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, $\psi_i$, contains identifiable attributes, such as name. The second table $\delta_i$, contains de-identified attributes, such as DNA.

The trail for data element $x$ is a vector $\left[v_{x,1}, \ldots, v_{x,|HOSP|}\right]$, 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 $\epsilon_i$ 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 $\kappa_i$, such that $F(F(F(x, \epsilon_i), \epsilon_j), \kappa_i), \kappa_j) = x$.

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,
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 telangiectasia (HT), Huntington’s disease (HD), Phenylketonuria (PK), sickle cell anemia (SC), and tuberous sclerosis (TS). Table 1 summarizes 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,
Table 1. Comparison of percentage of
samples re-identifiable prior to STRANON
and disclosed after STRANON, with 0 re-
identifications, for \( k = 5 \).

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

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.

In addition, we evaluated the protocol in a simulated en-
vironment consisting of 1000 patients and 100 hospitals.
Populations were generated according to a uniform distri-
bution 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 ob-
serve the TRANON-Force completely dominates TRANON-
Greedy, and it appears the former utilizes a superior heuris-
tic. This is expected, however, in some of our investigations
with the real world datasets (such as the CF dataset in Fig-
ure 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.

4 Discussion and Conclusions

The STRANON protocol addresses a data privacy chal-
lenge which arises in distributed systems. It enables a set of
independent data holders to collaborate in an encrypted sys-
tem 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 fed-
eral health data privacy regulations. We provided experi-
mental validation and showed significant quantities of data
can be disclosed with zero re-identification risk. Nonethe-
less, there are limitations to the STRANON protocol. For
example, for cryptographic purposes, STRANON is depen-
dent on a hash function incapable of preserving string sim-
ilarity. Thus, if a patient’s de-identified data is variable
across data collectors, improper trails can be constructed.
There exist several promising similarity-preserving alterna-
tives and in future research, we intend to evaluate their ef-
ficacy 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. Config-
urable security protocols for multiparty data analysis. In IEEE
[4] L. Sweeney. Weaving technology and policy together to main-