

The Effects of Location Access Behavior on Re-identification Risk in a Distributed Environment

Bradley Malin and Edoardo Airoldi

Institute for Software Research International, School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213 USA
{malin, eairloidi}@cs.cmu.edu

Abstract. In this paper, we investigate how location access patterns influence the re-identification of seemingly anonymous data. In the real world, individuals visit different locations that gather similar information. For instance, multiple hospitals collect health information on the same patient. To protect anonymity for research purposes, hospitals share sensitive data, such as DNA sequences, stripped of explicit identifiers. Separately, for administrative functions, identified data, stripped of DNA, is made available. On a hospital by hospital basis, each pair of DNA and identified databases appears unlinkable, however, links can be established when multiple locations' database are studied. This problem, known as trail re-identification, is a generalized phenomenon and occurs because an individual's location access pattern can be matched across the shared databases.

Data holders can not exchange data to find and suppress trails that would be re-identified. Thus, it is important to assess the re-identification risk in a system in order to develop techniques to mitigate it. In this research, we evaluate several real world datasets and observe trail re-identification is related to the number of people to places. To study this phenomenon in more detail, we develop a generative model for location access patterns that simulates observed behavior. We evaluate trail re-identification risk in a range of simulated patterns and our findings suggest that the skew of the distribution of people to places is one of the main factors that drives trail re-identification.

1 Introduction

DNA sequences are becoming an integral part of electronic patient medical records [1, 2]. Collections of detailed genomic data that are tied to clinical information are poised to yield significant healthcare breakthroughs [3], ranging from personalized medicine to drug discovery. However to share person-specific genomic data collections for research, data holders must adhere to legal regulations, such the Privacy Rule of the Health Insurance Portability and Accountability Act [4]. Though an individual's genome is unique, a database of DNA records, with no accompanying explicit demographic information or identifiers included, appears anonymous. But patients leave information behind at multiple institutions and the collections are autonomously controlled. As a result, the location-access patterns, or trails, of an individuals DNA can be extracted from shared databases. DNA trails are not necessarily re-identifiable, but publicly available information, such as hospital discharge databases, are available and

reveal identified individuals' trails. Uniqueness of an individuals discharge and DNA trails leads to re-identification.

Healthcare is one realm in which trail re-identification poses a privacy threat [5], but trails arise in many other environments [6, 7]. Though domains change the goal remains constant: share data such that the identity of sensitive information, can not be linked to the individual from which it was derived. Privacy protection methods have been proposed and adopted, such as [8] and [9], which advocate the removal or encryption of explicit identifiers associated with sensitive data. However, such methods do not prevent trail re-identification since the identities of the individuals are available in other shared or public databases.

Trail re-identification is a real concern. Inability to address the problem will prevent organizations, such as biomedical data holders, from sharing data [3, 10]. As an alternative to ad hoc protection methods, we propose formally evaluating the re-identification risk of a set of database entries prior to release.¹ The actual number of re-identifications can be measured as the number of shared database entries that are re-identifiable. Yet, when data can not be shared prior to re-identification evaluation, we must approximate the number of re-identifications that can be made. To do so we need to isolate the processes that influence re-identification, such as 1) the data generating process (e.g. How do people visit places?) and 2) the re-identification process (e.g. How are trails linked?) [12]. Then, for a given method of re-identification, substitute characteristics of location access patterns, as opposed to the actual patterns, to estimate of re-identification risk.

In this paper, we model the underlying processes governing trail re-identification to evaluate risk in a distributed environment. We have two goals. First, we tie together results from our previous case studies to conjecture how the number of people and locations in a system relates to the number of re-identifications that can be made. Second, we step back from specific cases and develop a statistical model to examine why different populations have varying degrees of re-identification in a distributed environment. Using this model, we then simulate several fundamental location visit strategies employed by individuals in the real world and assess the re-identification risk they entail.

The remainder of this paper is organized as follows. In the following section we review the formal basis and methods for trail re-identification. The methods are amenable to combinatoric proof, which suggests that the number of re-identifications scales with the number of subjects and locations. However, with real world populations, we demonstrate that such scaling does not exist. In addition, we show the power law feature of online environments, as well as how highskew populations are generated. Next, we simulate and perform linkage analysis on several types of simulated datasets corresponding to a range of distributions. Then, we investigate the relationship of trail re-identification risk to information theoretic principles. Finally, this work addresses limitations and extensions for future research.

¹ Provable solutions to guarantee trails can not be re-identified exist [11], but they require the use of third parties, which are not always practical due to trust or regulatory constraints.

2 Background

In this section we survey related research and provide an overview of several basic concepts for the trail re-identification problem.

2.1 Related Work

In the past, it was generally believed that person-specific data collections could be shared somewhat freely, provided none of the features of the data included explicit identifiers, such as name, address, or Social Security number. However, an increasing number of data detective-like investigations have revealed that collections of “de-identified” data, derived from ad hoc protection models, can often be linked to other collections that do include explicit identifiers to uniquely, and correctly, re-identify disclosed information by personal name [13–17]. Fields appearing in both de-identified and identified tables can link the two, thereby relating names to the subjects of the de-identified data. For example, Sweeney’s analysis of the fields $\{date\ of\ birth, gender, 5\text{-digit\ zip\ code}\}$, which, until recently, commonly appeared in both de-identified databases and publicly available identified data, such as voter registration lists, uniquely represented approximately 87% of the U.S. population [16].

Trail re-identification [5, 7] extends traditional re-identification and illustrates how the pattern of locations people visit, or trails, can be used for linkage. First, we provide an informal view of trail re-identification, which will be followed by a more formal presentation of the problem. The main premise of trail re-identification is based upon the observation that people visit different sets of locations where they can, and do, leave behind similar pieces of de-identified information. The de-identified data can consist of only one or very few fields. Each location visited collects and, subsequently, shares de-identified data on people who visited their location. In addition, locations also collect and share, in separate releases devoid of de-identified data, explicitly identified data (i.e. name, residential address, etc.), thereby naming some people. Individually, a single locations releases appear unrelatable, and thus identity and sensitive information appear unlinkable. However, when multiple locations share their respective data, this allows for trails, a characterization of the locations that an individual visited, to be constructed. Similar patterns in the trails of de-identified and identified data can then be used for linkage purposes.

The trail re-identification attack is related to other attack that have been studied in anonymous communications, such as the interaction attack [18, 19].

2.2 Elements of a Formal Re-identification Model

We now describe the problem in a more formal manner. Let L be a set of locations collecting data. At each location, data is organized as a database, which we model as a table of rows and columns. Each column corresponds to an attribute, which is a semantic category of information that refers to people, machines, or other entities. Each row contains attribute values specific to a person, machine, or other entity. A database is represented by $\tau(A_1, A_2, \dots, A_p)$, where the set of attributes is $A^\tau = \{A_1, A_2, \dots, A_p\}$ and each attribute is associated with its own domain of specific values. Each row

in the database is a p -tuple, which we represent in vector form $[a_1, \dots, a_p]$, such that each value a_i is in the domain of attribute A_i . We define the size of the database as the number of tuples and use cardinality, denoted with $|\tau|$.

A database τ is said to be *identified* if A^τ includes explicit identifying attributes, such as name or residential address, or attributes known to be directly linkable to explicit identifiers. If τ is not identified, then it said to be *de-identified*. Data holding locations attempt to protect the anonymity of sensitive data by stripping explicitly identifying attributes from sensitive data. In doing so, locations partition identified and de-identified data and make separate database disclosures. As such, in our model, each data holder releases a two-table vertical partition of its internal data by splitting τ into two tables $\psi(A_1, \dots, A_i)$ and $\delta(A_{i+1}, \dots, A_j)$, with attributes $A^\psi \subset A^\tau$ and $A^\delta \subset A^\tau$. For illustration, several tables are depicted in Figure 1.

2.3 The Trail Re-Identification Problem

Given the tables of a particular type (e.g. the sensitive data tables), we can construct a matrix X that is referred to as a trail matrix. The trail matrix X is the join of all locations' tables over a set of related attributes, such as when we trace an individuals DNA sequence from one location to another.² This matrix has a row for each distinct data element and $|L|$ columns, one for each location. Values in the matrix are drawn from $\{1, 0, *\}$. A "1" a cell denotes the data element for the row definitely visited the location corresponding to the column, while a "0" denotes a definite non-visit. A "*" is an ambiguous value and indicates that we are unsure if the data element was collected at the location. We use $X[x, :]$ to denote the trail of data element x in trail matrix X .

The basis behind trail re-identification is that there exist two different types of data collected at the set of locations in the environment. Thus two trail matrices, X and Y , can be constructed, and it assumed that both trail matrices are drawn from the same population of entities. An example of trail matrices are depicted in Table 2(a).

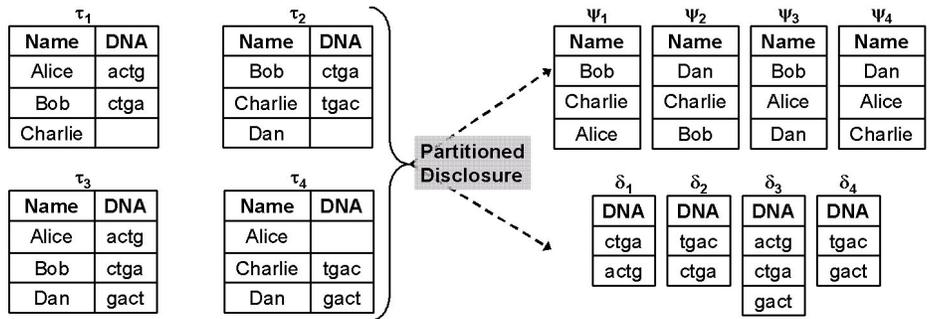


Fig. 1. Sample disclosures for four locations.

² This join can be constructed from traditional record linkage algorithms for tables with common attributes [20, 21].

The main distinguishing feature of trail re-identification algorithms is their characterization of data completeness. Trail matrices are said to be *unreserved*, if an entity's data is always collected and disclosed from a location. In some situations, a location can collect data of both types, but it undercollects (or underreports) data of one type (i.e. the data is not in the location's table). In this case, trail matrix X is said to be *reserved* to Y if the trail of each entity in matrix X , $X[x, :]$ can be transformed into the entity's corresponding $Y[y, :]$ in matrix Y by replacing only *'s with 0's and 1's. When this transformation can be performed, $X[x, :]$ is said to be a subtrail (represented with the \preceq symbol) of $Y[y, :]$. Similarly, y_Y is said to be the supertrail of $X[x, :]$, or $Y[y, :] \succeq X[x, :]$. Figure 2(a) provides an example of trail matrices where X is reserved to matrix Y . Notice $Y[actg, :] \preceq X[Alice, :]$ and $Y[actg, :] \preceq X[Charlie, :]$.

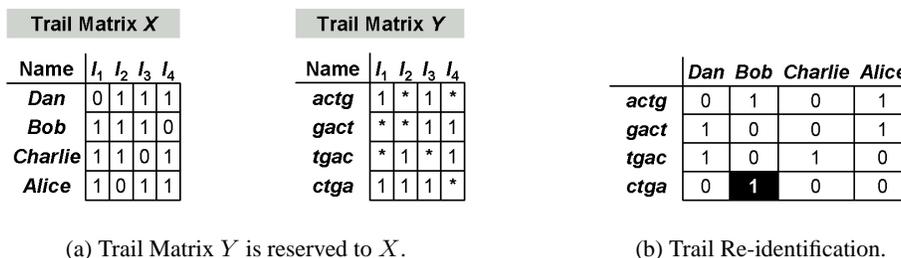


Fig. 2. (a) Trail matrices built from Fig. 1. (b) *Bob* is re-identified to *ctga* in the first iteration.

Recall, the goal of trail re-identification is to match the rows of two trail matrices to re-identify sensitive data to identity. In related research, [5, 7] introduced an algorithm called REIDIT (RE-Identification of Data In Trails) to perform such a task, such that every match is guaranteed to be a correct re-identification. Informally, REIDIT works as follows. First, we construct a $|Y| \times |M|$ matrix, called M , such that cell $M[i, j] = 1$ if $i_Y \preceq j_X$, and 0 otherwise. When we find a row or column that has only one cell $M[i, j] = 1$, we re-identify the corresponding data elements in the cell. We iterate this process until no more matches can be made. Figure 2(b) illustrates the initial matrix for Figure 2(a) and the first trail re-identification of *ctga* to *Bob* is made. In the next iteration *actg* will be re-identified to *Alice*, and so on.

3 Empirical Evidence: Lesson Learned

We assessed the feasibility of trail re-identification in several different domains. The first population we studied consisted of individuals visiting physical hospitals for treatment. The second population consisted of individuals visiting sites on the World Wide Web (i.e. a virtual world) for performing various functions, such as purchasing goods.

Healthcare Case Study. We analyzed the trails of DNA database records in a distributed healthcare environment. The observations were hospital discharge data for the

state of Illinois from 1990 to 1997 [22]. Trails were derived for eight different patient populations, each with a distinctive DNA-based disorder. In these populations, the entities were hospital patients and the locations were hospitals. The size of the populations ranged from 4 to 8,000 patients over 8 to 200 hospitals and the distribution of individuals to hospitals varies from uniform to approximately Gaussian, which are relatively low skew.

Internet Case Study. We studied the trails of IP addresses in a distributed online environment. The dataset used in this study was compiled by the Homenet project at Carnegie Mellon University, who provide families in the Pittsburgh area with Internet services in exchange for the monitoring and recording of the families' online services and transactions [23]. We studied URL access data collected over a two-month period that included 86 households and 144 individuals. Each individual was provided with a unique login and password for fine-grained monitoring. Overall, approximately 5,000 distinct website domains and 66,000 distinct pages were accessed. We analyzed the traffic at each domain with respect to the number of distinct visitors and discovered a generalized Zipf distribution, which represents high skew.

In both case studies, we found that re-identification rates correlate with the average number of people visiting a location. When we investigated this relationship in more detail, we found particular types of locations influence trail re-identification. For example, we ranked the popularity of each location by the number of distinct subjects visiting the location. When we measured trail re-identification from the least popular location to a location with a specific popularity, we found the re-identification rate correlated the average number of people per location. The result is shown in Figure 3, where we depict re-identification rates for three different populations. In Figure 3, the term "discovered" corresponds to the number of individuals' data that are observed given the set of locations that trails are constructed from. As we increase the number of locations considered, we increase the number of individuals that have their data discovered, but not necessarily re-identified.

The first two populations are derived from the healthcare case study. The first corresponds to a population afflicted with cystic fibrosis (CF) and the second to a population afflicted with phenylketonuria (PK). These two cases establish a comparison between the feasibility of trail re-identification on a population in which the number of subjects per location is relatively large (CF - approximately 6.60), with a population in which the average is closer to a single subject per location (PK - approximately 1.35). The third population corresponds to the online Homenet dataset, where the ratio of subjects per location is relatively small (approximately 0.017).

We observe that as the ratio of subjects per location grows large, such as in the CF dataset shown in Figure 3(a), we find evidence of an exponential relation between the number of locations considered (the X axis), and the number of people that are trail re-identifiable (the Y axis). As the ratio becomes negligible, as observed in the Homenet dataset in Figure 3(c), we find evidence of a logarithmic relation between the number of locations considered and the number of trail re-identifications. Furthermore, the PK dataset in Figure 3(b) supports this trend; in this case the ratio of people to locations is approximately 1, and we find evidence of a linear relation between the number of locations considered and the number of trail re-identifications.

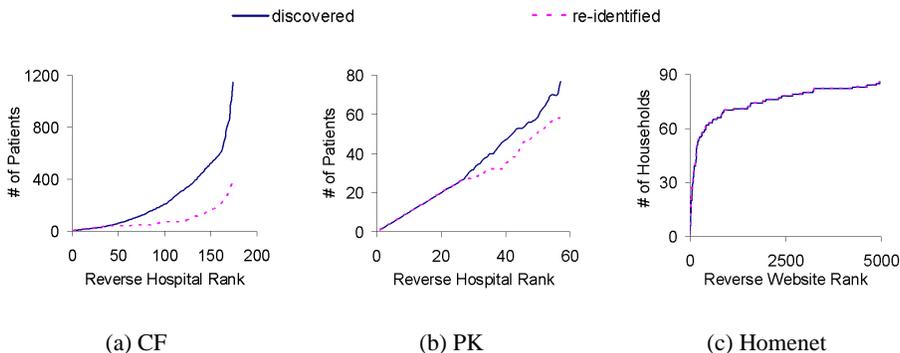


Fig. 3. Trail re-identification in unreserved systems for case studies. Number of locations increase from least-visited to most-visited.

The evidence from the case studies suggests that different types of location access patterns have an effect on trail re-identification. In the following section we study the degree to which specific types of access distributions influence re-identification.

4 Simulation Experiments and Results

There are many aspects of location-based information which influence trail re-identification. The main contributing components include the number of subjects, the number of locations, the distribution of subjects to locations, as well as the parameters controlling said distributions. In this research, we concentrate on the number of locations and the distributions guiding subject access to these locations. For our analysis, we fix the number of subjects to 1000. We simulate uniform and high skew distributions of subjects per location. We simulate both unreserved systems, i.e. neither trail matrix has *'s, and reserved systems, where one trail matrix has *'s. From an operational point of view, in the simulation of unreserved systems, we generate two equivalent trail matrices. In the simulation of reserved systems, instead, we generate trail matrices for an unreserved environment, and then we change all 0's in a matrix to *'s. For each distribution type and parameterization, these populations are allocations to sets of locations over the range of 3 to 40 locations.

Uniform Simulation. In this setting, subjects visit locations with uniform probability. We control the average number of subjects per locations, by specifying the probability that a subject visits each location, $p \in [0, 1]$. This sampling mechanism is from a location perspective. From a subject perspective, however, given that subjects act independently and there is no difference among locations, each subject's trail is a string of 0s and 1s, where the probability of observing a 1 at each location is also given by $p \in [0, 1]$. We perform different simulations by fixing p on a grid in.³

³ In theory, any number of points on the $[0, 1]$ interval will suffice.

Zipf Simulation. In this setting, subjects visit locations according to Zipf distributions, which lead to the desired high skew in the location access patterns. The set of available locations is denoted by L , and the population of subjects visiting those locations is denoted by S . The expected number of subjects who visit location $l_i \in L$ is equal to the mean of the corresponding distribution, e.g., equals $|S| \cdot r_i^{-\alpha}$, where r_i is the rank of l_i 's popularity, and α is a real number greater than zero. When α equals 1, then the distribution is a true Zipf and when $\alpha < 1$ the Zipf distribution is said to be in a generalized form. Given the high skew of the distribution, the log-log plot of “number of visitors” to “location rank” is linear, while the α coefficient serves as a dampening factor on the slope of the fitted curve. As with the uniform distribution, the Zipf is studied by varying the parameter α over the same interval $[0, 1]$, and sample points, as the p parameter of the uniform distribution. Note that the exponent of a Zipf distribution is allowed to vary in the larger interval, $\alpha \in (0, \infty)$, with $\alpha = 0$ corresponding to the case of a Zipf distribution that degenerates into a Uniform distribution, and $\alpha = 1$ corresponding to the case of moderate skew. Thus, our choice of studying the exponent in the smaller interval $[0, 1]$ allows us to explore how the re-identification risk changes as location access patterns smoothly change from uniform to skewed. An exponent larger than one would not add much to our study, beside adding coverage of different degrees of skewness, hence it is reasonable to truncate the range of α at 1. For example, the empirical evidence we presented in Section 3 supports (estimated) Zipf exponents as large as $\alpha = 0.6$. For each tested data point, such as $\langle |L| = 10, p = 0.3 \rangle$, we generate 100 populations. Populations that are guided by the Zipf distribution are generated using the formula described above.

4.1 Distribution Effect on Re-identification

The resulting 10-point plots for unreserved and reserved systems are depicted in Figures 4 and 5. In these plots the mean percentage and plus/minus one standard deviation⁴ for the 100 simulated populations are depicted. The x -axis corresponds to the parameter of the distribution in question and the y -axis corresponds to values of the mean percent of the population that is trail re-identified.

From the re-identification plots, though there is no direct way to compare the parameterizations of the uniform and Zipf distribution, there are several interesting observations that can be made. First with respect to both the unreserved and reserved systems, it is apparent that the uniform distribution consistently yields a larger number of re-identifications than the Zipf distribution. This can be seen by comparing the re-identification maximum, or peaks, in the left and right panels. Consider Figure 4, for example, in a situation with 10 locations, we re-identify a maximum of approximately 40% of the subjects distributed uniformly (which occurs when $p = 0.5$), as opposed to around 16% of the subjects that are distributed in Zipf high skew (which occurs when $\alpha = 0.4$). This finding is consistent across all systems as the number of the locations in consideration is increased.

Second, we consider a less readily observable feature that directly relates to the general success of re-identification, given a specific distribution for location access pat-

⁴ In Figure 4, the error bars are too small to be visible.

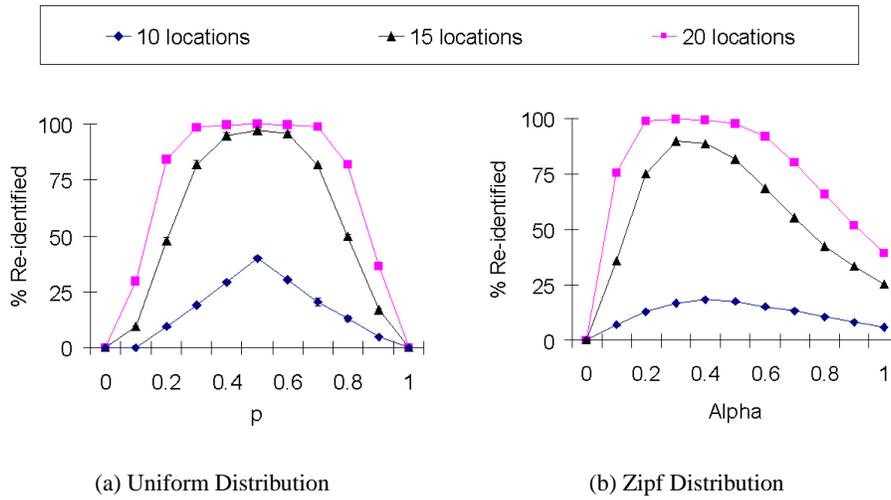


Fig. 4. Re-identification of simulated unreserved location access distributions.

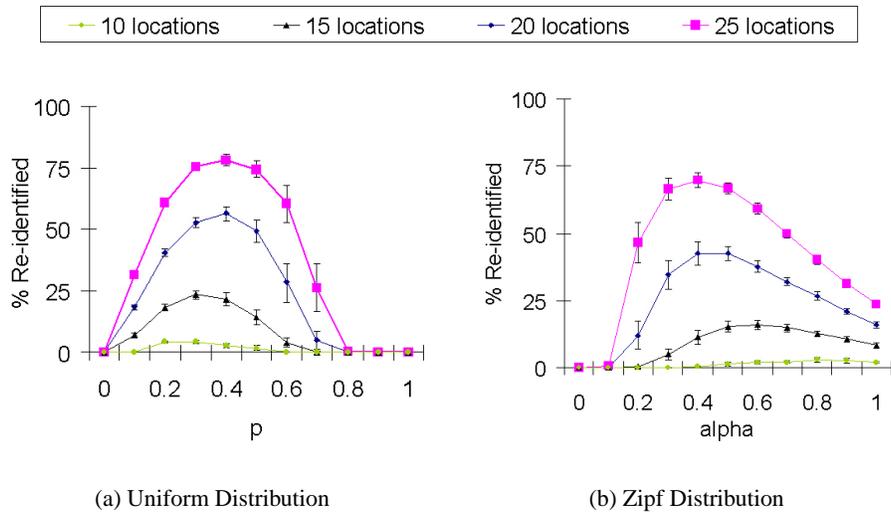


Fig. 5. Re-identification of simulated reserved location access distributions.

terns. To compare distribution archetypes, such as uniform vs. Zipf, we measure the area under the re-identification curve. This is calculated as the total area under the 10-point mean re-identification curve (average number of re-identifications in 100 simulated populations). The results of this calculation with respect to distributions and algorithm results are presented in Figures 6(a) and 6(b). Though the uniform distribution always yields the larger maximum number of re-identifications, the Zipf distribution is

almost always the more linkable when considering all parameterizations. This is obviously so in the case of the reserved system, where Figure 6(b) shows that the Zipf always dominates. Similarly, in an unreserved system, Zipf is both the initial and inevitable dominant. However, this analysis reveals an unanticipated and intriguing finding. In certain ranges, the uniform distribution is dominant to the Zipf! In Figure 6(a), this finding is observed between approximately 8 and 18 locations.

The flip in distribution linkage capability dominance occurs for two reasons. First, Zipf dominates when there are not many locations in consideration because it is more difficult to realize complete vectors of all 1's. Second, Zipf dominates as the number of locations increase because it is easier for lesser accessed locations, which is what the newly considered locations are, to convert an unlikely trail into an extremely unlikely trail.

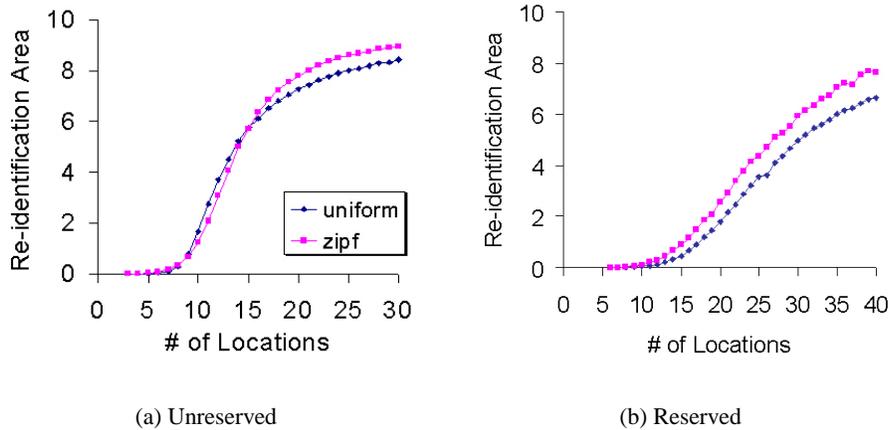


Fig. 6. Area under the mean re-identification curves for simulated populations.

4.2 Information and Re-Identifiability of a Distributed System

In this section we relate the re-identifiability of trails in a distributed system to the Shannon entropy of the set of trails. [24, 25]

Each trail is a Boolean vector of 0's and 1's, and, as such, we can compute its entropy as measure of information. If we consider all the possible trails with a given information score, we note that the more entropic a trail is, i.e., the more random looking an individual's location access pattern is, the larger is the set of trails that relate to it. Therefore, entropy is a measure that inversely relates to a notion of distinguishability of one trails from others. To what extent does this notion of distinguishability relate to the notion of distinguishability (via uniqueness and re-identifiability) we studied in the previous section? In other words, there are many random looking location access patterns

with high entropy, and fewer random looking location access patterns with low entropy, and we are interested in assessing to what extent we can relate the indistinguishability of trails according to their entropy score with the indistinguishability of trails from the standpoint of existing re-identification algorithms. If so, this would suggest that a low entropy systems leads to a low risk of re-identification.

For our purposes, let us assume we have the trail matrix that maps a population of subjects S to a set of locations L . Also, let f_l be the proportion of subjects in S that visit location l . Then, the entropy for location l , $H(l)$, equals

$$H(l) = -f_l \cdot \log (f_l) - (1 - f_l) \cdot \log (1 - f_l) .$$

Under the assumption that individuals decide whether to visit each location independently of other locations, the entropy of the set of location access patterns of the population S to the set of available locations L is given by $H(L) = \sum_{l=1}^{|L|} H(l)$.

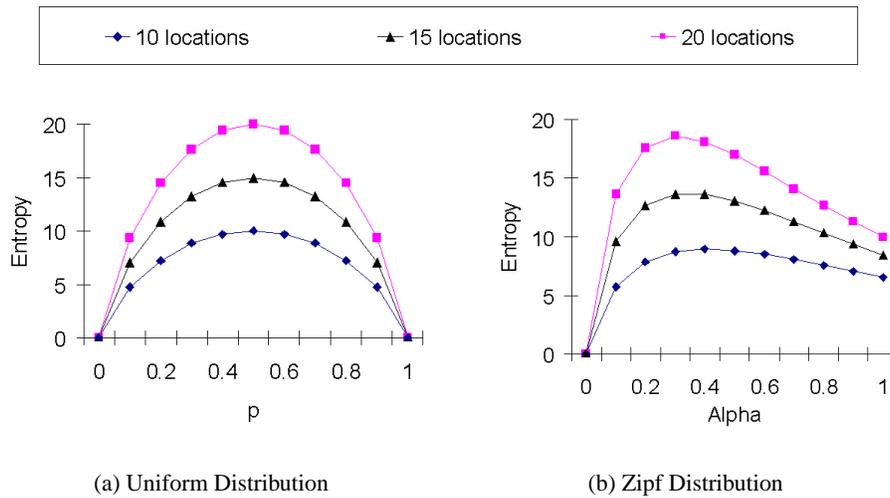


Fig. 7. Entropy plots corresponding to parameter values in the left and right panels of Figure 4.

In order to assess whether entropy and re-identifiability are capturing the same notion of distinguishability we need to compute a measure of correlation among the corresponding scores, as the number of locations changes. In an additional set of experiments, we observed that the entropy curves display a behavior that is similar to that of the percentage of people re-identified, displayed in Figures 4 and 5. In Figure 7 we report the results for the unreserved case.

Here we perform a formal correlation study of these two sets of behaviors by introducing a distance metric, σ , between two curves, which measures the absolute difference of their areas modulo a scaling factor. The scaling factor is proportional to the ratio between the peaks of the two curves. Let us denote the entropy curve by $E(i)$,

and the actual linkage curve by $R(i)$, where i is a point in the grid, G , for the interval $[0, 1]$ we used to generate the re-identification curves in Figures 4 and 5. Let $\max(R) = R(i^*)$ where $i^* = \arg \max\{R(i), i \in G\}$, and let $\max(E) = E(j^*)$ where $j^* = \arg \max\{R(j), j \in G\}$. The scaling factor is then $\frac{\max(R)}{\max(E)}$, and the distance metric, σ , is defined as follows,

$$\sigma(E, R) = \sum_{i=1}^{10} \sigma_i(E, R) = \sum_{i=1}^{10} \left| E(i) \frac{\max(R)}{\max(E)} - R(i) \right|.$$

Note that whenever $i^* = j^*$, i.e., whenever the entropy and re-identification curves peak at the same point $i^* = j^*$ on the grid G , it follows that $\sigma_i(E, R) = 0$. That is,

$$\sigma_i(E, R) = \left| E(i) \frac{\max(R)}{\max(E)} - R(i) \right| = \left| E(i^*) \frac{R(i^*)}{E(i^*)} - R(i^*) \right| = 0.$$

The resulting information from the shape metric is summarized in Figure 8. As values for shape tends toward 0, the curves converge. As expected, the curves tend toward convergence as the number of locations increase. Yet after convergence begins to come into the line of sight, a counter-intuitive phenomenon occurs. Specifically, after a certain number of locations are considered for a particular distribution, the E and R curves begin to diverge from each other. This is an artifact of the limits of re-identifiability. Notice that in Figure 4, when a lesser number of locations are considered the linkage curve has a well defined peak. This peak corresponds to the parameter at which the distribution is most amenable to linkage. But this peak is only discernible when less than all of the trails are linked. Thus, when the system is fully linked at multiple parameterizations of the distribution, the linkage curve plateaus at 100% at its peak, while the entropy continues to be well defined. This limit to linkage causes the observed linkage curve to be improperly matched to the entropy of the system. There is no divergence observed, but rather a limit to independent use of the entropy metric.

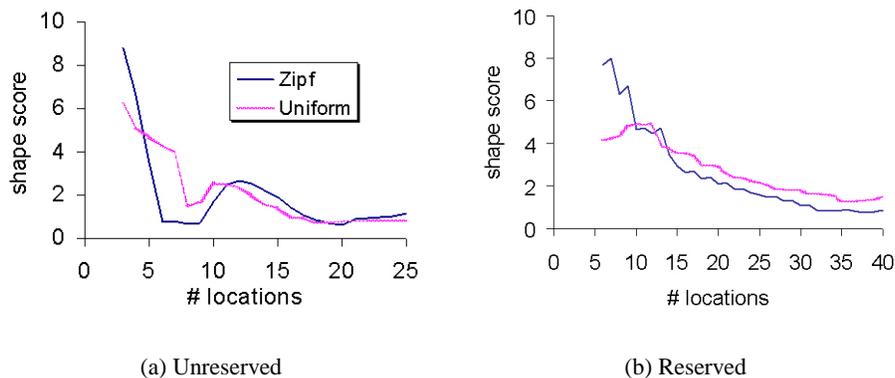


Fig. 8. Shape metric for similarity in simulated distributions and entropy.

The shape metric allows for the discovery of another notable feature that captures how the distribution type influence different trail linkage algorithms. Note that in the unreserved system, the uniform distribution converges earlier than the Zipf distribution. In contrast, when subject to the reserved system, the uniform distribution converges after the Zipf distribution. Ah, a paradox! At first consideration, one would expect that one distribution type, either uniform or Zipf, would converge earlier in both algorithms. However, this paradox results from how trails are generated under the two distributions as well as how the re-identification method leverages trails. First, consider the linkage algorithms. In an unreserved system, the re-identification method looks for a unique bit pattern because there are no *'s. So both 1's and 0's are contributing evenly to the re-identification process. This is why the re-identification curve for the uniform distribution is balanced and has no shift around the midpoint of p . In other words, the percent re-identified is approximately equivalent for $+/-x$ around the parameterization of $p = 0.5$. With respect to an reserved environment though, a * value in a trail functions as fuzzy bit, since it can be used as either a 0 or a 1. Thus, as p tends toward 1, trails with a lesser number of unambiguous values become more difficult to re-identify. As a consequence, the re-identification curve shifts away from high values of p which allow for trails with large amounts of 1's. The Zipf distribution should be hindered by this problem as well, but because it allows for locations to have different entropy values, the Zipf reveals more re-identifications. Thus, the total quantity of re-identifications the Zipf is capable of tends to be greater than the uniform. If one wanted to validate this claim, it is simple to observe that the average number of re-identifications, but not the maximum, for the Zipf is greater than the uniform.

5 Discussion

The above analysis provide a wealth of insight into the effects of location access patterns on the degree to which trail re-identification can be achieved in a distributed system. It also provides intuition into the relation between re-identifiability of a set of trails and the information they carry, as measured by the corresponding Shannon entropy. In this section we briefly address some findings of particular interest. After discussing revelations from our investigations, we consider some of the limitations and possible extensions of our framework. We conclude by presenting a conjecture that emerges from consideration of the empirical evidence we presented in Section 3.

5.1 Location Access Patterns and Re-identification

One of the more interesting findings of our experiments is that high-skew location access patterns yield higher *overall* re-identification when compared with low-skew location access patterns. This result holds despite the fact that low-skew distributions lead to a larger number of *peak* re-identifications, with respect to the parameter underlying the distribution of location access patterns, as well as for any given number of locations in the distributed environment. Further, this result holds in both situations where there is certainty about the information collected and released at the various locations, i.e. the

unreserved case, and in situations where there is uncertainty about the information collected and released at the various locations. This finding has immediate implications for the design of solutions to limit trail re-identification in disclosed databases. For example, one solution we could employ is to entrust an independent third party to identify the set of locations that contribute the most to the skewness of location access patterns, and prevent them from releasing some, or all, of their de-identified data. By doing so, we do not need to provide the third party with data per se, as is the case in prior solutions [11], but rather essential components of the distribution of people to places. Nonetheless, risk analysis is not a substitute for formal privacy protection to prevent trail re-identification, which can be subject to rigorous proof. Re-identification risk provides a proxy by which we can develop provable protection models.

Further, we find there is a strong correlation between the entropy of the system and re-identification. In particular, the lower the entropy in a the set of trails, the more individual trails can be re-identified. This correlation is stronger for distributed systems with more locations, but hold for smaller systems as well. With respect to minimizing risk, our experiments suggest that in order to predict the number of trail re-identifications that can be made, the distribution of location access patterns, or the entropy, should be modelled. In pursuing these strategies, it becomes crucial that the information which released is reliable. In fact, reliability of the information bears relevance to the expected quality of the estimates of both the parameters underlying the distribution of location access patterns, and the entropy of the set of trails of the population of interest.

5.2 Limitations and Extensions

An aspect of our analysis that requires further attention is the correlation between the entropy of a set of trails and the number of re-identifications that can be made. However intriguing, the fact that low entropy systems correlate with high re-identifiability, our experiments offer little intuition into what mechanism may link the two phenomena in a causal manner. We cannot explain “in what sense” low entropy location access patterns explain re-identifiability.

Though this research provides a theoretical investigation into how particular distributions of location access patterns influence trail re-identification, there are certain caveats of the simulation design which limit the extension of these results. First, the entropy computations are carried out under the assumption that individuals decide whether to visit each location independently. As a consequence, our simulations do not completely replicate the behavior of real world populations. This is because in the real world most entities are not random agents visiting locations independently. Rather they can play an active role in choosing which locations to visit. This manifests in the form of correlations between locations in the patterns of access. As a consequence of this dependence, the resulting location access patterns can be different than those obtained under the independent locations assumption. For example, individuals may tend to visit multiple locations in co-location patterns. As a result of such location access behavior, the re-identification capability of the synthetic populations used in this research may be inflated.

Second, the distributions used in this study consist of homogenous populations, such that location access to all locations adheres to a single distribution. However, we should

ask, “What is the effect of mixture models of populations on trail re-identification?” For instance, to what extent is re-identification facilitated when half the population is uniformly distributed while the other half is Zipf distributed? It is possible to speculate on the results, but it is a complex problem that is difficult to reason. As a result, another feasible direction for research into the fundamentals of trail re-identification is to study the effect of mixture models of distributions on re-identification.

5.3 A Conjecture: Re-identification Risk Through Subject-Location Ratio

The empirical evidence presented in Section 3 suggested we explore how different distributions of individuals’ location access patterns influences the number of re-identifications in a distributed environment. However, in the case studies it is the ratio of subjects per location that correlates strongly with the number of re-identifications. In particular we observed that: (i) as the ratio of subjects per location grows large, we find evidence of an exponential relation between the number of locations considered and the number of people that are trail re-identifiable, in the CF dataset; (ii) as the ratio becomes negligible, we find evidence of a logarithmic relation between the number of locations considered and the number of trail re-identifications, in the Homenet dataset; and (iii) when the ratio of subjects per location is approximately 1, we find evidence of a linear relation between the number of locations considered and the number of trail re-identifications, in the PK dataset.

The evidence from the case studies also suggests that the number of re-identifications can be explained by a simpler relation centered around the ratio of subjects per location. This may be due to statistical limiting phenomena that occur in the re-identification of individuals in a distributed environment. This will require further investigation. Specifically, if we denote the ratio of subjects per location with $\frac{|S|}{|L|}$, we conjecture that the number of re-identifications, R , can be expressed as $R \propto f(S, L)^{\frac{|S|}{|L|}}$. Therefore, if the exponent is greater than, equal to, or less than 1, the function may replicate the observed shape of the relations shown in Figure 3.

6 Conclusions

In this paper we proposed a model to estimate re-identification risk when an individual’s data is distributed across a set of locations. Specifically, we introduced methods and metrics for studying the effect of different location access behaviors on trail re-identification. We provided experimental evidence that implies the skew of the distributions of location access patterns is one of the main factors that influences re-identification. Though our models are based on simulation, this work provides a foundation for both basic and applied trail linkage research. One possible extension to this work is to study distributions with location dependencies, as well as mixture models of location access distributions.

Acknowledgements

The authors thank Kathleen Carley, Latanya Sweeney, and Stephen Fienberg for insightful discussions. This work was partially supported by the Data Privacy Laboratory at Carnegie Mellon University, NSF IGERT grant 9972762 in CASOS, NSF grant DMS-0240019, NIH Grant 1 R01 AG023141-01, ONR grant N00014-02-1-0973, and DOD grant MKIDS IIS0218466. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies of the U.S. government.

References

1. Altman, R.: Bioinformatics in support of molecular medicine. In: Proceedings of the American Medical Informatics Association Annual Symposium, Miami Beach, FL (1998) 53–61
2. Sax, U., Schmidt, S.: Integration of genomic data in electronic health records: opportunities and dilemmas. *Methods of Information in Medicine* **44** (2005) 546–550
3. Altman, R., Klein, T.: Challenges for biomedical informatics and pharmacogenomics. *Annual Review of Pharmacology and Toxicology* **42** (2002) 113–133
4. Department of Health and Human Services: 45 cfr (code of federal regulations), parts 160 - 164. standards for privacy of individually identifiable health information, final rule. *Federal Register* **67** (2002) 53182–53273
5. 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. *Journal of Biomedical Informatics* **37** (2004) 179–192
6. Karat, C., Brodie, C., Karat, J.: Usable privacy and security for personal information management. *Communications of the ACM* **49** (2006) 51–55
7. Malin, B.: Betrayed by my shadow: learning data identity via trail matching. *Journal of Privacy Technology* (2005) 20050609001
8. de Moor, G., Claerhout, B., de Meyer, F.: Privacy enhancing technologies: the key to secure communication and management of clinical and genomic data. *Methods of Information in Medicine* **42** (2003) 148–153
9. Gulcher, J., Kristjansson, K., Gudbjartsson, H., Stefansson, K.: Protection of privacy by third-party encryption in genetic research. *European Journal of Human Genetics* **8** (2000) 739–742
10. Lin, Z., Owen, A., Altman, R.: Genomic research and human subject privacy. *Science* **305** (2004)
11. Malin, B., Sweeney, L.: Composition and disclosure of unlinkable distributed databases. In: Proceedings of the 22nd IEEE International Conference on Data Engineering, Atlanta, GA (2006)
12. Airoidi, E.M.: A statistical theory of record linkage with applications to privacy. Technical Report CMU-ISRI-05-112, School of Computer Science, Carnegie Mellon University (2004) Revision, December 2005.
13. Bender, S., Brand, R., Bacher, J.: Re-identifying register data by survey data: an empirical study. *Statistical Journal of the United Nations ECE* **18** (2001) 373–381
14. Griffith, V., Jakobsson, M.: Messin’ with texas: deriving mother’s maiden name using public records. In: Proceedings of the Applied Cryptography and Network Security Conference, New York, NY (2005)

15. Malin, B., Sweeney, L.: Determining the identifiability of dna database entries. In: Proceedings of the American Medical Informatics Association Annual Symposium, Los Angeles, CA (2000) 537–541
16. Sweeney, L.: Uniqueness of simple demographics in the us population. Technical Report LIDAP-WP04, Data Privacy Laboratory, Carnegie Mellon University, Pittsburgh, PA (2000)
17. Willenborg, L., de Waal, T.: Statistical Disclosure Control in Practice. Springer, New York, NY (1996)
18. Danezis, G., Serjantov, A.: Statistical disclosure or intersection attacks on anonymity systems. In: LNCS 2119: Proceedings of the 6th International Workshop on Information Hiding. (2004)
19. Kesdogan, D., Agrawal, D., Penz, S.: Limits of anonymity in open environments. In: LNCS 2119: Proceedings of the 5th International Workshop on Information Hiding. (2002)
20. Winkler, W.E.: Matching and record linkage. In Cox et al., B., ed.: Business Survey Methods. J. Wiley, New York, NY (1995) 355–384
21. Winkler, W.: Data cleaning methods. In: Proceedings of the ACM SIGKDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation, Washington, DC (2003)
22. State of Illinois Health Care Cost Containment Council: Data release overview. State of Illinois Health Care Cost Containment Council, Springfield, IL (March 1998)
23. Kraut, R., Mukhopadhyay, T., Szczypula, J., Kiesler, S., Scherlis, B.: Information and communication: alternative uses of the internet in households. Information Systems Research **10** (2000) 287–303
24. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27** (1948) 379–423
25. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27** (1948) 623–656