Process-Driven Data Privacy

Weiyi Xia
Vanderbilt University
Nashville, TN, USA

Murat Kantarcioglu
University of Texas at Dallas
Dallas, TX, USA

Zhiyu Wan
Vanderbilt University
Nashville, TN, USA

Raymond Heatherly
Vanderbilt University
Nashville, TN, USA

Yevgeniy Vorobeychik
Vanderbilt University
Nashville, TN, USA

Bradley Malin
Vanderbilt University
Nashville, TN, USA

ABSTRACT

The quantity of personal data gathered by service providers via our daily activities continues to grow at a rapid pace. The sharing, and the subsequent analysis of, such data can support a wide range of activities, but concerns around privacy often prompt an organization to transform the data to meet certain protection models (e.g., k-anonymity or ε-differential privacy). These models, however, are based on simplistic adversarial frameworks, which can lead to both under- and over-protection. For instance, such models often assume that an adversary attacks a protected record exactly once. We introduce a principled approach to explicitly model the attack process as a series of steps. Specifically, we engineer a factored Markov decision process (FMDP) to optimally plan an attack from the adversary's perspective and assess the privacy risk accordingly. The FMDP captures the uncertainty in the adversary's belief (e.g., the number of identified individuals that match the de-identified data) and enables the analysis of various real world deterrence mechanisms beyond a traditional protection model, such as a penalty for committing an attack. We present an algorithm to solve the FMDP and illustrate its efficiency by simulating an attack on publicly accessible U.S. census records against a real identified resource of over 500,000 individuals in a voter registry. Our results demonstrate that while traditional privacy models commonly expect an adversary to attack exactly once per record, an optimal attack in our model may involve exploiting none, one, or more individuals in the pool of candidates, depending on context.

Categories and Subject Descriptors

K.4.1 [**Computing Milieux**]: Computers and Society—*Privacy*; K.6.0 [**Computing Milieux**]: General—*Economics*

General Terms

Algorithms; Security; Theory

CIKM'15, October 19–23, 2015, Melbourne, Australia.

DOI: http://dx.doi.org/10.1145/2806416.2806580.

Keywords

Privacy; Re-identification; Planning

1. INTRODUCTION

The quantity, quality, and diversity of personal data we shed through our daily activities continues to grow at a rapid pace. This data is collected by a wide range of organizations to assist in the optimization and refinement of the services they provide [1, 13]. At the same time, it is increasingly recognized that personal data has substantial worth beyond its initial use, such that it can be repurposed for a variety of endeavors, ranging from transparency in operations to basic research [23]. Despite the recognized value of personal data, organizations worry about how best to protect the rights of their constituents while maximizing the benefits [29].

One such right that our society tends to focus on is personal privacy. While privacy is an overloaded term that takes on many different forms [25], one quantitative, broadly concerned definition centers on the notion of anonymity. There are various regulations that encourage organizations to suppress identifying information from personal data prior to its dissemination. Several examples of regulations with explicit identity protections include the Privacy Rule of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) in the United States and the Data Protection Directive in the European Union. However, there is growing evidence that the resulting data is susceptible to intrusions [19]. One major type of intrusion is the re-identification attack, which happens when an adversary combines de-identified records with external resources to determine the identity of the corresponding individuals (e.g., [12, 18, 26, 27]).

While such attacks are possible, it is unclear if they are probable. This is important because laws and regulations do not require perfect protection, but rather that data be shared in a manner that makes it difficult to ascertain an individual's identity. Organizations are thus afforded an opportunity to achieve data protection using risk management techniques, but they are hampered from accomplishing this goal for several reasons. First, there is little historical data on re-identification attacks [6]. This may be due to the rareness of such events or that they transpired behind closed doors. Second, prior investigations assume privacy risks can only be managed by perturbing data according to a formal model (e.g., k-anonymity [24] or ϵ-differential privacy [4]). Yet, there are many other elements that contribute to privacy risks, such as deterrence mechanisms (e.g., data user agreements, the time and effort to gather the external information necessary to compromise the data, or penalties

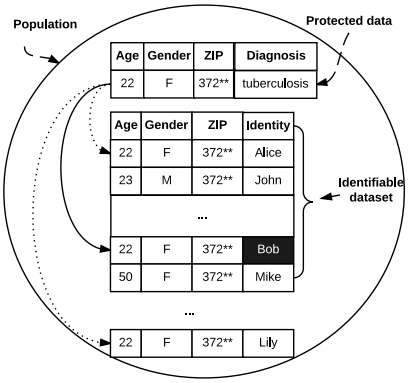

Figure 1: An example of a re-identification scenario.

for misusing data), that influence an adversary. Third, the typical adversarial model invoked in this setting assumes an attack is perpetrated in a pre-defined manner. For instance, under a k-anonymity framework it is assumed the adversary exploits an individual at random, with an expected success rate of $1/k$. Yet, an adversary could attempt to exploit all of the individuals in this group in the hopes that one will be the correct corresponding person. Under the ε-differential privacy framework, the publisher limits the number of times an adversary can query a statistical database to satisfy the given privacy budget.

To overcome these limitations, we introduce a principled approach to assess re-identification risk using a process framework that incorporates data- and penalty-based disincentives. This approach is based on a stochastic model, yielding a quantitative evaluation of the re-identification risk.

1.1 An Orienting Example

For context, Figure 1 depicts an example of the type of attack our re-identification risk framework is designed to handle. Here, Bob corresponds to the de-identified data. Now, there is a probability he is in the external dataset that contains identifying information and a set of attributes in common with the de-identified data. In the event the adversary accesses the external dataset, he will find that the record with a diagnosis of tuberculosis might correspond to Alice, Bob or neither (because the corresponding individual may not be in the dataset). The adversary will be successful if he exploits Bob; however, it is possible the adversary is not sufficiently motivated to start the attack and the adversary may choose to terminate his attack at any point in the process (because of insufficient expected payoffs). Our framework explicitly models the decisions the adversary must make and computes the probability that the adversary will reach a successful re-identification.

1.2 Contributions

The specific contributions of this paper include:

1. Re-identification Risk Model. This research proposes a novel re-identification risk framework that formalizes incentive and deterrence mechanisms (e.g., potential penalties and information uncertainty) that are present in the real world environments where a de-identified dataset is made available. This framework explicitly models the adversary as an optimal planning agent using a factored Markov decision process (FMDP). Given that the state space of the FMDP grows rapidly, we introduce a two-level linear programming algorithm to efficiently solve it.

2. Case Study. We evaluate risk in the scenario where an adversary attempts to leverage a public voter registry in a specific U.S. State to attack de-identified Census records. Under the traditional adversarial model, the adversary is assumed to randomly choose an individual from the group of indistinguishable individuals in the external dataset that matches the targeted record under attack. However, our findings illustrate that the adversary's behavior is highly dependent on the deterrence mechanism set in place. In particular, if the adversary's expectation of the probability of being detected for each exploitation of an individual remains constant across attempts and he will always be penalized if detected, he will choose to either attack 1) all the individuals in the correspond group or 2) none of the individuals. This result illustrates how traditional beliefs about risk can either underprotect or overprotect the data.

3. Sensitivity Analysis. We conduct a detailed sensitivity analysis on the parameters of the model to illustrate how the environment and policy decisions can influence the adversary's behavior. Specifically, we demonstrate how changes in the costs in each stage of the attack, penalty-based deterrence mechanism, and the probability of detection influences when the adversary will cease their attack. Our result demonstrates that the adversary's threshold is highly dependent on the deterrence mechanisms that are in place. This result is notable because it suggests that adversaries may be sufficiently deterred with a small amount of data manipulation, provided appropriate detection and penalization policies are instantiated.

We stress that our framework focuses on the application scenarios where record-level data needs to be shared. For example, in health care research setting, data sets containing patient information may need to be shared. In such scenarios, generalizing the data may be acceptable (e.g., instead of reporting exact birthday of individuals, we can only disclose the birth year) but adding noise may not be acceptable due to the semantic errors that may be introduced. Therefore, our proposed approach can be seen as complementary to recent developments in ϵ-differential privacy where synthetic data sets could be generated for preserving privacy and it is applicable for scenarios where differential privacy based approaches are not appropriate.

The remainder of this paper is organized as follows. In Section 2, we review related work on adversarial modeling and re-identification risk assessment approaches. In Section 3, we introduce the re-identification risk quantification framework. In Section 4 and 5, we present the FMDP and algorithms to solve it and compute the re-identification probability from the Markov process efficiently. In Section 6 and 7, we present an empirical analysis. In Section 8, we discuss the limitations of this work and provide future directions.

2. RELATED WORK

To provide context for our study, we discuss research in privacy preserving data publishing and adversarial modeling areas with a focus on where our work diverges.

2.1 Data Privacy Views

There are many different views on what constitutes a privacy violation when considering data publishing. These views argue that privacy can be compromised when a record is linked to the individual from whom it was derived (often referred to as *identity disclosure*) [10], the inference of a

sensitive value associated with the corresponding individual (often referred to as *attribute disclosure*) [16], the ability to detect if someone is a member of a dataset (often referred to as the presence/absence problem) [9, 20], or the degree to which viewing an individual's contribution to a dataset permits an adversary to gain knowledge about them (the basis of models like ε-differential privacy) [2, 5, 8].

In this research, we focus on the identity disclosure problem because this is the primary focus of current regulation. We note, however, that our framework is applicable to any of the above scenarios because we are not introducing a new method for manipulating data. Rather, we are showing how to reason about pressures (e.g., penalties for misuse) that are beyond the scope of what such data manipulation techniques offer.

2.2 Disclosure Risk Management

It has been suggested that assessing disclosure risk requires a holistic modeling of different types of adversaries [7]. Such models should account for the motivation, means, opportunity cost, consequence of attempt, and likelihood of success. In this vein, a recent study [17] presented the concept of a data environment that is composed of data, agents and infrastructure. However, such investigations have not provided a formal approach to risk quantification that accounts for the elements in the data environment. Rather, existing disclosure risk measures mainly focus on the uniqueness of records in the dataset and in the population. For instance, three popular disclosure risk metrics (prosecutor, journalist and marketer) [14] assume that the adversary is always motivated to attack and the extra information required for re-identification is always available. As a consequence, the risk level is only dependent on the data itself.

Our disclosure risk measure explicitly formalizes the three elements of the data environment around the adversary's decision making. We note that there have been several other investigations in applying game theoretic frameworks to analyze the adversary's best course of action and the corresponding disclosure risk [28, 31]. For instance, the adversary in Wan et al. [31] is formalized as an opponent of the data publisher in a Stackelberg game. To maximize payout, the adversary decides if they should attack by comparing the potential gain against the cost of committing an attack. Yet, this model oversimplifies the adversary's decision process of gathering, linking, and exploiting data. Moreover, in their formalization, there were no explicitly modeled penalties for detecting the misuse of the data.

To mitigate the disclosure risk, the publisher can adopt various protection methods (e.g., randomization and generalization) according to given formal protection models. To apply these formal protection models in practice, a protection threshold (e.g., the k value for $k$anonymity [24]) needs to be selected and supported by risk and utility analysis. A method based on Pareto-optimality has been proposed to find the solutions with an optimal utility at different levels of k [3]. However, the risk metric used in such methods to determine the threshold is based solely on the protected data. By contrast, our framework can assist in determining what the threshold should be based on both the data set and other elements in the data environment.

2.3 Adversarial Modeling and MDPs

To the best of our knowledge, MDPs have not been used to model adversaries in the privacy preserving data publication setting. Yet it has been proven to be a useful tool in modeling adversary's optimal planning in security problems. This is because the MDP representation captures an adversary's uncertainty on the outcome of a security related action [15]. Similarly, the adversary focused on privacy also exhibits uncertainty. For example, when the adversary chooses to exploit an individual, he is uncertain about whether or not he will be detected and pulished. Also, before the adversary chooses to access the external dataset, he may not be certain about number of identified individuals to which the de-identified record may be related. An important difference between our adversarial model and the one in Letchford and Vorobeychik [15] is that in the security scenario, the adversary terminates once he is caught, whereas in our model, the adversary may only pay a fine and continue to attack.

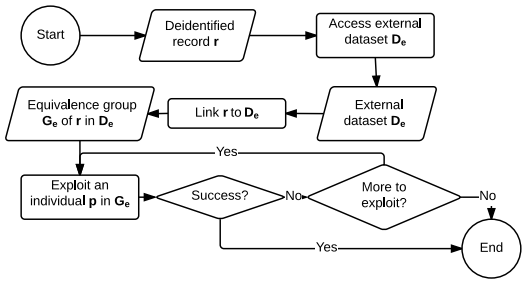

Figure 2: The re-identification attack process.

3. RE-IDENTIFICATION RISK QUANTIFICATION FRAMEWORK

Our framework quantifies the re-identification risk of publishing each record in a de-identified dataset. We assume the dataset is composed of person-level records in a relational form. We define re-identification risk as the composite of the probability that an adversary re-identifies a record and the harm it causes:

$$risk = P_{reid} \times L_{reid} \quad (1)$$

where P_{reid} is the re-identification probability and L_{reid} is the associated publisher loss. We assume L_{reid} is a predefined input, and focus on P_{reid}.

The re-identification probability is derived from the adversary's sequential decision process, outlined in Figure 2. The adversary begins with a de-identified record r. The adversary's first decision is to access an external table D_e or not. His second decision is whether to conduct a linkage attack, which yields an equivalence group of records G_e. This corresponds to the set of individuals with the same value as the target's published quasi-identifier (QI). At this point, each individual $\alpha \in G_e$ has a probability that they actually correspond to the targeted record r. This translates into a probability that an attack (e.g., confirmation of the patient's identity) on α will be successful. If the attack fails, the adversary can choose to exploit another individual from G_e. This process can repeat until the adversary decides to stop or he exhausts all of the records in G_e.

There are several notable aspects of this attack process. First, it should be recognized that this is a stochastic process. For example, the adversary may not know if the individual to whom the target record corresponds is in D_e. Therefore, the outcome of accessing the external dataset is

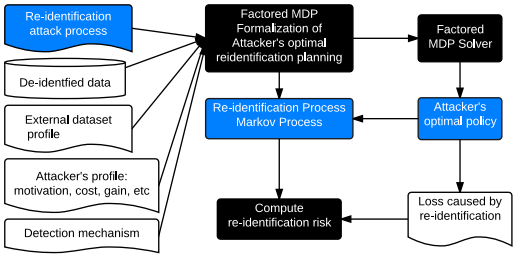

Figure 3: A general architecture of the re-identification risk quantification framework.

uncertain. Furthermore, the result of exploiting an individual is stochastic, with outcomes ranging from success to failure to being detected and punished. A second notable aspect of the attack process is that there is a cost and a reward associated with each action, for instance, the reward for a success, the cost of accessing D_e and the penalty if an attack is detected all determine the adversary's utility.

More precisely, we model the adversary as a planner using a factored Markov decision process (FMDP) [11]. In a FMDP, a state of the world is characterized by a collection of random variables (or factors). The adversary is modeled as a rational agent computing an optimal policy; i.e., an optimal action to choose in each state of the FMDP. Given such a policy, we can compute risk according to Equation 1.

In Figure 3, we show the general architecture of the re-identification risk quantification framework. The framework is composed of three modules (the black rectangles in Figure 3): 1) the FMDP formalization of the adversary's decision process, 2) the FMDP solver, and 3) the re-identification risk computation module. The FMDP formalization module takes four inputs: i) the attack decision process, ii) the de-identified data, iii) the external dataset profile, and iv) the adversary's profile. The factored MDP model is then solved by the FMDP solver module to determine the adversary's optimal policy. Finally, the risk computation module computes the quantified risk value given optimal attack policy and associated probability of successful re-identification attack. In the following sections, we dive into the details of each of the three modules.

4. RE-IDENTIFICATION AS AN FMDP

The FMDP model is a 4-tuple (X, A, R, P), where $X = \{X_0, ..., X_m\}$ is a finite set of random variables, each with a finite domain. In this model, A is a finite set of actions; R is the reward function $R(x, a)$, representing the reward for each action a taken in state $X = x$; and P is a Markovian transition function $P(X'_i|X_i^{parent}, a)$, which represents the probability distribution of the state variable X'_i in the next state given the value of a subset of state variables X_i^{parent} and action a (X_i^{parent} is the set of variables that X'_i is dependent on given the action is a). We denote the value of a state variable X_i in state x as $x[X_i]$. We assume that the FMDP has an infinite horizon, and time is exponentially discounted with a discount factor γ.

4.1 State variables

As summarized in Table 1, the FMDP model is based on eight state variables. Here, we take a moment to provide intuition into each of these variables. First, X_t is a binary variable that represents the termination of an attack. When $X_t = T$ (true), the corresponding state is absorbing, effectively ending the decision process. Next, we assume the existence of an attack detection mechanism, and the state of detection is indicated by a binary variable X_d. The following variable, X_s, indicates whether the exploit is successful (in which case the adversary obtains a positive reward). The next two variables are associated with data manipulation. X_a is a binary indicator of whether the external dataset D_e has been accessed, while X_l is a binary indicator of whether it has been linked to the published target record r. X_p maintains the number of times the exploitation has been detected and penalized. The final two variables, X_g and X_r keep track of the size of the equivalence group and the remaining unexploited individuals in the group. Thus, as the adversary attempts (unsuccessful) attacks on matched records, X_r decreases while X_g remains constant. This is because the original group size associated with linking is fixed. To keep our presentation compact, we represent each state x as a vector $[x_0, ..., x_m]$ in the FMDP model, where x_i denotes the value of the i^{th} variable in the list $[X_t, X_d, X_p, X_s, X_a, X_l, X_g, X_r]$.

Table 1: The state variables of the FMDP model.

Variable	Explanation
X_t, binary	If **T**, attack is terminated
X_d, binary	If **T**, exploit of an individual is detected
X_p, integer	Number of previous exploits penalized
X_s, binary	If **T**, target record r is successfully re-identified
X_a, binary	If **T**, external dataset D_e has been accessed
X_l, binary	If **T**, target record r has been linked to the external dataset D_e
X_g, integer	The size of the equivalence group of target record r in external dataset D_e
X_r, integer	The remaining number of unexploited individuals in the equivalence group for record r in external dataset D_e

Table 2: The actions of the FMDP model.

Action	Explanation
terminate	Abort the attack
access	Access the external dataset D_e
link	Link the de-identified record r_i to the external dataset D_e
exploit	Exploit a random individual in the equivalence group of record r in the external dataset D_e

4.2 Action set

There are four classes of actions in our system, which are summarized in Table 2. The adversary has the option of aborting the attack at any time by choosing the *terminate* action. The other three actions represent the adversary's operation in three different phases of the attack. The *access* action represents the accessing of the external dataset D_e. The *link* action represents the linking of the de-identified record r to the external dataset D_e. The *exploit* action represents a potentially harmful exploitation of an individual that is deemed to be related to the record r under attack. The particular type of exploitation may differ under various circumstances. For example, if the adversary's goal is to demonstrate the vulnerability of the system, the exploit may be to prove they can contact the individual and confirm the record is really associated with them [30]. Or, the adversary's goal may be to conduct direct marketing to the individual based on the sensitive information in the record (e.g., for a particular pharmaceutical). Regardless, an exploit is assumed to be successful if it is conducted against the individual to whom the record corresponds.

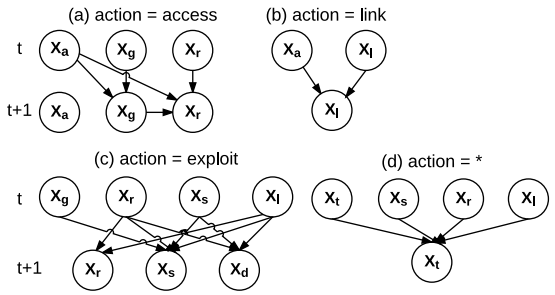

Figure 4: The dynamic Bayesian network (DBN) for each action of our FMDP model.

4.3 Reward

Reward functions are determined by several factors: the cost of taking an action, the loss to the adversary from detection (both negative rewards), and the gain from a successful attack. We formally define the reward function as:

$$R(x,a) = R_g(x[X_s], x[X_t]) + R_p(x[X_d], x[X_p]) - C_a \quad (2)$$

where C_a is the cost of action a (denoted by C_d, C_c, and C_e for *access*, *link*, and *exploit* actions, respectively). $C_a = 0$ for the *terminate* action. $R_g(x[X_s], x[X_t])$ represents the gain from a successful exploitation. $R_g(x[X_s], x[X_t]) = G$, if $X_s = T$ and $X_t = F$ and 0 otherwise.

We assume there is a maximum number of times, n_f, that the adversary will be subject to a penalty (e.g., a fine for law or contract violation) if he is detected. Note that this permits an analysis on the special case of $n_f = 1$, where the adversary is only penalized once. This is notable because it represents the real scenario where a data user is penalized for violating a contract, but is not prevented from continuing to exploit the data they have already received. We denote the cost related to the fine as $R_p(x[X_d], x[X_p])$. $R_p(x[X_d], x[X_p]) = -C_p$, if $X_d = T \wedge X_p < n_f$ and 0 otherwise.

4.4 State transition dynamics

We use a dynamic Bayesian network (DBN) $\tau_a = \langle G_a, P_a \rangle$ for each action a (except action *terminate*), as shown in Figure 4, to represent the transition function $P(X_i | X_i^{parent}, a)$. We denote the current state and the next state as x and x', respectively. If the action is to *terminate*, $x'[X_t] = T$.

If the action is to access, as the DBN shows in Figure 4(a), there are 3 state variables that may change in the following step: X_a, X_g and X_r. We highlight that if the external dataset D_e has not yet been accessed (i.e., $x[X_a] = F$), the adversary's belief of the equivalence group size in the next state $x'[X_g]$ is a probability distribution over a set of values, represented as $P(G_{r,D_e})$. Our experiments simulate $P(G_{r,D_e})$ under different levels of certainty and its influence on the adversary's behavior and re-identification risk.

In Figure 4(b), the *link* action sets $x'[X_l] = T$ when $x[X_a] = T$ (i.e., external dataset is available for linkage).

The prerequisite condition for the *exploit* action is $x[X_l] = T$, $x[X_s] = F$, and $x[X_r] > 0$. In other words, we can only exploit a record if 1) the equivalence group is non-empty, 2) the dataset has been linked to the record, and 3) the record has not already been re-identified. In this case, the number of remaining candidates in the equivalence group is decremented ($x'[X_r] = x[X_r] - 1$).

Moreover, the probability that the exploited individual is associated with the record is the probability of selecting an individual at random from the set of individuals in the population (with the same quasi-identifier) who have not been exploited. The number of individuals with the same quasi-identifier in the population who have not been exploited is the sum of the number of individuals outside (i.e., $\frac{1-prior_{r,D_e}}{prior_{r,D_e}} \times x[X_g]$) and inside (i.e.,$x[X_r]$) the external dataset d_e. Thus, the success probability of an exploitation can be formally represented as:

$$P_{suc}(x[X_g], x[X_r], prior_{r,D_e}) = \Big(\frac{1 - prior_{r,D_e}}{prior_{r,D_e}} \times x[X_g] + x[X_r]\Big)^{-1} \quad (3)$$

where $prior_{r,D_e}$ is the probability that the individual corresponding to the data is in the external dataset D_e.

$P(x'[X_d] = T)$ (i.e., the probability of being caught) denoted as P_{det} can be modeled in a number of ways. Since the probability an exploit is detected is very likely to increase with repeated attempts due to various factors (e.g., increased vigilance), we model the detection probability using a sigmoid function:

$$P_{det} = (1 + e^{-(h_0 + h_1 \times (x[X_g] - x[X_r]))})^{-1} \quad (4)$$

where $x[X_g] - x[X_r]$ corresponds to the number of exploit attempts the adversary has committed against records in the equivalence group. Note that this formulation allows us to model the special case, where the probability of detection does not increase over time by setting $h_1 = 0$.

Finally, regardless of the action, the transition of variable X_t is determined as follows (see Figure 4(d)): $x'[X_t] = T$ if $x[X_t] = T \vee x[X_s] = T \vee (x[X_l] = T \wedge x[X_r] = 0)$.

5. ALGORITHMS

5.1 Solving the MDP

Solving an infinite-horizon discounted MDP amounts to computing an optimal policy, $\pi(x)$, which prescribes an optimal action to take in each state [22]. Equivalently, it suffices to compute a value function, $V(x)$, which is the optimal discounted sum of rewards of an optimal policy.

A number of methods exist for solving an MDP. Linear programming (LP) is one such method, which computes the value function, $V(x)$, for every state x. An important limitation of the standard methods, including LP, is scalability. In particular, if we do not take advantage of problem structure, the runtime is polynomial in the number of states, which itself grows exponentially in the number of state variables. Approaches exist that leverage the structure of the factored MDP, but they are approximate, and require the pre-specification of a fixed set of basis functions over the state space. Next, we present a special-purpose method, which we call Two-Level LP, that takes advantage of our problem structure (including the factored state) and reports an exact answer.

5.1.1 Two-level Linear Programming

We designed the Two-level LP algorithm under the principle of removing all the "well-known" parts from the FMDP structure to save space and runtime. The algorithm constructs a two-level structure from the state space. The states in the FMDP model form a *sink cluster* sub-structure, which satisfies the following properties: a) there is no outbound and b) there is only one inbound state (i.e., x_{start} has only

one inbound edge). Based on the property of the FMDP, each sink cluster can be solved independently. The bottom-level of the Two-Level LP algorithm solves a LP and stores the value of the state x_{start} for each sink cluster. The top-level algorithm then constructs and solves a LP of the entire state space by replacing each sink state with its corresponding x_{start} and assigns the pre-computed $V(x_{start})$ to it.

Specifically, each sink cluster contains the descendant states of a state x_{start} in which the adversary has taken the action of access and link, but has not yet started exploitation, i.e., $x_{start} = [F, F, 0, F, T, T, s_i, s_i]$, $s_i \in (0, \max(G_{r,D_e}))$. Given two different group sizes s_1 and s_2, the two sink clusters with $x_{start} = [F, F, 0, F, T, T, s_1, s_1]$ and $x_{start} = [F, F, 0, F, T, T, s_2, s_2]$ do not overlap because the value X_g remains constant when $X_a = T$ and $X_l = T$. The resulting values of all the x_{start} states are used in the top-level LP to solve the values for the remaining states, such as the state in which the adversary is attempting to access the external dataset (i.e., $x = [F, F, 0, F, F, F, 0, 0)]$).

We make two performance improvements for Two-Level LP. First, we introduce a pruning strategy which leverages the fact that the value of the starting states for each cluster (i.e., $V(x_{start} = [F, F, 0, F, T, T, s, s])$) decreases as the size of the equivalence group $X_g = s$ increases. We omit the proof of this property due to brevity.

Thus, we sort the sink cluster by the value of $x_{start}[X_g]$ in ascending order. Specifically, if $V(x_{start}) = 0$ given $x_{start}[X_g] = s$, then all of the sink clusters with $x_{start}[X_g] > s$ will be pruned. Second, we use a result caching strategy. In doing so, the result from the bottom-level LP is cached and reused with multiple records. This happens when there is an overlap in the adversary's belief of the probability distribution interval of the equivalence group size X_g.

5.2 Computing Re-identification Probability

The re-identification probability P_{reid} is the sum of the probability of reaching each of the states with $x[X_s] = T$ and $x[X_t] = F$ in 1 to t_{max} time steps. Formally, P_{reid} is computed as:

$$P_{reid} = \sum_{t=0}^{t=t_{max}} \sum_{x \in x_{suc}} M^t[x_0, x] \quad (5)$$

In equation 5, $x_0 = [F, F, 0, F, F, F, 0, 0]$ represents the state in which the adversary has not accessed the external dataset yet, x_{suc} is the set of states with $x[X_s] = T$ and $x[X_t] = F$, x is an arbitrary state. M is the state transition $N \times N$ matrix of a Markov chain, where N is the number of states.

The state transition matrix M is obtained by replacing the action a in the transition dynamics function of the FMDP with $policy(x)$, i.e., $P(X_i' | X_i^{parent}, \ policy(x))$. However, there is one exception. Given the current state is x_0, in the FMDP model, $x'[X_g]$ is a probability distribution over a range of values due to the uncertainty of the adversary's belief, while, in the risk computation Markov chain, $P(x'[X_g] = g_{r,D_e}) = 1$, g_{r,D_e} is the actual group size in D_e. This is because the Markov chain already embed's the adversary's optimal policy, and consequently the adversary's belief in the group size no longer matters. Instead, what matters is the actual group size. We assume that g_{r,D_e} is an input to the risk framework.

The value t_{max} is the maximum number of time steps it takes for all the states to transit into a state where $X_t = T$ (i.e., an absorbing sink state). Formally: $\exists t_{max} > 0$ $\forall x_i, x_j \in x_i, i \in [0, N]$, if $x_j[X_t] \neq 0, M^{t_{max}}[x_i, x_j] = 0$.

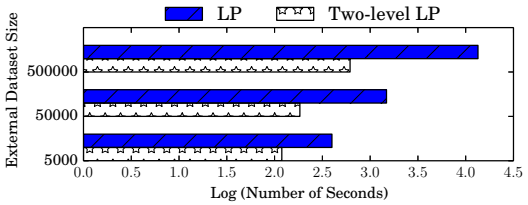

Figure 5: Runtime (log_{10}) of the FMDP solving algorithms for a dataset of 5000 de-identified records.

6. EXPERIMENTS

6.1 Dataset

Our experiments make use of three resources. First, we use the freely available North Carolina voter registration (NCVR) list [21] as the identified external dataset. This consists of 6,018,999 records without missing values over 18 fields. These include explicit identifiers (e.g., personal name and phone number), as well as quasi-identifiers (e.g., age, gender, race, and ethnic group). For the purposes of this study, we restricted the dataset to a set of four quasi-identifying attributes, {*Age*, *Race*, *Gender*, *5-Digit ZIP Code*}.

Second, we use the Adult dataset from the UCI Machine Learning Repository, as the de-identified dataset. This consists of 32,561 records with 14 fields each, based on a sample of the U.S. Census, without missing values. This dataset contains *Age*, *Race*, and *Gender*, but not *5-Digit ZIP Code*. As such, for each record in the Adult dataset, we synthesize and append a 5-digit NC ZIP code based on the population distribution in the US Census Bureau's 2010 Census Tables PCT12A-G. We also replaced a topcoded age value [90+] by a random value in the range of $[90, 120]$.

Third, we assume that both the de-identified and identified datasets are sampled from the entire population of NC. In this case, it should be noted that the total size of the NC population, according to the census is 9,553,967.

6.2 Equivalence Group Size Distribution

In the experiments, the probability distribution of the value X_g after the adversary takes the action to access the dataset $P(G_{r,D_e})$ is derived from the adversary's knowledge about the external dataset D_e or the population statistics. Here, we consider two scenarios. In the first scenario the adversary knows the target's equivalence group size when starting the attack. Specifically, $P(G_{r,D_e} = g_{r,D_e}) = 1$. We refer to this scenario as the *known group* model.

However, the adversary may not have such knowledge before accessing D_e. In this case, we assume the adversary knows only the total size of the external dataset, n, and the probability density of the target's record, i.e., the joint probability of the target's quasi-identifying values $P(r[QI])$, in the population. Assuming that the external dataset is sampled uniformly at random from the population, $P(G_{r,D_e})$ can be represented as a binomial distribution defined in Equation 6:

$$P(G_{r,D_e} = k) = B(k, n, P(r[QI])) \quad (6)$$

We refer to this mechanism as the *unknown group* model.

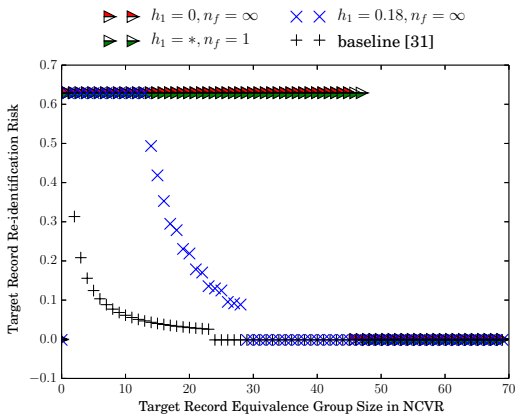

Figure 6: The equivalence group size for the target record in the NCVR dataset and the re-identification risk under the *known group scenario*.

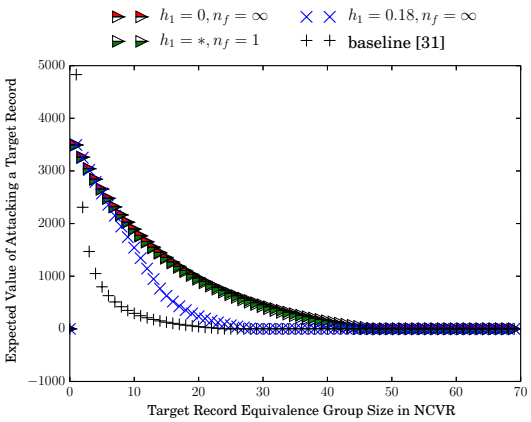

Figure 7: The equivalence group size of of the target record in the NCVR dataset and the adversary's expected payoff under the *known group scenario*.

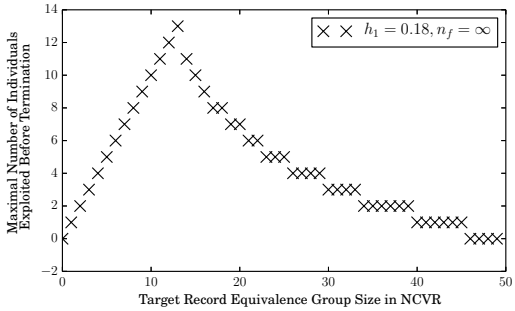

Figure 8: The size of the equivalence group of the target record in the NCVR dataset and the actual number of individuals the adversary exploits before terminating under the *known group* scenario.

7. RESULTS

7.1 Performance Analysis

We evaluated the runtime of the framework with 5000 randomly selected Adult records. In this analysis, we consider the *unknown group* scenario with a $P(G_{r,D_e})$ computed as Equation 6 in which the size of the external dataset is set to 3 different values: 5K, 50K, and 500K. We present the result of the *unknown group* scenario because the *known group* scenario yield FMDP with constant size state space, while *unknown group* scenario leads to increasing state space when the external dataset size n increases simply because of the interval of $P(G_{r,D_e})$ increase with n. The detection and penalty mechanism is set to $h_0 = -4.59$, $h_1 = *, n_f = 1$ (i.e., penalize only once and the probability of detection is 0.01 based on equation 4). The other parameters of the model were set to $prior_{r,D_e} = 0.63$, $C_d = 100$, $C_e = 10$, $G = 8000$ and $C_p = 10000$.

The algorithms were implemented in Python and all experiments were run on an Ubuntu server with 24 Intel(R) Xeon(R) CPUs at 2.4 GHz and 64 GB of RAM. The LP solver was the IBM ILOG CPLEX optimizer.

Figure 5 reports the runtime for the LP and Two-level LP algorithms. It can be seen that, as expected, the Two-level LP is always faster than the standard LP algorithm. The difference in speed is accentuated as the size of the external dataset grows. By the time there are 500K records in the external dataset, the runtime of the Two-level LP is approximately 21x faster (616 seconds vs. 13,444 seconds).

7.2 Case study

To perform a case study, we assume the Adult and NCVR records are random samples of the NC population. Thus, the prior probability that the individual corresponding to an Adult record is in the NCVR is the sample ratio, or $prior_{r,D_e} = 6,018,999/9,553,967 = 0.63$. The NCVR data is free; however, considering the effort to obtain it, we set the cost of accessing the external dataset C_d to \$100.

The cost to exploit, gain and penalty values were set to C_e = \$10, G = \$8000 and C_p = \$10000 for each record, respectively. We acknowledge these values may vary in practice. The goal is to simulate a case in which the adversary will attack at least a subset of the records. This allows us to examine how different deterrence mechanisms and uncertainty about the equivalence in the external datasets affects the adversary's behavior and the re-identification risk. Therefore, these parameters are selected from a range in which the adversary will attack some of the records.

7.2.1 Known Group Model

We compare the *known group* model to the risk model in [31]. This is the only available model for re-identification based on an adversary's optimal decision. In the baseline, the adversary makes a single decision on when to attack based on the total payoff:

$$Payoff_{baseline} = G * (\frac{prior_{r,D_e}}{G_{r,D_e}}) - p_{det} * C_p - C_d - C_l - C_e \quad (7)$$

If $Payoff_{baseline} > 0$, the adversary exploits a random individual and the risk of re-identification is $prior_{r,D_e}/G_{r,D_e}$. Otherwise, the risk is 0. The FMDP is configured under three detection and penalty settings: a) a constant detection probability with repeated penalties (i.e., $h_1 = 0$ and $n_f = \infty$), b) a one-time penalty (i.e., $h_1 = *$ and $n_f = 1$)[1] and c) an increasing rate of detection with repeated penalties (i.e., $h_1 = 0.18$ and $n_f = \infty$). In each setting, we set $h_0 = -4.59$. This yields a 0.01 detection rate for the first exploit, an increase to 0.012 for the next exploit, and so on.

The results are illustrated in Figure 6. There are three notable findings to highlight.

[1]The $*$ indicates that h_1 can be anything because only a single penalty is assigned.

Finding 1: The baseline risk never exceeds the FMDP models. This is because the baseline assumes that the adversary can only select one random individual, which is suboptimal. Thus, as can be seen in Figure 7, the baseline adversary's expected value drops at a faster rate than the adversary who acts according to the FMDP. Moreover, the adversary's success rate is also lower for the baseline. This is because the adversary only exploits one random individual from the equivalence group. This indicates that the baseline model often underestimates the re-identification risk.

Finding 2: When the detection probability is constant (i.e., $h_1 = 0$) and there is no upper bound on the number of times a penalty is levied on the adversary (i.e., $n_f = \infty$), the adversary either exploits 1) all records in the equivalence group or 2) no records.

Finding 3: When the probability of detection grows with repeated attempts (i.e., $h_1 > 0$) or there is an upper bound on the number of times a penalty is levied on the adversary (i.e., n_f is a finite value), the adversary exploits a subset of the equivalence group. In the scenario represented by Finding 2, the adversary chooses not to issue an attack when the equivalence group size is $\geq$ a threshold k, but the adversary exploits all the individuals in the equivalence group otherwise. Thus, the re-identification risk is either equal to the prior probability $prior_{r,D_e}$ or 0. This is because when the optimal action is to attack one individual in the NCVR equivalence group the subsequent optimal action is always to continue to exploit each of the remaining individuals provided that each exploitation has the same probability of being detected and the adversary will always be fined if detected.[2]

In the scenario of Finding 3, the adversary may terminate the attack before exhausting the candidates in the equivalence group. Thus, the risk can be any value between 0 and the prior probability $prior_{r,D_e}$. This is due to two possible reasons. First, if $h_1 > 0$, both the likelihood of detection and a successful re-identification are increasing when more individuals are exploited. Thus, the adversary stops when the increment in the expected penalty exceeds the increment in the expected payout, which can happen before the adversary exhausts all the candidates. Second, if n_f is finite, and the adversary was not detected in the previous exploitations, the expected payout can decrease when the number of the remaining cadidates reduces.

Similar to Finding 2, if the group size is $< k$, the adversary exploits all the individuals in the equivalence group. By contrast, if the group size is $\geq k'$, the adversary stops issuing an attack. For the group size in the range of (k, k'), the adversary's optimal action is to stop before reaching the last candidate in the group. The actual number of candidates the adversary will exploit before termination is shown in Figure 8. In this case, $k = 14$ and $k' = 29$.

These two findings are contradictory to what is expected by the baseline model. In particular, the records with equivalence group size $< k$ all have the same level of risk according to the FMDP model, while the records with smaller equivalence groups have more risk than those with larger equivalence groups according to the baseline model. The indication of this finding from the data protection perspective is that applying mechanisms, such as generalization, to increase the equivalence group size can only effectively reduce risk if the equivalence group size $\geq k$. In other words, increasing the equivalence group size to any value $< k$ will only harm the utility of the data without reducing the risk.

[2]We omit a proof of this claim due to brevity, but will make it available in a longer technical report upon acceptance.

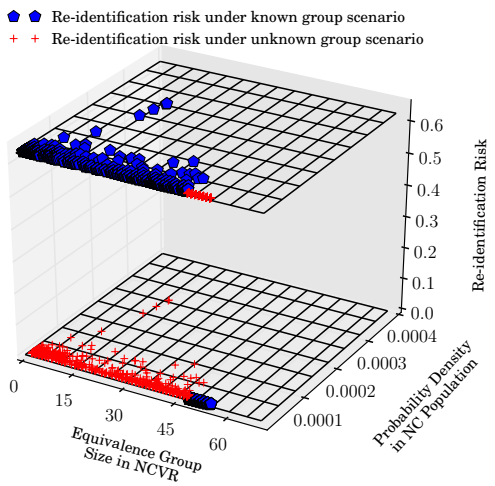

Figure 9: The equivalence group size, population probability density and the re-identification risk of the record with inconsistent risk values in the *known and unknown group* scenarios ($n_f = 1$).

7.2.2 *Unknown group Model*

The FMDP enables us to evaluate risk when the adversary is uncertain in the equivalence group size; i.e., the *unknown group* scenario. We assume that the adversary's belief of the group size is as in equation 6 with $n = 6018999$, with $p_{r[QI]}$ equal to the probability density of the corresponding target record in the NC census population. The other parameters are the same as defined in the *known group* model. Our result illustrates the following findings.

Finding 4: The unknown group scenario can yield *lower risk* than the known group scenario.

Finding 5: The unknown group scenario can yield *higher risk* than the known group scenario. These findings illustrate that uncertainty in the group size can change the action of the adversary. To make this observation more concrete, Figure 9 depicts the risk for the 1920 records that have exhibited different risk scores. 1118 of these records (or 58%) have a risk of 0.63 under the *known group* scenario and 0 under the *unknown group* scenario. The remaining 802 (or 42%) records have the exact opposite result. The former is due to the fact that the adversary underestimates the payoff by using the probability distribution of the equivalence group size. As a result, the adversary does not access D_e, while the actual group size is $<$ the threshold $k = 48$ and in the *known group* scenario the adversary will access D_e and attack. The latter is, on the other hand, due to adversary's overestimation of the expected payoff based on their inaccurate belief about the equivalence group size. These cases are counterintuitive because one may argue that even if the adversary decides to access D_e, he or she will not exploit and there is no risk because the actual equivalence group size is $\geq k = 48$. However, this is not always true because after the adversary obtains D_e, the cost C_d (i.e., the cost of accessing the external dataset) became a sunk cost. As a consequence, the payoff is computed without considering C_d and the threshold the adversary can tolerate increases

from 48 to 51. If the actual equivalence group size is between the two thresholds, the adversary with less knowledge (i.e., in the *unknown group* scenario) may be able to cause greater risk, even though the adversary does not necessarily obtain a higher payoff than the *known group* adversary.

Records resulting in different risk levels in the *known* and *unknown group* scenarios are not very common in this experiment setting. A majority of the records lead to the same risk (94%, or 30641 in total). The is due primarily to the fact that this analysis is dominated by records whose corresponding equivalence group size is larger than 55. Specifically, 64%, or 20891 in total, satisfy this situation. This is notable because, even if a positive payoff expected from $P(G_{r,D_e})$ leads the *unknown group* adversary to access the external dataset, the adversary never chooses to exploit such records, yielding a risk of 0.

7.3 Sensitivity Analysis

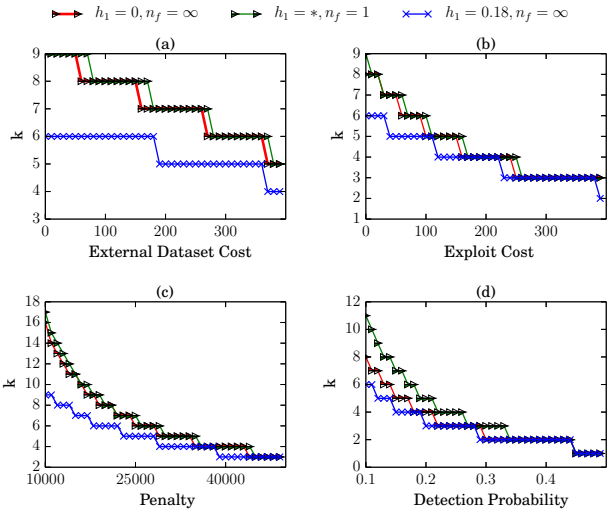

Figure 10: Sensitivity analysis on group size threshold (k) as a function of (a) external dataset cost C_a; (b) exploit cost C_e; (c) $Penalty$; and (d) detection probability P_{det}.

In this section, we investigate how the deterrence parameters influence the threshold k of the size of the equivalence group when the adversary walks away, such that the risk is 0 when equivalence group size is $\geq k$ under the three scenarios studied above. For this analysis, we assume a *known group* scenario and both the de-identified and external datasets cover the entire population, such that $prior_{r,D_e} = 1$. We vary 1) the cost to access data, 2) the cost to exploit the targeted individual, 3) the penalty levied when re-identification attempts are detected, and 4) the detection probability. In the analysis, we vary one variable at a time while holding all other variables constant to: C_p = \$20000, G = \$1000, C_e = \$10, C_l = \$0, C_a = \$100.

The result is unsurprising, but notable. Specifically, as illustrated in Figure 10, as the deterrence mechanism is ramped up, the expected payout is lower and the adversary tolerates less risk. For example, when the penalty is set to \$10,000, the adversary always attacks when the group size is smaller than 9 individuals. By the time the penalty is raised to \$50000, the adversary will only risk an attack if there is one individual in the group. This result clearly indicates that penalties and costs for access to data can quickly deter an adversary from committing an attack.

8. DISCUSSION AND CONCLUSIONS

This research provides a formal process-based approach to characterize the privacy risks for published data and opens a novel direction in the field of data privacy. It also introduces a scalable algorithm based on linear programming to solve the attacker's optimal planning problem. A core contribution of this approach is that it accounts for deterrence mechanisms beyond data manipulation methods. We demonstrated the feasibility through a case study in a real world scenario, where an adversary uses a publicly available population registry (with over 6,000,000 individuals) to attack a record subject to a data obfuscation mechanism.

Our results reveal that a broadly accepted adversarial model in which the adversary will randomly choose one individual that matches the record to attack can be suboptimal, and an adversary may try and exploit every individual in the corresponding equivalence class. In addition to penalization mechanism, our result demonstrated that the adversary's optimal decision depends on the information about the external resources they may use (e.g., voter registration lists) before they access them to mount an attack. This work provides strong evidence that the risk to such systems in the real world is heavily dependent on the amount of effort an adversary needs to exert and the expected payout they can receive based on their attack. This investigation further provides intuition into how data perturbation techniques can be complemented by alternative disincentive strategies (e.g., charging for access to data or levying fines for malicious behavior) to lower the risk inherent in data sharing.

Our approach has several limitations which can provide directions for future research in this area. First, if such a risk estimation procedure is to be put into practice, policy makers will need information about the nature of deterrence mechanisms, the existence and costs of external data resources, as well as the adversary's potential gain. Moreover, our work shows that knowing the prior probability that the corresponding target is in an external resource is critical to the model. Our model assumes that the external dataset is a random sample from a large population that also covers the protected data. Such information is not always readily available to the data publisher when evaluating risk. In the event the publisher believes they could underestimate such parameters, they may lobby for larger fines on misuse, thus deterring users with legitimate interests from accessing their resource. Thus, a future direction for research is in the development of approaches to estimate such parameters of the attack process. This may be possible, for example, by building a model for the detection rate based on existing detection mechanisms.

Second, there are limitations in the scope of the adversary's goals. Consider, the process model assumed an adversary targets only one record in the protected dataset at a time. It also assumes that the adversary has access to only one external resource to mount an attack. Perhaps more significantly, we assume that the success of an exploit will be confirmed. Yet, certain adversaries may be interested in multiple records in the protected data (or even the entire dataset) and may have access to multiple resources. Removing any of these assumptions will lead to an increase in the complexity of the adversary's decision problem. We note that the process model can be extended to account for these scenarios by introducing more state variables and actions. However, this will lead to an explosion in the state space.

Therefore, a future direction of research is to generalize the FMDP model while improving the scalability of the solver algorithm.

Finally, our empirical analysis was conducted on a specific type of data, namely the demographic information within the publicly available population registry. Such a process-based approach to privacy risk assessments is applicable to other types of data where the attack is not a linkage-based exploit, but focuses rather on other definitions of privacy, such as inferential disclosure. The adaptation of such a technique will depend on the extent to which the adversary's process for realizing their exploit can be represented.

9. ACKNOWLEDGMENTS

This work was sponsored by grants from the NIH (R01-HG006844, R01-LM009989, U01-HG006478, U01-HG006385), the NSF (CCF-0424422), the AFRL (FA8785-14-2-0180), and Sandia National Lab (contract 2191).

10. REFERENCES

[1] A. Bharadwaj, O. El Sawy, P. Pavlou, et al. Digital business strategy: toward a next generation of insights. *MIS Quarterly*, 37(2):471–482, 2013.

[2] L. Bonomi and L. Xiong. A two-phase algorithm for mining sequential patterns with differential privacy. In *Proc.* 22^{nd} *ACM Int'l Conf. on Inform. and Knowl. Management*, pages 269–278, 2013.

[3] R. Dewri, I. Ray, I. Ray, et al. POkA: Identifying pareto-optimal k-anonymous nodes in a domain hierarchy lattice. In *Proc.* 18^{th} *ACM Conf. on Inform. and Knowl. Management*, pages 1037–1046, 2009.

[4] C. Dwork. Differential privacy. In *Proc. Int'l Colloquium on Automata, Languages, and Programming*, pages 1–12. Springer, 2006.

[5] C. Dwork. The promise of differential privacy: A tutorial on algorithmic techniques. In *Proc. IEEE Annual Symp. on Foundations of Computer Science*, pages 1–12, 2011.

[6] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin. A systematic review of re-identification attacks on health data. *PLoS ONE*, 6(12):e28071, 2010.

[7] M. Elliot and A. Dale. Scenarios of attack: the data intruder's perspective on statistical disclosure risk. *Netherlands Official Statistics*, 14:6-10, 1999.

[8] L. Fan and L. Xiong. Real-time aggregate monitoring with differential privacy. In *Proc.* 21^{st} *ACM Int'l Conf. on Inform. and Knowl. Management*, pages 2169–2173, 2012.

[9] D. Freni, C. Ruiz Vicente, S. Mascetti, et al. Preserving location and absence privacy in geo-social networks. In *Proc.* 19^{th} *ACM Int'l Conf. on Inform. and Knowl. Management*, pages 309–318, 2010.

[10] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4), 2010.

[11] C. Guestrin, D. Koller, R. Parr, and S. Venkataraman. Efficient solution algorithms for factored MDPs. *CoRR*, abs/1106.1822, 2011.

[12] R. Jones, R. Kumar, B. Pang, and A. Tomkins. "I know what you did last summer": query logs and user privacy. In *Proc.* 16^{th} *ACM Conf. on Inform. and Knowl. Management*, pages 909–914, 2007.

[13] O. Kwon, N. Lee, and B. Shin. Data quality management, data usage experience and acquisition intention of big data analytics. *Int'l Journal of Inform. Management*, 34(3):387–394, 2014.

[14] D. Lambert. Measures of disclosure risk and harm. *Journal of Official Statistics*, 9:313–331, 1993.

[15] J. Letchford and Y. Vorobeychik. Optimal interdiction of attack plans. In *Proc. Int'l Conf. on Autonomous Agents and Multi-agent Systems*, pages 199–206, 2013.

[16] A. Machanavajjhala, D. Kifer, J. Gehrke, et al. l-diversity: Privacy beyond k-anonymity. *ACM Trans. on Knowl. Discovery in Data*, 1(1), 2007.

[17] E. Mackey and M. Elliot. Understanding the data environment. *XRDS*, 20(1):36–39, Sept. 2013.

[18] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *Proc.* 30^{th} *IEEE Symp. on Security and Privacy*, pages 173–187, 2009.

[19] A. Narayanan and V. Shmatikov. Myths and fallacies of "personally identifiable information". *Communications of the ACM*, 53(6):24–26, 2010.

[20] M. E. Nergiz, M. Atzori, and C. Clifton. Hiding the presence of individuals from shared databases. In *Proc. ACM SIGMOD Int'l Conf. on Management of Data*, pages 665–676, 2007.

[21] North Carolina Voter Registration Database, `ftp://alt.ncsbe.gov/data/`. Last accessed 4 Aug 2015.

[22] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

[23] L. Roderick. Discipline and power in the digital age: the case of the US consumer data broker. *Critical Sociology*, 40(5):729–746, 2014.

[24] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, 1998.

[25] D. Solove. A taxonomy of privacy. *University of Pennsylvania Law Review*, 154(3):477–560, 2006.

[26] M. Srivatsa and M. Hicks. Deanonymizing mobility traces: using social network as a side-channel. In *Proc. ACM Conf. on Computer and Communications Security*, pages 628–637, 2012.

[27] L. Sweeney. Uniqueness of simple demographics in the U.S. population. *Technical report, Carnegie Mellon University*, 2000.

[28] P. Tallon. An application of game theory to understanding statistical disclosure events. *UNECE/Eurostat Work Session on Data Confidentiality*, 2009.

[29] P. Tallon. Corporate governance of big data: perspectives on value, risk, and cost. *IEEE Computer*, 46(6):32–38, 2013.

[30] A. Tanner. Harvard professor re-identifies anonymous volunteers in DNA study. *Forbes*, 4 Apr 2013.

[31] Z. Wan, Y. Vorobeychik, W. Xia, et al. A game theoretic framework for analyzing re-identification risk. *PLoS ONE*, 10:e0120592, 2015.