Privacy Preserving Record Linkage

Xiaodong Lin

Recent interest in data linkage

- In recent years, interest in data linkage has grown enormously
- A lot of data is collected by many organizations
- Data warehousing and data integration
- Data mining of large data collections
- E-commerce and Web applications
- Sensor network and spatial data analysis

Data linkage techniques

- Deterministic linkage
  - Exact linkage (when a unique identifier is available. For example driver license number)
  - Rules based linkage
- Probabilistic linkage (Fellegi & Sunter, 1969)
- Machine learning approaches

  Supervised and non-supervised techniques
- These techniques assume that records are shared among data owners

Privacy preserving data linkage

- The private record linkage problem
  - Party "A" holds dataset A
  - Party "B" holds dataset B
  - Match common records between A and B, such that
    1. A and B remain private
    2. $A \cap B$ is shared
- Applications
  - Public health and biomedical research
  - Cooperation between government agencies
  - Sharing of intellectual property

Existing techniques

- Warehousing approach: de-identified data are centralized and linked. Mid-to-late 1990s
- Blindfolded record linkage (Churches and Christen, 2004). Allow approximate linkage of strings with typographical errors based on n-gram techniques
- Privacy-preserving data linkage protocols (O'Keefe et.al., 2004). Several protocols with improved security and less information leakage
- Blocking aware private record linkage (Al-Lawati et.al., 2005). Aapproximate linkage based on tokens and TF-IDF, and three blocking approaches

Naive three-party model

- Three parties
  - Two collaborating parties
  - A third party for matching
- All parties semi-trusted
  - Follow protocol precisely
    1. Provide accurate data
    2. Do not collude with other parties
  - However, all the parties are curious
    1. Dictionary attacks
    2. Frequency and statistical attacks

Naive three-party model

- Use one way hash function to encrypt data.
  - Hash function: mapping of string or numerical values to a fixed length string
  - Probability that two different source strings will produce same hash value is very small. For example, 10^{-24} for 160 bit hashing
  - Small changes in the original string cause major changes in hash value
- The third party "C" compares the Hashed value and shares the matched pairs

Possible problems

- Party "C" can mount a dictionary attach

  "C" may know that the hashed values are derived from surnames. He can exhaustively compute all the hash values from a surname list and compare with those given by A and B to find out the original values

- If a value in "A" is "Victoria" and value in "B" is "victoria", their hashing values are different

- Record with errors. For example, the value in "B" maybe "victora"

Secure n-gram similarity comparisons

- "A" and "B" agree on a secrete key that transforms the source value before . The resulting hashed values are now called keyed hashing value
- "A" and "B" must agree on a set of "pre-processing" rules and transformations to make the values alike
- Use different similarity measures to address the robustness problem w.r.t. record errors

  An n-gram is the set of sub strings of length n in a word string. For example, the bigrams in the word "peter" are "pe", "et", "te", and "er"

Secure n-gram similarity comparisons

- The corresponding similarity measure is defined as

$$\text{Similarity score} = 2 \times \left(\frac{|\text{bigrams(x)} \cap \text{bigrams(y)}|}{|\text{bigrams(x)}| + |\text{bigrams(y)}|} \right)$$

- The similarity score for "peter" and "pete" is $2 \times 3/(4+3) = 0.86$
- In order to compute the Dice coefficient, the power set of each bigram set needs to be calculated
- "A" and "B" sent the keyed hash values for the power set of their records to "C". "C" finds out which tuple among the power sets matches
- "C" then computes the similarity score

Protocol for blindfolded record linkage

- Compare each of the partially-identifying data elements and return a similarity score for each pair
- These similarity scores are then used to compute the matching weight
- The Fellegi-Sunter or the Winkler models can be used to classify the records as matches, possible matches and non-mathces
- Produce the linked data. "A" and "B" can do this themselves, or a new trusted fourth party can be created the link the data.

Protocol for blindfolded record linkage

- The communication cost for the protocol is very high
- To improve efficiency
  - Only pass those records with similarity scores over a pre-specified threshold
  - Use block-wise record linkage algorithms (Al-Lawati et al. 2005)
- Other similarity measures

  The TFIDF (Token Frequency / Inverse Document Frequency) distance metric

  The secure computation of this metric can be reduced to the secure computation of a scalar product (Cohen, Ravikumar and Fienberg 2004)

A very incomplete set of unsolved problems

- Efficiency. Even with the implementation of threshold and blockwise approach, the protocols are still inefficient
- Protocols need to use a third (or even a fourth) trusted third party
- Dealing with missing values
- Implementing other distance measures or linkage algorithms
- Connecting record linkage with database indexing on the fly, with or without privacy constraint

References

- Tim Churches and Peter Christen. Some methods for blindfolded record linkage. *BMC Medical Informatics and Decision Making*, 4(9), 2004.
- I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183 - 1210, 1969.
- Ali Al-Lawati, Dongwon Lee, and Patrick McDaniel. Blocking-Aware Private Record Linkage. *In ACM SIGMOD Workshop on Information Quality in Information Systems (IQIS)*, page 59-68, Baltimore, MD, USA,2005.
- W. Cohen, P. Ravikumar, and S. Fienberg. A secure protocol for computing string distance metrics. *In Proceedings of ICDM Workshop on Privacy and Security Aspects of Data Mining,* 2004

References

- Dusserre L, Quantin C, and Bouzelat H. A one way public key cryptosystem for the linkage of nominal files in epidemiological studies. *Medinfo* 1995, 8:644-7.
- Bouzelat H, Quantin C, and Dusserre L. Extraction and anonymity protocol of medical file. *Proc AMIA Annu Fall Symp* 1996:323-27.
- Quantin C, Kerkri E, Allaert FA, Bouzelat H, and Dusserre L. Security aspects of medical file regrouping for the epidemiological follow-up. *Medinfo* 1998, 9:1135-7.
- O'Keefe, C. M.; Yung, M.; Gu, L., and Baxter, R. Privacy-preserving data linkage protocols. *Workshop on Privacy in the Electronic Society (WPES'04)*. Washington, DC. 2004.