Malin B and Sweeney L. A secure protocol to distribute unlinkable health data. In *Proceedings of the 2005 American Medical Informatics Association Annual Symposium*. Washington, DC. 2005: 485-489. *Paper selected as Student Paper Competition Finalist.*

A Secure Protocol to Distribute Unlinkable Health Data

Bradley Malin MS and Latanya Sweeney PhD
Institute for Software Research International, School of Computer Science
Carnegie Mellon University, Pittsburgh, Pennsylvania

Health data that appears anonymous, such as DNA records, can be re-identified to named patients via location visit patterns, or trails. This is a realistic privacy concern which continues to exist because data holders do not collaborate prior to making disclosures. In this paper, we present STRANON, a novel computational protocol that enables data holders to work together to determine records that can be disclosed and satisfy a formal privacy protection model. STRANON incorporates a secure encrypted environment, so no data holder reveals information until the trails of disclosed records are provably unlinkable. We evaluate STRANON on real-world datasets with known susceptibilities and demonstrate data holders can release significant quantities of data with zero trail re-identifiability.

INTRODUCTION

The ability to share patient-specific health data is essential to facilitate research in biomedical informatics. At the same time, it is necessary to uphold patient privacy rights. For protection, state and federal regulations, including the Privacy Rule of the Health Insurance Portability and Accountability Act[1], require data holders to render personal health information anonymous prior to various disclosures. Until recently, anonymity was assumed when data was stripped of explicit identifying information, such as personal name or Social Security Number. However, an increasing number of investigations prove *ad hoc* de-identification methods do not guarantee the anonymity of health data, including genomic data records.[2-4] This paper rectifies a known vulnerability of current de-identification[4] methods and presents a computational method to provably anonymize data.

In a recent study, we reported existing genomic data privacy protection systems are open to several types of re-identification.[4] To counteract these attacks, formal methods, based on binning, generalization, and perturbation of DNA sequences are under development.[5,6] These methods strive to suppress unintended inferences of phenotype that genomic data can reveal. In general, the set of emerging protection techniques are a promising start to the design and evaluation of formal genomic data privacy protection models. Nonetheless, even when genomic records are not susceptible to such inferences, there remain additional re-identification threats.

In prior research, we illustrated de-identified records, such as DNA sequences, could be mapped to corresponding identities via unique patterns in location visits, or trails.[3] At the time we provided automated methods for achieving trail re-identification, but we offered no protection solution. To date, no solution has been offered, but trail re-identification remains a concern because significant portions of patient populations are at risk.

A fundamental challenge to the development of methods to prevent trails re-identification stems from a lack of support for communication between data holders. Specifically, open communication is hindered because it can comprise the anonymity of data the holders intend to protect.

We overcome the communication barrier and present the *S*ecure *TR*ail *ANON*ymizer, or STRANON, protocol. The protocol enables locations to cooperate such that de-identified records are not revealed until it is guaranteed that trail re-identification can not be achieved beyond a controlled parameter. STRANON is designed to maximize the number of distinct records released, as well as the number of locations releasing data. In this paper, we demonstrate how STRANON can facilitate in the construction of a research repository of de-identified data that is unlinkable to identified patients via trails, but remains distributed across data collectors.

BACKGROUND

Health Data Anonymity

Personal health information exists in a number of different formats. As a result, automated methods for anonymizing health records are approached from varying perspectives. For example, techniques designed to find and replace personally-identifying information in free-text include manually defined detection algorithms[7], natural language processing[8], and term trained classification systems[9].

In contrast, an alternative set of methods have been constructed to anonymize relational databases. The Datafly[2] and CellSecu[10] systems generalize and suppress values according to domain specific hierarchies. These methods were extended to binning, which has been used to generalize genomic data

sequences.[5] In addition, random perturbation methods for genomic data have been evaluated.[6]

The anonymization method specified in the STRANON protocol is related to generalization and suppression. Notably, it is similar to cell suppression strategies proposed by Vinterbo et al.[11] Yet, while their strategy is sufficient to anonymize a single data holder's database, it does not account for varying levels of knowledge distributed across locations.

Trail Re-identification

Consider a world where three patients *Ali*, *Bob*, and *Dan* visit hospitals H_1, H_2, and H_3. Every hospital records personally-identifiable (PI) data. In some cases, DNA data is collected as well. For certain purposes PI is always shared, such as for insurance processing or discharge record keeping. In Figure 1, the identified dataset disclosed by hospital H_x is represented as I_x. In addition, hospital H_x discloses dataset D_x in which DNA data is stripped of corresponding names.

D_1	D_2	D_3	I_1	I_2	I_3
dna_3	dna_1	dna_1	*Bob*	*Ali*	*Ali*
		dna_2	*Dan*	*Dan*	*Bob*

Figure 1. DNA (D) and personally-identifiable (I) datasets shared by three hospitals.

A recipient of the disclosed datasets constructs data-location visit matrices as shown in Figure 2. In these matrices, a trail is a row vector and each value corresponds to the presence or absence of data in a hospital's disclosed dataset. For example, dna_1[0,1,1] conveys dna_1 was observed in D_2 and D_3.

	H_1	H_2	H_3
dna_1	0	1	1
dna_2	0	0	1
dna_3	1	0	0

	H_1	H_2	H_3
Ali	0	1	1
Bob	1	0	1
Dan	1	1	0

Figure 2. Trails from disclosed datasets in Figure 1.

When a DNA trail x can be 0→1 bit flipped into an identified trail y, it is said to be a subtrail of y. Note, by the problem definition DNA is collected only with PI, so a patient's DNA trail can always be converted into the patient's PI trail by flipping 0's to 1's.

In Figure 3, the left matrix depicts which DNA are subtrails of which names. Previously we introduced he REIDIT-I algorithm, which iteratively searches for unique linkages in this matrix.[3] Since dna_1 is a subtrail of *Ali* only, they are correctly re-identified to each other. Both are removed from consideration in the next iteration, and in Figure 3's right matrix, dna_2 is re-identified to *Bob*. The link-remove process is iterated until no more re-identifications can be made.

To anonymize trails we must guarantee that such re-identifications are impossible.

	Ali	*Bob*	*Dan*
dna_1	**1**	0	0
dna_2	1	1	0
dna_3	0	1	1

→

	Ali	*Bob*	*Dan*
dna_1	0	0	0
dna_2	0	**1**	0
dna_3	0	1	1

Figure 3. First two iterations of REIDIT-I algorithm over the subtrail matrix. Re-identifications are shown in black cells.

METHODS

Hospitals can not share patient-specific health databases if they can not assure the anonymity of their data. Yet, visit patterns across hospitals must be accounted for to ensure anonymity. To solve this paradox, we designed STRANON to anonymize trails in a secure encrypted environment. The protocol consists of two components: a secure multiparty computation model, and a trail anonymization method.

Secure Multiparty Communication

Recently, we introduced a general framework for secure multiparty data comparison.[12] For STRANON, we define a specific implementation of the framework. From a high-level perspective, STRANON enables hospitals to submit encrypted data to a third party (TP). The TP analyzes the submissions, and responds to each hospital with encrypted feedback that only the hospital, not even the TP, can learn the plaintext contents of.

The protocol leverages commutative encryption.[13] Each hospital encrypts and decrypts data using keys that satisfy the property:

$$F(F(dna, y_1), y_2) = F(F(dna, y_2), y_1)$$

for any ordering of values y_i and a function F. Thus, if $dna_i = dna_j$, then $F(F(dna_i, y_1), y_2) = F(F(dna_j, y_2), y_1)$, and the TP can perform correct data comparisons without observing the real value dna_i. In addition, when key y_i is paired with an appropriate key z_i, such as in RSA[14], the original data value is recovered using the same function, so

$$F(F(F(F(dna, y_1), y_2), z_1), z_2) = dna.$$

The general procedure of the STRANON protocol is depicted in Figure 4. Let *HOSP* be the set of participating hospitals. Each $H \in HOSP$ maintains private key pair $\langle y_H, z_H \rangle$ and private dataset D_H. First, each hospital's dataset is encrypted by every hospital's key. Next, the encrypted datasets are sent to the TP, who runs the TRANON anonymization algorithm with protection parameter k and responds to each hospital with a return dataset of encrypted values it can disclose. Finally, each hospital decrypts every return dataset, and the values are disclosed.

As described, the protocol is insecure and leaks certain information, but elsewhere[12] we show the protocol can be secured. Specifically, it can be shown that 1) no set of hospitals can collude to learn the

contents or size of another location's dataset and 2) no hospital can deviate from the protocol without being detected.

1. **for each** $H \in HOSP$
   1.1. $M_H \leftarrow F(D_H, y_H)$
   1.2. **for each** $P \in HOSP$ $(P \neq H)$
      1.2.1. H sends M_H to P
      1.2.2. P sends $M_H \leftarrow F(M_H, y_P)$ to H
2. Each $h \in HOSP$ sends M_H to TP
3. TP executes $TRANON(M_1,\ldots M_{|HOSP|}, k)$ to generate encrypted datasets $N_1,\ldots N_{|HOSP|}$
4. **for each** $H \in HOSP$, TP sends N_H to H
5. **for each** $H \in HOSP$
   5.1. **for each** $P \in HOSP$ $(P \neq H)$
      5.1.1. H sends N_H to P
      5.1.2. P sends $N_H \leftarrow F(N_H, z_P)$ to H
   5.2. $N_H \leftarrow F(D_H, z_H)$
6. Each $H \in HOSP$ discloses N_H

Figure 4. General execution of the STRANON protocol.

Trail Anonymity

As mentioned earlier, the scenario we address is the construction of a de-identified data research repository. For such a repository, we assume only one copy of a data sample is needed. By the problem description, identified datasets are always disclosed. We do not want to inject false information into the system, so the trail anonymization algorithm, or TRANON, suppresses data from de-identified datasets.

TRANON notifies the TP which encrypted data can be shared by which hospital, such that trails of disclosed data can not be linked to their identities beyond a specified parameter. The privacy parameter in TRANON corresponds to the k in Sweeney's k-anonymity model.[15] For every de-identified trail there exist no less than k identified trails to which it could be re-identified. Transforming data to satisfy k-anonymity is a computationally challenging problem, so TRANON employs several greedy heuristics to maximize both the number of distinct samples released, as well as the number of hospitals which can release data. Pseudocode for the algorithm is provided in Figure 5. A less informal specification follows.

We call the dataset the TP sends back to a hospital the response. There are two data allocation procedures used to construct the responses. The first is a heuristic designed to boost the number of locations with non-null responses. In step 3, each hospital is allocated k encrypted samples (from its submission) to its response until either every hospital has been allocated k samples or no more hospitals can not be allocated any samples. Once a sample is added to a response it is prevented from being added to any other hospital's response. Samples are allocated to a hospital response using a maximum likelihood prediction. Simply, the samples selected have the lowest probability of being observed in any randomly chosen submission.

If a response can not be allocated k samples, then in step 4, the response is guaranteed to be allocated 0 samples. This is to make sure that no location releases a dataset of size less than k which would be in direct violation of the k-anonymity requirement. The reasons behind this are subtle and beyond the scope of this paper, but are addressed in an extended version.[16]

Second, in step 5, the hospital with the minimal sized submission is allocated the remaining encrypted samples from its submission. Again, these samples are removed from all other hospitals. This procedure continues until no more hospitals can be allocated samples for release. Once TRANON terminates, the TP tells each hospital which encrypted samples to release.

TRANON(Δ, k **)**
Input: $\Delta = \{D_1,\ldots,D_{|HOSP|}\}$, the set of encrypted datasets submitted to the central authority by hospitals $H_1, \ldots, H_{|HOSP|}$. k, an integer specifying the protection parameter to be applied.
Output: $R_1,\ldots,R_{|HOSP|}$, the encrypted datasets the third party sends to $H_1,\ldots H_{|HOSP|}$, respectively
Steps:
1. **for each** $D_i, D_j \in \Delta$
   1.1. **if** $|D_i| - |D_i \cap D_j| < k$, **then** $D_i \leftarrow D_i \cap D_j$
2. Let $USED \leftarrow \varnothing$
3. **for** $i \leftarrow 1$ to $|HOSP|$
   3.1. Let $H \leftarrow$ dataset of smallest size $|D_H| \geq k$, such that $X \notin USED$
   3.2. $USED \leftarrow USED \cup \{X\}$
   3.3. *Let* D_X^k be the k samples of D_X that occur in the least number of datasets in Δ
   3.4. **for** $p \leftarrow 1$ to $|HOSP|$
      3.4.1. **if** $p \neq i$, **then** $D_P \leftarrow D_P - (D_P \cap D_i^k)$
4. **for** $i \leftarrow 1$ to $|HOSP|$
   4.1. **if** $|D_i| < k$, **then** $D_i \leftarrow \varnothing$
5. **for** $i \leftarrow 1$ to $|HOSP|$
   5.1. $H \leftarrow$ dataset of smallest size $|D_H| > 0$
   5.2. **if** $p \neq i$, **then** $D_P \leftarrow D_P - (D_P \cap D_i)$
6. **return** $R_1 \leftarrow D_1,\ldots, R_{|HOSP|} \leftarrow D_{|HOSP|}$

Figure 5. Pseudocode for the TRANON algorithm.

Steps 2 through 5 of TRANON guarantee that for every trail constructed from the disclosed datasets, there exist at least k-1 equivalent trails. However, the released trails are not sufficiently protected because the hospitals possess undisclosed knowledge. Consider hospitals H_1 and H_2 with encrypted DNA submissions D_1 and D_2. Remove the intersection. If the remaining number of samples in D_1 is less than k, then H_1 can not release any of the remaining DNA samples. This is because, even if H_1's disclosed dataset has size greater than k, once H_2 remove DNA and identities he already knows are linked, he will be left with DNA samples that have trails mapped to less than k identities. Details regarding this concern are addressed in depth elsewhere[16], but are accounted for by Step 1. In combination, steps 1 through 5 of TRANON guarantee disclosed trails are unlinkable by any recipient.

Complexity**.** The computational complexity of TRANON can be calculated as follows. In Step 1, datasets are reduced based on the result of their intersection tests. This can be performed in O(|*HOSP*|log|*HOSP*|) comparisons. In Step 3, the data is each hospital is allocated k responses which are removed from all other hospitals. Assuming k is a relatively small constant, this requires O(|*HOSP*|^2) steps. In Step 4, datasets of an insufficient size are nulled, which is a simple linear scan in O(|HOSP|) steps. In Step 5, the second round of data allocation and removal is performed in O(|*HOSP*|^2). Taking the maximum over the steps, complexity is O(|*HOSP*|^2) and the algorithm can be executed in real time.

k-Trail Re-identification

For evaluation purposes, we describe an extension to the REIDIT-I algorithm. The original algorithm made unique re-identifications, so each DNA sample could be re-identified to one identity only. However, to evaluate privacy at varying levels of k we extend the REIDIT-I algorithm, which we call REIDIT-I-k, to re-identify a DNA sample to k identities.

Basically, for each de-identified trail, let T_d be the set of identified trails which d is a subtrail of. If $|T_d| < k$, then for all $t \in T_d$, add $<d, t>$ to the set of re-identifications. If $|T_d| = 1$, then re-identify $<d, t>$ and remove t from further consideration.

RESULTS

The trail anonymization algorithm was evaluated on real world datasets, derived from publicly available hospital discharge data from the state of Illinois[17], which in our prior studies were shown to be at risk to trail re-identification.[3] Seven populations diagnosed with single gene disorders were analyzed. The populations are cystic fibrosis (CF), Friedrich's Ataxia (FA), hereditary hemorrhagic teleganictasia (HT), Huntington's disease (HD), Phenylketonuria (PK), sickle cell anemia (SC), and tuberous sclerosis (TS).

		k=5		k=10	
	Samples	Re-identified	Dis-closed	Re-identified	Dis-closed
CF	1149	0.52	0.98	0.58	0.89
FA	129	0.92	0.76	1.00	0.19
HT	429	0.90	0.93	0.97	0.40
HD	419	0.84	0.88	0.97	0.28
PK	77	0.91	0.60	1.00	0.00
SC	7730	0.38	0.99	0.41	0.99
TS	250	0.93	0.78	1.00	0.46

Table 1. Fraction of samples re-identified without TRANON and disclosed via TRANON with zero re-identifiability.

Table 1 provides a summary of the datasets and a snapshot of results. For k=5 and k=10, it shows the fraction of samples that would be re-identified using REIDIT-I-k if hospitals do not communicate and the fraction of samples released with zero trail re-identifiability after running TRANON for k=5 and 10. To highlight some results, notice that for the TS cohort, 100% of the population is re-identifiable. However, if TRANON is used, then 45% of the samples can be disclosed and are provably unlinkable.

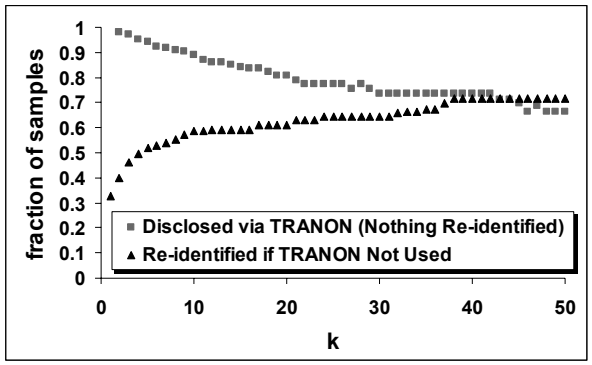

Figure 6. Fraction of samples in the CF dataset released via TRANON. (|*HOSP*|=174, |*S*|=1149)

We explore TRANON more in-depth with the CF cohort varying k from 1 to 50. This cohort consists of 174 locations and 1149 DNA trails. In Figure 6, the increasing trend corresponds to the number of samples re-identifiable via REIDIT-I-k if all data is disclosed. In contrast, the decreasing trend line displays the number of distinct samples that are released from the set of hospitals. Of particular interest is that the rate of change in the ability to disclose data is not as rapid as the ability to make re-identifications. By k=50, 70% of the system is initially re-identifiable, but TRANON can safely disclose almost the same quantity of data.

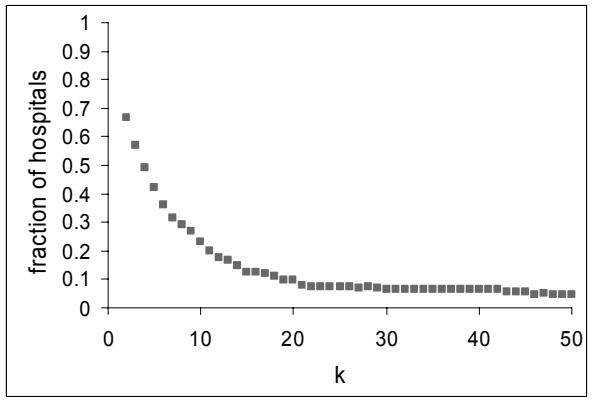

Figure 7. Fraction of hospitals in the CF dataset able to release via TRANON. (|*HOSP*| = 174, |*S*| = 1149)

In Figure 7 we show that TRANON is able to distribute data with success for the CF population. Note, at k=5, a typical protection level used by the

Census Bureau, TRANON can release 98% of the samples distributed across 80, or 40%, of the hospitals. By $k = 10$, the number of releasing hospitals remains above 50.

DISCUSSION AND CONCLUSIONS

STRANON is the first protocol to show how to simultaneously distribute data and guarantee trail re-identification can not be achieved. It does so in an encrypted space where no hospital reveals any samples until every trail resulting from disclosers are guaranteed to be sufficiently unlinkable. Moreover, we demonstrated it is applicable to real world populations.

One of the drawbacks to the STRANON protocol is its reliance on commutative cryptography. As a result, there is no fault tolerance built into the anonymization procedure. If a patient's DNA data is represented slightly differently at various locations, such as may occur from natural variation in DNA sequence samples, the encryption function obscures all observable similarities. Thus, record linkage methods for relating distributed samples are not applicable. This is a concern, but not insurmountable. Locations can design ontologies or standardized representations to minimize variation prior to encryption. Additionally, fuzzy representations, such as nucleotide generalizations, can help minimize missed linkages in the encrypted space. We intend to further develop these ideas in future research.

A second drawback to STRANON is it does not explicitly model the honesty of a location's behavior outside of the protocol. STRANON guarantees that every location will execute the protocol correctly before any information is revealed, however, it does not guarantee that the datasets submitted by the participating locations are truthful. This is a concern and a direction for future research. Specifically, we are interested in designing protocols to incorporate knowledge which permits the central authority to validate that submitted datasets are representative. Furthermore, we intend to design protocols to allow the participating locations to validate the honesty of the results sent by the central authority.

STRANON addresses the issue of unlinkability in patient-location visit patterns. It does not explicitly address the issue of phenotype inferences which may be applicable for linkage purposes. Thus, STRANON is a complement, not a substitute, for inference disclosure control.

Acknowledgements

This research was funded in part by the Data Privacy Laboratory at Carnegie Mellon University and NSF IGERT grant 9972762 in CASOS. The authors wish to thank the anonymous referees for useful comments.

REFERNCES

1. Federal Register, 45 CFR, 160–164. Standards for privacy of individually identifiable health information, Final Rule. Aug 12, 2002.
2. Sweeney L. Guaranteeing anonymity when sharing medical data, the Datafly system. *Proc AMIA Symp*. 1997: 51-55.
3. Malin B, Sweeney L. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *J Biomed Info*. 2004; 37(3): 179-192.
4. Malin B. An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. *JAMIA*. 2005; 12(1):28-34.
5. Lin Z, Hewitt M, Altman R. Using binning to maintain confidentiality of medical data. *Proc AMIA Symp*. 2002: 454-458.
6. Lin Z, Owen A, Altman R. Genomic research and human subject privacy. *Science*. 2004; 305: 183.
7. Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. *Proc AMIA Symp*. 1996: 333-337.
8. Ruch P, et al. Medical document anonymization with a semantic lexicon. *Proc AMIA Symp*. 2000: 729-733.
9. Taira RK, Bui AA, Kangarloo H. Identification of patient name references within medical documents using semantic selectional restrictions. *Proc AMIA Symp*. 2002: 757-761.
10. Chiang YC, et al. Preserving confidentiality when sharing medical database with the Cellsecu system. *Int J Med Info*. 2003 Aug; 71(1): 17-23.
11. Vinterbo S, Ohno-Machado L, Dreiseitl S. Hiding information by cell suppression. *Proc AMIA Symp*. 2001: 726-730.
12. Malin B, et al. Configurable security protocols for multi-party data analysis with malicious participants. *Proc IEEE ICDE*. 2005: 533-544.
13. Benaloh J, deMare M. One-way accumulators: a decentralized alternative to digital signatures. *LNCS 765: Proc Eurocrypt '93*. 1994: 274-286.
14. Rivest RL, Shamir A, Adleman L. A method for obtaining digital signatures and public-key cryptosystems. *CACM*. 1987; 21(2): 120-126.
15. Sweeney L. *k*-anonymity: a model for protecting privacy. *Int J Uncertainty, Fuzziness and Knowledge-Based Sys*. 2002; 10(5): 557-570.
16. Malin B, Sweeney L. A secure protocol to distribute unlinkable health data. Data Privacy Lab *Working Paper LIDAP-WP21*, CMU. 2005. Available online at http://privacy.cs.cmu.edu/dataprivacy/projects/trails/tranon.pdf.
17. State of Illinois Health Care Cost Containment. *Data release overview.* Springfield, IL. 1998.