Data and Collocation Surveillance Through Location Access Patterns

Bradley A. Malin
Data Privacy Laboratory, School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213-3890 USA
malin@cs.cmu.edu

Abstract
As technologies increase the ability to collect data, individuals leave behind personal fragments of information behind at a number of locations in both the physical and virtual world. When locations collect independent types of data (e.g. such as sets of unlinked facial images and identities) the data seems to be unrelated. However, when multiple locations’ data collections are taken into account, patterns in the locations data is left behind at, or data “trails”, allow for disparate data to be linked. In certain instances, this can lead to seemingly anonymous data being linked to the identity from which it was derived. This paper addresses how trail linkage, or re-identification via the REIDIT (RE-Identification of Data in Trails) algorithms can be applied to surveillance endeavors. The REIDIT algorithms are capable of uniquely re-identifying data, but are also extendible to tracking collocations of people based on location visit patterns. This paper builds upon previous results of trail re-identification in real world datasets by simulating re-identifiability in different types of well-defined location visit distributions. In conclusion, this work provides insight how surveillance through trails can be optimized through information theoretic metrics and an understanding of the distribution of data collection.

Introduction
Until recently, person-specific data was shared somewhat freely provided the data was “de-identified”. This implied that none of the columns included explicit identifiers, such as name, address, or Social Security number. Basically, it was believed that if data looked anonymous, it was anonymous. Yet, this practice has diminished in acceptance since it was demonstrated that such releases can often be linked to other datasets that do include explicit identifiers to re-identify people by name. [Sweeney, 2002] Much research in re-identification has been modeled from the perspective of privacy protection or its erosion, but this is derivative of the fact that when re-identification methods are successful they elicit political and social upheaval due to the demonstration that some privacy policies are insufficient or ill-defined. In actuality, re-identification methods themselves are simply methods for learning unique features or patterns in data and, in this respect, the concept of re-identification is applicable to many data driven environments for a variety of learning or inference problems. Specifically, in this paper we consider how a particular type of re-identification method can be useful for surveillance purposes.

One of the main reasons why re-identification methods are useful is the deluge of information that data capturing technologies have afforded. As technologies for collecting information infiltrate society, the ability to record and store personal information continues toward ubiquity. [Sweeney, 2001] Knowingly and unknowingly, individuals shed data to a number of data collectors both within, as well as beyond, the confines of one’s home. Information collection can be overt and apparent to the individual, such as when a consumer visits a retail store and completes a purchase with a personal credit card. Or data gathering can be less discernable, as when an individual’s image is captured by an unforeseen video surveillance system. [Jones, 2002]

This research addresses how surveillance can be achieved through the use of trail re-identification, by which an individual’s data is linked to identity through the pattern of locations where the data was collected. [Malin & Sweeney, 2004] In this paper, we address some of the more general aspects of trail re-identification, such as the extent to which trails allow for networks of individuals to be discovered, as well how different types and parameterizations of real world distributions of location access patterns affect re-identification. The distributions considered in this paper are based on actual observations made with respect to several different datasets and environments.

1 The author recognizes the previous and current members of the Data Privacy Lab, especially Dr. Latanya Sweeney, for their insightful discussions and support. This research was supported by a National Science Foundation IGERT Program grant and the Data Privacy Laboratory in the Institute for Software Research International, a department in the School of Computer Science at Carnegie Mellon University. The opinions expressed in this research are solely those of the author and do not necessarily reflect those of the National Science Foundation.
Trail Re-identification

A location may be modeled as a defined area in either the physical or virtual setting. For example, in the physical world, a location may be the area that a single video camera or sensor records information about, while in the virtual world a location could be a single website or a router on a network. In the general case, consider an environment consisting of a set of locations and two different types of data (A and B). To understand the power of trail re-identification, let us assume that, while locations can collect either of the two types of data, no single location is able to discern the relationship between the records of type A and B. Thus, the only knowledge that a single location has is the set of pieces of for data types A and data type B (which may be applied to simple aggregate analysis, such as for traffic load). When data is traceable, the data collected from disparate locations can be used to construct, what we term, “trails" of data. Each value in a trail represents the degree of belief that the data was actually observed at the location. Alone, there is no relationship between data of type A and type B, but the trails provide a bridge, or a common interface if you will, through which the independent data types can be related to one another. Trails are constructed over the set of locations collecting data, and thus set of trails is referred to as a track.

The ability to relate trails across tracks is dependent on both the completeness and multiplicity of the data. By completeness, we mean the degree to which the data types collocate with each other. In certain situations, data of type A is always collected with data of type B, and vice versa. When this phenomenon is observed for all locations, we say that trails are complete. In the case when one type of data may not be collected, but the other data type is, the under collected data trails are incomplete. For example, in an online scenario, when a webuser visits a website the IP address of their computer is always logged by the access log, while the name (or pseudonym) of the webuser may or may not be left behind. Furthermore, with respect to multiplicity, we consider the number of individuals which have access to a particular type of data. If a piece of data is controlled by one individual only, we say that trails across tracks are one-to-one. However, when multiple individuals can leave behind the same value of data, the data is one-to-many. Continuing with the online example, the one-to-one exists when every user surfs the web via their computer (one per person) only. The one-to-many scenario exists when multiple users browse through the same computer. Currently, our research does not address the scenario when both tracks consist of incomplete trails or the many-to-many environment.

The REIDIT (Re-Identification of Data In Trails) algorithms account for different aspects of completeness and multiplicity in Boolean trails. [Malin, 2002] [Malin & Sweeney, 2004] In these trails, a “1” or “0” represents the definite observation, or lack of, a piece of data at a location. The algorithms discover unique re-identifications, such that a single trail from one track is uniquely re-identified to data from the other track. In addition, they can discover clusters of data or individuals, in the form of collocation patterns of data or, when the data permits individuals. We briefly review them here for completeness. The REIDIT-Complete, or REIDIT-C, algorithm provides re-identification when both tracks consist of complete trails and are one-to-one. A trail from one track is linked to a trail from the other if there is one and only one trail it is equivalent to (i.e. all 1’s and 0’s are equal). The REIDIT-Incomplete, or REIDIT-I, algorithm, performs re-identification in the same environment as REIDIT-C except one track consists of incomplete trails. In this setting, in an incomplete trail a value of 1 represents the presence of data at a location, whereas a 0 is ambiguous. However, in a complete trail, both 0’s and 1’s are unambiguous. So, if an incomplete trail can be uniquely matched to a complete trail by converting only 0’s to 1’s, then a linkage is made. Finally, the REIDIT-Multiple, or REIDIT-M algorithm, performs re-identification in the same environment as REIDIT-I, except the tracks are one-to-many.

Both the REIDIT-C and –I algorithms can be generalized to re-identify distinct or fuzzy clusters for learning collocation information about individuals. To extend the algorithms for cluster discovery, instead of searching for unique linkages of trails between tracks, the algorithms search for exclusive subsets of trail linkages. In comparison, the design of the REIDIT-M directly permits the re-identification of collocations of individuals, though not necessarily as one might expect. Thus, the REIDIT-M algorithm learns the relationships of individuals based on access to the same piece of data. For example, in previous research, we were able to re-identify households, or cohabitation, of individuals with access to the same IP address. [Malin & Sweeney, 2003]

Data Distribution Effects on Re-identifiability

In theory, the re-identification limits of the REIDIT algorithms scale exponentially, however, such limits are rarely observed in real populations. In previous research, the REIDIT algorithms were evaluated on both hospital visit patterns for patients with genetic diseases and in the online environment. [Malin & Sweeney, 2003] [Malin & Sweeney, 2004] Our findings suggested that re-identifiability is dependent on several factors, including the number of pieces of data in a track, the number of locations of a trail, and the distribution of data to locations. The latter implied the probability an individual visits a location is influenced by features such as their geographic proximity as
well as the locations’ capabilities, such as specialization in a service. In general, we observed that in the physical environment of health a location’s ability to capture an individual’s data was guided by a uniform to Gaussian distribution, and in the online environment data was distributed according to a high-skew Zipf distribution. This latter finding confirms previous observations of online web usage observed in other studies. [Brin & Page, 1998] [Breslau, et. al., 1999]

In this paper, we consider the influence of the location distribution, and the number of locations, on the re-identifiability of a system. We simulate populations with the number of subjects fixed to 1000 individuals. For these synthetic populations, two types of distributions are generated, the first according to a uniform distribution, and the second according to a Zipf distribution. A subject’s trail in a uniform distribution is controlled by a single parameter \( p \), which is the probability that an individual will visit a location. For our experiments we sample \( p \) from the range \([0, 1]\) at equidistant intervals of 0.1. Similarly, populations are guided by a general form of the Zipf distribution. The probability that an individual visits a location, \( f_i \), is inversely proportional to the location’s rank (as determined by its frequency) \( r_i \) via the equation \( Z \times f_i = r_i^{-\alpha} \), where \( \alpha \) is a constant between \([0,1]\) and \( Z \) is the total number of observations (i.e., the total number of visits made over all locations). As with the uniform distribution, the Zipf is studied by varying the parameter \( \alpha \) over the same interval \([0,1]\), and sample points, as the \( p \) parameter of the uniform distribution. For each tested data point, we generate 100 populations. Each population is subjected to either the REIDIT-C or REIDIT-I algorithm. Examples of the resulting 10-point plots for REIDIT-I are depicted in Figures 2. In these plots the mean percentage and +/- one standard deviation of mean for the 100 simulated populations are depicted in the lower of the two plotted curves. The x-axis corresponds to the parameter of the distribution in question, while the left y-axis corresponds to values of the mean percentage re-identified. For completeness, and dispel confusion, the upper curve corresponds to entropy (which will be addressed in a moment).

![Figure 2. Simulated population re-identification with REIDIT-I at 15 and 25 locations. Left plots uniform and right plots Zipf distributions. Lower and upper curves correspond to % of population re-identified and entropy, respectively.](image)

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 REIDIT-C and REIDIT-I re-identification algorithms, it is apparent that the uniform distribution consistently yields a larger maximum number of re-identifications than the Zipf. 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 re-identifiability of a distribution type. To compare distributional types, we consider the area under the re-identifiability curve. This is calculated as the total area under the 10-point mean re-identifiability curve (average re-identifiability of 100 simulated populations). Though the uniform distribution always yields the larger max re-identifiability, the Zipf distribution is almost always the more re-identifiable when considering all parameterizations. This is obviously so in the case of REIDIT-I re-identification, where the figure to the right of Figure 3 shows that the Zipf always dominates. Similarly, under REIDIT-C, 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! This finding is observed between approximately 8 and 18 locations.

![Figure 3. Area under the re-identified curve over all parameterizations of the uniform (uni) and Zipf distributions. Re-identification plots for left) REIDIT-C and right) REIDIT-I.](image)
The flip in distribution re-identifiability dominance occurs for two reasons. Initially, Zipf dominates when there are not many locations in consideration because it is more difficult to realize complete vectors of all 1’s. Then later on, 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 that is subsequently re-identified.

Experimental analyses demonstrate that as the number of locations in the system increase, the re-identifiability curves converge towards the measure of Shannon entropy. This is an interesting finding, since it suggests that simulation may not be necessary for constructing re-identifiability curves. Yet, in the current work, we generated trails with locations independent of each other. In the real world, trails tend to be more complex and such issues as degree of collocation will need to be studied further before such a metric, or variant of it, is advocated.

Discussion and Application of Results

The above analyses provide a wealth of insight into the capabilities of the REIDIT re-identification algorithms. Most interesting is the finding that Zipf distributions yield higher overall re-identifiability in comparison to uniform distributions. This is especially so in light of the fact that uniform distributions always provide the potential for a larger number of re-identifications at a given number of locations. Given this finding, it has profound implications if one was to consider designing a system of locations that individuals leave information behind at, which relates directly to risk management theory. If information is always to be released such that it is susceptible to REIDIT-I, then the Zipf distribution is always the better choice. Regardless of the parameterization of the Zipf, it will always yield more re-identifications than corresponding uniform distribution.

When information in trails is less certain, and subsequently the relations between trails across tracks, (i.e. when data is under collected), then designing a system where location access is in the form of uniform distribution may be the best choice. Note that the word may is used because it is at this point where the majority of the risk occurs. In a REIDIT-C environment, if the system falls into worst case location access scenario, such that the parameterization of the distribution maximizes re-identification, then the uniform distribution will reveal more re-identifications. If there is some doubt as to whether the parameterization will yield max re-identifiability, then one is actually better off in the uniform system. This is because of the finding that the average number of re-identifications is lesser in the uniform than in the Zipf distribution. It appears that the question of which distribution will yield more re-identifications is a matter of how confident one predicts the parameter of the distribution of the way with which subjects access locations.

References