Sanitization Models and their Limitations

R. Crawford, M. Bishop, B. Bhumiratana, L. Clark, and K. Levitt
Dept. of Computer Science
University of California at Davis
One Shields Ave.
Davis, CA 95616-8562
United States of America
{crawford,bishop,bhumirbh,clarkl,levitt}@cs.ucdavis.edu

ABSTRACT
This work explores issues of computational disclosure control. We examine assumptions in the foundations of traditional problem statements and abstract models. We offer a comprehensive framework, based on the notion of an inference game, that unifies various inference problems by parameterizing their problem spaces. This work raises questions regarding the significance of intractability results. We analyze common structural aspects of inference problems via case studies; these emphasize why explicit policies are needed to specify all social context and ethical values relevant to a problem instance.

Categories and Subject Descriptors
C.2 [Computer-Communication Networks]: General—*security and protection*; E.m [Data]: Miscellaneous; K.4.1 [Computers and Society]: Public Policy Issues—*privacy*

General Terms
Security.

Keywords
Data sanitization, inference problem, disclosure control, closed world assumption

1. INTRODUCTION
Increasingly, entities in modern society are recognizing the downsides of exposing their information to others' access—how we wish we could delete our email addresses from all those old posts accessible forever on Listserv archives, or somehow limit access to those who would not use that data against us! But it is impossible to retrofit role-based access controls [6] throughout the Internet, and impractical even to devise policies to implement this enforcement mechanism prospectively to cover most corporate Intranets.

As an alternative, consider what might happen if newly-published data itself served as the enforcement mechanism for one of the key principles of the 1973 Code of Fair Information Practices [32], namely to prevent information "obtained for one purpose from being used ... for other purposes without ... consent." This is the goal of *sanitizing* data—to alter it so that it remains usable for beneficial purposes, while minimizing its use for harmful purposes. Sanitization attempts to provide additional safeguards beyond traditional access control. Access to the sanitized data, without access to the original data, should not enable an adversary to cause significant harm.

In the context of statistical queries to databases, the inference problem is a long-studied research area [10]. Numerous related problems exist, including the *database inference* problem (whether any classified information can be inferred from unclassified data), the *query audit* problem (how to prevent or detect that query responses disclose sensitive data), the *privacy-preserving data mining* problem (how data can be altered to protect individual privacy, yet still be useful for data mining), and variants in census data and medical records analysis. Some of these are specific to particular application areas. But all these variants contain the same difficult core problem—how to control the *inferences* that can be drawn from particular data.

Sanitizing IP traffic before sharing it is a well-known problem in network security [18], and we offer it as a tangible case study that highlights why explicit privacy and analysis policies are a necessary part of any inference problem's specification. Initially, we provide one simple formalization of the sanitization problem, and describe our results (applicable to IP traffic and to different applications with similar semantic structure or syntactic constraints). Next, we offer a more comprehensive framework to unify our understanding of various inference problems. We show how certain unwarranted assumptions and flawed problem statements are inherent to large portions of the classical research core. Hence, we argue that what historically have been viewed as fundamental inference problems and results may be little more than arbitrary and isolated peripheral objects. In the remainder of this work, we examine the more important structural aspects common to many inference problems.

2. A SANITIZATION FRAMEWORK

We review one simple framework for the sanitization problem [3,4]. The sanitization problem involves three entities: a *collector*, an *analyst*, and an *adversary*. The *collector* captures raw data, then sanitizes it (to keep sensitive aspects of the data confidential) before sending it to the *analyst* for analysis. *Sanitization* means to perturb, delete, and/or add enough information that the sensitive data cannot be inferred. For now, we oversimplify by saying the goal of the *adversary* is to recover as much of the original, sensitive raw data as possible. Several assumptions may be used:

1. The adversary may have partial "upstream" read or write access at the source of the raw data. Analogous to a known (or chosen) plaintext attack in cryptography, the adversary attacks by remembering or creating specific markers in the unsanitized traffic. If these markers remain recognizable after sanitization, they may help the adversary infer the original, raw data from the sanitized data. We assume the sanitizer is informed of all such markers via an explicit threat model.

2. The adversary may be able to infer sensitive aspects of the raw data directly from supposedly non-sensitive (and hence unsanitized) components of the data that were sent to the analyst. Alternatively, the adversary may have access to auxiliary data or metadata—either public or private—such as organizational information, the role of specific systems, or certain correlation functions, that when combined with the sanitized data, allows her to infer the raw information. This is the database inference problem in a different context. Again, we assume the sanitizer is informed of all such inference paths via an explicit threat model.

3. There may be various degrees of public transparency at various stages of the process. Most importantly, the collector may "publish" its sanitized data, making it equally available to both the analyzer and the adversary. Many situations appear to relax this requirement by providing the data to analysts under non-disclosure contracts. However, even with such agreements, it is still possible for sanitized data to leak to the adversary. The analyst's network or data may be compromised via insider misuse or outsider attack, for example. We assume the worst case, of equal access by analyzer and adversary.

This model may seem to harbor hidden contradictions. Consider the situation in which multiple, mutually-distrusting collectors feed data to one mutually-trusted analyst. But if all collectors trust the analyzer never to reveal their sanitized data, then why not simply give their raw data directly to the analyst? Besides guarding against inadvertent leakage by the analyst, legal or contractual requirements may forbid the collectors from sending unsanitized data to the trusted analyzer. Moreover, the analyst may not be *fully* trusted for all *purposes*, or for all time!

Conversely, suppose multiple collectors trust each other to keep raw data secret, but do not trust the external analyzer with their raw data. Then why do the collectors not share raw data and perform the analysis themselves? The external analyst may have more resources or expertise, access to more aggregate data (via collectors who *do* trust it), or be more cost-effective. The same arguments hold for the case of a single collector.

The collector's goal is to sanitize the data in a way that maximizes the efficiency and accuracy of the analyzer's task, while minimizing the efficiency and accuracy of the adversary's attempt to infer the raw data. To formalize this, the collector develops (and updates) an explicit threat model of the adversary.[1] When coupled with the specifics of the data to be sanitized, and the inferences that the collector wishes to remain confidential, this yields an explicit *privacy policy*.

Another way to look at these policies is that the privacy policy suggests the original data that should be changed. The analysis policy suggests the original data that should not be changed. If the privacy policy inherently conflicts with the analysis policy, then sanitization cannot proceed without violating a policy, so the two policies must somehow be reconciled. Otherwise, one could sanitize the data in various degrees. In one extreme, one could sanitize every datum except those needed to conduct the analysis. In the other extreme, one could sanitize precisely those data that the privacy policy requires to be sanitized. A range of possibilities lie between these extremes. The threat model describes what the adversary can do, and will affect the selection of the sanitization technique accordingly.

Previous formal approaches to related problems presented privacy policies, but often those policies were based on unexamined or untenable assumptions. We extend that work by incorporating context-specific threat models into privacy policies, and by elevating the issue of analysis to a similar and explicit policy level. Just as a privacy policy is needed to specify which inferences to prohibit, so also an explicit analysis policy [4] is needed to specify which inferences to preserve. In some application domains (such as IP traffic collection), the data may have been primarily "obtained for one purpose" [32]—exactly such an analysis!

Using formal extensional semantics, we can express a privacy policy (analysis policy, respectively) as a relation between the raw data, the sanitized data, and a threat model (analysis capability model, respectively). Intensional semantics offers more insight: a privacy policy expresses constraints on the adversary's ability to derive certain inferences[2] from the sanitized data, given the adversary's external knowledge. An analysis policy expresses the maximum permissible deviations between results of analyzing the sanitized data, and the results if analysis had been performed on the raw data.

---

1. Threat models of adversaries in a particular problem domain would be useful to all collectors, and so could be shared.

2. A word of caution: every privacy policy will prohibit the adversary from inferring all the original, raw data. But while this clause might seem to provide a "bottleneck" through which all other prohibited inference paths must flow, real-world risks are not always so mathematically well-behaved. For example, we might sanitize a one-record medical database so that a life insurance company cannot recover its raw data, which states that Mr. Smith has heart trouble. But if our sanitized record instead convinces the life insurance company that Mr. Jones has AIDS, our sanitization has "framed" an innocent person by promoting an inference that put him at risk. Ignoring such inferences via a "closed world assumption" may be legally or morally unacceptable.

In this manner, the twin goals of safety and efficacy may be pursued even if adversary and analyst are the same entity, or have equal access to the sanitized data. The sanitizer may even be able to insert completely spurious "confounders" into the data to mislead adversaries, as well as hints that lead friendly observers to the right conclusions (even if for the wrong reasons).[3]

Formalizing privacy and analysis policies is difficult, partly because most *implicit* policy constraints originate informally. As such, they tend to be ill-defined or subject to varied interpretations. Moreover, attempting to define informal semantics formally is not only an error-prone process, but also may ignite social strife as certain interpretations are favored over others. The benefit of this organizational "values-clarification" process is more accurate assessment of—and control over—the risks and rewards of releasing sanitized data for analysis.

3. SANITIZING NETWORK DATA

In this section, we sketch certain issues and results that arose in the application area of sanitizing IP traffic. Our purpose is to provide tangible examples for subsequent discussion, and to show that such issues are common across many application domains. For details of our work in sanitizing network traffic, interested readers are referred to other papers [3,4], which raise additional issues not covered here.

In what follows, we focus on IP addresses because, for many organizations, they are a key point of conflict between the competing interests of privacy and analysis. Moreover, attempting to sanitize them demonstrates how the state of the art is inadequate for handling even this most basic data type. These difficulties arise in every application domain whose syntax is constrained by a finite *namespace* (set of names), which carries semantic *connotations* (in addition to referential denotations).

Two kinds of data anonymization prove useful in sanitizing IP addresses. *Pseudo-anonymous* transformations map all instances of a particular raw identifier to the *same* unique identifier in the target namespace. An example is replacing all occurrences of "John" by "Paul". The advantage of pseudo-anonymous transformations is that an analyst may correlate data related to an identifier without knowing what the raw identifier is. For example, given a set of network traces, an analyst can determine if two connections are between the same hosts. The disadvantage is that the adversary may be able to deduce private information from that knowledge.

Fully-anonymous transformations map each instance of a particular raw identifier to a *different* identifier in the target namespace. An example is replacing the first occurrence of "John" with "Paul", the second "John" with "George", and the third "John" with "Ringo".

Either mapping may be done explicitly, using a table holding each raw identifier and its corresponding sanitized identifier, or implicitly via hash functions, in which case the inverse mapping from sanitized identifier back to raw identifier is not apparent. As a research vehicle, we implemented a prototype sanitizer, *tcpsani*, that allows the user to configure hybrid modes of sanitization [3].

3.1. Finite Namespaces

IP addresses as identifiers are drawn from a namespace of finite size. This has two crucial implications. First, any pseudo-anonymous mapping on a finite namespace must be a permutation. Therefore, if a mapping implementation is not parsimonious and "over-reserves" target space for a block of IP addresses in the original namespace, the target namespace will overflow when all possible input names are sanitized.

Second, by the pigeonhole principle, any fully-anonymous mapping of n+1 name occurrences to a namespace of n names is impossible. This means that, if a long conversation between two network nodes is to be sanitized fully-anonymously, the target IP address namespace will eventually become exhausted and repetitions of sanitized names will occur.

3.2. Risks of Aggregated Analysis

Consider the situation in which a single analyzer aggregates data from several different collectors. The desired analysis may require that all collectors pseudo-anonymously map certain raw IP addresses to the same target namespace addresses, to ensure consistency of identity and to prevent conflation. Prior approaches using hashing to accomplish these results have required that all collectors use the same hash function for the entire namespace.

This is unacceptable for collectors who trust the analyzer, but who do not trust each other. The obvious insider attack is to guess an IP address, hash the guess, and compare the result to the sanitized data. Explicit maps can mitigate this problem by allowing mutually distrusting collectors to share—and hence risk—only portions of a codebook. This feature could also be implemented by binding specialized hash functions to particular input IP address regions, and sharing only some of those hash functions. Both methods allow fine-grained control over privacy risks caused by sharing sanitization functions.

One property whose preservation benefits analyzers is that of IP address prefixes. In its default configuration, *tcpsani* implements only byte-aligned prefix preservation. *Tcpsani*'s sanitization modules could be augmented with code to implement prefix preservation on the fly. This would provide the functionality of *tcpdpriv* [25]. But if prefix preservation is the sole objective, a better alternative is CryptoPan [16], an elegant and efficient specialized hash. CryptoPan also has the virtue that multiple collectors can implement the same prefix-preserving permutation for aggregation by a single analyzer merely by sharing a small secret key, rather than sharing large explicit maps.

CryptoPan and similar work (for example, Peuhkuri [26]) require a high degree of trust among different collectors; perhaps this is why they have not been more widely accepted as "the solution" to the tension between IP address sanitization and the desire for aggregated analysis. For example, given a prefix-preserving hash key, any trusted collector (or its rogue insiders) can invert any target IP address by a sequence of 32 chosen-plaintext attacks. These attacks are performed offline—there is no need to inject them into a monitored traffic stream. Hence, if any aggregated CryptoPan-sanitized dataset is ever made available to an adversary, that data

3. An analysis policy should explicitly state what circumstances, if any, would cause the "good" analysis inference ends to justify the means. We address issues of responsibility for distorted inferences in a later section.

set will remain vulnerable to these insider attacks for all eternity. The sharing of completely identical, explicit maps (even if they do not preserve prefixes) by multiple collectors likewise carries this perpetual vulnerability if the aggregate sanitized dataset is published.

Thus, regardless of how mutually-identical aggregate sanitization is implemented, a collector is perpetually vulnerable to any other entity with access both to the permutation key/map and to the permuted data. In light of recent security breaches at many commercial analyzers of credit information, it is worth noting that a collector who trusts an aggregating analyzer with its data today, must also trust that analyzer in the future not to fall prey to an adversary masquerading as a new collector who wants to sanitize its data using the common, historical sanitization function.

It is intriguing to consider how the problem of sanitizing data aggregated from multiple collectors is, in some sense, a dual of the query audit problem. Given a central database and multiple queriers, the query audit problem asks how the database management system should respond to each query so that the aggregated results reveal *no more* sensitive information than the sum of the individual query results. By way of contrast, the aggregate sanitization problem asks how to perform a globally-uniform sanitization *confidentially* on multiple distributed writes to a central database, so that analyzing the aggregate results *will* reveal more than the sum of analyzing each separately-confidential database write. Results from multiparty-secure computation theory, where the tasks of the parties are not symmetric, may prove helpful; in particular, the approach in [21] appears promising. (Although Zhong et al. [33] offer insight, they solve a different problem.)

3.3. Risks of *Not* Sharing an Aggregate Threat Model

Privacy policies must be revised as we discover new vulnerabilities in our data, and threat models must be updated as the adversaries' capabilities and external knowledge evolve. In economics, the common good is harmed when companies "externalize" their costs (e.g., health risks of toxic pollution) onto others. The realm of sanitization is vulnerable to a similar dynamic, in which collectors may externalize disclosure risks onto others.

Consider credit card records. If the adversary is an identity thief, then the only information to be concealed may be the name of the credit card holder, the credit card number, and the expiration date. Raw purchase transaction data might be disclosed for market research analysis. But if the adversary is a private investigator trying to determine whether *any* card holder is having an affair, the information to be concealed may include that raw purchase transaction data: the merchants from whom purchases are made (e.g., jewelers, florists, hotels), their locations, and the amounts.

Suppose one credit card company's privacy policy does not consider private investigators to be a threat. Analyzing a sanitized version of the raw transaction data then reveals that 7% of (pseudo-anonymized) cardholders seem to be having affairs, and the onset of this anomalous behavior causes them suddenly to carry a balance near their credit limit, whereas previously they had always paid their balance in full each month. This correlation (between affairs and monthly payoff behavior) then becomes "external knowledge" that increases the capabilities of the private investigator threat class for the next attack iteration. Subsequently, another credit card company—which *does* consider private investigators a threat—must respond by updating its own privacy policy to require sanitization of monthly balance data.

4. INFERENCE GAMES

Having examined the problem of sanitizing network traffic, we now broaden the scope of our inquiry to encompass other problems. We consider all these problems as inference games, and offer a unifying framework for understanding such games. Informally, we define an *inference game* as an attempt by several players to draw various inferences from various data sources.

Inference games constitute an extremely broad category of games. When keeping score in such games, the scoring dimensions are not limited to traditional notions of individual privacy, corporate or government-classified secrecy, or accuracy of statistical analysis. The only requirement for an interesting game is that some inferences, by certain players, accrue higher scores than others. Inference scores are relative to the sanitizer's values and preferences because the sanitizer's job is to limit, enable, or promote particular inferences. For example, if a privacy-penetrating inference by an adversary has a negative score, this does not mean that the adversary loses points; rather, the sanitizer loses points if it allows that inference. Moreover, since the players in inference games are not limited to one (bad) adversary and one (good) analyzer, it may be helpful to view all players as different *interpreters*, with varying degrees of access to various data, and we use this term to encompass both adversaries and analysts.

For example, suppose a teenager travels to a statewide science fair. During the trip, she asks people to take pictures of her with her camera. These pictures constitute a sanitized record of her trip: the teenager has selected particular scenes—and staged, posed, and perturbed them to some extent—while rejecting (*i.e.*, suppressing) other photo opportunities. The inference game she plays is the following. Her parents, and her friends at school, will have equal access to view the pictures. She wants her parents to infer that she was focused on science throughout her trip, but she wants her friends at school to infer that she had a romantic relationship with a boy from another school who attended the fair. Which inference corresponds to the "true" situation?

Truth has relevance in our framework only insofar as it is reflected in the inference scoring metric. This is because, in real-world contexts, it is not truth, but *beliefs* (confidence levels of inferences) that trigger the risks or probabilities of actions with *consequences*. If our teenager's friends infer that she is a "science nerd", they may start to ostracize her as a consequence; whereas if her parents infer that she behaved irresponsibly, they may limit her future privileges as a consequence.

Accurate game models that incorporate *all relevant context* are crucial. Although the parents and school friends have equal access to the sanitized data, these players begin the game with different initial conditions—specifically, different inferential "working hypotheses", and different external or auxiliary data. Moreover, the teenager may send different hints to different players via private channels. Clearly, an accurate model of this game has no "rule" requiring that all information be equally available to all players.

To simplify assumptions, in this paper, we exclude from the scope of our exploration those games where the significance of the "out of band" data privately transmitted by the sanitizer to the interpreters dominates the "equal access" data made public via sanitization.

4.1. Unifying Framework for Inference Games

In two-player non-zero-sum games, the two players often have different initial conditions, different capabilities, and different goals. So far, this description applies to the games played by an adversary and an analyzer. But what distinguishes sanitization from much of game theory is that, given certain constraints, the sanitizer must *generate* a game that the analyzer always wins, and the adversary always loses.

Our focus is further limited by the frequent assumption or constraint that the analyzer is so dumb, that it may not even know its data has been sanitized; whereas the adversary is so smart, that it may even be able to simulate the analyzer. In such cases, it may be acceptable to approximate the situation as a game played directly between the sanitizer and the adversary.

All the "classical" inference problems are special cases of a more general family of inference games. We introduce a generalized framework for these inference games, in which any particular game is distinguished by its rules of play, and by its system of characterizing and scoring possible outcomes of play.

4.2. Terminology for Evaluating Sanitization Methods

For brevity in our discussion, we often will use the simplistic phrase "score points". But readers should bear in mind that such "points" are a conceptually sloppy, albeit convenient, linguistic abbreviation for characterizing the impact of a sanitization method along various dimensions of value. These values may not be commensurable; in particular, ethical values may not be reducible to a one-dimensional numeric score.

For our purposes, a *metric* is a measure that quantifies a sanitization method's performance along a particular value dimension, totally ordering those values. A *policy* is a method that may partially or totally order points in (a typically multi-dimensional) space, and characterizes regions in that space qualitatively—for example, as "acceptable", "preferable", "minimally protective of privacy", "too distorted for useful analysis", and so forth. Thus, a policy may determine whether a group of incommensurable operational metrics constitutes a metric space.

As a tangible example, one privacy metric might account for all risks to an individual arising from disclosures of private medical data, and a second privacy metric might consider risks from financial data disclosure. A privacy policy might employ these two metrics to evaluate overall risks to individual privacy along both medical and financial dimensions. An analysis policy might determine whether a sanitization method introduces unacceptable distortion into an analysis that computes average life expectancy and average salary. The policy resulting from the composition of these privacy and analysis policies must reconcile any conflicts by specifying for which regions of value space "the good of the many" in having relatively accurate analysis outweighs "the good of the few" individuals whose outlying attributes on the salary and life expectancy axes might make them more vulnerable to privacy risks.[4]

If a policy imposes a total order covering all game-relevant dimensions of value, then it is a *utility function* as defined in the standard game theory literature. Such a utility function is completely different from the "utility metrics" mentioned occasionally in the literature on classical inference problems. We refer to the latter as *analysis metrics*, because they quantify how well or poorly a sanitizer performs, measured along the dimensions of value that matter to an analyzer.

Finally, a *strategy* is a plan for how to extend a policy (or group of policies) to cover more than one iteration of an inference game.

4.3. Scoring Inference Games

Our framework begins with the notion of an analysis metric: how useful is a particular sanitized dataset for the purposes of an analyzer? We score the result of this sanitization accordingly. In a symmetric fashion, we consider the adversary: how useful is this sanitized dataset for the purposes of an adversary? Again, we assign an appropriate score(s), this time to quantify the benefit(s) to the adversary.

Since, by definition, anything that assists the adversary to achieve its goals harms privacy (and/or secrecy, depending on the semantics of the particular game), then every adversary metric is the negative reflection of a privacy metric, and *vice versa*.[5]

Various notions of computational *costs* can be included in our inference game framework. The sanitizer, analyzer, and adversary all have limited computational budgets, yet the most cost-effective way to allocate the budget depends on how the opponents (and allies) allocate their budgets.

A major limitation of the inference game model is that it lacks an accurate feedback channel so the sanitizer can learn how the adversary scored. This can be alleviated somewhat by inserting "honeypot records" in sanitized data, or by periodically sampling or monitoring real-world data subjects to see whether they incurred damages that may be ascribed to the sanitization. But in general, we must impute an adversary's score by simulation, based on what we know or suspect about its capabilities and tactics from other known privacy breaches.

We elaborate further on our framework in subsequent sections, by emphasizing how its generalized playing field differs from those of classically-defined special case inference problems.

---

4. Clearly, such fundamental value judgments should never be entrusted to an automatic policy composer.

5. To emphasize this, we might call an adversary metric a "disutility metric", but such terminology is too similar to the "utility function" mentioned above. Moreover, such usage would imply that privacy and analysis utility are diametrically opposed. Such zero-sum game instances do occur, but positive-sum inference games (in which privacy and analysis utility are somewhat independent) also occur frequently.

4.4. Questionable Assumptions

Initial simplifying assumptions help exploring hard problems. But such assumptions become problematic when they are unexamined, when they become an impediment to further progress, or when they are unconsciously built into the underlying problem statement. Two crucial assumptions permeate the literature on inference problems: the *closed world assumption* (CWA), and the *uniform analysis metric* (UAM).

In relation to inference problems, CWAs restrict researchers' attention and concern to items that appear *explicitly* in the original (raw) dataset.[6] But even if an adversary cannot infer original data, it can still cause damage, by inferring joint probability distributions of (still-indeterminate) raw data, by making probabilistic inferences about a class of raw data, or even by making completely wrong inferences (as in our example in footnote 2 above).

The UAM postulates that all attributes are equally valuable for analysis purposes. Clearly, this might be true for certain particular analyses. In contrast with CWAs, which limit the scope of problems, the UAM has been used to limit the scope of *solutions*. The implicit claim is that the UAM applies to a broad class of problems, but we are unaware of any attempt to justify that claim explicitly. Absent such justification, we question whether fundamental intractability results based on the UAM apply to any broad class of real-world problems. We discuss this issue more fully in our Open Problems section.

We emphasize that both the UAM and CWA occur often in the literature. Further, they are seldom recognized as assumptions, but instead are implicit. Hence they constitute an ongoing conceptual hurdle for extending the work to many real-world domains.

4.5. What Are Legitimate Inference Goals and Targets?

Computer scientists tend to approach inference problems by scoring their sanitization method (either implicitly or explicitly) solely on the dimension of "attribute distortion". That is, the sanitizer loses points to the degree that an adversary is able accurately to infer particular "sensitive" raw attributes from the sanitized data, and gains points to the degree that an analyzer is able to access the raw attributes it needs without undue distortion (that is, without the data of interest being perturbed, swapped, generalized, or suppressed).

Exceptions to the above are primarily in one application area—census data. Here, a sanitizer loses points exactly as in the above rule, yet it gains points not so much by revealing particular raw attributes, but rather by enabling the analyzer to draw statistical inferences from the sanitized attributes that are similar to those inferences it would have drawn, had the analyzer been able to access the raw data. But for other applications, computer science researchers tend to treat information not represented *explicitly* within the raw data set (call this a "higher-level" inference target) as an illegitimate inference goal. We assume that such "higher-level" inference targets are legitimate goals for the analyzer, and that, by symmetry, "higher-level" inference targets likewise may be legitimate goals for the adversary.

But the opposite notion remains a pervasive (and unexamined) assumption. In particular, the Multi-Level Secure Database Inference Problem has a strong tradition of relying on a *closed world assumption* [9,22]. Among other things, this assumption postulates that all inference targets of concern are represented explicitly as attributes in the *same* database. Conveniently, this limits the problem's scope to a great extent, but poses a problem. Consider the ultimate in secure database design, where all classified data is partitioned between two separate and unconnected databases. All "Low" data is in database 1, and all "High" data is in database 2. Now, by the closed world assumption, it is impossible for an adversary to infer any "High" data from "Low" data using any method covered by the term "database inference" because the two databases are separate. Yet in our "ordinary language" sense of the term "inference", it might be possible to infer virtually all the "High" data from the "Low" data.

An example will clarify this. Suppose a sanitizer is given raw medical records that include as attributes NAME, STREET_ADDR, CITY, ZIP, GENDER, and DATE_OF_BIRTH. To protect patient privacy, the Sanitizer first suppresses NAME, STREET_ADDR, and CITY entirely, because they constitute sensitive (High) data. But among the remaining, supposedly non-sensitive (Low) attributes, the compound attribute of {ZIP, GENDER, DATE_OF_BIRTH} may serve as a quasi-identifier that uniquely can identify perhaps 87% of the population of the United States. Therefore, the sanitizer perturbs or suppresses these attributes as well. Next, the sanitizer is given exactly the same raw dataset, except this time, someone else has already deleted the attributes, NAME, STREET_ADDR, and CITY. If our sanitizer operates by the closed world assumption, then no further sanitization is needed to protect patient privacy, because no directly sensitive (High) inference target appears explicitly in the raw dataset.

Therefore, our approach explicitly rejects the closed world assumption. Instead, we allow inference targets that are *not* represented explicitly as attributes within the raw dataset to be considered legitimate scoring goals for players in the game.

Returning to the multi-level secure database inference problem, with its notion of explicit government-issued classification labels, one might argue that classifying attributes is an end in itself, rather than a means to the end of preventing certain inferences by an adversary. Therefore, ascribing particular inference goals to an adversary is irrelevant. But, if knowledge is power, then classifying attributes is a means to deny power to adversaries. Those adversaries may include other nations, internal bureaucratic rivals, and investigative reporters from one's own nation. For all these cases, whoever does the classifying must *already* have an implicit threat model in mind[7]. Given the existence of this (possibly implicit) threat model, it makes sense to develop and elaborate it by ascribing particular types of inference goals to the adversary—otherwise, how else can the classifier have any confidence that the low-

6. The phrase "closed world assumption" is overloaded. It also refers to an assumption that applies to database languages and logics that is different from the assumption that pertains to inference problems.

7. Unless the rule is to classify absolutely every piece of data,

level raw data it denies to the adversary would correspond to denying higher-level inferences leading to undesirable power?

In general, much previous work on the database inference problem has focused on a method-oriented privacy (secrecy) policy, whose expressiveness was severely limited by the explicit attributes available in a particular database. Instead, we advocate a goal-oriented privacy (secrecy) policy that incorporates an explicit threat model, and ascribes potential inference goals to that adversary.

4.6. Semantically Rich Scoring of Methods

We have introduced a framework for inference games in which all scoring (both positive and negative) is explicit, and may be based on arbitrary, higher-than-attribute-level inference targets. Some of these inferences may have particular raw attribute values as their ultimate target. But we do not limit the inference targets—or the scoring—to the raw data.

Our generalized scoring framework stands in stark contrast to other possible frameworks. A simplistic scoring function to implement the Uniform Analysis Metric would deduct one point for each attribute value that is altered or suppressed by a sanitization method. But many Analyzers have no interest in accessing certain attributes, and for other attributes, certain perturbed values may have only minor impact on the ultimate analysis results (inferences). Moreover, in some applications, two perturbed attribute values might offset each other, thus achieving the desirable effect of minimal or zero distortion in the analysis result. If this seems unlikely, consider that such offsetting perturbations are exactly why the *swapping* method has been so successful in supporting statistical analysis of census data. Thus, many real-world applications warrant very different analysis metrics than the simplistic "change one attribute, lose one point" scoring system of the Uniform Analysis Metric.

Consider previous secrecy policies and metrics for the classical multi-level secure database inference problem. The fine-grained scoring gradations of our framework may seem irrelevant to this problem, because the problem needs no subtle scoring nuances; allowing certain attribute values to be inferred by someone with insufficient classification level is absolutely forbidden. Thus, either a sanitization method satisfies the classification policy, or it does not, so the only possible scoring outcomes are either 0 or 1.

First, consider the detection phase that searches for inconsistencies in classification, for example, a situation where "Low" attributes allow one to infer the value of a "High" attribute. But if there is any possibility that not all classification inconsistencies will be remedied (perhaps due to constraints on time, budget, or qualified personnel resources), then a nuanced scoring system can help the sanitizer prioritize its limited resources. In particular, if one pair of "Low" attributes discloses a single "High" attribute, whereas another pair of "Low" attributes discloses multiple "High" attributes, then (other things being equal) it seems more important to address the second pair. As another example, suppose our MLS system has not two, but three levels of classification. In this system, it is worse for a "Low" attribute to disclose a "High" attribute rather than a "Medium" one. Yet a simple Boolean secrecy policy cannot express such gradations and distinctions.[8]

Next, consider the correction phase of the database inference problem, in which classification levels of attributes are altered to remedy detected inconsistencies. But there may be many possible ways to produce a consistent classification. A "High" attribute may be downgraded to "Low", or *vice versa*. Both attributes could be reclassified at a "Medium" level. When faced with a choice of several "Low" attributes, only one of which must be upgraded to resolve an inconsistency, a framework with fine-grained scoring can help the sanitizer evaluate different possible perturbations or suppressions (re-classifications) that satisfy a secrecy policy.

4.7. Iterated Games and Greedy Strategies

We now extend our model to discuss iterated games. Previously, we examined sanitization in a non-iterated context. The sanitizer "dealt" a single round to the analyzer and adversary, and the game was scored based solely on that non-interactive round. In retrospect, it is apparent that all sanitizers practiced a *greedy* strategy of maximizing their score in the current round. Obviously greed is the optimum strategy for a single-round game. But this strategy may not be optimal if scoring is cumulative and extends across a sequence of iterated rounds. Naively, as one cannot foresee the future and may not know how many rounds will be played, greed may seem at least as good as any other strategy.

But if a sanitizer can reasonably foresee that future rounds will involve a different analysis on the same raw dataset, or the same analysis on an updated or extended dataset (that is, attributes pertaining to some or all of the same entities), or an adversary armed with new auxiliary information, then the sanitizer may adopt a heuristic strategy that "errs on the side of privacy" and defers some potential analysis points in the current round, in order to achieve a better cumulative score (position in multi-dimensional value space) after a series of rounds.

The strategic context is further complicated by the existence of other collectors and sanitizers, whose data may pertain to some or all of the same entities that appear in our own sanitizer's raw data. These other sanitizers may have very different privacy or analysis policies (even if, by chance, they share the same metrics as our sanitizer). Hence the sanitized data they release in the future, if accessed by our adversary, may boost its auxiliary or external knowledge, and thus retroactively endanger the privacy of entities in our current dataset.

Therefore, even if our sanitizer plays only a single round—without explicit iteration—there exists the potential for future "virtual rounds" of scoring revisions, as new external data becomes available to its adversary. Perhaps our sanitizer also can gain points in such virtual rounds, if a particular game's rules allow its analyzer likewise to operate on that new data.

---

8. Since its is unsafe to limit the problem scope under the closed world assumption, it is possible that not all classification inconsistencies will be remedied, because it is possible that not all such inconsistencies will be detected. One could argue that it makes sense to allocate some sanitization resources to detecting vulnerable inference targets not represented in the database, rather than to remedying known, but minor, classification inconsistencies.

4.8. The Query/Audit Problem

The query/audit problem has several definitions. The classical versions limit their scope to statistical databases, but even then, their rules may differ regarding which statistical queries are allowed. The version we use requires that a database respond to all queries accurately until the answer to the current query, in combination with answers to previous queries, would allow the questioner to deduce sensitive information. Then the database refuses to answer the current query.

Our approach models the query-audit problem as an iterated sanitization game. In each round, a query is issued against the same raw dataset, and the rules require the sanitizer (called an *auditor* in this context) to practice a greedy analysis strategy. The auditor must maximize the analysis score in each current round by answering without perturbing or suppressing any values, until it receives a query whose correct, raw-data answer—in conjunction with the data disclosed by all the previous queries and responses—would violate its privacy policy. In that case, the auditor must suppress its answer, responding by giving an explicit refusal to answer the query. This "all or none" sanitization method is required by the rules of this game.

These sanitization rules governing the auditor raise two important points. First, suppose we assume that the auditor is a *rational* game player, and that its query-response behavior is not constrained by the above rules. Rather, the auditor's behavior is a rational strategy, constrained solely by the scoring incentives of the game. We then ask what scoring incentives would motivate it to behave in such a seemingly greedy manner.

By thus transforming a game's rules into equivalent scoring incentives, we can "reverse engineer" its social context. This technique can serve as an important check, because it helps ensure that a particular game—as governed by its specific rules and policies—is played ***only in an appropriate real-world context***. The policy makers for the auditor (sanitizer) can compare this game-inherent scoring system with the values they place on various privacy and analysis inferences. If the implied scoring incentives are not congruent with their values, then their sanitizer has been told to play the wrong game. The policy makers need to change the rules (typically those rules constraining the sanitization method), or play a different game entirely. Transforming a game's rules into equivalent scoring incentives serves a social purpose analogous to the technique of developing a prototype to implement proposed software requirements. Making the implications of a given set of requirements (rules) visible (in a "What-You-Specify-Is-What-You-Got" manner) often helps policy makers revise, revisit, and/or reconcile their requirements.

Obviously, for those contexts that are accurately modeled by the query/audit problem's scoring incentives, the short-term gains from analysis inferences must be extremely lucrative—otherwise, the auditor would practice at least a 1-round lookahead strategy as a precaution to protect privacy. In our earlier discussion of the multi-level secure database inference problem, its scoring incentives were dominated by penalties for violating the secrecy (privacy) policy. But in stark contrast, the scoring incentives for the query/audit problem are so dominated by analysis points, that explicit violations of the secrecy (privacy) policy can occur.

This brings us to the second important point: Rules constraining a sanitizer's ability to alter its responses may constitute a vulnerability. This is possible if the adversary knows the rules of the game—in this case, if it knows that the auditor will refuse to answer a query ***if and only if*** that answer, in conjunction with previously disclosed data, would violate the privacy policy. By exploiting the auditor's greedy strategy—and its transparent sanitization method—the adversary can set a trap whereby the auditor's refusal to answer actually reveals as much information as its correct response would have revealed.

For example, in queries regarding a particular subset of individuals, suppose the adversary asks for the COUNT of records in that subset, and then asks for the SUM of their SALARY attributes. The auditor will answer both queries, since they reveal very little about any particular individual. Next, the adversary asks for the MAXIMUM among those SALARY attributes. If the auditor refuses to answer, this refusal tells the adversary that every individual in that subset has the same SALARY = SUM/COUNT [19]. In effect, by failing to look ahead even one move when evaluating whether the current query's answer might lead to a privacy violation, the auditor fails to recognize positions where the adversary has placed it in "check', and thus allows a "checkmate" to occur.

4.9. Disclosing Sanitization Meta-Data: Perturbation and Generalization

Sanitization mechanisms may be grouped into two broad classes. *Generalization* either omits a sensitive datum (*suppression*), categorizes it with other data (*aggregation*), or replaces it by less specific values (for example, by rounding). Examples of this include cell suppression techniques and aggregation techniques [5,8,17]. *Perturbation* replaces a sensitive item by some other legal value from the domain of that datum; for example, by adding a random number to it. Perturbation techniques have been widely studied in databases [2].

Misconceptions abound regarding an alleged advantage of generalization over perturbation as a sanitization method. But to address them, we need a deeper understanding of the relationship between those methods. At this stage, our informal characterizations of those terms are adequate.

We begin with what an adversary is told about the sanitization procedure: *How much information regarding your raw data can leak by disclosing particular sanitization meta-data?* In the context of the generalization-vs.-perturbation issue, that question can be rephrased as: Why should the sanitizer let the adversary know when the sanitizer is lying, or telling less than the whole truth?

Analyzing the query/audit problem using our inference game paradigm helped to highlight this issue. But the underlying question is not new. It is a somewhat similar question, for sanitization, that Kerckhoffs' Principle answered long ago for cryptography. But sanitization is a different problem than cryptography. In particular, we must consider the differential benefits—to the adversary, to the analyzer, and to other sanitizers—of disclosing sanitization meta-data. Moreover, it is not obvious that all sanitization procedures will decompose cleanly into a (presumably public) algorithm, and a nontrivial (secret) key.

Duncan et al. [15] explore one sanitization context that does admit a clean decomposition, namely, the use of additive multivariate noise to sanitize data, when the analyzer's goals involve specific estimation problems, such as finding regression coefficients. The authors' purpose is not to satisfy a fixed privacy and analysis policy, but rather to help policy makers assess trade-offs when developing or reconciling such policies. Given well-defined application domains with specific inference goals for the adversary and the analyzer, they develop metrics for privacy and analysis, then plot graphs of privacy risk versus analytic usefulness, as the sanitization procedure varies its additive noise parameter.

In this work, the authors explore how the privacy vs. analysis trade-off curves vary if they disclose the value of their additive noise parameter. Thus, they offer policy makers a tool to assess the differential benefits—to adversary and analyzer—of disclosing a key item of sanitization meta-data.

With this background, we return to the relationship of generalization and perturbation. In essence, generalization means that the adversary knows, with absolute certainty, the extent to which sanitization may have perturbed each raw datum.

Typically, generalization discloses this sanitization meta-data via an explicit syntactic tag applied to a datum, such as Birth_Year = "195*" or Birth_Year = "1950–1959". Alternatively, if such an enhanced syntax is not allowed, generalization might replace the last digit of all Birth_Year values by "0", and then publish to all interpreters the enhanced semantics, namely that a "0" value in the last digit of Birth_Year really means the individual may have been born in that year, or in any of the following 9 years.

But in the above, the sanitizer (generalizer) could achieve the same result by changing the last digit in Birth_Year to *any* digit, and then publishing the semantic interpretation that *any* least significant digit in Birth_Year must be interpreted in the context of this same new semantics. Using this method, the syntax of generalized data appears indistinguishable from perturbed data. In fact, for this example, if we could "subtract" the meta-data publication, we could convert this from a generalization method to a perturbation method—without altering the algorithm that transformed Birth_Year.

Similarly, many perturbation methods can be converted to generalization methods, without any software changes. Perturb the data, and then publish, for each and every sanitized datum, a set of possible raw pre-image data values.[9]

For example, consider numeric data perturbed by additive noise, as follows:

$$\text{Sani\_Birth\_Year} = (\text{Raw\_Birth\_Year } \mathit{div}\ 10) * 10 + \mathit{rand}(0, 9)$$

where $\mathit{rand}(x, y)$ returns a random integer in the $[x, y]$ interval.

9. For perturbation methods where those pre-image values are not semantically "contiguous", opinions will differ regarding whether such meta-data publication does or does not transform them into generalization methods.

If we add a post-processing step and publish the above equation as sanitization meta-data, that would convert this perturbation method into the same generalization method as our previous example.

Note that a sanitizer need not publish *tight* bounds on its meta-data to implement generalization. For example, a sanitizer might perturb a numeric field by adding a value between +1 or –5, but publish a looser bound ranging between +20 or –6. Later, in a subsequent "virtual round" of scoring, the sanitizer might choose to disclose the tighter bound. This is much easier—and safer—than publishing a re-sanitized version of the same dataset. Moreover, a sanitizer might choose to publish meta-data that is not even true; this raises interesting strategic possibilities.

Armed with this deeper understanding that a generalization is a perturbation augmented by disclosure of certain meta-data, we are now prepared to explore issues regarding the "quality" or "distortion" of sanitized data, from the analyzer's viewpoint.

4.10. Meta-Data, Analysis Distortion, and the UAM

In this section, we compare how an analyzer's inferences may be distorted by perturbation and generalization.

Assume, simplistically, that the analyzer's initial condition is a "blank slate," meaning it has neither external data sources, nor preconceived ontological categories or inferential "curves" to which it will attempt to fit the sanitized data points. Supposedly, the strength of generalization is that its "truthfulness" relative to some attribute is preserved [29] even though an inference based on that data may not be as precise as that inference based on the raw data. For example, if the raw attribute for John's height is 7 feet, and sanitization generalizes that to "at least 6 feet", then a resulting inference that "John is not short" remains true (for the usual meaning of "short").

But the integrity of such inferences is not preserved if the uncertainty introduced by generalization crosses the boundaries of an ontologically significant category. Sanitizing John's height to "at least 3 feet" introduces the possibility that "John might be short". Thus, generalization allows the modal logic inference "*X* must be *false*" to become "*X* might be *true*". Clearly, generalization cannot guarantee modal logic consistency between inferences drawn from the raw and sanitized data, even in a vague "blank slate" context.

Although generalization increases uncertainty about a datum, it allows a careful interpreter to bound the uncertainty of inferences based on that datum. In contrast, unless sanitization meta-data is published, perturbation methods do not tag sanitized data with "error bars". If an analyzer tries to draw inferences other than those guaranteed by the analysis policy, its inferences may be based on values whose perturbations cause significant distortion. Hence, neither the probability nor the precision of those inferences are guaranteed. But by publishing sanitization meta-data, perturbation methods are more versatile than generalization. For example, semantic meta-data easily can specify joint probability distributions among groups of sanitized data; but syntax-based generalization cannot match that capability.

Yet, for example, Meyerson and Williams [24] argue that, "with perturbations in data, only 'probably true' inferences may be

drawn." They imply that generalizing data preserves the capability "of rigorously proving to a judge that a certain trend is indeed occurring." But this claim depends on the analysis policy's published bounds on uncertainty, not on the method (generalization or perturbation) used to implement that policy. Even in a "blank slate" context, generalization might not preserve the capability of proving that a certain trend is occurring, nor can it guarantee that a countervailing trend is not occurring. The method used to implement that policy is secondary.

In practice, very few adversaries and analyzers start in a "blank slate" condition; rather, they operate under (different) contextual constraints, and with various sources of external knowledge. Analysis outcomes in medical, military, terrorism, and legal contexts are constrained by various decision thresholds such as reasonable doubt, preponderance of evidence, and unacceptable risk. In these contexts, the analysis may be performed to refute or confirm some provisional hypothesis. Whether an inference game player must cross a "burden of proof" threshold in order to score for or against a particular hypothesis is critical, because the absence of evidence to support that position will be treated as evidence that support is absent. The standard that the analysis must meet to confirm or refute that hypothesis is determined by the social context of the application.[10] Thus, the inferential power of evidence—and of its absence—is context-specific. Inferences based on "evidence of absence" can be crucially important in law, medicine, and even literature (as when the dog that did *not* bark in the night provided a crucial clue to Sherlock Holmes [12]).

There is a risk that generalization techniques may cause a distorted *under*emphasis. For example, in clinical trials of drugs, if even a small number of detrimental side effects are suppressed or overgeneralized, resulting inferences may have fatal real-world consequences. The analysis policy must preclude such outcomes, and the sanitized data must conform exactly to that analysis policy.

Similarly, there is a risk that generalization may cause a distorted *over*emphasis. For example, Sweeney [29] describes a Google search for information on a particular person. The unavailability of certain records regarding that person's charitable donations might cause a reasonable person to conclude either that the subject was stingy, or that the subject heavily favored the one charitable organization that the search did reveal.

Whether perturbation techniques are better than generalization techniques depends on each specific real-world application context, the characteristics of each specific adversary and analyzer, and the content of each specific dataset to be sanitized.

Sanitizing data using methods of generalization offers obvious benefits [29]. The drawbacks are less apparent. To illustrate them, we examine generalization in the context of sanitizing IP addresses in network traffic.

---

10. Such analysis outcomes may include not only the first-order results (to confirm or refute a particular hypothesis), but also the *meta-analysis* outcomes of whether, and how, to continue searching for further evidence, or whether, and how, to formulate a plan to develop a new hypothesis.

Suppose we generalize every IP address by "blanking out" the lowest-order bit. This preserves *most* of the locality of the addresses present in the original data, and might be considered "minimal distortion" by some analysis metrics. The cost of doing so requires that the IP address' lowest-order bit can have *three* values: 0, 1, and the blank, so we need new hardware to handle the ternary value of that "bit". As an alternative, in the software we might set that low-order bit to the same value for all sanitized IP addresses. In either case, the analyzer must know whether we are transforming the *syntax* of the data that it will see, and if so how. The analyzer will need to understand how to interpret the *semantics* of the sanitized data appropriately.

But changing software is costly too. If modifying the analyzer's software to understand the blank value for a "bit" is acceptable, then most likely modifying the software to increase the size of the IP address space its data structures can represent is also acceptable. Doing so allows us to overcome the problem of a limited namespace, and eases other types of constraints.

Similar considerations apply to the syntax of less-constrained data. For example, suppose previously the analyzer understood ASCII integers to represent a person's age in years. But the sanitized data does not say that Tom's age is 27. It says that Tom's age lies in the *range* of "20–30" years. It might be easy to modify the analyzer to understand dashes in the syntax. But if the analyzer was designed to measure complex statistical relationships, such as those between one's age and weight, the analyzer would need major changes to properly *interpret* the syntax of the sanitized data. Even if the analyzer were designed simply to count the number of people in different age brackets, those ontological brackets exist only at the semantic level, through internal logic and data structures. The analyzer does not understand the age bracket syntax, unless modified to do so. Worse, suppose the analyzer understands age brackets, but expects them to be "18–25", "25–35", and so forth. The *ontology* of the sanitizer's data intervals does not match what the analyzer is expecting. This disconnect renders the analyzer useless. Clearly, embedding the meta-data semantics into the sanitized data syntax can be a disadvantage. But it also constitutes the only inherent advantage of generalization, since the resulting data are more "idiot-proof" (and perhaps more lawsuit-resistant) than perturbation (where the meta-data semantics must reside in a separate file or a commented header).

To determine whether a specific generalization or perturbation method is "better", a sanitizer must know the relative importance of particular inferences. To express those inferences adequately requires not only a privacy policy, but also an *explicit* analysis policy.

Some authors [11,24] adopt the UAM as an *implicit* analysis policy. This might be formulated as: "any datum whose disclosure the privacy policy does not prohibit is *equally* valuable under the analysis policy". In addition, Meyerson and Williams [24] and others imply that the proper metric for characterizing the *aggregate* analysis value of a sanitized data set is the sum of the number of individual unsanitized data elements it contains.

But in real-world problem contexts, many data items of minor privacy concern (for example, whether someone has dial-up or DSL access to the Internet) may have negligible value for certain kinds

of analysis (such as for analysis of medical records), yet have considerable value for other kinds of analysis (such as mining purchase transaction data for market research purposes). Furthermore, the inferences that can be drawn from an aggregate set of *m* attributes often depend more on which particular attribute subsets the aggregate contains, rather than on the number *m*.

Following Sweeney [28], if an adversary could pick only a few attributes to identify an individual, the adversary's privacy-penetrating inference would be far more accurate if it chose as attributes gender, zipcode, and birth date, rather than weight, height, zipcode, and birth date. Despite height alone partitioning people into more than the 2 equivalence classes of gender, voter registration data (which is public information) records gender, not height and weight.

As not all attributes or subsets of attributes have equal value for detrimental, privacy-penetrating inferences, not all attributes (or subsets) have equal value for drawing *beneficial* inferences.

Dinur and Nissim [11] develop a highly-nuanced privacy metric, yet use an analysis policy based on "noise". They model the adversary as a 2-phase query-audit game. First the adversary queries the database (adaptively), and the auditor returns sanitized results. After this phase, the adversary emits a list of all the data elements whose original values it intends to guess. In the second phase, the auditor/sanitizer reveals the entire contents of the raw database to the adversary, except for those data elements the adversary will try to infer. The privacy metric is the probability with which the adversary is able to infer the specified data.

Although this privacy metric seems to have some drawbacks, suppose we also allow the analyzer to play this game. A more useful analysis metric than "noise" might be the probability with which the analyzer is able to infer the data elements it has specified. Such an adaptive querying model may be useful for characterizing beneficial data mining performance on sanitized data.

Finally, Sweeney [30] suggests balancing warranties with privacy protection; she states that a developer "should provide a guarantee related to the utility of the algorithm (a *warranty*) and a guarantee of privacy protections the algorithm provides (a *privacy statement*)." We do not assume such a guarantee can be made; indeed, the essence of our approach is to determine *what* guarantees, *if any*, can be made for a given analysis algorithm and a privacy policy, and moreover how either (or both) must change in order to provide whatever guarantees are desirable.

4.11. Actionable Inferences

Earlier, we stated that in real-world contexts, it is not *truth*, but rather *beliefs* (confidence levels of inferences) that can transform abstract risks or probabilities into actions that have consequences. To provide privacy protection against detrimental consequences, it is essential that a sanitization method be based on an accurate threat model. Specifically, a privacy policy must assess what beliefs and confidence levels will trigger damaging action by an adversary.

Terminology can shape—or distort—our thinking. *Identification risk* and *re-identification risk* have become problematic terms. Such research focuses on what seems necessary—but is not sufficient—to protect individuals from the consequences of inference. In this section, we argue that exclusive focus on a single, isolated aspect of risk, coupled with the failure to develop an adequate threat model, can result in privacy policies and sanitization methods that may not reduce detrimental consequences, but may instead *amplify* those detrimental consequences, and shift them onto innocent individuals.

To understand the limits of the focus on identification risk, we examine what is required to trigger damaging action by an adversary. We begin by introducing the notion of predicate risk.

We define *predicate risk* as a change in uncertainty (typically but not necessarily a reduction). It is a newly-changed probability (or confidence level) that an interpreter (in this case, an adversary) assigns to a particular predicate, as a result of viewing a dataset.[11] Identification risk might be considered one type of predicate risk, but for clarity in our discussion, we exclude that possibility.

As an example, an anonymized medical record's unsanitized diagnosis attribute says "Ms. D" has cancer. Previously, the interpreter either had assigned that predicate an undefined probability (as it did not know this particular anonymized tuple existed), or it had inferred an average cancer risk for the population that it would apply to every unknown entity. But based on this record, the interpreter assigns a 100% probability to the predicate (Ms. D, Cancer).

In this example, suppose the interpreter is a life insurance company. Although the cancer predicate risk for this particular tuple is significant (100%), it is uncoupled from the real world because the identity of "Ms. D" is not known. It might seem that this predicate risk will not become actionable unless and until the interpreter can couple it with some identification risk. That is, if the life insurance company can infer (with some degree of confidence) which entity in the real world is denoted by "Ms. D", then it can translate the predicate risk of cancer into actual consequences, by raising insurance rates for Ms. D, or denying her coverage.

It is clear from this example that identification risk is a useful concept. But it is a mistake to conclude that sanitizing the data merely to reduce Ms. D's identification risk below some threshold is sufficient to ensure that the adversary will not take detrimental action.

Now, suppose that we sanitize the data solely by hiding the diagnosis attribute. Even though patients' names and addresses are revealed, an insurance company (adversary) would have no reason to raise rates or deny coverage to anyone, because all cancer predicate risk has been eliminated from the data. Instead, if we sanitize solely to reduce a vulnerable individual's identification risk, this does nothing to reduce the underlying predicate risk. Thus, in the absence of policy safeguards, that predicate risk may be shifted onto other individuals. For example, if sanitization causes the adversary to infer—incorrectly—that Ms. D corresponds to Robin in the real world, then the privacy policy has done nothing to *reduce* the real-world consequences of data disclosure, but has merely *shifted* those consequences onto the wrong person. Alternatively, suppose that the sanitization method does not alter any diagnosis attribute, but it perturbs ZIP codes and birthdates so that,

11. This informal definition is adequate for our purposes. Readers desiring a formal definition may choose from a variety in the literature (see for example [9,20]).

based on these quasi-identifier fields, Ms. D is indistinguishable from *k* people in the real world. If *k* is a small number, the insurance company could raise rates or deny coverage to *all* of those *k* indistinguishable individuals. Thus, in comparison to disclosing the raw data, sanitizing the data has caused this adversary to *multiply* its damaging consequences by a factor of *k*.

In effect, such sanitization schemes first redistribute identification risk, and then assume the adversary will transform that risk redistribution into damage reduction. But sanitization methods that attempt to defuse a threat solely by diffusing risk are successful only for certain types of adversaries. Hence it is vital that the threat model be explicit, and accurate.

To promote conceptual clarity, we note that, in general, nothing forbids an interpreter from taking action based strictly on a predicate risk that is uncoupled from any identification linking that predicate to a specific entity, or even to a location. For example, if the United States infers that a nuclear missile recently was stolen from a former Soviet country by an unknown party, it is likely the United States would commence a worldwide search. This interpreter (the United States) takes action because it infers *what* predicate occurred, even though it knows nothing about who, where, or why, and has only a general idea of when.

Rather than the often misleading notion of identification risk, it is safer and more accurate to define *entity access risk* as the ingredient that makes predicate risk actionable. For example, to access a particular anonymous individual, an interpreter might publish its attributes, and offer a reward for information regarding its identity or location. Alternatively, previously-anonymous individuals have been apprehended via sting operations that enticed them by offering customized rewards that appealed to their known attributes. Finally, in many contexts, probabilistic entity access may be sufficient to trigger action. For example, an interpreter may target a class of individuals reachable through a particular access channel because the interpreter infers that the probability density of various undesirable predicate risks in that class is high. Thus, individual identification is merely one *means* to entity access—but it is not a prerequisite for entity access.[12]

We have shown in this section that a research focus on identification risk, sometimes called "re-identification risk", is necessary but not sufficient to protect individuals from the adverse consequences of inference. The adversary determines what is actionable, and from an adversary's viewpoint, predicate risk can become actionable when coupled with sufficient entity access risk, even in the absence of identification risk.

5. CASE STUDY: *K*-ANONYMITY AND *L*-DIVERSITY

An examination of *k*-anonymity offers tangible examples of some issues mentioned previously, as well as some issues unique to *k*-anonymity. In what follows, it is important to distinguish between criticism of *k*-anonymity's suitability as a *policy* in particular real-world contexts, and criticism of generalization as the *method* chosen to implement a given *k*-anonymity-based privacy policy (after the analysis policy is fixed for some context).

We begin with the suitability of *k*-anonymity as a *policy*. *K*-anonymity seems to provide all sanitized entities with a uniform level of privacy protection, because it renders every entity indistinguishable from at least *k*–1 other entities in that sanitized dataset. But is a uniform level of protection always the best privacy policy, and does *k*-anonymity provide that?

From a public policy standpoint, one could argue plausibly that both the very young and the very old warrant more protection than others. Elders tend to be more trusting, and vulnerable to financial scams. Moreover, because elders tend to live off their lifetime savings (rather than current income), their ability to recover from a significant theft is more limited. The prevalence of medical problems among the elderly—both as an isolated factor and in conjunction with financial vulnerability—similarly argues for additional privacy protection.

Younger people would have many years in which to recover from a breach of privacy. But given current dynamics and trends in data collection and (lack of) protection, it seems only prudent to assume that young people will incur far more cumulative exposure to privacy risks during the course of their lives than have previous generations. As noted earlier, a "greedy analysis" strategy of releasing all data that is "safe" today, may be recognized as a reckless privacy policy for this subpopulation in a future inference game.

Similarly, privacy risks and damages should be assessed to identify other extra-vulnerable subpopulations.

As we have pointed out, privacy protection is commonly equated with the notion of identification risk. But it would be a mistake to equate *any* form of risk with privacy protection. A reputable threat assessment in computer security analyzes both risk probabilities and the consequent *damages* that would result from exploiting various vulnerabilities. Likewise, a privacy policy regarding personal security ought to consider damages, not merely risk probabilities.

Let us assume the above policy issues have been settled. Now we apply any *k*-anonymity algorithm that generalizes attribute values to achieve the desired policy result—namely, that a sanitized entity is indistinguishable from exactly *k*–1 others in the sanitized dataset based on its quasi-identifier (QID) attributes.

In [20], the authors point out one significant concern with *k*-anonymity generalization: if all *k* members of a QID-anonymous group have the same value for one of their sensitive (unsanitized) attributes (e.g., diagnosis = cancer), then the homogeneity of this attribute may negate much of *k*-anonymity's protection. As an example, an insurance company might take action if sanitized data reveals that all individuals in ZIPCODE = "8520*", with BIRTHDATE = "May **, 1947", have cancer. As a solution to this problem, the authors offer an elegant family of fixes they call *l-diversity*, and provide guarantees that are not merely provable, but also quite useful in the real world. But there are two problems.

One problem is that diversity of attribute values in the *l*-diversified dataset does not guarantee similar diversity of consequences when *interpreted* by an adversary in the real world. Suppose, in the

12. Nor does identification guarantee access. As an example, if an adversary wishes to harm the President of the United States, learning the identity of the individual holding that office will open very few new access channels.

above example, the diagnosis attribute had a different value for each entity in that *k*-anonymous group. If every diagnosis implies a gloomy prognosis (for example, AIDS, Ebola, metastasized Melanoma, and so forth), then all these diverse attributes imply the same *high-level* inference target—namely, major medical costs over a short life expectancy. In an insurance context, a prudent threat model would expect an adversary to take action to mitigate its otherwise significant financial loss. But the old closed world paradigm from the database inference problem does not take into account consequences arising from the adversary's higher-level inferences.

It might appear that *l*-diversity can remedy this problem, merely by simulating one aspect of the adversary, and mapping values of the diagnosis attribute to their corresponding high-level inference targets. Then, by appropriate *k*-grouping of sanitized records, one could ensure diversity from the adversary's viewpoint.

This approach acknowledges the central role a specific threat model must play when designing an adequate privacy protection system. But this method may not fully alleviate the problem. As noted earlier, for many realistic adversaries in an insurance context, the threshold for triggering action may be probabilistic entity access (perhaps via targeting a location-based class of individuals), coupled with a high probability *density* of various undesirable predicate risks in that class.

If an adversary obeys this economic logic, how can we rehabilitate *k*-anonymity and *l*-diversity? One option is to call for intervention at the policy level, because decisions regarding the redistribution (or even multiplication) of costs onto innocent people should be recognized as matters of ethics, not low-level implementation details. We emphasize that, in the absence of explicit additional policy safeguards, *k*-anonymity—like many other sanitization methods and policies—can redistribute and even amplify risks and damages.

But suppose policy makers want to maintain privacy protection for members of such *k*-groups, and want to maintain the diagnosis attribute unsanitized, for purposes of medical analysis.Another alternative is to recognize that the problem has two roots. One root is the nature of the adversary, which implementors cannot change. But the other root is the initial decision to implement the sanitization method of *generalization* rather than perturbation to achieve *k*-anonymity (and *l*-diversity). Various forms of perturbation have different strengths and weaknesses compared to generalization, so perhaps a perturbation-based technique can improve the situation.

A third, obvious fix is to increase the value of *k* (at least for vulnerable *k*-groups) or shift this vulnerable class of individuals to a higher degree of generalization (for example, by substituting the wildcard character for not merely one, but the two least significant digits in their ZIP codes). When the adversary interprets the entire dataset, this dense cluster of predicate risk will then be diffused throughout a larger region of attribute space, and its high predicate risk will tend to revert to the mean.

The problem with this fix is that it accepts the implicit value judgement that it would be bad to disclose a homogenous cluster of sensitive attributes having the same (or similar) detrimental value. By this reasoning, because the predicate risk—when interpreted by an insurance adversary—could have detrimental consequences for the vulnerable individuals; therefore, we must alter the sanitization to protect their privacy. But one key theme throughout this paper has argued for elevating analysis to the level of policy, because value judgements about both privacy and analysis are matters of ethics. Hence they should be addressed—and reconciled if necessary—explicitly via policies.

In this particular example, the existence of (for example) a cancer cluster is a matter of significant concern to public health officials and medical researchers. Other people may be at risk if the cause is a public health hazard such as an unknown toxic waste dump. Moreover, if the individuals in this cluster were aware of its existence, they might rightfully be more concerned about how to protect their children's health, rather than how to protect their own privacy and insurance coverage.

We are not arguing for the primacy of public health over privacy rights, or vice versa, in this hypothetical example. But we *do* argue that sanitization designs and decisions must be grounded in specific application domains and social contexts. Ethical aspects of situations must not be abstracted out of the problem; rather, they should be incorporated into formal approaches via explicit policies and shared ontology libraries.

6. OPEN PROBLEMS

This section describes some open, fundamental problems in data sanitization. They underlie much of the work being done, yet are addressed only in the context of the particular work, or are captured by (often implicit) assumptions in the work. Here, we make them explicit, because they are basic to many application domains.

6.1. Privacy, Analysis Policy Languages

Consider first how to express privacy and analysis policies formally. A language for policy expression will meet the following criteria, perhaps using tools that transform one expression of the policy into different forms or languages:

- It should express the policy in terms that a non-technical policy maker can understand; in essence, the language needs to be a "what you say is what you get" language.
- It needs to allow a direct comparison of privacy and analysis policies to detect and identify conflicts. This suggests the policy expression should be declarative rather than procedural, because policies make statements; they do not describe procedures. That raises foundational question of how to represent inferences as distinct from more general forms of computation.
- The expression of the policies needs to lead to an efficient sanitization function. Ideally, that function can be generated automatically from the policy expression.
- Finally, the expression of the policy must allow changes to the threat model to update the privacy policy automatically.

As each application will occur in a particular context and environment, significant portions of such policies will be specific to the particular domain for which they are generated. One question is whether the common features of these policies will be sufficient to warrant a single, common language to express these policies. Clearly, some augmentation of such a language will be necessary,

but if the common body of those policies is too small, then domain-specific languages may seem more appropriate. But we suspect this assessment has stymied progress, because it has balkanized the problem. A more productive approach might be to develop application-specific ontology libraries. This would allow research costs, results, and insights to be shared across domain boundaries. In any case, perhaps common characteristics of threat models may lead to a better understanding of how to express these policies across multiple domains. This would lead to a modular form of the threat model, with different components expressing differing details of many threats, but a common set of threats making up the core of the model. As many different environments have similar threats, such as inferencing and aggregation attacks, perhaps a general view of an adversary would provide something to focus the threat model around.

A second issue related to policy language is the need to measure the efficacy of the sanitization. First, does the sanitization allow the adversary to determine information that violates the privacy policy? Second, does the sanitization prevent analysis that are required by the analysis policy? Answers to these questions form a continuum, not a binary set. Hence they can be measured, but deriving a *meaningful* metric depends upon context.

6.2. Sanitization and Game Theory

As noted, the provable results pertaining to classical special-case inference games tend to be impossibility (or at least intractability) results. And such results appear isolated, giving scant insight into "neighboring" issues. Given the above framework, a new approach to data sanitization is to formulate new families of inference games, whose rules, scoring incentives, dominant tactics, and ontological structures will define boundary conditions between regions in the inference game space.

By thus mapping the contours of the game space, games within certain regions then may be proven to have tractable solutions, and policy makers may learn the extent to which certain game parameters may (or must) be varied if certain types of solutions are desired. Producing even a rough "Guide to Winnable Games" would provide a significant service not only to policy makers, but also to sanitizers, collectors, analyzers, and the subjects of data collection, who could then decide which games they consider sufficiently safe and worthwhile to play with (selected) other parties, and which games are too risky to play in any form other than as solitaire.

Thus, one goal of future research should be to determine what kind of new games define useful boundaries in the problem space.

6.3. Justifying the UAM

In studying the literature, we found that the intractability results we had regarded as fundamental depend on the UAM. Clearly the UAM does apply in certain situations. Hence certain instances of the inference problem are indeed computationally intractable. By introducing our inference game framework, we have shown how to unify and parameterize several large problem spaces. Yet until boundaries of parameterized regions can be delineated (for example, by our suggested contour-mapping research program), it seems premature to make any large-scale claims about NP-hardness.

Similarly, we question the scope of existing UAM-based intractability results. To extend these beyond a few isolated problem instances, explicit justifications must address and overcome several issues.

One argument for the CWA and the UAM begins with the premise that we know nothing about the adversary or the analyzer, and concludes that, therefore, the CWA and the UAM are the most accurate assumptions we can make regarding the respective inference goals of the adversary and the analyzer. But the conclusion does not follow from the premise. Indeed, the CWA and the UAM are completely arbitrary assumptions.

An interesting experiment would run a Monte Carlo simulation with many different real-world analyzers. We suspect that several other analysis metrics would beat the UAM, for two reasons. First, all real-world analysis has a purpose that has real-world relevance, and certain attributes tend to serve as the *actionable* "links" to the real world, even if those attributes are neither explicit identifiers nor quasi-identifiers. (This especially is true for data mining.) Second, if privacy requires that attribute X be perturbed, then the UAM penalizes a sanitizer if it perturbs attribute Y, X's multivariate partner. This discourages multivariate analysis.

If the analyzer's "high-level" inference goals include terms of the form A–B, we can sanitize attributes A and B by adding the same noise to each. This changes 2 attributes without distorting the analysis. This example is a significant real-world case. Many medical analyses compute elapsed time, and the *date of death* is very important for studying disease and treatment dynamics (for example, did a particular treatment delay death?) as well as for privacy. Real-world sanitization practitioners have used this simple method for over 10 years [14]. But neither the UAM nor the CWA apply to such problems.

For many attributes, and for many interpreters, small perturbations in "syntacs" produce no significant changes in "semantix". That claim proves itself by inspection. But the UAM insists we lose one point by changing one attribute. Next, consider a person's SALARY, represented not as a single attribute, but rather as a bitstring, with one bit per attribute. We name these attributes b_0, ..., b_n, with b_0 being the most significant bit and b_n the least significant bit. But the UAM insists that all these bits are equally significant, and a CWA insists a sanitizer bears no responsibility for considering the impacts of sanitization on some "high-level" inference goal, denoted SALARY, derived from these low-level attributes. Instead, under these assumptions, analysis is harmed equally and in direct proportion to the number of these one-bit attributes we change.

Logically, from these two assumptions, we must conclude that perturbing the single bit b_0—rather than the two bits b_{n-1} and b_n—is better for analysis. Given the context of the attributes, and their semantics, this is almost certainly incorrect.

The UAM and the CWAs risk ignoring the things that really matter, in favor of the things that are easily counted.

7. CONCLUSION

As more aspects of life are routed into digital channels, previous *de facto* privacy protections are bypassed and undermined. Researchers have recognized this. Some have sketched a vision of

personal data moving rapidly between different security contexts. For example, Dragovic et al. [13] suggest maintaining and updating an extremely dynamic threat model, and sanitizing the data accordingly. But our work shows how hard it is to develop a *realistic* threat model for particular relatively slow-changing, or even *static*, situations. And once we have an adequate threat model, we have shown that it is nontrivial to sanitize data, to satisfy even a simple analysis policy.

The goal of this paper has been to strengthen the foundation of the data sanitization problem as an information security problem by explicitly characterizing the gap between the core sanitization problem and various formal approaches to this problem, and by showing how the classical special-cases of sanitization problems can be subsumed by a unifying inference game paradigm. Current formal approaches develop abstract problems under various assumptions that essentially remove or ignore many contextual elements in which the problems arise. The result is that the solid foundational work is applicable in a mathematical context, or in real-world contexts that do not reflect the environment in which the problem arose.

One avenue that might prove fruitful is to study inference methods using *ontologies*. In particular, Sowa's inference graph and conceptual structure work [27] seems ideal for expressing privacy and analysis policies in a semi-executable form, as noted by Thuraisingham [31] and Delugach and Hinke [7] over a decade ago in database work. Given the subsequent popularity of ontologies in other problem areas, significant expertise and software have developed in the ensuing decade. Perhaps sufficient resources now exist that ontologies can be applied productively to policy languages for specific application domains.

Formulating a common ontological language would promote sharing of results both within and across application domains. By thus unifying research efforts among formerly isolated disciplines, this might serve to catalyze significant progress toward a general solution to the inference problem, leading to metrics for testing the trade-off between analysis policies and privacy policies—with results immediately applicable to many areas including the database inference problem [23].

Perhaps the negligible role of ontologies arises because, even within one particular application domain, such ontologies have been expressed via specialized, often proprietary, languages. But a semantics-preserving language for exchanging ontological content, called Common Logic [1], is under consideration as a possible ISO standard. Expressing privacy and analysis policies in terms of application-domain specific ontologies in this (or some other) language may allow inferencing to be performed. In this manner, ontologies and threat models may be shared usefully within an application domain. Research results, including inference heuristics, would become portable *between* domains, and the component problems could be subsumed as special cases under a "grand unified theory."

Acknowledgement: The support of National Science Foundation grant CCR-0311671 to the University of California at Davis, and a gift from Intel Corporation, are gratefully acknowledged.

8. REFERENCES

[1] *Information Technology—Common Logic (CL)—A Framework for a Family of Logic-Based Languages*, Final Committee Draft ISO/IEC FCD 24707, Reference No. ISO/JTC 1/SC 32N1377 (Dec. 2005); available at http://cl.tamu.edu.

[2] J. Achugbue and F. Chin, "The Effectiveness of Output Modification by Rounding for Protection of Statistical Databases," *INFOR. Canadian Journal of Operational Research and Information Processing* **17** (3) pp. 209–218 (Aug. 1979).

[3] M. Bishop, R. Crawford, B. Bhumiratana, L. Clark, and K. Levitt, "Some Problems in Sanitizing Network Data," *to appear in the Proceedings of the 15th IEEE International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE 2006)* (June 2006).

[4] M. Bishop, B. Bhumiratana, R. Crawford, and K. Levitt, "How to Sanitize Data," *Proceedings of the 13th IEEE International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE 2004)* pp. 217–222 (June 2004).

[5] P. Chu, "Cell Suppression Methodology: the Importance of Suppressing Marginal Totals," *IEEE Transactions on Knowledge and Data Engineering* **9** (4)pp, 513–523 (July 1997).

[6] M. Covington, M. Moyer, and M. Ahamad, "Generalized Role-Based Access Control for Securing Future Applications," *Proceedings of the 23rd National Information Systems Security Conference* pp. 1–10 (Oct. 2000).

[7] H. Delugach and T. Hinke, "Using Conceptual Graphs To Represent Database Inference Security Analysis," *Journal of Computing and Information Technology* **2**(4) pp. 291–307 (Dec. 1994).

[8] D. Denning, "Restricting Queries that Might Lead to Compromise," *Proceedings of the IEEE 1981 Symposium on Security and Privacy* pp. 33–40 (Apr. 1981).

[9] D. Denning, "A Preliminary Note on the Inference Problem in Multilevel Database Management Systems," *Proceedings of the National Computer Security Center Invitational Workshop on Database Security* (June 1986).

[10] D. Denning, P. Denning, and M. Schwartz, "The Tracker: A Threat to Statistical Database Security," *ACM Transactions of Database Systems* **4**(1) pp. 76–96 (Mar. 1979).

[11] I. Dinur and K. Nissim, "Revealing Information While Preserving Privacy," *Proceedings of the 22rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 202–210 (2003).

[12] Sir A. C. Doyle, "Silver Blaze," in *The Annotated Sherlock Holmes*, Volume 2, Clarkson N. Potter, Inc., New York, NY pp. 261–281(1967).

[13] B. Dragovic and J. Crowcroft, "Information Exposure Control through Data Manipulation for Ubiquitous Computing," *Proceedings of the 2004 Workshop on New Security Paradigms* pp. 57–64 (2004).

[14] G. Duncan, T. Jabine, and V. de Wolf, *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*, Panel on Confidentiality and Data Access, National Academy Press, Washington DC 20055 (1993).

[15] G. Duncan, S. Keller-McNulty, and S. Stokes, "Database Security and Confidentiality: Examining Disclosure Risk vs. Data Utility through the R-U Confidentiality Map," Technical Report 142, National Institute of Statistical Sciences, Research Triangle Park, NC (Mar. 2004); available at http://www.niss.org/technicalreports/tr142.pdf

[16] J. Fan, J. Xu, M. Ammar, and S. Moon, "Prefix-Preserving IP Address Anonymization", *Computer Networks* **46** (2), pp. 253-272 (Oct. 2004).

[17] M. Fischetti, J. Salazar, "Solving the Cell Suppression Problem on Tabular Data with Linear Constraints," *Management Science* **47** (7) pp. 1008–1027 (July 2001).

[18] D. Frincke, "Balancing Cooperation and Risk in Intrusion Detection," *ACM Transactions on Information and System Security* **3**(1) pp. 1–29 (Feb. 2000).

[19] K. Kenthapadi, N. Mishra, and K. Nissim, "Simulatable Auditing," *Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* pp. 118–127 (2005).

[20] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-Diversity: Privacy Beyond *k*-Anonymity," *Proceedings of the 22nd IEEE International Conference on Data Engineering* p. 24 (April 2006).

[21] B. Malin, E. Airoldi, S. Edoho-Eket, and Y. Li, "Configurable Security Protocols for Multi-Party Data Analysis with Malicious Participants," *Proceedings of the 21st International Conference on Data Engineering* pp. 533–534 (Apr. 2005).

[22] D. Marks, "Inference in MLS Database Systems," *IEEE Transactions on Knowledge and Data Engineering* **8**(1) pp. 46–55 (Feb. 1996).

[23] D. Marks, L. Binns, and B. Thuraisingham, "Hypersemantic Data Modeling for Inference Analysis," *IFIP Transactions* **A-60**: *Proceedings of the IFIP WG11.3 Working Conference on Database Security VII* pp. 157–180 (Aug. 1994).

[24] A. Meyerson and R. Williams, "On the Complexity of Optimal *K*-Anonymity," *Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 223–228 (2004).

[25] G. Minshall, "Tcpdpriv", Release 1.1.10 (Aug. 1997); available at http://ita.ee.lbl.gov/html/contrib/tcpdpriv.html

[26] M. Peuhkuri, "A Method to Compress and Anonymize Packet Traces", *Proceedings of the First ACM SIGCOMM Workshop on Internet Measurement* pp. 257-260 (Nov. 2001).

[27] J. Sowa, *Conceptual Structures: Information Processing in Mind and Machine*, Addison-Wesley Publishing Co., Reading, MA (1983).

[28] L. Sweeney, "*k*-Anonymity: a Model for Protecting Privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* **10**(5) pp. 557–570 (2002).

[29] L. Sweeney, "Achieving *k*-Anonymity Privacy Protection Using Generalization and Suppression," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* **10**(5) pp. 571–588570 (2002).

[30] L. Sweeney, "Privacy-Enhanced Linking," *ACM SIGKDD Explorations Newsletter* **7**(2) pp. 72–75 (Dec. 2005).

[31] B. Thuraisingham, "The Use of Conceptual Structures for Handling the Inference Problem," *IFIP Transactions* **A-6**: *Results of the IFIP WG 11.3 Workshop on Database Security V: Status and Prospects* pp. 333–362 (1991).

[32] W. Ware, "Records, Computers, and the Rights of Citizens: Report of the Secretary's Advisory Committee on Automated Personal Data Systems," DHEW Publication (OS)73-94, U.S. Dept. of Health, Education and Welfare (July 1973).

[33] S. Zhong, Z. Yang, and R. Wright, "Privacy-Enhancing k-Anonymization of Customer Data," *Proceedings of the 24rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 139–147 (2005).