Composition and Disclosure of Unlinkable Distributed Databases

Bradley Malin Latanya Sweeney
Data Privacy Laboratory, School of Computer Science
Carnegie Mellon University, Pittsburgh, PA USA 15213-3890
{malin, latanya}@privacy.cs.cmu.edu

Abstract

An individual's location-visit pattern, or trail, can be leveraged to link sensitive data back to identity. We propose a secure multiparty computation protocol that enables locations to provably prevent such linkages. The protocol incorporates a controllable parameter specifying the minimum number of identities a sensitive piece of data must be linkable to via its trail.

1 Introduction

Current technology permits the collection, storage, processing, and transfer of personal information with minimal monetary or computational constraints. Yet, the dissemination of personal information must be performed in a manner that upholds an individual's legal rights to privacy. For example, in healthcare, various regulations, such as the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA), require data holders to render personal health information "sufficiently de-identified" prior to various disclosures [1]. However, ad hoc de-identification methods, such as the removal of explicit identifying information, provide false assurances and do not guarantee the anonymity of health data.

For this paper we concentrate on a susceptibility detailed in prior research [3], in which de-identified records, such as DNA sequences, are linked to corresponding identities via patterns in location visits, or trails. This problem manifests because data holders are not legally permitted to openly discuss the contents of their respective databases. In this research, we introduce a solution to provably prevent trail re-identification while adhering to defined policy constraints, such as the HIPAA Privacy Rule. We formalize a secure multiparty protocol by which a set of data holders can work with a third party, such that no recipient of disclosed de-identified data, including the participating data holders and the third party, can achieve re-identification beyond a configurable parameter. The protocol makes use of an anonymization algorithm with a heuristic based on probabilistic intuition to maximize specified utility functions for the disclosed data. We evaluate effectiveness on real world data with known susceptibilities and demonstrate the protocol supports disclosure of substantial quantities of data with exact guarantees of privacy protection.

2 Data Privacy Framework

Consider a set of hospitals $HOSP$. Each hospital H_i collects data on a population of patients S and maintains a private database T_i of person-specific information. For disclosure, H_i partitions T_i into two tables. The first table, ψ_i, contains identifiable attributes, such as name. The second table δ_i, contains de-identified attributes, such as DNA.

The trail for data element x is a vector $[v_{x,1}, \ldots, v_{x,|HOSP|}]$, such that $v_{x,i} = 1$ means a patient visited H_i; $v_{x,i} = 0$ means the patient did not visit H_i; and $v_{x,i} = *$ means visit status is ambiguous. We assume identified data is always collected, but de-identified data is only sometimes collected. Thus, a patient's de-identified trail is the same as the identified trail, but with *'s. Trail re-identification occurs when we can correctly and discriminantly link a patient's identified data trail and de-identified data trail.

To render trails unlinkable, hospitals must compare databases without revealing their contents. We designed a specific implementation of a secure multiparty computation framework for list analysis [2]. The framework, based on commutative cryptography, allows a third party to analyze encrypted lists and provide personalized responses. Each hospital H_i encrypts data x using encryption key ϵ_i and a function F that satisfies $F(F(x, \epsilon_i), \epsilon_j) = F(F(x, \epsilon_j), \epsilon_i)$, for any subset and ordering of the keys. Moreover, each encryption key is paired with a decryption key κ_i, such that $F(F(F(F(x, \epsilon_i), \epsilon_j), \kappa_i), \kappa_j) = x$.

We call the protocol Secure TRail ANONymization, or STRANON, pseudocode provided in Algorithm 1. First, each hospital encrypts every hospital's de-identified database with its encryption key. Next, the fully encrypted databases are sent to the *sTTP*. Upon reception of all databases,

Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)
8-7695-2570-9/06 $20.00 © 2006 IEEE

Algorithm 1 STRANON ($HOSP$, $sTTP$)

for each $H_i \in HOSP$ **do**
  $M_i \leftarrow F(\delta_i, \epsilon_i)$
  for each $H_j \in HOSP, i \neq j$ **do**
    H_i sends M_i to H_j
    H_j sends $M_i \leftarrow F(M_i, \epsilon_j)$ to H_i
  end for
end for
Each $H_i \in HOSP$ sends M_i to *sTTP*
sTTP executes $TRANON(\Psi, \{M_1, \ldots, M_{|HOSP|}\}, k)$ to generate encrypted datasets $\{N_1, \ldots, N_{|HOSP|}\}$
sTTP sends $N_1, \ldots, N_{|HOSP|}$ to $H_1, \ldots, H_{|HOSP|}$
for each $H_i \in HOSP$ **do**
  for each $H_j \in HOSP, i \neq j$ **do**
    H_i sends N_i to H_j
    H_j sends $N_i \leftarrow F(N_i, \kappa_j)$ to H_i
  end for
  H_i discloses $\phi_i \leftarrow F(N_i, \kappa_i)$
end for

the *sTTP* runs an anonymization algorithm, referred to as *TRANON*, and then responds to each hospital with a dataset of encrypted values to disclose. Finally, each hospital decrypts every return dataset, and the values are disclosed.

Algorithm 2 TRANON-Greedy(Ψ, $F(\Delta)$, k)

Input: $\Psi = \{\psi_1, \ldots, \psi_{|HOSP|}\}$, the set of disclosed identified databases; $F(\Delta) = \{F(\delta_1), \ldots, F(\delta_{|HOSP|})\}$, the set of encrypted databases sent to the central authority by the hospitals; k, an integer specifying the protection parameter.
Output: $F(\Phi) = \{F(\phi_1), \ldots, F(\phi_{|HOSP|})\}$, the set of encrypted databases to return to $H_1, \ldots H_{|HOSP|}$.

$F(\phi_1) \leftarrow \{\}, \ldots, F(\phi_{|HOSP|}) \leftarrow \{\}$
let $F(\Delta) \leftarrow$ *Contributor-Clean*(Ψ, $F(\Delta)$, k)
repeat
  $H_p \leftarrow \arg\min_{H_i \in HOSP} |F(\delta_i)| \geq k$
  $F(\phi_p) \leftarrow F(\delta_p)$
  for each $H_i \in HOSP$ **do**
    $F(\delta_i) \leftarrow F(\delta_i) - (F(\delta_i) \cap F(\phi_p))$
  end for
until $\forall H_i \in HOSP$: $|F(\delta_i)| \equiv 0$
return $F(\Phi)$

We developed two algorithms to be executed for *TRANON*. The first algorithm we present is called *TRANON-Greedy*, pseudocode shown in Algorithm 2. First, the algorithm executes a method called *Contributor-Clean*, pseudocode in Algorithm 3, which guarantees that data holders can not use private knowledge for re-identification purposes. Next, the algorithm iteratively chooses the location with the smallest dataset larger than size k, and allocates exclusive rights of the elements to the location. The third party will tell this location, but no other location, to disclose these elements.

Algorithm 3 Contributor-Clean(Ψ, $F(\Delta)$, k)

Input: See *TRANON-Greedy*.
Output: $F(\Omega) = \{F(\omega_1), \ldots, F(\omega_{|HOSP|})\}$, encrypted databases with cleaned pairwise non-intersections.

for each $H_i \in HOSP$ **do**
  if $|\psi_i| < k$ **then**
    $\psi_i \leftarrow \{\}$
  end if
end for
for all $H_i, H_j \in HOSP$ **do**
  if $\max(|\psi_i - \psi_i \cap \psi_j|, |\psi_i| - |F(\delta_j)|, |F(\delta_i) - F(\delta_i) \cap F(\delta_j)|) < k$ **then**
    $F(\omega_i) \leftarrow F(\omega_i) \cap F(\omega_j)$
  end if
end for
return $F(\Omega)$

Since *TRANON-Greedy* uses a greedy heuristic, we derived algorithm designed to mitigate its deficiencies. We call the algorithm *TRANON-Force*. It functions in the same manner as *TRANON-Greedy*, except it makes two passes over the datasets. First, it attempts to exclusively allocate k elements to each location. Then, it attempts to allocate the remaining elements to the set of locations allocated elements in the first round.

Output from both *TRANON-Greedy* and *TRANON-Force* satisfy the k-anonymity requirement for trails. The reader can verify that no location, including the third party, can perform trail matching to link decrypted data to less than k identities via their trails, unless they were aware of such knowledge beforehand.

3 Experiments and Results

The algorithms were evaluated on datasets derived from publicly available hospital discharge databases from the state of Illinois [4]. Seven populations diagnosed with single gene disorders were analyzed [3], including cystic fibrosis (CF), Friedreich's Ataxia (FA), hereditary hemorrhagic teleganictasia (HT), Huntington's disease (HD), Phenylketonuria (PK), sickle cell anemia (SC), and tuberous sclerosis (TS). Table 1 summarizes of the number of samples, hospitals, and presents a snapshot of re-identifiability prior to STRANON, and disclosure capabilities of STRANON. *TRANON-Greedy* and *TRANON-Force* exhibited similar results, so only results for *TRANON-Greedy* are presented for $k = 5$.

TRANON-Greedy permits disclosure of significant quantities of data in the face of re-identification. For instance, at $k = 5$, 90% of the elements in the HT dataset would have been re-identified (i.e. a DNA linked to less than 5 identities) if all locations were permitted to disclose all of their data. However, After execution of *TRANON-Greedy*,

Dataset	# of Samples	# of Locations	% Re-identified Before STRANON	% Disclosed After STRANON
CF	1149	174	52%	98%
FA	129	105	92%	33%
HD	419	159	84%	88%
HT	429	172	90%	93%
PK	77	57	91%	60%
SC	7730	207	37%	99%
TS	220	119	93%	78%

Table 1. Comparison of percentage of samples re-identifiable prior to STRANON and disclosed after STRANON, with 0 re-identifications, for $k = 5$.

we are able to disclose 93% of the samples with zero re-identifications. Similar findings are observed for the other databases. A more detailed plot of CF dataset, with results for k from 2 to 50, is shown in Figure 1.

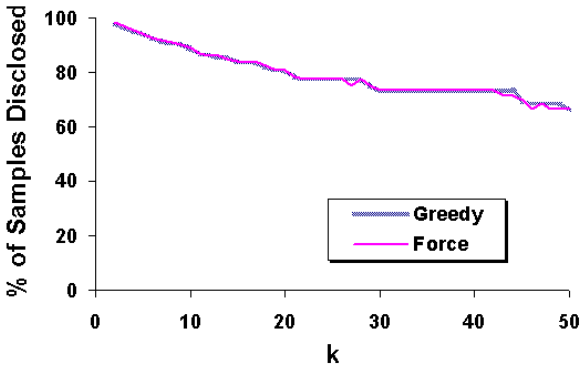

Figure 1. Percent of samples disclosed from the CF databases.

In addition, we evaluated the protocol in a simulated environment consisting of 1000 patients and 100 hospitals. Populations were generated according to a uniform distribution with $Pr(\text{patient } s \text{ visits } H) = 0.5$ for all patients and hospitals. Twenty-five simulations were run for each level k from 2 to 100, where vertical bars correspond to +/- 1 standard deviation. In Figure 2 we depict the number of de-identified elements disclosed for varying levels of k. We observe the *TRANON-Force* completely dominates *TRANON-Greedy*, and it appears the former utilizes a superior heuristic. This is expected, however, in some of our investigations with the real world datasets (such as the CF dataset in Figure 1), we observe that at times *TRANON-Greedy* appears to dominate *TRANON-Force*. In the future, we expect to study the affect of data distribution on the two algorithms.

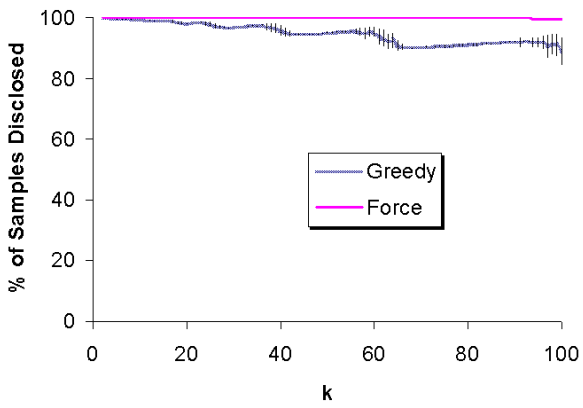

Figure 2. Mean percent of elements disclosed in simulated populations.

4 Discussion and Conclusions

The STRANON protocol addresses a data privacy challenge which arises in distributed systems. It enables a set of independent data holders to collaborate in an encrypted system to provably prevent location-visit patterns from playing a role in re-identification. Moreover, the protocol is applicable within current data privacy policies, such as recent federal health data privacy regulations. We provided experimental validation and showed significant quantities of data can be disclosed with zero re-identification risk. Nonetheless, there are limitations to the STRANON protocol. For example, for cryptographic purposes, STRANON is dependent on a hash function incapable of preserving string similarity. Thus, if a patient's de-identified data is variable across data collectors, improper trails can be constructed. There exist several promising similarity-preserving alternatives and in future research, we intend to evaluate their effectiveness within the protocol.

References

[1] Dept of Health and Human Services. Standards for privacy of individually identifiable health information, final rule. *Federal Register*, 45 cfr, 160-164. aug 12, 2002.

[2] B. Malin, E. Airoldi, S. Edoho-Eket, and Y. Li. Configurable security protocols for multiparty data analysis. In *IEEE ICDE*, pages 533–544, 2005.

[3] B. Malin and L. Sweeney. How (not) to protect genomic data privacy in a distributed network. *J Biomed Info*, 37(3):179–192, 2004.

[4] L. Sweeney. Weaving technology and policy together to maintain confidentiality. *JLME*, 25:98–110, 1997.