

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 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 appears unrelated. However, when multiple locations' collections are account for, 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 the re-identification of seemingly anonymous data to the named identity from which it was derived. This paper addresses how trail 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, and are also extendible to tracking collocations of people based on location visit patterns. The research presented in this paper builds upon previous experimental results of trail re-identification with 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.

Contact:

Bradley Malin
Institute for Software Research International
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3820

Tel: (412) 268-1097

Fax: (412) 268-6708

Email: malin@cs.cmu.edu

Keywords: Surveillance, distributed data, re-identification, pattern analysis, network analysis, clustering, information theory, network simulation

Acknowledgements: 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.

Data and Collocation Surveillance Through Location Access Patterns

Bradley A. Malin

Introduction

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] However, the collection of mass amounts of data does not necessarily imply the collecting location possesses the ability to relate different kinds of data collected.

In previous research, we introduced the REIDIT (*RE*-Identification of *Data in Trails*) algorithms which learn relationships between seemingly disparate types of data through location patterns. [Malin, 2002] [Malin & Sweeney, 2004] These algorithms exploit the fact that with increased data collection, individuals leave similar information behind at multiple locations. Subsequently, patterns in the locations data is left behind at, or data "trails", can be employed to bridge a link between data derived from the same individual or group of individuals. Initially, these methods were developed in the data privacy community to link seemingly anonymous data to named identities (thus the term "re-identification") and formally model privacy protection, or the lack of such. [Sweeney, 1997] Yet, re-identification methods, REIDIT included, are basically tools for learning unique features or patterns in data. As a result, the concept of re-identification is applicable to many data driven environments for a variety of learning or inference problems.

In this paper, we demonstrate that REIDIT is applicable for surveillance purposes. 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.

Trail Re-identification

The REIDIT algorithms re-identify information through patterns in the locations data is left behind. A location is an area of data collection, which can be modeled as in either the physical or virtual setting. For example, in the physical world, a location can be the area a single video camera, or sensor, records information about, while in the virtual world it can 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 individual records of type *A* and *B*. The only knowledge that a single location has is the set of data pieces for types *A* and *B*. When data is traceable the data collected from multiple 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 observed at the location and the trails provide a bridge, or a common interface, through which the independent data types can be related to one another. Trails are constructed over the set of locations collecting data and a 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. First, by completeness, we mean the degree to which the data types collocate with each other. When data of type *A* is always collected with data of type *B*, and vice versa, is observed for all locations the trails are called *complete*. In the case when one type of data is not always collected, the trails constructed from such information are termed *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. Second, 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 situation 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 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 the core features of unique re-identification via the REIDIT algorithms. 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, a value of 1 in an incomplete trail represents the presence of data at a location, whereas a 0 is ambiguous. In contrast, both 0's and 1's are unambiguous in a complete trail. When 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 REIDIT-M directly permits the re-identification of collocations of individuals, though not necessarily as one might expect. 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 prior 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 implies that the probability an individual visits a location is influenced by features including, their geographic proximity as well as the locations' capabilities, such as the specialization in a service. In general, we observed that in the physical environment of health populations a location's ability to capture an individual's data was guided by a uniform to Gaussian distribution, whereas 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 others research'. [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. 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 α 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). From a statistical basis, the parameterization of the Zipf affects the skew of the distribution. The distribution becomes increasing skewed as $\alpha \rightarrow 1$ and approaches uniformity with $p=1$ as $\alpha \rightarrow 0$. 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.

Populations are simulated with the number of subjects fixed to 1000 individuals. Additionally, 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-C are depicted in Figure 2. In these plots the mean percentage as well as +/- one standard deviation of mean for 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).

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 is observable, even by visual inspection, by considering

the maximum re-identifiability of the distribution type. For example, when considering 10 locations, REIDIT-C re-identifies 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 skew (which occurs when $\alpha = 0.4$). This finding is consistent across all systems as the number of the locations in consideration is increased.

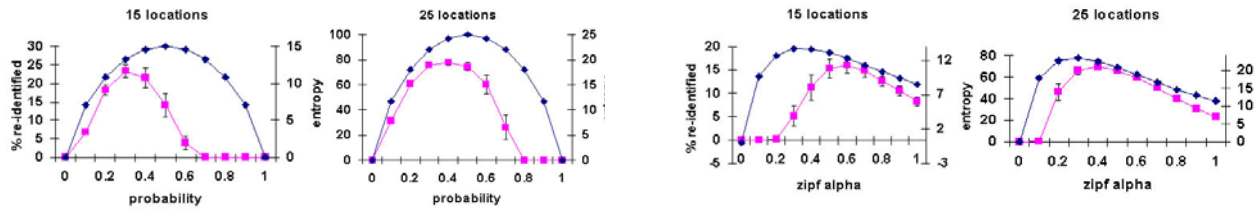


Figure 1: Simulated population re-identification with REIDIT-C at 15 and 25 locations. *Left plots*) uniform and *right plots*) Zipf distributions. Lower and upper curves correspond to % of population re-identified and entropy, respectively.

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.

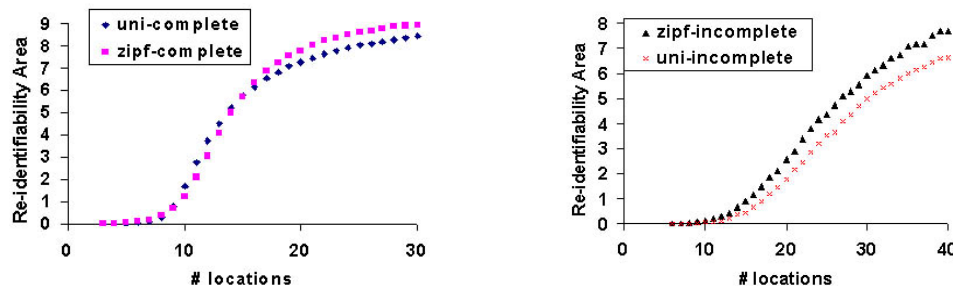


Figure 2: Area under the re-identified curve over all parameterizations of the uniform (uni) and Zipf distributions. Re-identification plots for *left*) REIDIT-C and *right*) REIDIT-I.

The flip in distribution re-identifiability dominance occurs for two reasons. Initially, the Zipf dominates when there are not many locations in consideration because it is more difficult to realize complete vectors of all 1's. Then, as the number of locations increase, the Zipf regains dominance 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.

The synthetic trails generated for the experiments are Boolean vectors of 0's and 1's. As such, it seems feasible that each trail can be likened to a measure of information available on a subject. Continuing along this line of thought, it is plausible that the trail re-identifiability of a system is related to the Shannon entropy, as defined in information theory. [Shannon & Weaver, 1949] From a general standpoint, the entropy provides a characterization of the total amount of randomness in the distribution of 1's and 0's for a variable. For our purposes, let S be the set of subjects distributed over a set of locations L . Also, let g_l be the fraction of subjects in S that visit location l . Given this information, the entropy for a single location $H(l)$ is equivalent to $-g_l \log(g_l) - (1 - g_l) \log(1 - g_l)$. In the synthetic populations, each location is allocated subjects independently; thus the entropy measure for the whole system of trails is computed as the sum of entropies for each location. Both the entropy of the system and the re-identifiability of populations over different distributions produce response curves in terms of how re-identifiability is influenced. As stated above the entropy is the upper line of the graphs in Figure 1, and the y-axis to the right provides the scale. The 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 regarding risk management theory and choices regarding the design of a system of locations for capturing data on a population of individuals. If information is always to be released such that it is susceptible to REIDIT-C, then the Zipf distribution is always the better choice. Regardless of the parameterization of the Zipf, it will always yield less 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-I 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

- [Breslau, et. al., 1999] Breslau, L, Cao, P., Fan, Phillip, G., and Shenker, 1999, "Web caching and Zipf-like distributions: evidence and implications." In *Infocom*.
- [Brin & Page, 1998] Brin, S., and Page, L., 1998, The anatomy of a large-scale hypertextual web search engine. In *Proc 7th World Wide Web Conference*.
- [Jones, 2002] Jones, M., 2002, "All eyes on oceanfronts's new surveillance system." *The Virginian Pilot*. Sept 10.
- [Malin, 2002] Malin, B. 2002, "Compromising privacy with trail re-identification: the REIDIT algorithms." Tech Report CMU-CALD-02-108, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.
- [Malin & Sweeney, 2003] Malin, B. and Sweeney, L., 2003, "Compromising online anonymity with trail re-identification." Data Privacy Working Paper #14. Carnegie Mellon University, Pittsburgh, PA.
- [Malin & Sweeney, 2004] Malin, B. and Sweeney, L., 2004, "How (not) to protect genomic data privacy in a distributed network: Using trail re-identification to evaluate and design anonymity protection systems." *Forthcoming in the Journal of Biomedical Informatics*. Earlier version available as Tech Report CMU-ISRI-04-115, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.
- [Shannon & Weaver, 1949] Shannon, C.E. and Weaver, W., 1949, "The mathematical theory of communication." University of Illinois Press. Urbana, IL.
- [Sweeney, 1997] Sweeney, L., 1997, "Weaving technology and policy together to maintain confidentiality." *Journal of Law, Medicine, & Ethics*, 25: 98-110.
- [Sweeney, 2001] Sweeney, L., 2001, "Information explosion." In: Zayatz, L., et. al.(eds): *Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies*. Urban Institute, Washington, DC.