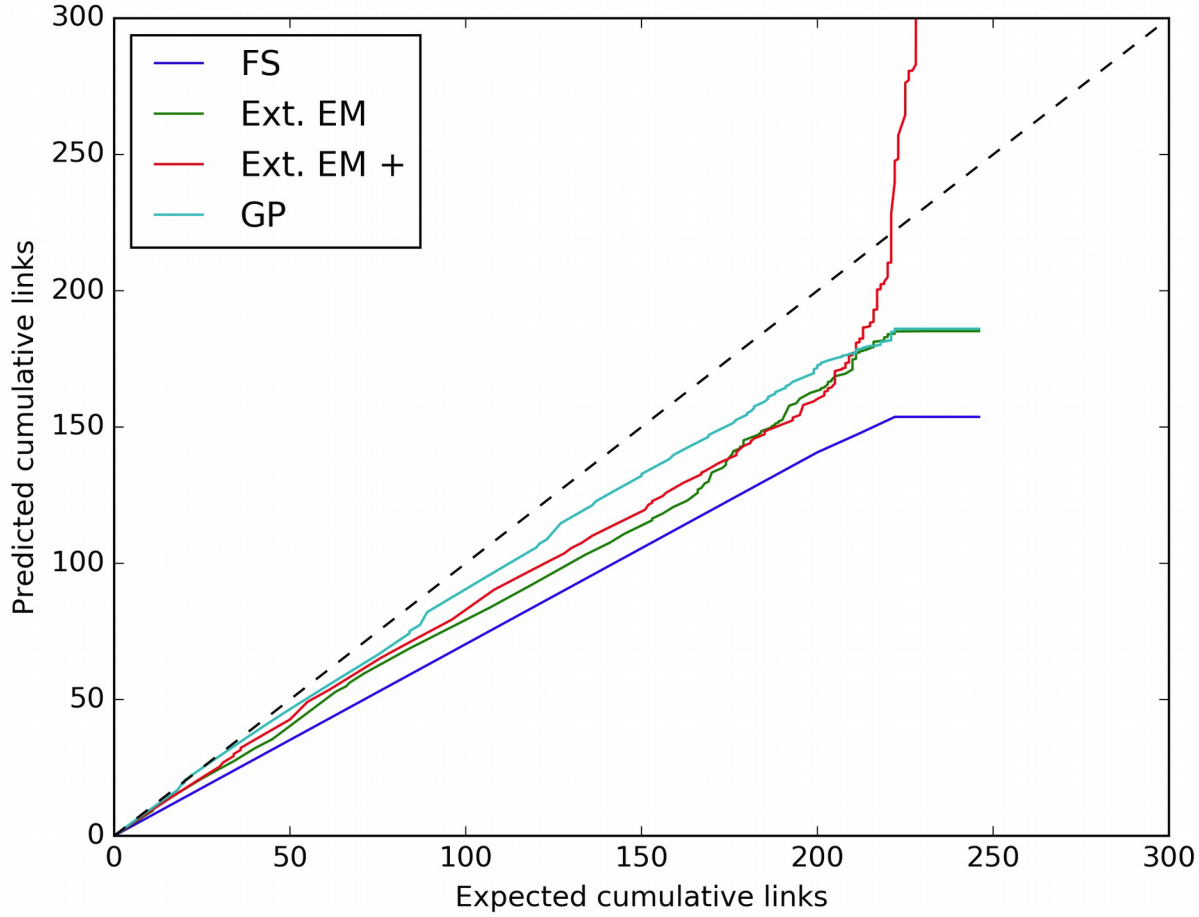
For Fellegi-Sunter 69.2% of the most probable matches are matches

For all other strategies 100% of the most probable matches are matches

Fellegi-Sunter underestimates the identification risk from an intruder using the extended latent model

Without known match statuses we cannot generate precision recall curves

We can use the match probabilities as risk measures



Plot of cumulative match probabilities against expected cumulative matches – record pairs ordered from most probable matches to least probable matches

For Fellegi-Sunter match probabilities are significantly underestimated (69.2% matches versus estimate of 0.487 for most probable matches)

All other strategies also underestimate match probabilities, but to a lesser degree than FS and less so for higher probability matches

Fellegi-Sunter underestimates the identification risk from an intruder using the standard latent model

Attribution risk depends on the posterior beliefs that can be generated for sensitive attribute values

- Prior versus posterior
- Differential privacy

Posteriors can be generated via aggregation over match probabilities, or via the fitted model $\mathcal{M}$

	Jensen-Shannon divergences			
	Fellegi Sunter	Extended EM	Extended EM +	Generating process
Aggregation over match probs	0.573	0.546	0.328	0.508
$\mathcal{M}$	N/A	0.333	0.199	0.113

In terms of the Jensen-Shannon divergence $\mathcal{M}$ provides a better estimate of the joint distribution

Summary

The extended latent model approach generates match probabilities and an estimate of the joint distribution

Match probabilities → identification risk

The joint distribution → attribution risk

Both match probabilities and the estimated joint distribution are better than when using Fellegi-Sunter

Additional slides that might help in answering questions

Let x be the complete data vector equal to $\langle \gamma , g \rangle$, where g_j = (1,0) iff the jth record pair $r_j \in M$ and g_j = (0,1) iff $r_j \in U$.

Let the configuration of the evidence on the values of the variables for the jth record pair be denoted δ^j.

$$p = \frac{|M|}{|M \cup U|}$$

$$Pr(\gamma^j | M) = \prod_{i=1}^{n} m_i^{\gamma_i^j} (1 - m_i)^{1-\gamma_i^j}$$

$$Pr(\gamma^j | U) = \prod_{i=1}^{n} u_i^{\gamma_i^j} (1 - u_i)^{1-\gamma_i^j}$$

Estimation

- Let $\gamma_i^j = 0$ if attribute i differs for the jth record pair r_j, and $\gamma_i^j = 1$ if attribute i matches for the jth record pair r_j.

$$m_i = Pr\left(\gamma_i^j = 1 | r_j \in M\right)$$

$$u_i = Pr\left(\gamma_i^j = 1 | r_j \in U\right)$$

$$p = \frac{|M|}{|M \cup U|}$$