Decision Support Systems 51 (2011) 10–20

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

An entropy approach to disclosure risk assessment: Lessons from real applications and simulated domains

Edoardo M. Airoldi [a], Xue Bai [b,*], Bradley A. Malin [c]

[a] Department of Statistics, Harvard University, Cambridge, MA 02138, USA
[b] Department of Operations and Information Management, School of Business, University of Connecticut, Storrs, CT 06269, USA
[c] Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37203 USA

ARTICLE INFO

Article history:
Received 5 March 2010
Received in revised form 15 November 2010
Accepted 15 November 2010
Available online 24 November 2010

Keywords:
Disclosure
Distributed systems
Information theory
Information privacy
Risk analysis

ABSTRACT

We live in an increasingly mobile world, which leads to the duplication of information across domains. Though organizations attempt to obscure the identities of their constituents when sharing information for worthwhile purposes, such as basic research, the uncoordinated nature of such environment can lead to privacy vulnerabilities. For instance, disparate healthcare providers can collect information on the same patient. Federal policy requires that such providers share “de-identified” sensitive data, such as biomedical (e.g., clinical and genomic) records. But at the same time, such providers can share identified information, devoid of sensitive biomedical data, for administrative functions. On a provider-by-provider basis, the biomedical and identified records appear unrelated, however, links can be established when multiple providers' databases are studied jointly. The problem, known as trail disclosure, is a generalized phenomenon and occurs because an individual's location access pattern can be matched across the shared databases. Due to technical and legal constraints, it is often difficult to coordinate between providers and thus it is critical to assess the disclosure risk in distributed environments, so that we can develop techniques to mitigate such risks. Research on privacy protection has so far focused on developing technologies to suppress or encrypt identifiers associated with sensitive information. There is a growing body of work on the formal assessment of the disclosure risk of database entries in publicly shared databases, but less attention has been paid to the distributed setting. In this research, we review the trail disclosure problem in several domains with known vulnerabilities and show that disclosure risk is influenced by the distribution of how people visit service providers. Based on empirical evidence, we propose an entropy metric for assessing such risk in shared databases prior to their release. This metric assesses risk by leveraging the statistical characteristics of a visit distribution, as opposed to person-level data. It is computationally efficient and superior to existing risk assessment methods, which rely on ad hoc assessment that are often computationally expensive and unreliable. We evaluate our approach on a range of location access patterns in simulated environments. Our results demonstrate that the approach is effective at estimating trail disclosure risks and the amount of self-information contained in a distributed system is one of the main driving factors.

1. Introduction

Modern society is marked by mobility amidst ubiquitous technology [53]. The increasing ease of data collection, sharing, and processing has led to the capture and subsequent reuse of substantial quantities of personal information. While many individuals desire the more efficient and effective services that the application of their information can support, there are concerns that the reuse could violate expectations of privacy [4]. As an example, let us consider the healthcare domain, where the apparent sensitivity of medical information heightens privacy concerns [6].

Collections of detailed genomic data tied to clinical information are poised to yield significant healthcare breakthroughs, ranging from personalized medicine to drug discovery [10,45,48]. Research in this area is often sponsored through federal funding agencies, such that various policies now require researchers to share “de-identified” genomic data for validation of findings and reuse in general [51,52]. At the same time, hospitals often share identifiable resources, devoid of genomic data, for administrative purposes, such as discharge data for public health and policy evaluation [49,59]. While the genomic and discharge data appear unrelated, a patient can leave information behind at multiple institutions, where the data is independently managed and shared. And, the location–access patterns, or *trails*, of an individual's data can be constructed as a bridge between the disparate resources, which leads to “re-identification” [42]. Notably, this is a problem that will escalate in the healthcare domain as the cost of

* Corresponding author.
E-mail addresses: airoldi@fas.harvard.edu (E.M. Airoldi), xue.bai@business.uconn.edu (X. Bai), b.malin@vanderbilt.edu (B.A. Malin).

0167-9236/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2010.11.014

genome sequencing decreases and participation of patients in research studies increases.

Yet, it should be recognized that the problem of trail disclosure is poised to become a real concern not only in healthcare, but in many other domains [37]. Though domains change the problem remains constant: there is a risk that trails derived from shared data will lead to disclosure. Inability to address the problem could prevent organizations from sharing data on a broad scale.

1.1. Overview of the approach

In this paper, we propose a novel approach to formally evaluating the risk of trail disclosure in distributed database systems prior to release. This approach adapts the concept of *self-information* in information theory [28] and explores the relationship between the measure of self-information, also known as *entropy*, of a database system to the rate of disclosure of data entries in the system [2,40]. Entropy based measures have already been applied for estimating information loss [68], and recently for estimating disclosure risk for microdata [8], yet, to the best of our knowledge, this is the first study in which the problem of trail disclosure is characterized from a statistical and predictive perspective. The entropy approach to trail disclosure risk is computationally efficient and superior to existing risk measures in the sense that it estimates the risk based on statistical characteristics of a trail distribution and requires minimum access to the individual data entries, which in most cases leads to computationally inefficiencies. In our approach, the actual number of disclosures can be measured as the number of shared database entries that are re-identifiable.

The goal of our work is to approximate the number of disclosures that can be made on an arbitrary given database, prior to its disclosure. To do so we need to isolate the processes that influence disclosure, such as 1) the data generating process (e.g., How do people visit places?) and 2) the disclosure process (e.g., How are trails linked?). We then substitute statistical characteristics of location access patterns for the actual patterns in other shared databases to estimate the disclosure risk.

Our approach can be summarize as follows: First, we explore several real environments in which trail re-identification has been studied and discover how the location access patterns relate to the number of trail disclosures. Based on the empirical evidence on location access patterns, we then characterize the underlying process that governs trail disclosure. Next, we propose an entropy metric to capture the relationship between distributions of location access and the degrees of disclosure in shared databases. We test the proposed metric in a simulated environment using several fundamental location visit strategies employed by individuals in the real world, and finally, we assess the disclosure risk they entail. The empirical evidence from the earlier cases, as well as simulated experimental results, suggest that the entropy metric is effective in estimating the risk of trail disclosure in database entries. The statistical characteristics of the distribution of people to places is shown as one of the main factors that drives trail disclosure.

1.2. Outline

The remainder of this paper is organized as follows. In Section 2, we briefly review relevant literature on disclosure risk assessment. Next, in Section 3 we review the formal basis and methods for trail disclosure. In Section 4, we propose a metric based on information theory to assess the trail disclosure risk. In Section 5, we revisit several real applications in which trail disclosure has been studied to demonstrate the fact that the disclosure rates can be characterized as a function of the access patterns of people to places. In section 6, we perform linkage analysis on several types of simulated datasets corresponding to a range of distributions of location access patterns, and analyze the relationship between the entropy scores of each distribution and the corresponding disclosure rate. We evaluate our metric in Section 6.3. Finally, we summarize our work, address limitations and extensions for future research in Section 7.

2. 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. This notion, for instance, set the precedent for the “Safe Harbor” standard of the *de-identification* designation in the Privacy Rule of the Health Insurance Portability and Accountability Act [14]. 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 include explicit identifiers to uniquely, and correctly, disclose private information by personal name [7,30,41,50,64,69]. 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 instance, the fields {*date of birth*, *gender*, *5-digit zip code*} have commonly appeared in both de-identified databases and publicly available identified data (e.g., birth records, death records, marriage records, or voter registration records) and are estimated to uniquely represent over 60% of the U.S. population [29,64].

2.1. Perspective on privacy and data protection

Given the potential for re-identification of individuals in shared data [38,65], many privacy protection methods have been proposed and adapted by researchers from multiple disciplines, including statistics, computer science, medical informatics, and information systems [9,13,18–20,25–27,31,54,58]. Some of these methods have focused on developing various technologies, including cell suppression [3,11], encryption [31,47,56], data perturbation [18,19,26], and microaggregation [46,74] to limit the disclosure of explicit identifiers associated with sensitive data. Others have focused on developing sophisticated database operational mechanisms to restrict access to confidential data [1,27,33], or hybrid approaches that combine both data entry manipulation and database access restrictions [5,9,24,54]. Trail disclosure extends traditional data disclosure problems as it illustrates how the pattern of locations people visit, or trails, can be used to facilitate standard data linkage and disclosure [37,42]. Existing privacy protection methods are not designed to address trail disclosure problems.

2.2. Disclosure risk assessment

To apply data protection methods in a formal manner, various measures have been proposed for estimating the risk of data disclosure. For instance, [19] proposed a decision-theoretic framework for risk assessment in masked data that includes the intruder's objectives and strategy for compromising the database and the information gained by the intruder. In [20], this model was extended into a *risk-utility* (or R-U) confidentiality *map* to study the trade-off between utility and disclosure risk in shared data, notably tables of aggregated counts (also known as tabular data). [16] used the posterior probability for re-identification of data entries in a given database as a proxy of the disclosure risk measure and [66] applied this measure to empirically compare statistical disclosure control methods. [67] created a set of maximum, minimum and weighted disclosure risk measures based on inversion and change factors on the magnitude of masking modification in tabular data. [62] studied issues of using the probability of a correct match in data records as a measure of disclosure risk and explored the nature of this probability and its estimation. The focus on risk of disclosure is an issue that is shared by recent literature on microdata (i.e., individual records) more generally (e.g., see [15,21,57,73]).

Although these measures have been valid in their specific applications, the majority have focused on the risk of re-identification of data entries in tabular or microdata with standard attributes, not on the amount of entries susceptible to re-identification via trails that occur in shared databases. Furthermore, for all the previously proposed measures, the accuracy of the measurement depends on the reliability of the estimation of threshold disclosure rates and the data sharing method that is used, which in most cases are strongly affected by random fluctuations or are computationally intractable.

2.3. Distinguishing characteristics of the trails problem

The distinguishing aspects of the trail disclosure problem are that it is a *tragedy of the commons* situation combined with a lack of coordination among the locations. The tragedy of the commons [32] is marked by a domain in which there are a set of consumers and a shared limited resource. Each consumer rationally utilizes the resource to maximize their benefits; however, when the actions of all consumers are considered jointly, the result fails to maximize the benefits for all entities. In the context considered in this work, the shared resource is the environment into which data is shared. Separately, each location's partitioned pair of identified and de-identified records can be disclosed in a provably unlinkable manner and each location attempts to disclose as much data as possible. Yet, when the collection of locations' shared databases are studied, a vulnerability arises that is due to the traceability of an entity's information (e.g., an individual's name has a low likelihood of changing) [42]. As we trace an individual across databases, we accumulate the list of locations visited, which serves as a way to compare seemingly incomparable information (e.g., names and DNA).

Thus to maximize the benefits (i.e., data sharing) without compounding costs (i.e., re-identification), the locations need to withhold certain amounts of information. In previous work, we introduced an algorithm to facilitate coordination when the locations are permitted to exchange information [38]. Unfortunately, information exchange for coordination is not always possible, due to legal or regulatory constraints. To overcome such a problem, we demonstrated how locations could use a third party to facilitate coordination without revealing information [39,43,44]. Yet, these approaches require certain assumptions of trust in encryption technology, as well as the inclusion of third parties, which is not always practical.

3. Identity disclosure by trails

We begin this section by providing an informal view of identity disclosure by trail re-identification. We then introduce a formal presentation of the disclosure problem.

3.1. Informal problem description

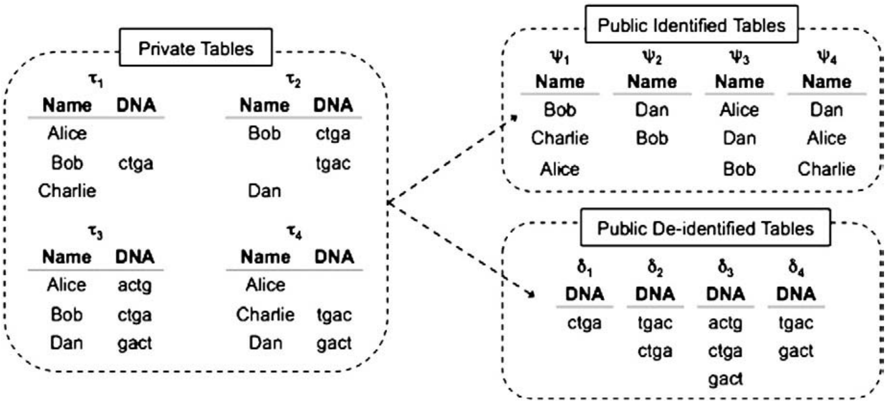
The main premise of trail disclosure is based upon the observation that people visit different sets of locations where they can, and do, leave behind similar pieces of de-identified information. The de-identified data can consist of only one or very few fields. Each location visited collects and, subsequently, 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 location's releases appear unrelated, 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.

3.2. Formal problem description

A formal presentation of the disclosure problem considered in this work is as follows: Let $L = \{l_1,..., l_n\}$ 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 $\tau(A_1, A_2, \ldots, A_p)$, where the set of attributes is $A^\tau = \{A_1, A_2, \ldots, A_p\}$ 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 $[a_1, \ldots, a_p]$, 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 [12]. 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 $\psi(A_1, \ldots, A_i)$ and $\delta(A_{i+1}, \ldots, A_j)$, with attributes $A^\psi \subset A^\tau$, $A^\delta \subset A^\tau$, and $A^\psi \cap A^\delta = \emptyset$. For illustration purposes, consider the four databases depicted in Fig. 1.

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 individual's DNA sequence from one location to another. In this example we have four locations $L = \{l_1, l_2, l_3, l_4\}$. The corresponding trail matrices are depicted in Fig. 2(a). For



τ_1

Name	DNA
Alice	
Bob	ctga
Charlie	

τ_2

Name	DNA
Bob	ctga
	tgac
Dan	

τ_3

Name	DNA
Alice	actg
Bob	ctga
Dan	gact

τ_4

Name	DNA
Alice	
Charlie	tgac
Dan	gact

ψ_1	ψ_2	ψ_3	ψ_4
Name	Name	Name	Name
Bob	Dan	Alice	Dan
Charlie	Bob	Dan	Alice
Alice		Bob	Charlie

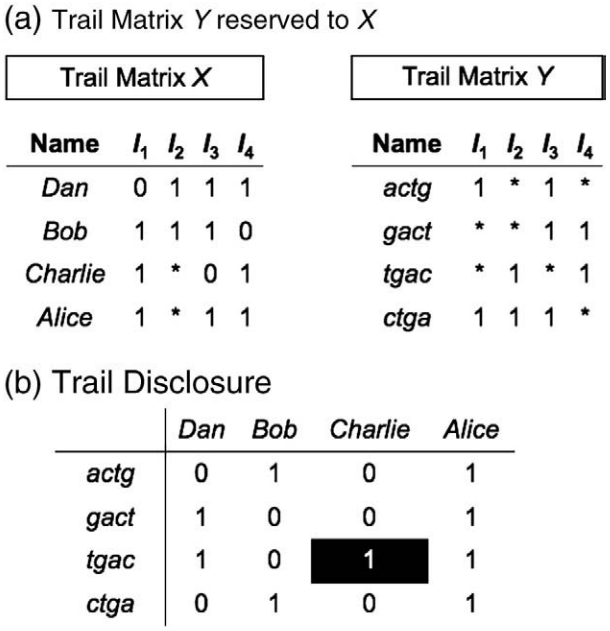
δ_1	δ_2	δ_3	δ_4
DNA	DNA	DNA	DNA
ctga	tgac	actg	tgac
	ctga	ctga	gact
		gact	

Fig. 1. Sample disclosures for four locations.

large databases, the trail matrix could be constructed using exact joins, or might be constructed from traditional record linkage algorithms for tables with common attributes [23,70,72]. The 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" in 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 matrix X.

The basis behind trail disclosure 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 are drawn from the same population of entities. As a result, it is assumed that each entity's trail in X and Y can be transformed into a common trail by flipping *'s into 0's and 1's. When this transformation can be performed, $X[x,:]$ is said to be relatable (represented with the $\preceq$ symbol) to $Y[y,:]$ (and similarly from Y to X). Fig. 2(a) provides an example of trail matrices Notice $Y[actg,:] \preceq X[Alice,:]$ and $Y[tgac,:] \preceq X[Charlie,:]$.

The goal of trail construction is to explore the patterns of trail distribution and the disclosure risk. We use the ReIDIT-I (Re-Identification of Data In Trails — Incomplete) algorithm introduced by [42] to conduct trail disclosure experiments and our model validation section. Although alternative methods have been proposed for extracting information from anonymized public sources [17,36,71], we are motivated to use ReIDIT-I because it is computationally tractable and guarantees every match to be a correct disclosure [38]. Briefly speaking, ReIDIT-I involves the following steps. First, it constructs a $|Y|\times|X|$ matrix, called M, such that cell $M[i,j]=1$ if $Y[i,:]\preceq X[j,:]$, and 0 otherwise. When it finds a row or column that has only one cell $M[i,j]=1$, it re-identifies the corresponding data elements in the cell. This process is iterated until no more matches can be made.[1] Fig. 2(b) illustrates the initial matrix for Fig. 2(a) and the first trail disclosure of *tgac* to *Charlie* is made. In the next iteration *gact* will be re-identified to *Dan*, and so on. Detailed algorithm steps can be found in [38].


(a) Trail Matrix *Y* reserved to *X*

Trail Matrix *X*

Name	l_1	l_2	l_3	l_4
Dan	0	1	1	1
Bob	1	1	1	0
Charlie	1	*	0	1
Alice	1	*	1	1

Trail Matrix *Y*

Name	l_1	l_2	l_3	l_4
actg	1	*	1	*
gact	*	*	1	1
tgac	*	1	*	1
ctga	1	1	1	*

(b) Trail Disclosure

	Dan	*Bob*	*Charlie*	*Alice*
actg	0	1	0	1
gact	1	0	0	1
tgac	1	0	1	1
ctga	0	1	0	1

Fig. 2. (a) Trail matrices built from the partitioned tables in Fig. 1. (b) In the first step of the ReIDIT-I algorithm, *Charlie* is re-identified to *tgac*.

4. Disclosure risk

In this section we introduce an entropy metric for assessing the risk of trail disclosure.

Self-information in information theory [28] is defined as a measure of the information content associated with a probabilistic event. The underlying intuition is that the smaller the probability of an event, the larger the self-information associated with receiving the information that the event indeed occurred. In the case of trails of location access, the occurrence of a trail is understood as a probabilistic event. A trail that occurs more rarely is associated with a larger amount of self-information if occurred. We speculate that the larger amount of self-information contained in a trail, the more re-identifiable the trail is, given that the trail occurred in the shared databases. We are interested in examining to what extent we can relate the distinguishability of a trail to the amount of self-information associated with the system. Next, we present an entropy metric for evaluating the risk of trail disclosure of database entries.

The expected amount of self-information contained in a random event is called *entropy* [60,61]. The lower the entropy, the lower the expected amount of self-information that is associated with the system. The mathematical definition of entropy is as follows: Let X be a discrete random variable with possible values $\{x_1,...,x_n\}$. Let $p(x_i)$ be the probability of the outcome value $X=x_i$. The entropy of X is calculated as:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\cdot log_b p(x_i). \tag{1}$$

In the case of location access patterns, each trail is represented as a Boolean vector of 0's and 1's such that we can compute its entropy in the distributed database system. If we consider all the possible trails with a given entropy score. The entropy score in our context of trail disclosure is a measure that relates to re-identifiability. We argue that a database system with a low entropy score leads to a low risk of disclosure in database entries. The intuition for our argument is that, for a given amount of information that is required to re-identify data entries in a database system, if the expected amount of self-information contained in the system is low, the likelihood of disclosure is low.

Let us assume we have the trail matrix that maps a population of subjects S to a set of locations L as defined earlier. Let $p(l_i)$ be the probability of subjects in S that visit location l_i. Then, the entropy for location l_i, $H(l_i)$, equals:

$$H(l_i) = -p(l_i)\cdot log(p(l_i)) - (1-p(l_i))\cdot log(1-p(l_i)). \tag{2}$$

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 $H(L)=\sum_{i=1}^{n} H(l_i)$. Next, we compute the entropy score and the corresponding disclosure rate for each simulated system in two sets of experiments.

5. Empirical evidence from applications in the wild

The trail re-identification phenomenon has been assessed in several different environments. In this section, we review two particular cases studies from which we will abstract a generative model. The first pertains to healthcare, and recounts how individuals visit locations in the physical world. The second pertains to the Internet and illustrates how individuals visit virtual domains.

[1] The REIDIT-I algorithm trades between re-identification discovery and efficiency. All of the re-identifications made by the algorithm are correct, but, to run efficiently, the algorithm does not guarantee discovery of all re-identifications in the system [42].

5.1. Case description: physical trails

The first case is based on the study in [42], which utilized publicly-available hospital discharge databases from the State of Illinois for the years 1990–1997 [63]. In such databases, patient demographics, hospital identity, and diagnosis codes are among the attributes stored with each hospital visit. The demographic attributes for patients include date of birth, gender, five-digit zip code of residence, as well as an identifier for the hospital visited. As mentioned earlier, these demographics can be highly distinguishing for patients and thus we can track them from one hospital to the next. In the case study, the patient subpopulations diagnosed with particular DNA-based disorders (which were documented in the diagnosis codes) were extracted from the discharge databases. The demographics associated with these patients were found to 99% unique and thus, the trails for these patients were assumed to represent unique individuals.

In [42], it was shown that the distribution of individuals to hospitals varies from uniform to Gaussian. To substantiate the empirical evidence presented in this paper, we consider the hospital-visit trails of two particular subpopulations, *Cystic Fibrosis* and *Phenylketonuria*. In these datasets, the entities were patients and the locations were hospitals. The size of the populations was 1149 and 77 patients over 174 and 57 hospitals, respectively.

This study is generative, in that according to the National Association of Health Data Organizations reports, 44 of 50 states have legislative mandates to gather hospital-level data on each patient visit [49]. In the Illinois databases, for example, there are approximately 1.3 million discharges per year with compliance for greater than 99% of all hospital discharges in the state.

5.2. Case description: virtual trails

The second case is based on [37], in which the trails of IP addresses in a distributed online environment were considered. The dataset used in this study was compiled by the Homenet project at Carnegie Mellon University, who provided families in the Pittsburgh area with Internet services in exchange for the monitoring and recording of the families' online services and transactions [35]. Trails were built from 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 5000 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 power-law distribution of people to places, specifically a Zipf distribution, which represents high skew [22].

5.3. Empirical findings

Across the case studies we found that disclosure 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 disclosure. For example, we ranked the popularity of each location by the number of distinct subjects visiting the location. When we measured trail disclosure rates from the least popular location to a location with a specific popularity, we found the disclosure rate correlated the average number of people per location. The result is shown in Fig. 3, where we depict disclosure rates for the three populations. The first two plots are derived from the healthcare case study, where the leftmost plot corresponds to a population afflicted with Cystic Fibrosis and the middle plot to the population afflicted with phenylketonuria. These two cases establish a

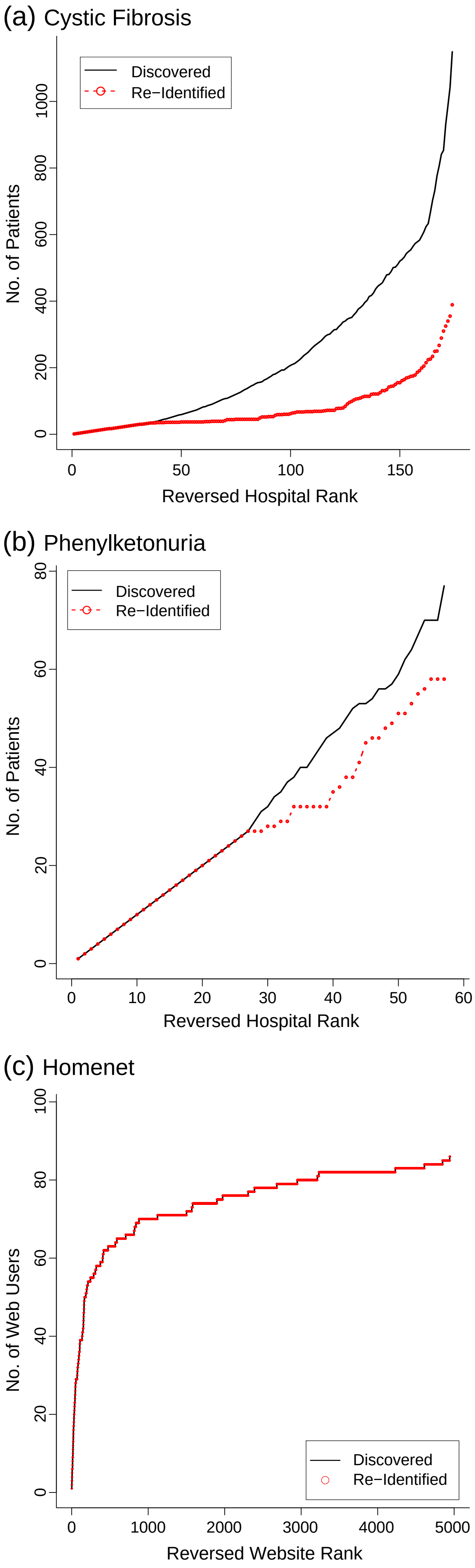


Fig. 3. Trail re-identification in three case studies. The number of locations increases from least-visited to most-visited.

comparison between the feasibility of trail disclosure on a population in which the number of subjects per location is relatively large (Cystic Fibrosis — approximately 6.60) with a population in which the average is closer to a single subject per location (Phenylketonuria — approximately 1.35). The rightmost plot 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 Cystic Fibrosis dataset shown in Fig. 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 Fig. 3(c), we find evidence of a logarithmic relation between the number of locations considered and the number of trail re-identifications. Furthermore, the Phenylketonuria dataset in Fig. 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 disclosures.

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

6. A controlled simulation study

The case studies suggest there are many aspects of location-based information which influence trail disclosure. 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. Our experiments focus on the number of locations and the distributions guiding subject access to these locations. We study the disclosure rate and the corresponding entropy score in different configurations of locations access distribution patterns and the system parameters that govern such distributions.

6.1. Design of experiments

For our experiments, we fix the number of subjects to 1000. We simulate uniform and high skewed distributions of subjects per location. Additionally, we simulate a special case in which neither trail matrix has *'s, which is called a *completely specified* environment and a scenario where one trail matrix has *'s, which we call an *incompletely specified* environment. From an operational point of view, in the simulation of completely specified systems, we generate two equivalent trail matrices. In the simulation of incompletely specified systems, instead, we generate trail matrices for a completely specified 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.

6.1.1. 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 0's and 1's, 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 Fig. 4.[2]

6.1.2. Heavy-tailed simulation

In this setting, subjects visit locations according to Zipf distributions, which lead to the highly skewed 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 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.

For each tested data point, such as $<|L| = 10, p = 0.3>$, we generate 100 populations. Populations that are guided by the Zipf distribution are generated using the formula described above.

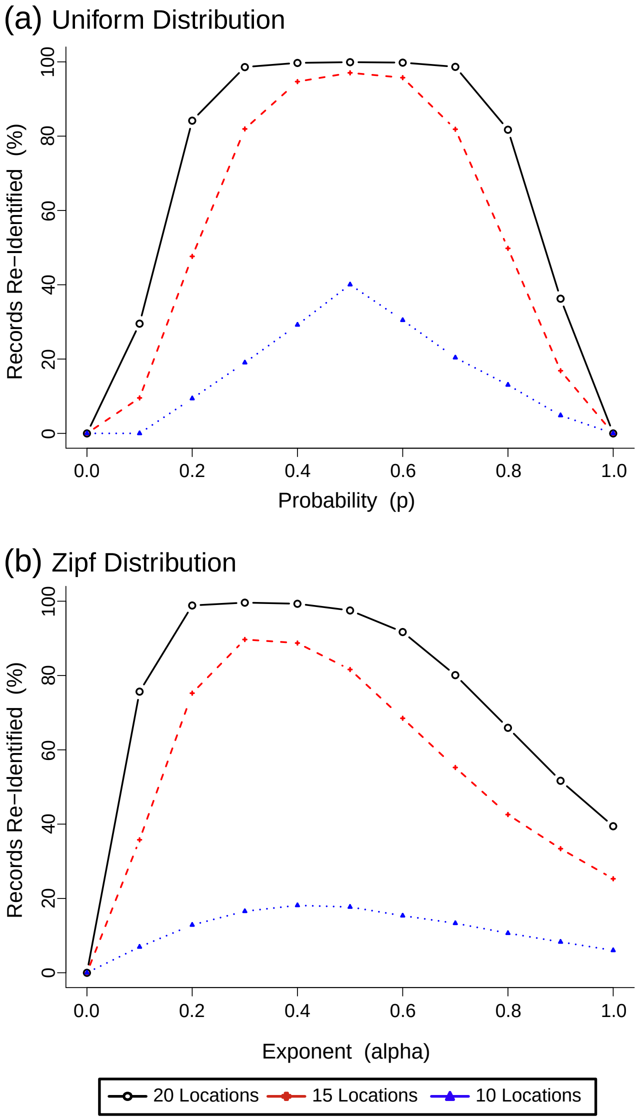


Fig. 4. Disclosure of simulated unreserved location access distributions.

6.2. Entropy score and trail disclosure rate

The resulting 10-point plots for unreserved and reserved systems are depicted in Figs. 4 and 5. In these plots the mean percentage and

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

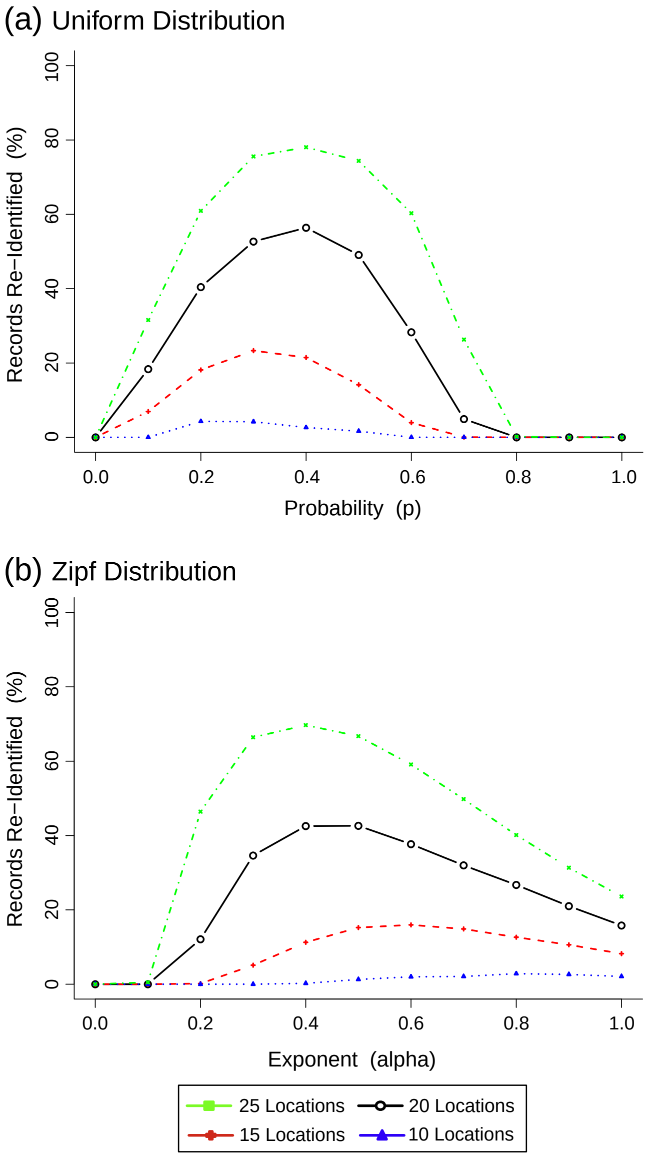


Fig. 5. Disclosure of simulated reserved location access distributions.

plus/minus one standard deviation[3] 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 disclosure 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 disclosures than the Zipf distribution. This can be seen by comparing the disclosure maximum, or peaks, in the left and right panels. Consider Fig. 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 disclosure, given a specific distribution for location access patterns. To compare distribution archetypes, such as uniform vs. Zipf, we measure the area under the disclosure curve. This is calculated as the total area under the 10-point mean disclosure curve (average number of disclosures in 100 simulated populations). The results of this calculation are presented in Fig. 6(a) and (b). Though the uniform distribution always yields the larger maximum number of disclosures, 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 Fig. 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 Fig. 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 the number of locations in consideration is small 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.

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 Figs. 4 and 5. In Fig. 7 we report the results for the unreserved case.

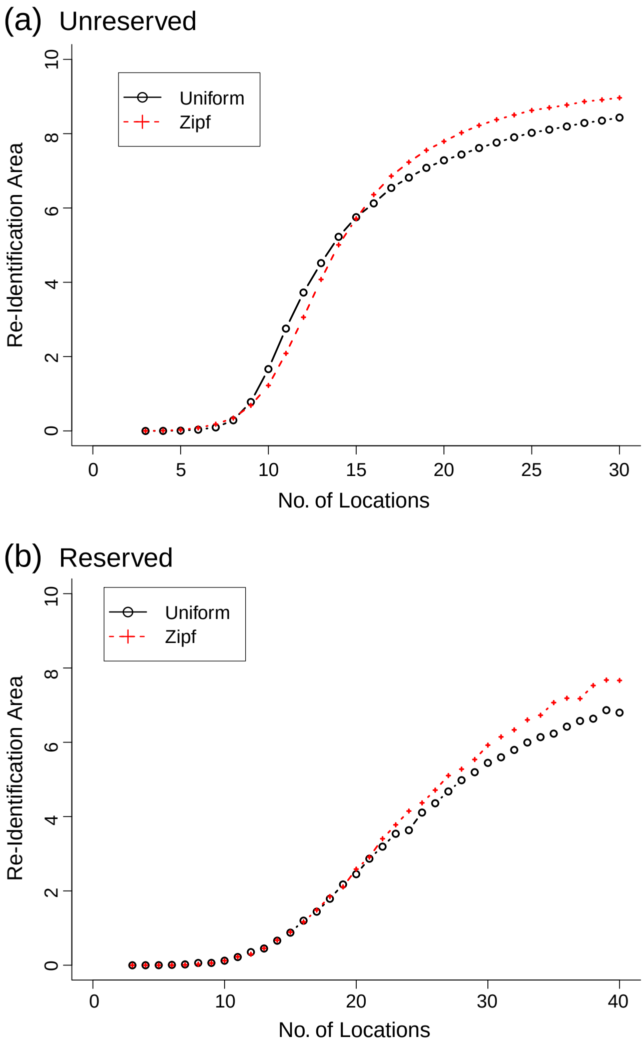


Fig. 6. Area under the mean disclosure curves for simulated populations.

[3] In Fig. 4, the error bars are too small to be visible.

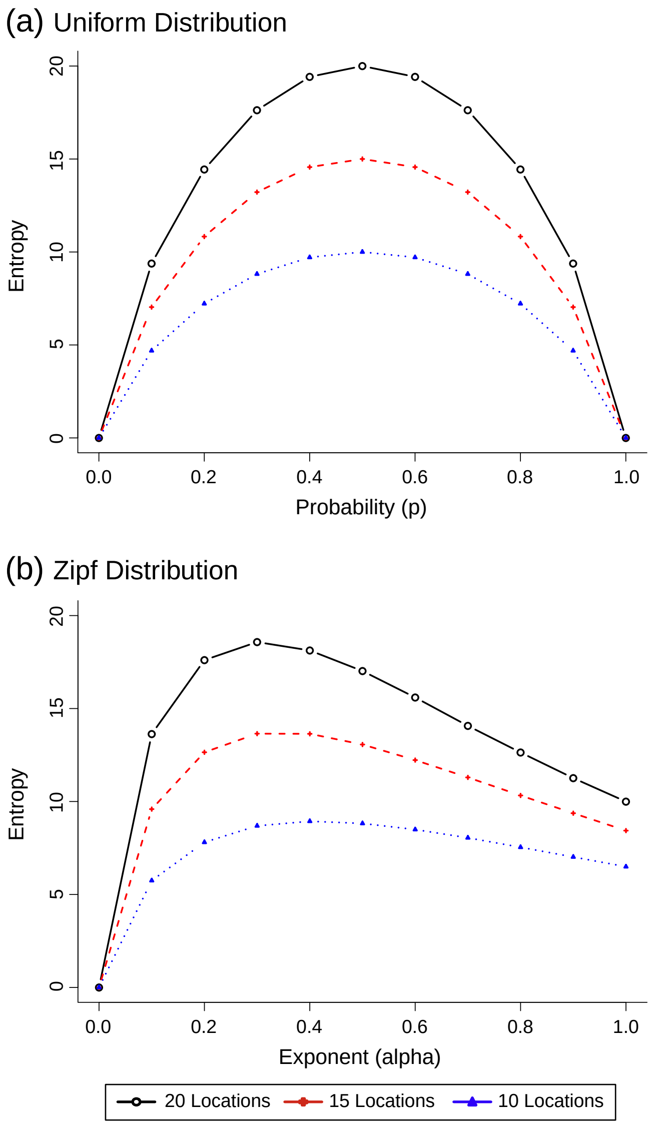


Fig. 7. Entropy plots corresponding to parameter values in the left and right panels of Fig. 4.

6.3. Evaluation

In order to assess whether the entropy metric is able to capture the notion of distinguishability, in this section, we introduce the distance measure *shape score* σ to measure the difference and correlation between and disclosure rate and the corresponding entropy score of a system. Let us denote the entropy score of a population that visit the set of available locations L by $E(i)$, and the actual disclosure rate of the system $R(i)$, where i is a point in the grid, G, for the interval $[0,1]$ we used to generate the disclosure curves in Figs. 4 and 5. Let $\max(R) = R(i^*)$ where $i^* = \text{argmax}\{R(i), i \in G\}$, and let $\max(E) = E(j^*)$ where $j^* = \text{argmax}\{R(j), j \in G\}$. The scaling factor is then $\frac{\max(R)}{\max(E)}$, and the distance metric, σ, is defined as follows,

$$\sigma(E,R) = \sum_{i=1}^{10} \sigma_i(E,R) \tag{3}$$

$$= \sum_{i=1}^{10} \left| E(i) \frac{\max(R)}{\max(E)} - R(i) \right|. \tag{4}$$

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

$$\sigma_i(E,R) = \left| E(i) \frac{\max(R)}{\max(E)} - R(i) \right| \tag{5}$$

$$= \left| E(i^*) \frac{R(i^*)}{E(i^*)} - R(i^*) \right| = 0. \tag{6}$$

We compute the shape scores σ between two curves. The scaling factor is proportional to the ratio between the peaks of the two curves. The resulting information from the shape score is summarized in Fig. 8. As values for shape scores tend 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 Fig. 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

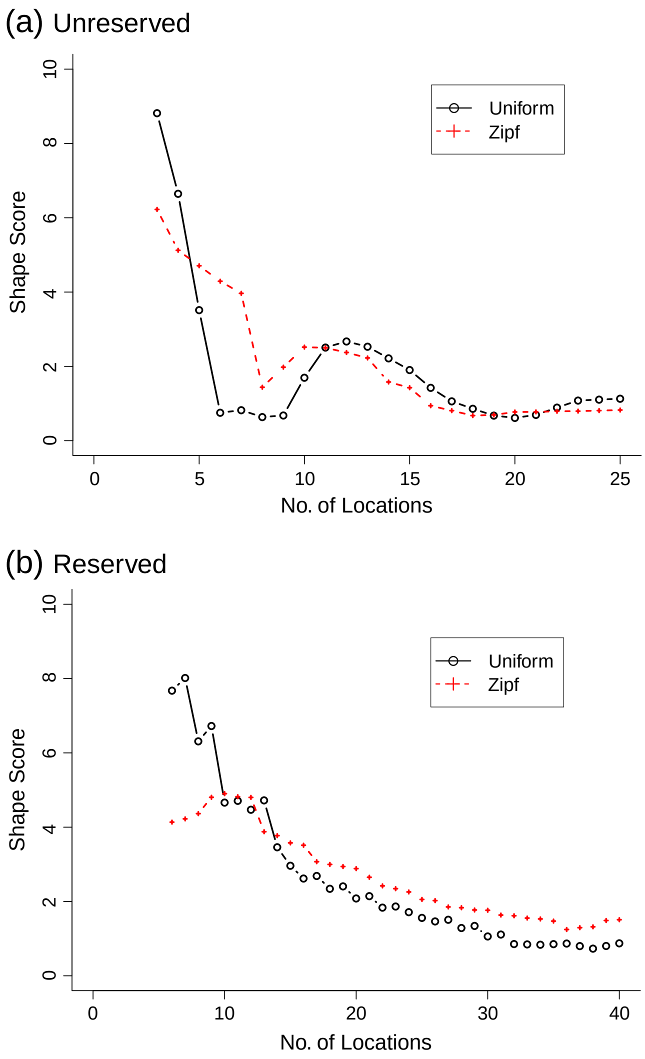


Fig. 8. Shape scores for similarity in simulated distributions and entropy.

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 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.

The shape score allows for the discovery of another notable feature that captures how the distribution type influences 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 disclosure method leverages trails. First, consider the linkage algorithms. In an unreserved system, the disclosure method looks for a unique bit pattern because there are no *'s. So both 1's and 0's are contributing evenly to the disclosure process. This is why the disclosure 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 a reserved environment though, a * value in a trail functions as a 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 disclosure 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 disclosures. Thus, the total quantity of disclosures 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 disclosures, but not the maximum, for the Zipf is greater than the uniform.

7. Discussion

The above analysis provides a wealth of insight into the effects of location access patterns on the degree to which trail disclosure 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 entropy, and especially the extent to which the amount of re-identifiability can be predicted from the statistical characteristics of a database before its release. 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.

One of the more interesting findings of our experiments is that high-skew location access patterns yield higher *overall* disclosure 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* disclosures, 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 incompletely specified 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 disclosure 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 a certain portion 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 [39,44], 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 disclosure, which can be subject to rigorous proof. Disclosure risk analysis provides a proxy by which we can develop provable protection models.

Furthermore, we find there is a strong correlation between the entropy of the system and disclosure. 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 disclosures that can be made, the distribution of location access patterns, or the entropy, should be modeled. In pursuing these strategies, it becomes crucial that the information which is 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.

7.1. 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 disclosures that can be made. Our experiments suggest that low entropy systems correlate with high re-identifiability, but they 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 disclosure, 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 [34,55]. 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 disclosure 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 disclosure?" For instance, to what extent is disclosure 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 about. As a result, another feasible direction for research into the fundamentals of trail disclosure is to study the effect of mixture models of distributions on disclosure.

7.2. Concluding remarks

In this paper we proposed a novel approach to formally evaluating the risk of trail disclosure in distributed database systems. Specifically, we introduced an entropy metric for assessing the effect of different location access distributions on trail disclosure when an individual's data is distributed across a set of locations. We provided case-based and controlled experimental evidence that implies the characteristics of the distributions of location access patterns is one of the main factors that influence disclosure. Though our model is based on empirical observations and simulation, this work provides a foundation for both basic and applied trail linkage and data disclosure research, in general. Our work also provides a compelling exploration into the relation between the disclosure of a set of trails and the information they carry. From a practical point of view, we are able to turn our intuitions into a quantification of the disclosure risk that can be predicted from the statistical characteristics of databases before their release.

Acknowledgments

The authors thank Kathleen Carley, Latanya Sweeney, and Stephen Fienberg for insightful discussions. This work was partially supported by the National Science Foundation under grants nos. DMS-0907009 and IIS-1017967, by the National Institutes of Health under grant no. R01-GM096193, and by the Army Research Office Multidisciplinary University Research Initiative under grant no. 58153-MA-MUR all to Harvard University, and by the National Institutes of Health under grant no. R01-LM009989 to Vanderbilt University. Earlier funding was provided by the Data Privacy Laboratory and by the Center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Institute of Health, the Office of Naval Research, the National Science Foundation, or the U.S. government.

References

[1] N. Adam, J. Worthmann, Security-control methods for statistical databases, a comparative study, ACM Computing Surveys 21 (4) (1989) 515–556.

[2] E.M. Airoldi. A statistical theory of record linkage with applications to privacy. Technical Report CMU-ISRI-05–112, School of Computer Science, Carnegie Mellon University, October 2004. Revision, December 2005.

[3] A. Amiri, Dare to share: protecting sensitive knowledge with data sanitization, Decision Support Systems 43 (1) (2007) 181–191.

[4] N. Awad, M. Krishnan, The personalization privacy paradox: an empirical evaluation of information transparency and the willingeness to be profiled online for personalization, MIS Quarterly 30 (1) (2006) 13–28.

[5] X. Bai, R.D. Gopal, M. Nunez, D. Zhdanov, Managing information security risk in business processes, International Journal of Decision Science 1 (1) (2010) 55–65.

[6] G. Bansal, F. Zahedi, The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online, Decision Support Systems 49 (2) (2010) 138–150.

[7] S. Bender, R. Brand, J. Bacher, Re-identifying register data by survey data: an empirical study, Statistical Journal of the United Nations ECE 18 (2001) 373–381.

[8] M. Bezzi, An entropy based method for measuring anonymity, Security and Privacy in Communications Networks and the Workshops, 2007. Third International Conference on SecureComm 2007, Sept. 2007, pp. 28–32.

[9] C. Boyens, R. Krishnan, R. Padman, On privacy-preserving access to distributed heterogeneous healthcare information, HICSS, 2004.

[10] F. Collins, Has the revolution arrived? Nature 464 (2010) 674–675.

[11] L. Cox, Network models for complementary cell suppression, Journal of the American Statistical Association 90 (432) (1995) 1453–1462.

[12] T. Dalenius, Finding a needle in a haystack or identifying anonymous census records, Journal of Official Statistics 2 (3) (1986) 329–336.

[13] G.J. de Moor, B. Claerhout, F. de Meyer, Privacy enhancing technologies: the key to secure communication and management of clinical and genomic data, Methods of Information in Medicine 42 (2003) 148–153.

[14] 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 (157) (2002) 53182–53273.

[15] J. Domingo-Ferrer, V. Torra, Disclosure risk assessment in statistical microdata protection via advanced record linkage, Statistics and Computing 13 (4) (2003) 343–354.

[16] J. Domingo-Ferrer, J.M. Mateo-Sanz, V. Torra, Comparing sdc methods for microdata on the basis of information loss and disclosure risk, Proceedings of ETK-NTTS 2001, Luxemburg: Eurostat, Eurostat, 2001, pp. 807–826.

[17] S. Dreiseitl, S.A. Vinterbo, L. Ohno-Machado, Disambiguation data: extracting information from anonymized sources, Proceedings of AMIA Symposium 8 (2001).

[18] G.T. Duncan, D. Lambert, Disclosure-limited data dissemination (with discussion), Journal of the American Statistical Association 81 (1986) 10–28.

[19] G.T. Duncan, D. Lambert, The risk of disclosure of microdata, Journal of Business and Economic Statistics 7 (1989) 207–217.

[20] G.T. Duncan, S.E. Fienberg, R. Krishnan, R. Padman, S.F. Roehrig, Disclosure limitation methods and information loss for tabular data, in: J. Theeuwes, P. Doyle, J. Lane, L. Zayatz (Eds.), Confidentiality, disclosure and data access: theory and practical applications for statistical agencies, North-Holland, Amsterdam, 2001.

[21] E. Elamir, C. Skinner, Record-level measures of disclosure risk for survey microdata, Journal of Official Statistics 22 (3) (2006) 525–539.

[22] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the Internet topology, Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, 1999, pp. 251–262.

[23] I.P. Fellegi, A.B. Sunter, A theory for record linkage, Journal of the American Statistical Association 64 (328) (1969) 1183–1210.

[24] E. Fernandez-Medina, J. Trujillo, R. Villarroel, Access control and audit model for the multidimensional modeling of data warehouses, Decision Support Systems 42 (3) (2006) 1270–1289.

[25] S.E. Fienberg, Privacy and confidentiality in an e-commerce world: data mining, data warehousing, matching and disclosure limitation, Statistical Science 21 (2006) 143–154.

[26] S.E. Fienberg, U.E. Makov, R.J. Steele, Disclosure limitation using perturbation and related methods for categorical data (with discussion), Journal of Official Statistics 14 (1998) 485–512.

[27] R. Garfinkel, R.D. Gopal, P.B. Goes, Privacy protection of binary confidential data against deterministic, stochastic, and insider threat, Management Science 48 (2002) 749–764.

[28] S. Goldman, Information Theory, Prentice Hall, New York City, NY, 1953.

[29] P. Golle, Revisiting the uniqueness of simple demographics in the U.S. population, Proceedings of the ACM Workshop on Privacy in the Electronic Society, 2006, pp. 77–80.

[30] V. Griffith, M. Jakobsson, Messin' with Texas: deriving mother's maiden name using public records, Proceedings of the Applied Cryptography and Network Security Conference, New York, NY, 2005.

[31] J. Gulcher, K. Kristjansson, H. Gudbjartsson, K. Stefansson, Protection of privacy by third-party encryption in genetic research, European Journal of Human Genetics 8 (2000) 739–742.

[32] G. Hardin, The tragedy of the commons, Science 162 (3859) (1968) 1243–1248.

[33] S. Jajodia, Database security and privacy, ACM Computing Surveys 28 (1) (1996) 129–131.

[34] E. Johnson, W. Moe, P. Fader, S. Bellman, G. Lohse, On the depth and dynamics of online search behavior, Management Science 50 (3) (2004) 299–308.

[35] R. Kraut, T. Mukhopadhyay, J. Szczypula, S. Kiesler, B. Scherlis, Information and communication: alternative uses of the internet in households, Information Systems Research 10 (2000) 287–303.

[36] M.D. Larsen, D.B. Rubin, Iterative automated record linkage using mixture models, Journal of the American Statistical Association 96 (March 2001) 32–41.

[37] B. Malin, Betrayed by my shadow: learning data identity via trail matching, Journal of Privacy Technology (2005), 20050609001.

[38] B. Malin, A computational model to protect patient data from location-based re-identification, Artificial Intelligence in Medicine 40 (3) (2007) 223–239.

[39] B. Malin, Secure construction of k-unlinkable patient records from distributed providers, Artificial Intelligence in Medicine 48 (2010) 29–41.

[40] B. Malin, E.M. Airoldi, The effects of location access behavior on re-identification risk in a distributed environment, Privacy Enhancing Technologies, number 4258 in Lecture Notes in Computer Science, Springer-Verlag, 2006, pp. 413–429.

[41] B. Malin, L. Sweeney, Determining the identifiability of dna database entries, Proceedings of the American Medical Informatics Association Annual Symposium, Hanley & Belfus, Los Angeles, Ca, 2000, pp. 537–541.

[42] B. Malin, L. Sweeney, 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 (3) (2004) 179–192.

[43] B. Malin, L. Sweeney, A secure protocol to distribute unlinkable health data, Proceedings of the American Medical Informatics Association Annual Symposium, American Medical Informatics Association, Washington, DC, 2005, pp. 485–489.

[44] B. Malin, L. Sweeney, Composition and disclosure of unlinkable distributed databases, Proceedings of the 22nd IEEE International Conference on Data Engineering, Atlanta, GA, 2006.

[45] T. Manolio, Collaborative genome-wide association studies of diverse diseases: programs of the NHGRI's office of population genomics, Pharmacogenomics 10 (2) (2009) 235–241.

[46] J. Mateo-Sanz, J. Domingo-Ferrer, Practical data-oriented microaggregation for statistical disclosure control, IEEE Transactions on Knowledge and Data Engineering 14 (1) (2002) 189–201.

[47] E. Meux, Encrypting personal identifiers, Health Services Research 29 (2) (1994) 247–256.

[48] A. Motsinger-Reif, E. Jorgenson, M. Relling, D. Kroetz, R. Weinshilboum, N. Cox, and D. Roden. Genome-wide association studies in pharmacogenomics: successes and lessons. Pharmacogenetics and Genomics (in press).

[49] NAHDO, Nahdo inventory of statewide hospital discharge data activities, National Association of Health Data Organizations, Falls Church, VA, 2000.

[50] A. Narayanan, V. Shmatikov, Robust de-anonymization of large sparse datasets, Proceedings of the IEEE Security and Privacy Conference, 2008, pp. 111–125.

[51] National Institutes of Health, Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (gwas), August 2007.

[52] National Institutes of Health, NIH final statement on sharing research data, February 2003.

[53] E. Ngai, A. Gunasekaran, A review for mobile commerce research and applications, Decision Support Systems 43 (1) (2007) 3–5.

[54] M. Nunez, R. Garfinkel, R.D. Gopal, Stochastic protection of confidential information in statistical databases: a hybrid of query restriction and data perturbation, Operations Research 55 (2007) 890–908.

[55] Y. Park, P. Fader, Modeling browsing behavior at multiple websites, Marketing Science 23 (3) (2004) 280–303.

[56] C. Quantin, H. Bouzelat, F. Allaert, A. Benhamiche, J. Faivre, L. Dusserre, How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure, International Journal of Medical Informatics 49 (1) (1998) 117–122.

[57] S. Samuels, A Bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment, Journal of Official Statistics 14 (4) (1998) 373–383.

[58] R. Sarathy, K. Muralidhar, Secure and useful data sharing, Decision Support Systems 42 (1) (2006) 204–220.
[59] J.A. Schoenman, J.P. Sutton, A. Elixhauser, D. Love, Understanding and enhancing the value of hospital discharge data, Medical Care Research and Review 64 (4) (2007) 449–468.
[60] C.E. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (July 1948) 379–423.
[61] C.E. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (October 1948) 623–656.
[62] C. Skinner, Assessing disclosure risk for record linkage, PSD '08: Proceedings of the UNESCO Chair in data privacy international conference on Privacy in Statistical Databases, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 166–176.
[63] State of Illinois Health Care Cost Containment Council, State of Illinois Health Care Cost Containment Council, Data release overview, Springfield, IL, March 1998.
[64] L. Sweeney, Uniqueness of simple demographics in the us population, Technical Report LIDAP-WP04, Data Privacy Laboratory, Carnegie Mellon University, Pittsburgh, PA, 2000.
[65] L. Sweeney, k-anonymity: a model for protecting privacy, International Journal of Uncertain Fuzz Knowledge-Based Systems 10 (5) (2002).
[66] V. Torra, J.M. Abowd, J. Domingo-Ferrer, Using Mahalanobis distance-based record linkage for disclosure risk assessment, Privacy in Statistical Databases, 2006, pp. 233–242.
[67] T.M. Truta, F. Fotouhi, D. Barth-Jones, Assessing global disclosure risk in masked microdata, WPES '04: Proceedings of the 2004 ACM workshop on Privacy in the electronic society, ACM, New York, NY, USA, 2004, pp. 85–93.
[68] L. Willenborg, T. de Waal, Elements of statistical disclosure contro, Springer, New York, NY, 1996.
[69] L. Willenborg, T. de Waal, Statistical Disclosure Control in Practice, Springer, New York, NY, 1996.
[70] W.E. Winkler, Matching and record linkage, in: B.G. Cox, et al., (Eds.), Business Survey Methods, J. Wiley, New York, NY, 1995, pp. 355–384.
[71] W.E. Winkler, Methods for record linkage and bayesian networks, Technical Report Statistics-2002–05, Statistical Research Division, U.S. Census Bureau, Washington, DC, 2002.
[72] W.E. Winkler, Data cleaning methods, Proceedings of the ACM SIGKDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation, Washington, DC, 2003.
[73] W. Yancey, W. Winkler, R. Creecy, Disclosure risk assessment in perturbative microdata protection, Inference Control in Statistical Databases, number 2316 in Lecture Notes in Computer Science, Springer-Verlag, 2002, pp. 49–60.
[74] D. Zhu, X.B. Li, S. Wu, Identity disclosure protection: a data reconstruction approach for privacy-preserving data mining, Decision Support Systems 48 (1) (2009) 133–140.

Dr. Edoardo M. Airoldi is an Assistant Professor of Statistics at Harvard University. He is also a member of the Center for Systems Biology in the Faculty of Arts and Sciences at Harvard University. He received a PhD degree in Computer Science from Carnegie Mellon University. He was postdoctoral fellow at Princeton University, in the Lewis-Sigler Institute for Integrative Genomics, and the Department of Computer Science. His research interests include statistical methodology and theory for the analysis of complex networks and random graph dynamics, with application to the social and biological sciences.

Dr. Xue Bai is an Assistant Professor of Management Information Systems in the School of Business at University of Connecticut. Her research interests include developing data mining and machine learning methods for text classification, sentiment extraction, online marketing analysis and clinic diagnosis. Another of her research interests is in the area of risk management of information and data quality related issues in business processes.

Dr. Bradley Malin is an Assistant Professor of Biomedical Informatics in the School of Medicine, Vanderbilt University. His research focuses on the construction and evaluation of data privacy models for personal information that is collected, stored, and shared in large complex systems. His research aims to design technology that is accountable to organizational, social, and legal regulations. He is particularly interested in clinical and genetic information captured in electronic medical records and shared for research purposes.