Privacy-Preserving Record Linkage

Rob Hall and Stephen E. Fienberg

Problem Setup

Two agencies wish to perform some statistical calculation on the union of their data:

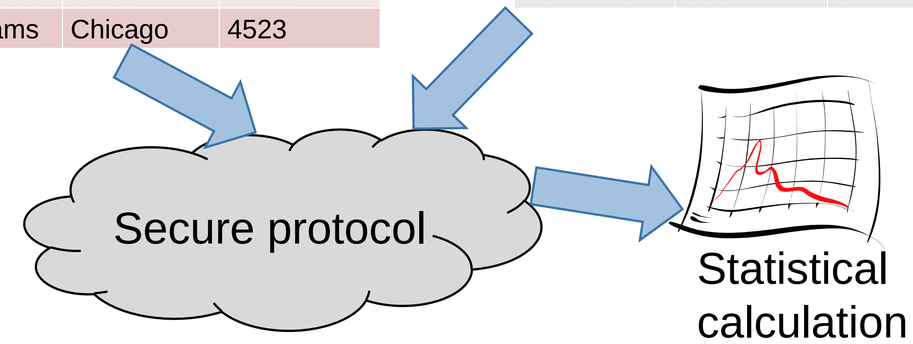
1. **In the absence of unique identifiers for the records.**
2. **Without revealing their information to each other**

Internal Revenue Service

Name	City	Tax
A. Smith	Pittsburgh	28723
B. Johnson	Boston	12356
C. Williams	Chicago	4523

Census Bureau

Name	City	# Children
Adam S	Pgh	0
Bob Jonson	Boston	2



(1) Requires “record linkage.”
(2) Requires “secure multiparty computation.”

We survey some recent approaches to the problem.

State of Current Techniques

Match Variables: Well Understood

- Edit distance: via secure set intersection, secure vector product, bloom filter etc.
- Thresholding: via Yao’s protocol.
- Protocols still slow or making use of a third party.

Statistical Modelling: Not Done

- Current techniques assume a match if e.g., edit distance is below some cutoff.
- Reasonable heuristic but must understand its statistical properties.

Blocking: Partially Understood

- Blocking attempts to strip out obvious non-matches in order to save computation (i.e., via a heurstic).
- Proposed methods: release permuted data (e.g. differential privacy).
- Weakens privacy guarantee compared to cryptographic model, may allow better expected computation time as privacy guarantee is weakened (needs more analysis)

Composition of Protocols: Well Understood

- For an end-to-end secure protocol we require that only the regression output is released.
- **Cannot reveal:** match variables, match decisions, matched data… **any intermediate values**.

Traditional Record Linkage

Internal Revenue Service

Name	City	Tax
A. Smith	Pittsburgh	28723
B. Johnson	Boston	12356
C. Williams	Chicago	4523

Census Bureau

Name	City	# Children
Adam S	Pgh	0
Bob Jonson	Boston	2

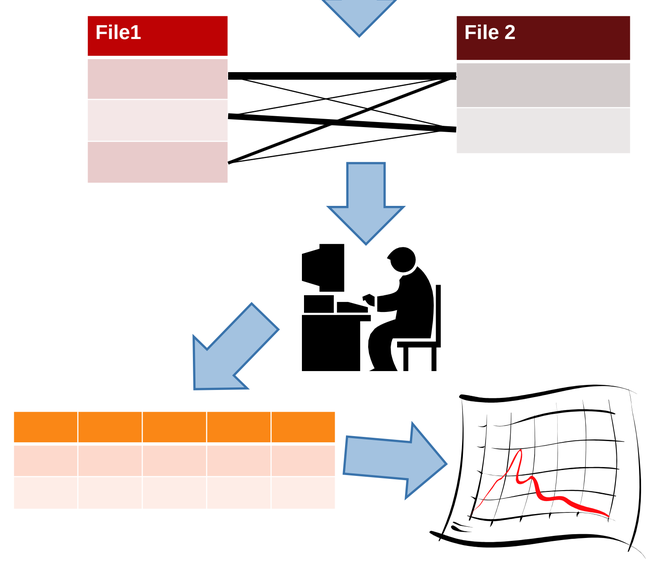
For all pairs, compute “match variables”

$$m_{i,j} = \begin{pmatrix} 1\{\text{city}_i = \text{city}_j\} \\ \text{edit-dist}(\text{name}_i, \text{name}_j) \\ \vdots \end{pmatrix}$$

Compute models of match variables for the matches and non-matches (i.e., using EM)

$$p_\theta(m_{i,j}|(i,j) \in \text{non-matches})$$
$$p_\theta(m_{i,j}|(i,j) \in \text{matches})$$

Estimate matches and non-matches, via LRT



Clerical review of edge-cases

Regression against integrated data

Challenges and Current Work

Propagation of Error

- Integrated data never sanity checked, so cannot be regarded as error-free.
- Need to understand the distribution of errors of the linkage technique.
- Need to understand how errors in the linked data propagate into errors in the statistical analysis.

Efficiency

- Understanding when it is appropriate to relax the privacy guarantee (e.g., for blocking)
- Understanding the impact of blocking on computation time for varying privacy guarantees.

Statistical Modeling of Linkage

- In principle is possibly via Yao’s construction.
- Need to find a more efficient or heuristic way in the cryptographic model.

Carnegie Mellon CyLab
CONFIDENCE FOR A NETWORKED WORLD