A Data Privacy Taxonomy

Ken Barker, Mina Askari, Mishtu Banerjee, Kambiz Ghazinour, Brenan Mackas, Maryam Majedi, Sampson Pun, and Adepele Williams

Advanced Database Systems and Applications Laboratory
University of Calgary, Canada
kbarker@ucalgary.ca

Abstract. Privacy has become increasingly important to the database community which is reflected by a noteworthy increase in research papers appearing in the literature. While researchers often assume that their definition of "privacy" is universally held by all readers, this is rarely the case; so many papers addressing key challenges in this domain have actually produced results that do not consider the same problem, even when using similar vocabularies. This paper provides an explicit definition of data privacy suitable for ongoing work in data repositories such as a DBMS or for data mining. The work contributes by briefly providing the larger context for the way privacy is defined legally and legislatively but primarily provides a taxonomy capable of thinking of data privacy technologically. We then demonstrate the taxonomy's utility by illustrating how this perspective makes it possible to understand the important contribution made by researchers to the issue of privacy. The conclusion of this paper is that privacy is indeed multi-faceted so no single current research effort adequately addresses the true breadth of the issues necessary to fully understand the scope of this important issue.

1 Introduction

Owners of data repositories collect information from various data suppliers and store it for various purposes. Unfortunately, once the supplier releases their data to the collector they must rely upon the collector to use it for the purposes for which it was provided. Most modern database management systems (DBMS) do not consider privacy as a first-order feature of their systems nor is privacy an explicit characteristic of the underlying data model upon which these systems are built. Since privacy is poorly defined technically in the literature, even when a DBMS vendor builds privacy features into their systems, they must do so using their own understanding of what constitutes privacy protection. Research efforts that address an aspect of privacy often fail to formally define what they mean by privacy or only consider a subset of the many dimensions of privacy so they often miss critical issues that should be addressed to fully understand the privacy implications.

This paper specifically addresses the key "definitional" problem. We identify four key technical dimensions to privacy that provide the most complete definition to appear to date in the literature. The key players when considering privacy issues in any data repository environment include the data provider, the data collector, the data users, and the data repository itself.

A.P. Sexton (Ed.): BNCOD 2009, LNCS 5588, pp. 42–54, 2009.
© Springer-Verlag Berlin Heidelberg 2009

Our key contributions are: (1) A clear description of the key components that make up the privacy research space with respect to data repositories. (2) An easily understandable taxonomy of these features and how they relate to each other. (3) A demonstration the taxonomy's expressiveness in both the real-world and academic literature. (4) A tool for researchers who need to clearly define where in the total space their work contributes and what other work is more comparable to their own. (5) A tool for graduate students to understand the scope of this domain and to be able classify work appearing in the literature.

Before describing our privacy taxonomy it is important to provide some terminology that is either new or used in different ways elsewhere in the literature. The *provider* is the individual or organization providing data that is to be stored or used. The provider may or may not be the owner of the data (see Section 2.2). The *collector* is the individual or organization that initially collects, uses, or stores data received from the provider. The collector is the individual or organization that ultimately solicits the data from the provider even if this is through another party who actually acquires the data (i.e. via outsourcing). A *third-party* is any individual or organization that acquires the provided data from the collector. (Note that there is no distinction made between authorized or unauthorized data release to a third-party.)

2 Defining Privacy

Article 8 of the Hippocratic Oath states: "And about whatever I may see or hear in treatment, in the life of human beings – things that should not ever be blurted out outside – I will remain silent, holding such things to be unutterable." This well-known standard is the corner stone of privacy standards upon which doctor-patient confidentiality is established. Unfortunately, modern database technology is far from being a subscriber to such a high standard and, in fact, much of it has actively engaged in violating Article 8 in an attempt to expose or discover information that is deliberately being withheld by data providers.

The U.S. Privacy Act (1974) articulates six principles including: ensuring an individual can determine what data is collected; guarantee that data collected is only used for the purpose for which it is collected; provides access to your own data; and the information is current, correct and only used for legal purposes [1]. The Act provides explicit clauses to collect damages if privacy is violated but it also permits explicit statutory exemption if there is an important public policy need. The Organisation for Economic Cooperation and Development (OECD) have developed a list of eight principles including: limited collection, data quality assurance, purpose specification, limited use, security safeguards, openness, individual participation and accounatiblity to guide providers and collectors to undertake best practices. Several countries have adopted these principles either explicitly in law or through various ethical codes adopted by practitioners that utilize private data [3].

These have been collected into a set of ten principles by Agrawal *et al.* [2] in their development of the seminal work on Hippocratic databases. The ten principles mirror those indicated by the governmental definitions but are described so they can be operationalized as design principles for a privacy-aware database. The principles are: (1) a

requirement to specify the purpose; (2) acquiring explicit consent; (3) limiting collection to only required data; (4) limiting use to only that specified; (5) limiting disclosure of data as much as possible; (6) retention of the data is limited and defined; (7) the data is accurate; (8) security safeguards are assured; (9) the data is open to the provider; and (10) the provider can verify compliance to these principles.

This paper's first contribution is to identify four dimensions that collectively operationalize the abstract privacy definitions above into a tool that can be used to categorize and classify privacy research appearing in the literature. Some aspects, such as the requirement for legal recourse, are beyond the scope of a technical paper because they belong rightly in the venue of legal and/or legislative domains. However, any privacy-aware system must be cognizant of the legal and legislative requirements to ensure systems can conform to and facilitate these transparently. Thus, although legal recourse rightfully belongs in a different domain, it is reflected technologically as a requirement to protect logs and provide auditability.

The four dimensions identified in this work are *purpose*, *visibility*, *graunlarity*, and *retention*; to which we now turn.

2.1 Purpose

Providers have various motivations for providing data to a collector. A patient may provide highly personal information to a health care provider to ensure that they receive appropriate medical care. If that patient is a student, then some aspect of the same information may be provided to a teacher to ensure that a missed exam is not unfairly penalized. The provision of such data is for very different purposes so the provider is likely to want to provide different levels of detail (we will return to this aspect in Section 2.3). However, in both scenarios the student/patient is releasing the information for a very specific purpose so it is critical that a privacy-aware system explicitly record and track the purpose for which a data item is collected.

Figure 1 depicts the *purpose* dimension of our privacy definition along the x-axis. There are a number of discrete points along this dimension that represent increasingly general purposes as you travel along the axis. The origin of the purpose axis could be

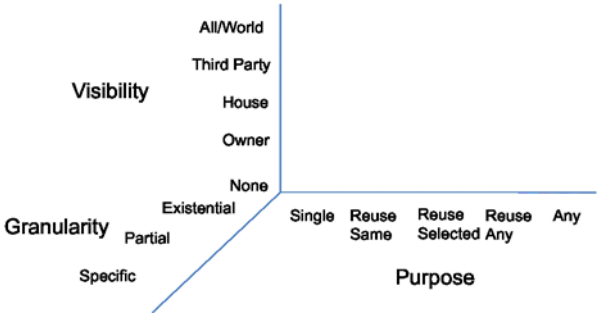

Fig. 1. Key Contributors to Data Privacy in a Data Repository

thought of as a provider providing data for no purpose whatsoever. This point is vacuous on this axis because it would mean that the data could not be used by the collector under any circumstances, for any reason, so there would be no need to collect it. In fact, we would argue that if the collector conforms to the spirit of the various acts and principles summarized in Section 2, they should refuse to collect any data that does not have an explicit purpose identified. Furthermore, a privacy-preserving database system should not permit the inclusion of any data that does not explicitly define the purpose of any data item to be collected or stored.

The first point beyond the origin is the most restrictive and explicit statement of data collection. This point identifies data that a provider supplies for an explicit *single* purpose and for a single use. The data can only be used for this one time single purpose and any other use would be considered a violation of privacy. The next point (labeled *Reuse Same*) allows the collector to reuse the data for the same purpose more than once so it represents a relaxation of privacy. Reuse of data can also permit a data item to be used for a selected set of explicitly defined other purposes (*Reuse Selected*). This scenario might occur when a patient provides medical information to a health-care provider for treatment and for the purpose of collecting payment from an insurance company. The next point on the x-axis (Reuse Selected) is intended to capture the case where the data collected can be used for purposes related to the primary purpose for which the data was provided. Thus, the *Reuse Any* point could be used when a patient is willing to let their data be used for un-foreseeable related purposes such as a disease research study or to monitor efficiency in the hospital itself to improve the hospital's operations. The final point on the x-axis captures the case where a provider does not care how the data might be used. The *Any* point is the least privacy protective point and probably should be discouraged as a matter of normal data collection. However, there are scenarios that could be envisioned where the data is either of such low value or the information already exists in the public domain, where this might be desirable.

2.2 Visibility

Visibility refers to who is permitted to access or use data provided by a provider. This dimension is depicted in Figure 1 on the y-axis. Once again the origin represents the case where the data should not be visible to anyone. As with the purpose dimension, this point should be a clear indication that the data should not be collected or retained because it should not be made available so there is no reason to acquire the data.

The first point on the y-axis introduces important new terminology as we use the term *Owner* to indicate that only the "owner" has access to the data provided. We have deliberately used the term *provider* to this point because the literature is extremely conflicted over who "owns" data in a data repository. This is a critical semantic because two possible definitions could be imposed:

1. Provider is owner: if the provider retains ownership of their private information then this point would indicate that only the provider would have access to their data at this point.
2. Data repository is owner: if the data repository becomes the owner of the data stored on the system then subsequent use of it (i.e. permission to access by others) will be that of the repository.

OECD, US Privacy Act (1974), and the Hippocratic Oath are silent on this critical issue largely because they were developed in a time when it was more difficult to quickly transmit private information in such a way that its original ownership could become obscured. Most data collectors consider the data stored in their systems to be their property and this has consequences on modern high speed networks. Thus, we argue that the ownership of data should be retained by the original provider no matter how many transitive hands it may have passed through. The implication of this on our visibility dimension is that data stored in a repository should only be accessible by the data provider at the *owner* point on the y-axis.

The next point on this dimension permits the repository to access the data provided. We have introduced the term *House* indicating that the the data repository housing the data can view the data provided. The current psyche of most data collectors is that they can view data provided to them by default. However, since storage is now considered a "service", and is often sold as such, it is critical that this distinction be made explicit.

Two examples will help illustrate the need for the distinction between the owner and the house. Google™ [1] is known to utilize data provided to it so they can provide a "value add" for their users. Conversely, iomega™ [2] offers a product called iStorage™that permits online storage of subscriber data on their servers[3]. Google™ would be placed at the house point on the y-axis since they will utilize and access the data as they see fit because they have not made any promises to the provider of how their data might be used. In effect, they are behaving like an "owner" in the traditional sense. However, iomega™offers a service to securely protect a providers data in a convenient way so they have made, at least an implicit promise, to only allow the provider (as owner) to access their data. In short, Google™visibility is the house while iomega istorage™should only be visible to the owner.

Third-parties are data users that are authorized to do so by the house. Typically, such access is permitted under certain conditions and the third-party is required to conform to some kind of an agreement before access is granted. This might occur when the house wants to "outsource" a portion of their data analysis or by virtue of an explicit agreement to share collected data when the provider initially agreed to provide their information. Thus, third-parties are those that have access to data provided to the house but are covered by an explicit agreement with the house.

The final point on the y-axis is the *All/World* extreme where the data is offered to anyone with access to the data repository. Web search tools are the best known examples of such systems in that their fundamental business model is to provide data to those posing queries on their data repositories. This represents the least amount of privacy protection so it is depicted at the extreme point on the y-axis.

2.3 Granularity

The *granularity* dimension is intended to capture characteristics of data that could be used to facilitate appropriate use of the data when there could exist multiple valid

[1] http://www.google.ca/

[2] http://www.iomega.com/

[3] iomega™ is only one of many such storage/backup providers widely available currently (see http://www.iomega.com/istorage/).

accesses for different purposes. For example, a medical care-giver could require very specific information about a particular condition while an insurance company may only need a summary statement of the costs of the patients care. Once again we begin at the origin where no information is to be made available of any kind. We will return to why the origin is needed for each of these dimensions when we consider how to use the taxonomy shortly but the same caveat should be applied to this point on the granularity dimension as on the others where this could suggest that the data should not be collected or stored at all.

The first non-origin point is necessary to capture scenarios where privacy could be compromised by undertaking a set of existence tests. Thus, a query could be written to probe a data repository that may not return any actual data value but a response could indicate the existence of data and this may be sufficient to reveal private data. This actually occurs in a number of different ways so we only mention two here to justify the data point. A query which asks for the "count" of the number of occurrences of a particular data item may not provide the questioner with a value for an attribute but if the system returns a "0" response, information is clearly revealed. A sequence of such queries can narrow down the search space to a unique value without actually releasing private information. This may also occur in more complicated scenarios. For example, by posing queries that include generating the difference between two similar queries could result in information being leaked by virtue of returning results for known data values (i.e. individuals may permit their data to revealed) and subtracting these from the full set, thereby revealing information about the existence of data of a particular type. Techniques that perform output filtering as a privacy protection mechanism may be susceptible to such attacks. There are many other possible existence tests that may unwittingly reveal private information so there is a non-zero chance of privacy compromise when these kinds of queries are permitted. We have identified this point on the z-axis as *existential* because existence testing can be done in some way.

The extreme point on the granularity dimension occurs when *specific* data is released by the data repository. This is actually the most common scenario in that most queries will operate on the actual data stored in the data repository. A query for the data item returns its actual value so the specific accurate data is released in an unobscured way. This point clearly depicts the least amount of privacy protection.

The final point considered on this dimension also provides the best intuiton for the name *"granularity"*. Many systems attempt to protect data privacy by altering it in some non-destructive way so privacy is preserved but a value is still useful to the query poser. Several techniques including aggregation or summarization and categorizing have been used to protect data in some way. Classic examples of this have appeared in the literature such as changing "age" to a category such as "middle age" or an actual income value to "high income." Thus, the system is providing *partial* information but not specific data. We have grouped these techniques under a single heading because all of them, with respect to privacy, attempt to obscure the data in some way to make them resistant to various kinds of attacks. The taxonomy captures the nature of the protection provided by grouping related approaches into classes that ultimately exhibit similar privacy characteristics, although they may be implemented in very different ways. For example, the data ordinalization represented by the age example is in the same class (i.e. partial)

as one that provides partial information using anonymization of quasi-keys [5] or its varieties. Many techniques have been developedto break down such methods including statistical attacks, inference, or query attacks all aimed at improving the confidence the attacker has about the actual value stored. In fact, the entire field of data mining is specifically intended to extract information from a data repository when the provider has specifically not included that information to the collector. Obviously, these techniques can (and will) be applied whenever partial results are applied so the point provides less protection than existential testing but more than simply giving the querier the specific value.

2.4 Retention

The privacy characteristics identified by the OECD and the 1974 US Privacy Act implicitly require that data be collected for a particular purpose and remain current. These documents imply that the data should be removed after it has been used for its intended purpose or it has become dated. Thus, a privacy definition must also include an explicit statement of its *retention* period so any privacy model developed must explicitly indicate for all data collected how long it can be used under any circumstances. Data that has passed its "best before" date must be deleted or the retention period must be refreshed by a renewed agreement with the owner.

2.5 Summary and Nomenclature

In the interest of brevity, a point on the taxonomy is described by a tuple $Pr =< p, v, g, r >$ where $p \in \{none, single, reusesame, reuseselected, reuseany, all\}$; $v \in \{none, owner, house, third-party, all-world\}$; $g \in \{none, existential, partial, specific\}$; and $r \in \{\infty, < date >\}$ where $< date >$ is an explicit date indicating when this data is to removed and ∞ indicates no epiration is specified. Thus, a privacy statement such as: $Pr_1 =< all, third-party, partial, \infty >$ would describe a privacy system that permitted third party(s) (visibility) to access modified non-specific data (i.e. partial) (granularity) for any purpose and there would be no requirement to remove this data from the repository.

3 Applying the Privacy Definition

This taxonomy provides the most complete definition of privacy to appear in the literature to date. The lack of a complete definition has meant researchers have generally provided a very brief, often intuitive, definition for privacy and then used the term liberally in their work as though it was a complete definition. Although many useful contributions have been made recently, the lack of a generally accepted definition that can be used in developing privacy research has undoubtedly slowed progress as researchers often "talk past" each other. The dimensional taxonomy provided here addresses this gap by providing a framework that can be used by researchers to place their work relative to an absolute definition (Section 3.1 provides some examples of how to use this framework). It can also help researchers understand how their work can legitimately be compared to other work thereby avoiding the miscommunication that often happens when a term carries different connotations (see Section 3.2).

3.1 Examples of Possible Taxonomizations

Figure 1 can be used in several ways to precisely define the aspect of privacy that is of interest to a particular research project or system. Before describing where on this taxonomy other work should be placed, we first illustrate how the taxonomy can be used in a more abstract way by providing a few examples.

Example 1 Online Storage as a Service: A relatively recent business model has developed where a company provides online storage to clients so they can access the data from anywhere. This is a valuable service as it provides access globally but there is a clear implication, at least on the provider's part, that the data is stored by the company for the provider's exclusive use[4]. The best "real life" analogy would be a safety deposit box. Thus, this scenario explicitly describes a point on all three dimensions where the visibility is only provided to the owner (y-axis), for the repeatable unique purpose (reuse same on the x-axis) that returns the specific data (z-axis) to the owner when requested. Furthermore, there is an implication that the data should only be retained for one year as this is the length of time that the provider agreed to pay for this service (we assume that the current year is 2008 for this example). Thus, this point on our taxonomy would be: $Pr_2 =< reusesame, owner, specific, 2009 >$ and represents the most explicit description of privacy in that it is a point (see Figure 2).

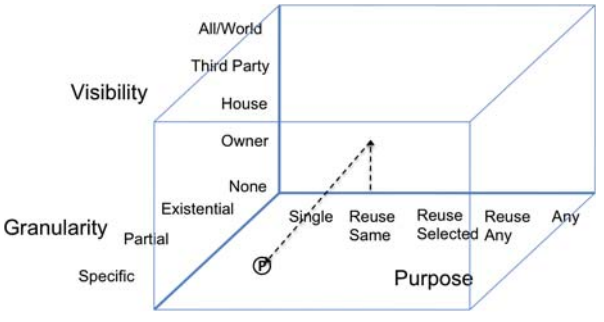

Fig. 2. Privacy point indicating expecation for online storage for clients

The degree of freedom is most restricted when all axes are specified explicitly by a privacy definition. If only one non-retention dimension is unspecified, a line is described in our 3-dimensional taxonomy. To illustrated this we consider two further examples. The first does not specify any aspect of one dimension while the second only partially specifies one of the dimensions and we use these to illustrate the flexibility/utility of our taxonomy.

Example 2 Facebook Social Network: Facebook[5] and other similar forms of social network provide users with a mechanism to present "individual" data that is largely

[4] This is similar to iomega iStorage™ mention above.
[5] http://www.facebook.com/

accessible openly. The implied purpose for such social networks from the provider's perspective is to provide a public interface about themselves and as such the provider permits any such posted data to be revealed for "any" purpose so the *reuse any* point on the x-axis best describes its purpose. The data provided is not obscured in anyway so it is clearly *specific* in terms of its granularity. However, the provider is able to control to some varying extend how visible this data is made. For example, a group of "friends" can be defined that provides increased access restrictions to the specific data stored. Thus, the privacy description on the visibility dimension is not a specific point but rather something that must be defined by the provider so it represents a range of possible values on the y-axis but is restricted to the line defined by the x and z axes as depicted in Figure 3 (labeled (f)). Thus: $Pr_3 =< reuseany, \phi, specific, \infty >$.

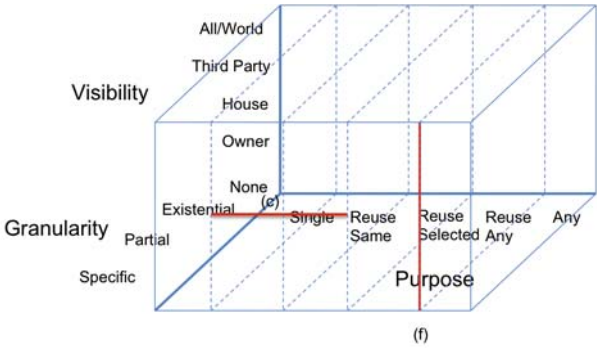

Fig. 3. Privacy point indicating expecation for online storage for clients

We will use ϕ to indicate when a dimension is unspecified or irrelevant for a particular dimension. Recall that our goal is to develop a method to properly capture privacy definitions and we will see in Section 3.2 that some authors do not explicit define a particular dimension or view as it is irrelevant to their research agenda.

Example 3 Online Credit Card Transactions: A primary payment mechanism for goods and services on the Internet is via credit cards. This is also probably the single greatest risk associated with online commerce from a user's perspective so it is clear that this information must be kept private. However, there is no way to meaningfully obscure or "roll-up" a credit card number so only specific data can be used meaningfully in a transaction. The online vendor is the house in this scenario so the visibility is clearly at this level. An argument could be made that a credit-card broker or some trusted-party could act as an intermediary between the vendor and the purchaser but this simply means that the broker is now acting as the house in that it is the holder of the private information. Such trusted third-parties clearly can increase security pragmatically; but from a definitional viewpoint, which is our purpose here, it really only changes who the "house" is but does not change the definitional relationship. In short, the house must have specific information to facilitate online transactions. The purpose may vary however. A credit card could be used for the purpose of a single purchase, it could be used

for the same purpose repeatedly (eg. making monthly payments automatically); or for various selected purposes such as purchasing books a intermittent times into the future. The final point to consider is retention. It is easy to see that if the card is being used for a single transaction, then it should only be retained for this single purchase so the retention period should be "now". However, if it is to be retained for a recurring purpose, then it must be retained until the commitment period expires. Thus, the privacy description on the visibility dimension is not a specific point but rather something that must be defined by the provider so it represents a range of possible values on the x-axis but is restricted to the line defined by the y and z axes as depicted in Figure 3 (labeled (c)). Thus: $Pr_4 =< \{\text{single}, \text{reusesame}, \text{reuseselected}\}, house, specific, \{now, \infty\} >$.

We will use set notation to indicate when only a subset of a particular dimension is defined by the system. This will provide us with substantial flexibility to correctly describe systems and research efforts quite accurately either diagrammatically or with our tuple notation.

When only one of the dimensions is defined by a system, there are two degrees of freedom so a plane is described. To illustrate this we now turn to an all too common modern scenario.

Example 4 Online Data Collection for a Service: Many internet vendors often offer either software or a service in exchange for information about yourself. This may or may not involve also paying for that software or service in addition to the data collection. The privacy agreement typically indicates how this data can be used by the collector and usually limits its use to the company or its affiliates but this is not necessarily the case. Furthermore, for legal reasons the collector may need to make the data collected available to the provider as well to ensure its correctness. Thus, the visibility is essentially undefined or unconstrained. The collector will use this data in many different ways including using the specific data or by aggregating it in various ways so the granularity is often also undefined. They generally limit the purpose to only ones specified by them but they rarely limit how they can use it once it is collected so we somewhat generously identify the point on the x-axis as "reuse any". The retention period, if it is defined at all, is unlimited so this scenario can be defined as: $Pr_5 =< reuseany, \phi, \phi, \infty >$.

Figure 4 illustrates this scenario and demonstrate the lack of specificity and increased degree of freedom indicates that user has a much greater privacy exposure than a system such as the one depicted in Figure 2.

3.2 Placing Literature on the Taxonomy

The taxonomy helps us understand privacy pragmatically in various scenarios such as those described above. It is clear that the framework is able to capture several different scenarios and it highlights the orthogonality among the dimensions described. However, we also feel that the taxonomy is very valuable in classifying research that has appeared in the literature. It also forces us to be more precise about these contributions and leads to a deeper understanding of the strengths and limitations of a piece of work. We have tested the taxonomy's utility by classifying much of the literature on privacy using it.

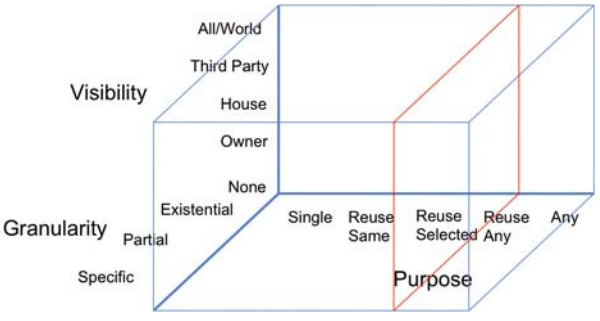

Fig. 4. Privacy point indicating expecation for online storage for clients

Table 1. Applying Privacy Taxonomy to Various Research Systems

Reference	Purpose	Visibility	Granularity
k_anonimity [5]	{RSpec, Rany, any, ϕ}	3^{rd} Party	Partial
Domiongo-Ferrer [4]	ϕ	Provider $\rightarrow$ Owner, House $\rightarrow$ respondant, 3^{rd} party $\rightarrow$ user	ϕ
Hippocratic [2]	Reuse Same <purpose/data item>	House $\rightarrow$ Access based on G and P	Roll up to facilitate privacy

However, we will limit our discussion here to three contributions that are particularly interesting or represent a seminal contribution to privacy in the data repository area and because each highlights a particular feature of the taxonomy that we believe is valuable. We understand that there are many excellent contributions that could have been included but space limitations prevent a full treatment. The reader is encouraged to test the taxonomy by placing their favourite contribution on it to convince themselves of the utility of the taxonomy. Furthermore, the purpose of this section is not to comment on the "value" of the contributions made by these researchers but rather to frame it. We feel that the greatest challenge in understanding and determining how to make future contributions in this area is the need to provide a clear statement of where these contributions are and this requires a clear understanding of the many dimensions identified in this paper. Table 1 provides a summary of some key contributions appearing in the literature and we briefly provide a rationale for this classification.[6]

The **Hippocratic database** [2] proposal is arguably the first that explicitly addresses the need to manage data privacy within a data repository. We do not consider the utility of the model described by the research but rather focus only on those aspects of privacy are explicitly identified by the work.Agrawal *et al.* [2] use the classical definition found in medical domains where data collected can be reused for the purpose for which it is collected. Thus, this point is identified on the x-axis as "reuse same" in our taxonomy.

[6] In all cases, the retention value is undefined (i.e. ϕ) so this column is omitted from the table.

The "House" can view, and may be able to grant access to others, based on a predefined granularity and purpose. Thus, it appears that this point can best be defined as the "House", though there is a strong implication that the "House" become the owner; so it can subsequently grant access to third-parties but this is not made explicit in their initial paper. Granularity is less explicitly defined in the work but it appears that the work suggests that a "roll-up" or "aggregation" mechanism could be used to further facilitate privacy while allowing for "provider" privacy that permits "collector" utility. Finally, retention is not explicitly mentioned in the data model though the paper indicates that this is a critical aspect of privacy protection. However, the lack of retention's explicit inclusion in the model means that we must assume that for all practical purposes, the data retention period is undefined: $Pr_H =<$reuse same, House, Partial,$\phi >$.

k-anonimity [5] is specifically developed to protect privacy by eliminating any form of identifying "keys" that may exist in the data. The obvious requirement to eliminate identifying information is assumed to have occurred so any primary or candidate key data is eliminated. However, the remaining data (or *quasi-keys*) may be sufficient to uniquely identify individuals with respect to some of their attributes so k-anonimity has been developed to ensure that the characteristics of the accessible data protects "provider" privacy. The assumption is that the data being protected with k-anonomity is either undefined or can be used for any purpose including those not necessarily anticipated by either provider or collector. This is in fact the primary motivation for the work on k-anonimity. The second motivation for this data "cleaning"[7] for the purpose of privacy is so it can be safely released to "third-parties" for subsequent analysis. Thus, the visibility assumed by this work is easily identified as 3^{rd}-party. This approach is really intended to provide a form of pre-computed aggregation so the "partial" point best describes the granularity dimension. The retention dimension is not discussed at all but the work implicitly assumes that the data will be retained by both the house and third-partiies indefinitely. Thus: $Pr_k =<$any, 3^{rd}-party, Partial,$\phi >$.

Domiongo-Ferrari [4] makes a valuable contribution by describing a "three-dimensional" framework that is an abstraction of our visibility dimension. The work specifically identifies privacy issues as they relate to our provider, house, and third-party. Thus, our work includes the definitional aspect of Domingo-Ferrari so it is included within our taxonomy as illustrated in Table 1. The contribution provides additional insight along this dimension but also demonstrates the unidimensional privacy definitions often assumed by researches when addressing a specific aspect of the issue. This results in the absence, in this particular case, of any issues associated with purpose, granularity, or retention.

4 Conclusions and Ongoing Work

The abstract claimed, and the paper shows, that the unavoidable conclusion of this paper is that privacy is indeed multifaceted so no single current research effort adequately addresses the true breadth of the issues to fully understand the scope of this important issue. The paper provides a thorough discussion of the privacy research space and

[7] "anonymity" might be a more accurate term but this can only be said to be as strong as "k-anonymous" so we use the weaker term "cleaning" imprecisely and with apologies.

demonstrates that many of the issues are orthogonal. The first contribution is an explicit characterization of the issues involved and their relationship. The paper demonstrates its applicability in a number of different "real-world" scenarios before using the taxonomy to understand the place a few key contributions. A critical value in the paper is a tool that can be used to frame future research in the area. Researchers can use it to identify precisely what aspect of privacy they are considering in their work and to more acurately compare their contributions to other work in the field. Some of the points in the taxonomy may not co-occur as they contradict one another. However, we argue that the importance is it produces a better understand of where contributions fit within the breadth of ongoing privacy research. Furthermore, we have demonstrated that the taxonomy is widely applicable to much of the current literature and have tested it against a wide-range of the literature currently available.

To create a complete system, capable of protecting user privacy, requires that all identified aspects be considered. This paper provides the most complete definition to appear in the literature to date and should be used in developing and using modern data repositories interested in protecting privacy. Fundamental to realizing this vision is the need to develop a data model that has privacy as first-order feature so we intend to build the aspects of this taxonomy into a new privacy-aware data model. In tandem with this formal underpinning we are also working to implement the model into the DBMS architecture at all levels from the query processor through to the cache management system. Thus, we are developing a novel data directory that incorporates this data model and extending SQL (and an XML derivative) to incorporate these privacy features into the query language.

References

1. The privacy act of 1974, September 26 (2003) (1974)
2. Agrawal, R., Kiernan, J., Srikant, R., Xu, Y.: Hippocratic databases. In: VLDB 2002: Proceedings of the 28th International Conference on Very Large Databases, VLDB Endowment, Hong Kong, China, vol. 28, pp. 143–154 (2002)
3. Bennett, C.J.: Regulating Privacy: Data Protection and Public Policy in Europe and the United States. Cornell University Press (April 1992)
4. Domiongo-Ferrer, J.: A three-dimensional conceptual framework for database privacy. In: Jonker, W., Petković, M. (eds.) SDM 2007. LNCS, vol. 4721, pp. 193–202. Springer, Heidelberg (2007)
5. Sweeney, L.: k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557–570 (2002)