

# How to Calculate the Information Privacy

Sabah S. Al-Fedaghi  
Department of Computer Engineering  
Kuwait University  
[sabah@eng.kuniv.edu.kw](mailto:sabah@eng.kuniv.edu.kw)

## Abstract

*This paper introduces a new theoretical formalism to specify private information. It utilizes single-referent linguistic assertions in defining 'private information' in terms of 'atomicity' and identification. Atomicity refers to the private information of a single person, in contrast to referring to compound information (e.g., X and Y are in love). Compound private information is shown to be reducible to atomic ones (X and someone are in love, and Y and someone are in Love). The purpose of this reduction is to isolate "centers" of privacy for each person who is the proprietor of the information. We show that this discretion of information privacy enables us to perform set theory operations on pieces of private information about a person in order to calculate the relative amount of his/her information privacy. Additionally, the new definition of private information can be utilized in several privacy-related areas such as information ethics and databases.*

**Keywords:** Privacy, Private Information, Personal Information, Sensitivity, Private Information Ethics.

## 1. Introduction

Several types of privacy have been distinguished in the literature including 'physical privacy', privacy of personal behavior, privacy of personal communications and privacy of personal data [6][10]. Information privacy refers to information proximity. According to Clarke, it is "the interest an individual has in controlling, or at least significantly influencing, the handling of data about themselves" [6]. It involves personal (private) information such as credit data, medical and government records, etc. 'Personal information', is said, to denote information about identifiable individuals in accessible form [16]. It means "any information concerning a natural person which, because of name, number, symbol, mark, or other identifier, can be used to identify that natural person" [7]. The Canadian privacy legislation, *Personal Information Protection and Electronic Documents Act* (PIPEDA) defines *personal information* as "information about an identifiable individual" [13]. There are many such 'definitions' for personal or private information. They are usually closely linked with the notions of identification and

de-identification. According to the U.S. Health Insurance Portability and Accountability Act of 1996 [12], "Individually Identifiable Health Information" refers to "health information that identifies the individual or can reasonably be used to identify the individual." Under the HIPAA Privacy Rule, one aspect of "de-identification" is that the health data not include eighteen identifiers of persons which could be used alone or in combination with other information to identify the subject. These identifiers include: names, telephone numbers, fax numbers, email addresses, social security numbers, URLs, etc. The EU Data Protection Directive [9] extensively uses the terms "person-identification" and "identifiable/non-identifiable data". The P3P Specification Working Group specifies "identified data" is "information in a record or profile that can reasonably be tied to an individual." [3] These descriptions of 'private information' are very general. We claim that our new formalism provides a systematic foundation for defining private information and its related notions. In our discussion, we prefer to use the name 'private information' rather than 'personal information' because the second term implies 'ownership', as in 'personal property' while 'private information' has different 'connotations' as will be described in details later.

In the research field, we also observe that informal definitions of private information are utilized. We mention here a representative sample of works related to private information. Sweeney's pioneering work [14] on anonymizing private information aims at detecting information that can personally identify any person. The purpose is removing personal identifying information from the text, so that the integrity of the information remains intact even though the identity of the persons remains confidential. Sweeney also introduced an algorithm and software program called 'Scrub Extractor' that automatically extracts names, addresses, and other identifying information from the free text documents. Other authors in the area of medical textual information have worked on morpho-syntactic aspects of the 'term formation' in medical language [5]. Taira et al. [15] presented a methodology that manually tags all references to patient identifiers and context information. The scheme searches for logical relations that are characterized by a predicate and an ordered list of one or more arguments. It examines candidate words and their potentiality for taking on the role

of the PATIENT within the logical relations, and hypothesizes associations between words (e.g., admitted) and the semantic constraints on concepts that can fill their roles (e.g., patient names). “Personal information” in these works may be understood as referring to a whole text, the paragraph, a sentence, a phrase that embeds identity or may be only a word that denotes the identity. While an identifier can be pinpointed, ‘private information’ seems to be an ambiguous textual context of this identifier. Furthermore, the interpretation of the term ‘personal information’ is problematic. Marx [17], for example, thinks of ‘information about persons’ as involving ‘concentric circles’ of five types of information: individual information (e.g., owning a four door car), private information (e.g., an unlisted phone number), intimate information (e.g., selectively revealed information), sensitive information (e.g., strategic information in a conflict situation), and unique and core identity information (e.g., a unique identity in being attached to an “embodied” individual).

Another important area of research in this direction is the notion of ‘k-anonymity’ [14]. The k-anonymization of a relational table assumes that a table with a prime key that refers to a person, is the personal information. Its main concern is anonymizing entries in the table in order to block any attempt to reach “identifiability” that stems from these entries. Systems that use such techniques aim at protecting individual identifiable information and simultaneously maintaining the entity relationship in the original data. The table structure that includes the name of the relation, the attributes, and the tuple’s values, somewhat fixes the meaning of ‘private information’ in these works. The definition of “private information” seems to be understood through association of the attributes with the identifying key of the relation. Thus the tuple (23, 10:00, John, diabetes, Jim) in the HOSPITAL schema, APPOINTMENT (APPT-NO, APPT-TIME, PATIENT, CLINIC-TYPE, DOCTOR) includes the private information: *John’s medical appointment is about diabetes*. It also includes the private information: *Dr. Jim has a patient’s appointment at 10:00*. The last assertion may be claimed to be public information, not private information. How about the assertion, *John has an appointment with the diabetes doctor Jim*? What is ‘private’ about such information? Clearly, even in this limited relational table, ‘private information’ is also defined informally.

This paper introduces a systematic approach to ‘private information’. Systematization here means concentrating on a well-defined characterization (atomicity and identity) of privacy information and applying the resultant definition to privacy-related notions such as anonymization, confidentiality, database privacy rules, etc. The next section introduces this new definition. In section 3, we fix the relationship between persons and their private information in terms of a type of ‘ownership’ called ‘proprietaryship’. In section 4, we propose deferring the notion of sensitivity to

further study since there are progressive levels of sensitivity that can be applied to private information. In section 5, we concentrate on ‘compound private information’ that refers to more than one person. Section 6 introduces categories of atomic private information of a certain person according to ‘ownership’ and to possessing it. Section 7 applies the mechanism of calculation to private information. In section 8, we demonstrate some applications of the new approach.

## 2. Information Privacy

The classic treatment of an assertion (a judgment), divides it into two concepts: subject (referent) and predicate that form a logical relation. We take an approach such that a linguistic ‘assertion’ may or may not have a well formed internal structure. *Newton is a genius*, *Newton genius*, *genius Newton*, *Newton genius is*, *Newton is x*, and *y Newton x*, are, according to our conceptualization, assertions as long as something is said about a person. The ‘assertion’ should have a recognizable referent. Thus, *xyx xxx zzzz* is nonsense, while *xyx xxx Newton zzzz* is an assertion because we can recognize ‘Newton’ in the string of data. The important thing is that *xyx xxx Newton zzzz* says something about the identified person, Newton. Eventually, for us, even a linguistic expression with one word such as *Newton* is an assertion in which the non-referent part is empty.

Our theory includes a universal set of private information agents,  $\mathbf{Z} = \mathbf{V} \cup \mathbf{N}$ , of two fundamental types of entities: *Individual* (persons) and *Nonindividual*. *Individual* represents the set of individuals  $\mathbf{V}$  of natural persons and *Nonindividual* represents the set of non-individuals  $\mathbf{N}$  (e.g., *company*, *government agency*, etc.) in  $\mathbf{Z}$ .

**Definition:** Linguistic assertions are categorized according to the number of their referents as follows:

- (i) *Zero* (privacy) assertion: an assertion that has no referent that signifies a single individual  $V \in \mathbf{V}$ .
- (ii) *Atomic* assertion: an assertion that has a single referent that signifies a single individual  $V \in \mathbf{V}$ .
- (iii) *Compound* assertion: an assertion that has more than one referent that signifies individuals in  $\mathbf{V}$ .

“Atomic” in this definition refers to the ‘subject’ of the statement and not the composition of the statement that expresses that fact. Thus, *V1 is tall and handsome*, *V1 is tall*, and *V1 is handsome* are all atomic pieces of information, even though the first contains structurally the second and third statements. Linguistically, referents of type *Individual* include the typical proper names, personal pronouns, and definite noun phrases (e.g., *The tall man over there*). A single referent does not necessarily imply a single occurrence of a referent. Thus, *The man wounded himself* has one referent. Notice that in general, any

information identifying an individual such as a number on a piece of paper, is atomic, if it uniquely identifies that individual. However, without loss of generality, we use, in our examples, the predicate-form of assertions.

**Examples:** *John is shy* and *John is in love* are examples of atomic (private) assertions because each has one (identifiable) referent. On the other hand, *spare part ax123 is in store 5*, is a zero (privacy) assertion because it does not involve any individual (human). *They are in love* is a compound (private) assertion because it has two referents. Any atomic assertion has two components: referent-part and (may be empty) zero-part. For example in *John is shy*, ‘*John*’ is the referent-part and the predicate *is-shy* is the zero-part.

An atomic assertion is said be atomic private information assertion of  $V \in \mathbf{V}$ , if  $V$  is the sole identifiable individual in  $\mathbf{V}$ . Compound information is said be compound private information of  $\mathbf{V1} \subseteq \mathbf{V}$  if  $\mathbf{V1}$  is a set of identifiable individuals. For simplicity’s sake, we may refer to ‘atomic/compound private information assertions’ as simply ‘private information’. Information is assumed to be a true assertion, hence, the two basic ingredients of private information are: identification and truth. The issue of false assertions in this conceptualization of information privacy is a very interesting aspect that will not be discussed in this paper.

### 3. Possession and Proprietorship of Private Information

Private information is also related to its possessor. A single piece of atomic private information may have many possessors; where its proprietor (i.e., the referent in the private information) may or may not be among them. A *possessor* refers to any entity in  $\mathbf{Z}$  that knows, stores or owns the information. Individuals in  $\mathbf{V}$  can have private information of other individuals; and companies and government agencies in  $\mathbf{N}$  can possess a great deal of private information about individuals.

We call the relationship between individuals and their own atomic private information *proprietorship*. Proprietorship of private information is different from the concepts of possession, ownership, and copyrighting. If  $p$  is a piece of atomic private information of  $V \in \mathbf{V}$ ; i.e.,  $V$  is its referent, then  $p$  is *proprietary* private information of  $V$  and  $V$  is its *proprietor*. So, every piece of private information has its proprietor. A proprietor of private information may or may not be its possessor and vice versa. Individuals can be proprietors or possessors of private information; however, non-individuals can only be possessors of private information.

The notion of proprietorship here is different from the legal concept of ownership. The ‘legal owning’ of a thing is

equated with exclusive possession of this thing with the right to transfer this ownership of the thing to others. Proprietorship of private information is non-transferable in the absolute sense. Others may possess or (legally) own it but they are never its proprietors (i.e., it cannot become their proprietary data). Also, proprietorship of private information is different from the concept of copyrighting. Atomic private information of  $V$  is proprietary information of  $V$ , while others (e.g., other individuals, companies) can only possess it. Compound private information is proprietary information of its referents: all donors of pieces of atomic private information that are embedded in the compound private information. It is also important to notice the difference between “proprietorship” and “knowing” of private information. ‘Knowing’ here is equivalent to a type of possession of private information. Atomic private information of  $V$  is proprietary information of  $V$  but it is not necessarily “known” by  $V$  (e.g., secret medical tests of employees). Possession-based “knowing” is not necessarily a cognitive concept. It is possible that a non-individual  $N$  (e.g. government agency) in  $\mathbf{N}$  “knows” atomic private information of  $V$ . “Knowing” is varied in its scope, thus, at one time there may be a piece of atomic private information that is “known” only by a limited number of entities then it may later become “known” by more entities.

### 4. Sensitive Private Information

According to our definition of atomic private information, every assertion about an identified individual is his/her atomic private information. Clearly, much of this atomic private information is trivial. It may be claimed that it is simply impractical to count every piece of information that refers to a person as private information. They are simply impossible to count and it is unreasonable to give all of them various values conferred upon private information. However, the unaccountability of pieces of private information should not discourage us from developing a theory of private information since a potentiality for infinite objects have not presented an obstacle in many scientific fields such as mathematics and linguistics. The argument here is, in a way, similar to claiming that a moral theory is impractical because it does not exclude many human actions with negligible impacts. Our definition of private information is designed to answer framework questions that provide better understanding of the ways in which private information is theoretically applied.

Here, we can introduce the notion of sensitive private information. However, while identifiably is a strict measure of what is private information, ‘sensitivity’ is a notion that is hard to pin down. It is “context dependent and thus global measures of sensitivity cannot be adopted” [11]. The sensitivity *thresholds* of applicability are a pragmatic

concern. However, we briefly discuss the issue of sensitivity of private information next.

Private information ‘sensitivity’ refers to its degree of ‘privacy-ness’. Why is some private information more sensitive than other private information? We can pursue an approach that involves a linguistic inquiry to discover the ‘tendencies’ of different types of private information to ignite different levels of sensitivity. For example, why printing a name is, generally, a less sensitive matter than printing a health history, which in turn, is, generally, less sensitive than printing sexual orientation. There are degrees of sensitivity. For example, one review of a tattoo artist said, "He's getting old and having problems with his eyesight" [8]. After the artist hired a lawyer and threatened to sue, the Website changed the wording to, "His eyesight is not what it used to be." "Sensitivity" in the context of private information refers to a special category of private topics that may disturb people. This description of sensitive private information is related to the typical definition where sensitivity of information refers to the impact of disclosing information. In this paper, we build the foundation for analysis that deals with such discernment of private information.

## 5. Compound Private Information

The ‘protection’ of atomic private information applies naturally to the corresponding compound information. If the atomic private information *V1 is in love* is blocked in the compound private information *V1 and V2 are in love*, then the possessor of the remaining part, *Someone and V2 are in love*, has no private information of V1. That is the reason we have concentrated on atomic private information in this paper. Atomic assertions are “pure” private information while compound information is not proprietary information of the individual, rather, it is shared privacy; thus, control over it is shared among its proprietors.

**Example:** The intimate relationship that is expressed by the compound private information *V1 and V2 are in love* is not just a pair of two pieces of atomic private information. The ‘intimacy’ sense emerges from tying the identities of V1 and V2 with a special type of predicate: love. Therefore, its components are two atomic assertions and the “pure” compound information. At the conceptual level, these categories of information should be recognized explicitly in a database as a different type of private information than merely a collection of embedded atomic private information. For example, releasing private information about the wife (e.g., name, age, etc.) may cause difficulties even with the husband’s approval (e.g., in case of divorce). Similarly, releasing birth records of individuals may result in releasing private information of others such as the mothers’ maiden names. Notice that recognizing “compound” interests is practiced in many fields. For

example, in certain legal procedures all participants in the relationship (e.g., marriage) have to approve explicitly of any deal related to this relationship (e.g., buying a house).

Suppose that we have the compound private assertion  $p(V_1, V_2, \dots, V_n)$  where  $V_1, V_2, \dots, V_n$  refer to different identifiable individuals. Privacy-reducibility of  $p(V_1, V_2, \dots, V_n)$  refers to producing  $n$  atomic private assertions  $p(V_1), p(V_2), \dots, p(V_n)$ , where each  $p(V_i), i = 1, 2, \dots, n$ ; has a single referent to the identifiable individual  $V_i$ .

**Proposition:** Any compound private assertion is privacy-reducible to a set of atomic private assertions.

**Justification:** Suppose that we have the compound private assertion  $p(V_1, V_2, \dots, V_n)$  where  $V_1, V_2, \dots, V_n$  refer to different identifiable individuals. The privacy-reduction process involves producing  $n$  atomic private assertions  $p(V_1), p(V_2), \dots, p(V_n)$ , where each  $p(V_i), i = 1, 2, \dots, n$ ; has a single referent to an identifiable individual. The process of producing each  $p(V_i)$  can be described as follows: For each  $V_j$ , produce its atomic private assertion by replacing all  $V_i, i$  is not equal to  $j$ , in  $p(V_1, V_2, \dots, V_n)$  by non-identifiable description of  $V_i$ .

For example, *John and Mary are in love* can be privacy-reducible to *John and someone are in love* and *Someone and Mary are in love*. Reducing a compound assertion to a set of atomic assertions refers to isolating the privacy aspects of the compound assertion. This means that, if we remove the atomic element from the compound assertion, then the remaining part will not be a “purely” privacy-related assertion with respect to the individual involved. However, it is obvious that privacy-reducibility of a compound private assertion causes a loss of ‘semantic equivalence’ since the identities of the referents in the original assertion are separated. Semantic equivalency here means preserving the totality of information: the atomic assertions and their link. This is not a main concern in our approach; however, this issue is important in certain applications, such as preserving information in a private database. Suppose that a hospital database includes the information *V1 is V2’s Kidney donor*. The database would include the two atomic assertions *V1 is a kidney donor* and *V2 had kidney transplantation*. These two assertions can be stored in the two different private databases of V1 and V2. A control mechanism facilitates any access to these databases separately. We can connect the private information in the two databases by creating a third database that includes V1 and V2 to facilitate the fact that *V1 is V2’s Kidney donor*. Nevertheless, this is not a good solution in terms of privacy, since we have recreated, practically, the original compound private database in addition to creating two atomic private databases. Is it possible to develop a method that avoids storing a third private database when connecting private databases? The answer is positive, as follows.

**Proposition:** Every compound private assertion is semantically reducible to a set of atomic private information and a set of zero information meta-assertions.

**Justification:** Let  $p(V_1, V_2, \dots, V_n)$  be a compound private assertion, where  $V_1, V_2, \dots, V_n$  refer to different identifiable individuals. According to the previous proposition, we can produce  $n$  atomic private assertions  $p(V_1), p(V_2), \dots, p(V_n)$ , where each  $p(V_i), i = 1, 2, \dots, n$ ; has a single referent to the identifiable individual  $V_i$ . Let assertion- $i$  denote  $p(V_i)$ . We introduce a statement of the type: (assertion-1, ..., assertion- $n$ ) + zero private information, where zero-private information facilitates links between the atomic assertions  $p(V_1), p(V_2), \dots, p(V_n)$ . Since the zero private information is non-private information, then the atomic assertions and their link are preserved.

**Example:** Suppose we have the compound private information, *John saw Mary's uncle, Jim*. The privacy-reducibility process produces the following three atomic private assertions:

Assertion-1: *John saw someone's uncle.*

Assertion-2: *Mary has an uncle.*

Assertion-3: *Jim is the uncle of someone.*

Additionally, we can introduce the zero-information meta-assertion: *Assertion-1, assertion-2, and assertion-3 are assertions of one compound private information*, from which it is possible to reconstruct the original compound assertion. The methodology of syntactical construction is not of central concern in this work. Alternatively, we can introduce the zero-information assertion: *Assertions 1, 2, and 3 are about x who saw y's uncle, z.*

**Example:** In *V1 is V2's Kidney donor* we have:

(1) Atomic private information database1: Includes fact1: *Identity: V1*; fact2: *V1 is a kidney donor*

(2) Atomic private information database2: Includes fact1: *Identity: V2*; fact2: *V2 had kidney transplantation.*

(3) Zero-private information database3: fact1: (database1: fact2, database2: fact2).

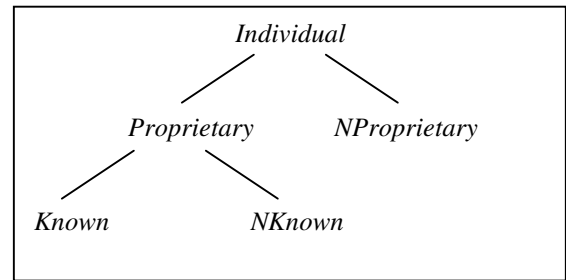
Notice that each of database1 and database2 is composed of a collection of pieces of atomic private information. The control mechanism facilitates any access to these databases separately. For example, access to database1 requires the consent of  $V_1$ , while access to database2 requires the consent of  $V_2$ . Database3 includes no private information of either  $V_1$  or  $V_2$ . It even does not include any information about the type of the relationship between facts in database1 and database2. Constructing the original compound private information requires access to the three databases. Thus, we have succeeded in isolating the private information of each individual and at the same time have preserved the possibility of recovering any compound private information. We assume in our examples about databases that identities are used as keys.

**Example:** Consider the compound assertion *John, Jim and Alice hate each other*. It embeds the atomic assertions  $x$  hates someone, where  $x$  can be John, Jim, or Alice. These

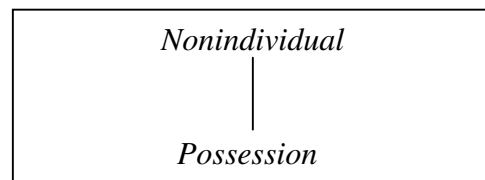
compound assertions can be represented by three atomic assertions and a meta-assertion that states that they are components of a single compound assertion. This type of compound private information can be reconstructed uniquely. The atomic assertions  $\{John\ hates\ someone, Jim\ hates\ someone, Alice\ hates\ someone\}$  and the information that these atomic assertions are from one compound assertion, produce the assertions  $\{John\ hates\ Jim, Jim\ hates\ John, John\ hates\ Alice, Jim\ hates\ Alice, Alice\ hates\ John, Alice\ hates\ Jim\}$  which leads eventually to the original compound assertion. As we stated previously, this paper concentrates on the reduction process in order to identify privacy 'centers' and leaves the issue of the reconstruction process for further study.

## 6. Private Information Types

Let  $V$  and  $Z$  be two entities of type *Individual* and *Nonindividual*, respectively. The definitions of *Individual* and *Nonindividual* can be accomplished utilizing a simplified class definition as shown in Figure 1 and 2. We use the familiar notation '.' to indicate the hierarchy of objects such as *Individual.Proprietary.Known*.



**Figure 1. An Individual can have two types of private information: Proprietary and NProprietary.**



**Figure 2. A Nonindividual can only possess (and not be proprietor of) private information. The set of possessed information is Possession. Notice Possession is called NProprietary if the entity is Individual.**

Figures 1 and 2 include the following sets:

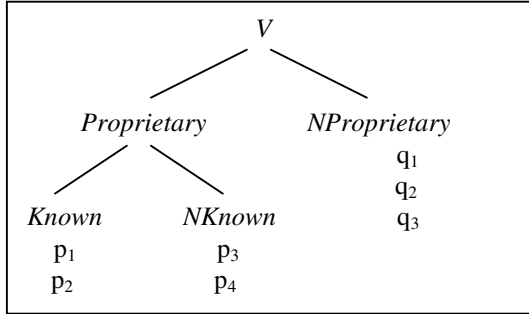
**1. Proprietary** is the set of pieces of atomic private information of an individual. *Proprietary* has two components:

(a) **Known**: This is the set of pieces of atomic private information that is in the possession of others. Notice that there may be private information in *Known* that is unknown to the proprietor (secret medical tests) even though it is his/her own private information.

(b) **NKnown**: This is the set of pieces of atomic private information that is only known by the proprietor. This type of private information is usually not specified explicitly; however, it appears in certain circumstances.

2. **V.NProprietary** is the set of pieces of private information of the other individuals that is in the possession of an individual, however he/she is not its proprietor.

**Example**: Suppose that an individual *V* has the following set of private assertions  $V.Proprietary = \{p_1, p_2, p_3, p_4\}$ , such that what is known by company XYZ is  $\{p_1, p_2\}$ . Furthermore, assume that *V* knows private assertions of others  $\{q_1, q_2, q_3\}$ . The total private assertions can be represented in Figure 3 as an instance of ‘data structures’.



**Figure 3. Private assertions about or in the possession of V.**

Let *Z* be of type either *Individual* or *Nonindividual*. *Z.Possession* is the set of pieces of private information that are in the possession of *Z*. If *Z* is the individual *V* then *V.Possession* is exactly the set *V.NProprietary*. To simplify notation, we may write *Possession* (no object name and no dot preceding it) to refer both to *Z.NProprietary* and *Z.Possession* when the object under consideration is immaterial. We will follow this notational convenience for other subcomponents that will be introduced in the following sections. Also, we will eliminate intermediate names when possible.

There are a huge number of pieces of atomic private information associated with each individual. They are easily recognizable in their linguistic form through atomicity (i.e., one referent). In a limited data environment, it is possible to factor these private assertions as in the case of private information collected by companies. Every relational database can be divided into a private information database and other types of databases. For example, the tuple (123, John, 80, Physics, Senior), in the relation: STUDENT(ID, CREDITS, MAJOR, STATUS), has four atomic assertions:

*John’s Id is 123, John passed 90 credits, John’s major is Physics, and John’s status is senior.* The relation ORDER (SUPPLIER, PART-ID, QUANTITY) has no private information.

Other types of atomic private information such as *Individual.NKnown* are potentially uncountable. Of course, no one expects to count the number of pieces of this type of private information. Nevertheless, sets in this approach have enabled us to identify different types of information related to privacy.

The next section goes further into exploring the potential of this formalism for the purpose of exposing different characterizations of private information. It applies the mechanism of ‘counting’ to pieces of private information. At least, and at the theoretical level, this demonstrates that information privacy lends itself, in principle, to rigid analysis. We claim that such a result is a positive characterization in comparison with the general informal definitions of private information discussed in the introduction.

## 7. Calculated Privacy

An interesting result comes from taking the unions of some sets of private assertions introduced previously.

**Proposition:**  $\cup \{V.Known, \text{ for all } V \text{ in } \mathbf{V}\} =$

$$\cup \{Z.Possession, \text{ for all } Z \text{ in } \mathbf{Z}\}.$$

That is, with regard to the total population, since ‘we’ include ‘others’ and ‘others’ include ‘us’, the private information about us known by others is the private information in our possession about others. Notice *V.NProprietary* is a kind of *Z.Possession* whenever *Z* is an individual. This result indicates that reducing the accumulation of private information by all agents (reducing *Z.Possession*), say by 50%, has an equal effect to the reduction of releasing (*V.Known*) private information of all individuals to others by the same amount. Reducing *V.Known* requires individuals to take protective measures to block their private information from reaching to others. Reducing *Z.Possession* requires a self-imposed policy not to collect private information about others. *Z.Possession* also includes types *Individual* and *Nonindividual*, while *V.Known* is related only to entities of type *Individual*.

**Example:** Suppose that we have three individuals let  $\mathbf{V} = \{V1, V2, V3\}$  and *Nonindividual*  $\mathbf{N} = \{company-XYZ\}$ . Let the atomic assertions for each individual be as follows:

$$V1.Proprietary = \{p_1, p_2\}$$

$$V2.Proprietary = \{q_1, q_2\}$$

$$V3.Proprietary = \{t_1, t_2\}$$

Assume that:

$$V1.NProprietary \cap V2.Known = \{q_1\}$$

$$V1.NProprietary \cap V3.Known = \{t_1\}$$

$$V2.NProprietary \cap V1.Known = \{p_1\}$$

$V2.NProprietary \cap V3.Known = \{t_1\}$   
 $V3.NProprietary \cap V1.Known = \{p_2\}$   
 $V3.NProprietary \cap V2.Known = \{q_1\}$   
 $XYZ.Possession \cap V1.Known = \{p_1\}$   
 $XYZ.Possession \cap V2.Known = \{q_1\}$   
 $V1.NProprietary = \{q_1, t_1\}, V2.NProprietary = \{p_1, t_1\},$   
 $V3.NProprietary = \{p_2, q_1\}, XYZ.Possession = \{q_1, p_1\}.$   
 $V1.Known=\{p_1, p_2\}, V2.Known=\{q_1\}, V3.Known=\{t_1\}.$

$\cup\{V.Known, \text{ for all } V \text{ in } \mathbf{V}\} =$

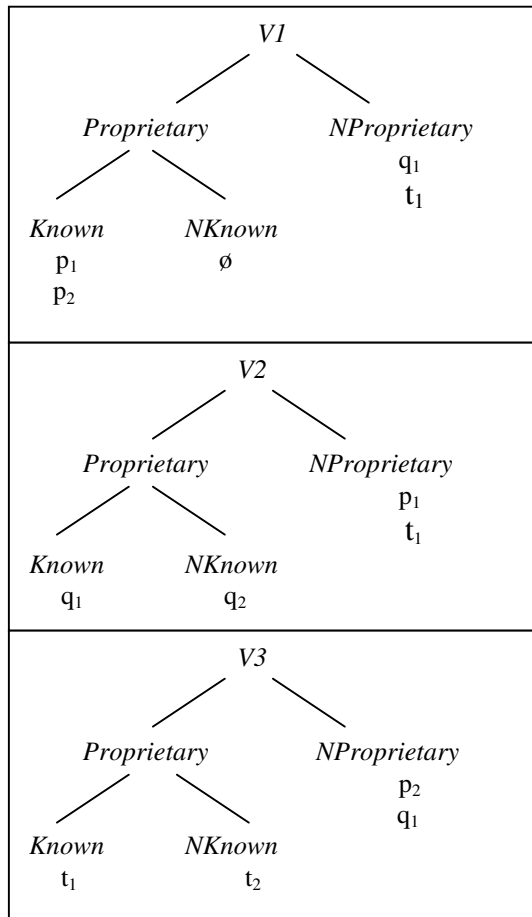
$\cup\{Z.Possession, \text{ for all } Z \text{ in } \mathbf{Z}\} = \{q_1, t_1, p_1, p_2\}$

The given situations with respect to V1, V2, and V3 are shown in Figure 4.

To find a measure of information privacy, it is necessary to look at the following two possible types of factors that contribute to any measure of privacy as:

**Protective measures:** What individuals can do to protect their private information.

**Preventive measures:** What others can do to reduce collecting private information.



**Figure 4. The private databases of V1, V2, and V3.**

The first obvious preventive measure deals with the individual's private information that is unknown by anyone.

**Principle 1:**  $NKnown(V)$  should be maximized

*This principle is the “controlling-access”-version of privacy protection. Everyone should minimize releasing information to others. Releasing no private information means  $V.Known = \emptyset$ , where all private information about  $V$  is only known to  $V$ . Notice that if  $V.Known = \emptyset$ , then there is no known compound assertion of  $V$ . That is “no known atomic information” of  $V$  implies “no known compound information” of  $V$ . Because, if the compound information is known, then its atomic information is known. This is one sense of privacy-reducibility of compound information to atomic information.*

The system may play a role in making it possible for individuals to maximize their  $NKnown$ . For example, cash transactions may provide complete privacy in the context of purchasing, if all cash money notes paid for purchasing are not traceable. Anonymous e-money (digital cash) aims at preserving the untractability of monetary transactions in order to give individuals the option stated in Principle 1. Principle 1 is based on the assumption that the less known about private information, the better.

Next, we move from the factor of ‘amount’ of private information to the factor of how many ‘others’ know this private information. An entity in  $Z$  is said to be a collector with respect to an individual  $V, Z \neq V$ , if  $Z.Possession \cap V.Known$  is not an empty set.

**Principle 2:** For any Proprietor  $V$ , the number of collectors should be minimal.

Principle 2 is based on the observation that the fewer the number of others that know about the private life of an individual, the more one has privacy. There is a potential conflict between principles 1 and 2. Is it ‘better privacy’ to have a few others who know a great deal of your private information, or to have many others know a small portion of it? Most people share almost all private information with few people (e.g., husband, wife, mother, father, etc.) with no privacy complaints.

**Principle 3:** For any individual  $V$  and a possessor  $Z, Z \neq V$ ,  $Z.Possession \cap V.Known$  should be minimal.

Principle 3 is clearly favorable to a privacy policy where: the less each collector knows, the better. Principles 2 and 3 represent the “disinterest”-version of protecting privacy. Everyone should minimize collecting private information about others. The aim is to minimize:

- (a) The number of private information collectors and the total collected private information.
- (b) The amount possessed by each collector.

For any individual, the priorities of the three principles are important policy decisions. It seems reasonable that principle 1 should be his/her first concern. Also, it seems reasonable that principle 2, minimizing the number of collectors, is more important than minimizing the amount of

collected private information (principle 3). These assumptions will be applied in the next example.

**Example:** Suppose that  $v = \{V1, V2, V3\}$  such that  $V1.Proprietary = \{x, y, z\}$ ,  $x$  is  $V1$ 's name,  $y$  is  $V1$ 's phone number, and  $z$  is  $V1$ 's bank account. Consider the following situations shown in table 1. The first column shows  $V1.NKnown$ , the private information known only by  $V1$ . The second column shows private information of  $V1$  in possession of  $V2$ . The third column shows private information of  $V1$  in possession of  $V3$ . In situation (1),  $V2$  has  $x$ , the private information of  $V1$  while  $V3$  has  $y$ . In (2), both have  $x$ . In (3),  $V2$  has  $x$  and  $y$ . In (4),  $V2$  has  $x$  while  $V3$  has no private information of  $V1$ . In (5)  $V2$  has  $x$  and  $y$  while  $V2$  has  $y$ . Clearly,  $V1$  is a data source and  $V2$ , and  $V3$  are collectors. Which situation achieves more privacy for  $V1$ ?

**Table 1. Possible distributions of pieces of private information of  $V1$ .**

	$NKnown(V1)$	$V2.NProprietary \cap V1.Known$	$V3.NProprietary \cap V1.Known$
1	z	x	y
2	y, z	x	x
3	z	x, y	
4	y, z	x	
5	z	x, y	y

Applying the assumed priorities, Examples of resultant judgments when using the given principles are:

- (2) is better than (1) by principles 1 and 2
- (3) is better than (1) by principle 2
- (4) is better than (1) by principles 1 and 2
- (1) is better than (5) by principle 3
- (2) is better than (3) by principle 1 (Note: Overruling principle 2)
- (4) is better than (2) by principle 2
- (2) is better than (5) by principles 1 and 3
- (4) is better than (3) by principles 1 and 3
- (3) is better than (5) by principle 2
- (4) is better than (5) by principles 1,2 and 3

A privacy measure  $\rho$  with respect to any data agent  $V$ , can now be defined through the value:

$$\rho = \frac{(|V.NKnown| + 1)}{(|\text{possessors}| * \sum |Z.Possession \cap V.Known| + 1)}$$

for every private information collector  $Z$ . The equation divides the number of pieces of atomic private information known only by their proprietor, by the total number of pieces of atomic private information known by others about the proprietor. The multiplication is used in the

denominator to make the equation sensitive to the number of collectors as indicated by principle 2. The larger the number of collectors, the less the value of sigma is. "One" is added in the denominator to avoid dividing by zero. "One" is added in the numerator to distinguish among different values of the denominator when  $|V.NKnown|$  is zero as will be shown in the following example. The range of sigma is between  $[1/(\text{very large number}), \text{very large number}]$ . The lower range (minimum privacy) represents the situation when every thing is known about the individual, i.e.,  $(|V.NKnown| = 0)$ , and the upper range (maximum privacy) represents the situation when nothing is known about the individual, i.e.,  $|\text{possessors}| = 0$ .

In the example, for case (1):

$$|V1.NKnown| = 1$$

$$\sum |Z.Possession \cap V1.Known| = 2$$

where,  $(V2.NProprietary \cap V1.Known) = \{x\}$  and,

$$(V3.Possession \cap V1.Known) = \{y\}.$$

Hence,  $\rho = 1/(2*2+1) = 2/5$ . To cover most cases of distribution of the pieces of private information of  $V1$ , we add the other cases shown in table 2. Table 3 shows the calculations of  $\rho$  for all cases.

**Table 2: Other possible distributions of pieces of private information of  $V1$ .**

	$NKnown(V1)$	$V2.NProprietary \cap V1.Known$	$V3.NProprietary \cap V1.Known$
6	z	x, y	x, y
7		x, y, z	
8	z	x, y, z	x, y, z
9	y, z	x, y, z	x
10	z	x, y, z	y, x
11	x, y, z		

**Table 3: Calculations of  $\rho$ .  $\Sigma$  denotes  $\sum |Z.NProprietor \cap V.Known|$**

	$ V.NKnown(V) $	$ \text{Collectors} $	$\Sigma$	$\rho$
1	1	2	2	2/5
2	2	2	2	3/5
3	1	1	2	2/3
4	2	1	1	3/2
5	1	2	3	2/7
6	1	2	4	2/9
7	0	1	3	1/4
8	0	2	6	1/13
9	0	2	4	1/9
10	0	2	5	1/11
11	3	3	0	4

Thus, the cases according to the ascending value of  $\rho$  are: 11, 4, 3, 2, 1, 5, 6, 7, 9, 10, 8. Situation 11 achieves complete privacy to V. The next best situations are 4, 3, etc.

Again, the aim here is not to enumerate private pieces of information and apply calculations to them. Rather, the objective is to set up a theoretical model that precisely identifies different types of relationships that should be taken into consideration in any information privacy study.

It is typically claimed that privacy is hard to define, “even [in] explicit textual information” [4] and that privacy metrics are ad hoc in nature. Privacy measurability as reflected in our equation, is an attempt at quantifying information privacy in order to develop an informational privacy theory. This measurability is a development in the right direction, even if the methods of private information auto-identifying and the practical applications of the measure are not easily apparent. The equation of privacy measure reflects the common sense notion, that privacy is a function of how many ‘others’ know portions of your private information and the amount known by each of them. It expresses that your informational privacy can be measured as:

$\frac{\text{The 'amount' of PI not known by others}}{\text{The 'amount' of PI known by others}}$
---

where PI denotes private information. That is, your informational privacy is proportional to how much of your PI has not been released to others and inversely proportional to the amount of your private information in the possession of others. This is a reasonable approximation of a common sense notion. Consider the following definition of informational privacy given by Floridi [10]: Informational privacy is “freedom from epistemic interference or intrusion, achieved thanks to a restriction on facts about S that are unknown or unknowable.” Floridi’s ‘epistemic interference or intrusion’ decreases the quantity of informational privacy  $\rho$ , in another words, decreases the ‘quantity’ of freedom in Floridi’s definition.

One of the most important weaknesses of our approach is that we have counted pieces of information without regard to the quality of each piece of information. For example, one piece of information might be a credit card number and another might be a skin color. Both would be given equal weight in the formula. However, our preliminary investigation of this point has shown that there is some measure of ‘quantifying’ sensitivity of private information. Atomic private information can be further refined to self-assertions that refers to the proprietor him/herself (e.g., *John has Cancer*) and assertions that refer to a thing associated with the proprietor (e.g., *John’s house is burning*). The later assertion, can be further refined to

self-assertion through separating embedded zero-privacy assertions (e.g., *John has a house* and *A house is burning*). Self-atomic-assertions are ‘pure’ atomic assertions that express actions ‘of’ the proprietor (e.g., has, feel, walk, etc.) It seems that the sentivity factor can be associated with the action (verb), the rest-of-the assertion or both. For example, in an actual alleged defamation case related to the assertion *V stranded passengers*, the sensitivity stems from the verb. In another actual alleged defamation case related to the assertion *V has a mean streak*, the sensitivity comes from the non-verb portion. Identifying the source of sentivity in these simple linguistic units can be complemented with semantic ranking (e.g., verbs: He ‘taught’ vs. ‘molested’ juveniles; non-verbs: He engaged in ‘discussion’ vs. ‘sex’).

## 8. Applications

Additionally, our approach seems to have some utility in the whole domain of privacy. Below, we give two examples of utilizing the private information definition in information ethics and databases fields.

Information Ethics (IE) has encompassed issues that stem from connecting technology with such topics as privacy, intellectual property rights, information access, intellectual freedom, etc. According to Froehlich, the issues in information ethics were raised as early as 1980 and the field of information ethics “has evolved over the years into a multi-threaded phenomenon, in part, stimulated by the convergence of many disciplines on issues associated with the Internet” [18]. Floridi [10] has proposed to base IE on the machine-independent concept of information as its basic phenomenon. According to this IE, information itself, in some form or role, is recognized to have an intrinsic moral value. Being an information entity is the minimal condition of possibility of moral dignity and hence of normative respect. Floridi claims that his thesis can contribute a new ethical perspective on the notion of informational privacy. Mathiesen criticized such a theory of IE maintaining, “a theory of information ethics will need to specify the relation between persons and information such that information can be of ethical import” [19]. Utilizing the premise that information has intrinsic moral value and our definition of private information, we can construct a theory for private information ethics that provides a framework for organizing private information ethical issues.

Private Information Ethics (PIE) concerns with the moral consideration of private information because private information’s ‘wellbeing’ is manifestation of proprietors’ welfare. PIE recognizes private information itself to have an intrinsic moral value. So the ‘human being’ enters the agent/action/patient ethical discourse from all three aspects: as an agent, as an actor, and as a receiver, since the action involves his/her private information. In PIE, the “patient” is

private information. In this way, PIE is a step backward from the abstract impartiality of Floridi's IE. We consider this as a positive aspect since it regains the center of basic ethical claims to humans and, at the same time, keeps the IE thesis of making information as the primary object of the ethical discourse. An example of PIE is developing moral justification for lying about private information. We claim that dividing lying-based state of affairs into private and non-private information types eliminates moral difficulties that produce a non-common sense conclusion in situations that involve identifiable individuals. Privacy provides a universal requirement that supports lying about private information in order to avoid harm. This thesis has been applied to a current situation where customers, who are lying about their private information, are met with moral outrage and loss of credibility. Our ethical analysis has neutralized relying on psychological feelings of compulsion, and authoritative force to pressure these consumers to be truthful when submitting proprietary private information. In an organization, the private information acquisition method should be oriented such that it depends on facts not on the hope that the customer believes that *Lying is wrong* [1].

In the database area, a formalism to conceptualize 'private information databases' can be developed based on atomic assertions. The relationship between individuals and their own atomic private information is identified through the notion of *Proprietorship*. Suppose that company XYZ maintains a private information database. Its whole database includes information privacy databases of three departments and the database for employees working in XYZ. Each *Possession* of a department contains the private information about its employees and maybe other departments' employees. The sub-schema of the finance department, for example, has the financial private information of all employees in XYZ. It is possible to enforce global privacy constraints on information in *Possession*. Additionally, each individual has his/her information privacy database. His/her database *Proprietary.Known* tells him/her what, for example, department 1 "knows" about him/her. Thus, the database automatically inserts any private information in the appropriate employee's *Proprietary.Known*. This provides a verification of the claim that the employee is aware that his/her private information in the possession of department-1. If XYZ is a hospital, then each doctor has in his *NProprietary* medical (say, not financial) private information of his/her patients. He/she can also use his/her *Proprietary* as a repository of his/her private information, which no one can access except him/herself. The database in this case, represents a 'bank' of private information that maintains its different types including people's 'safe boxes' and their different 'assets' [2].

## 9. Conclusion

This paper introduces a theoretical foundation for information privacy including a new definition that precisely identifies private information assertions. Based on such a definition, a formula for evaluating the informational privacy is also proposed. Additionally, it is shown that the new definition of private information can be utilized in several privacy-related areas such as information ethics and databases. It can also be applied to many notions such as anonymization, misinformation, confidentiality, etc.

## 10. References

- [1] Al-Fedaghi, S. S., "Lying about Private Information: an Ethical Justification", 16th Annual Conference International Information Management Association (IIMA), September 22 – 24, 2005, Dublin, Ireland.
- [2] Al-Fedaghi, S. S., Fiedler G., and and B. Thalheim B., "Privacy Enhanced Information Systems", Proceedings of The 15th European-Japanese Conference on Information Modelling And Knowledge Bases: Tallinn, Estonia, 2005.
- [3] The Platform for Privacy Preferences 1.1 (P3P1.1) Specification, [http://www.w3.org/TR/2005/WD-P3P1-20050104/Overview.html#def\\_identity](http://www.w3.org/TR/2005/WD-P3P1-20050104/Overview.html#def_identity)
- [4] Andrew Senior, Sharath Pankanti, Arun Hampapur, Lisa Brown, Ying-Li Tian, Ahmet Ekin, Blinkering Surveillance: Enabling Video Privacy through Computer Vision, IBM Research Report, RC22886 (W0308-109) August 28, 2003 Computer Science, [http://domino.watson.ibm.com/library/cyberdig.nsf/papers/9A2991BD5F0041EC85256ED8006D84D2/\\$File/rc22886.pdf](http://domino.watson.ibm.com/library/cyberdig.nsf/papers/9A2991BD5F0041EC85256ED8006D84D2/$File/rc22886.pdf)
- [5] Baud R.H., C. Lovis, P. Ruch, AM. Rassinoux, "A Toolset for Medical Text Processing", Medical Informatics Europe, Hannover 2000, MIE 2000, IOS Press.
- [6] Clarke, R., "Introduction to Dataveillance and Informational privacy, and Definitions of Terms", 1999. <http://www.anu.edu.au/people/Roger.Clarke/DV/Intro.html>.
- [7] Department of Motor Vehicles, Privacy and Security Notice, Internet Office, New York State, 2002, <http://www.nydmv.state.ny.us/securitylocal.htm>.
- [8] Eden, Eric, "Libel & Defamation in the Information Age", 1995, <http://www.skepticfiles.org/hacker/cud709.htm>
- [9] EU Directive 1995/46/EC, On the protection of individuals with regard to the processing of personal data and on the free movement of such data, Official Journal of the European Communities, No. L 281, 23.11.1995.

- [10] Floridi, di Luciano, "Information Ethics: On the Philosophical Foundation of Computer Ethics", ETHICOMP98 The Fourth International Conference on Ethical Issues of Information Technology, 1998. <http://www.wolfson.ox.ac.uk/~floridi/ie.htm>.
- [11] Fule P., and J. Roddick 2004, "Detecting Privacy and Ethical Sensitivity in Data Mining Results", Twenty-Seventh Australasian Computer Science Conference, Dunedin, New Zealand, 2004. <http://crpit.com/confpapers/CRPITV26Fule.pdf>
- [12] HIPAA, Glossary of Common Terms, Health Insurance Portability and Accountability Act of 1996, <http://healthcare.partners.org/phsirb/hipaaglos.htm#g3>
- [13] PIPEDA, "PIPEDA Overview - What", 2004, <http://privacyforbusiness.ic.gc.ca/epic/internet/inpfb-cee.nsf/en/hc00005e.html>
- [14] Sweeney, Latanya, "Replacing Personally-Identifying Information in Medical Records, the Scrub System", In: Cimino, J., ed. Proceedings, *Journal of the American Medical Informatics Assoc.* Washington, DC: Hanley & Belfus 1996:333-337.
- [15] Taira, R. K., Alex A. T. Bui, and Hooshang Kangarloo, "Identification of Patient Name References within Medical Documents Using Semantic Selectional Restrictions", Proceedings of the AMIA 2002 Annual Symposium, p. 757.
- [16] Wacks, R., "Privacy in Cyberspace", In: Birks, P. (editor), *Privacy and Loyalty*, Clarendon Press, Oxford, New York, 1997, p. 91-112.
- [17] Marx, G. T., "Varieties of Personal Information as Influences on Attitudes Toward Surveillance", In: K. Haggerty and R. Ericson (editors), *The New Politics of Surveillance and Visibility*, 2005. <http://web.mit.edu/gtmarx/www/vancouver.html>
- [18] Froehlich, T., "A brief history of information ethics", 2004, <http://www.ub.es/bid/13froel2.htm>
- [19] Mathiesen, K., "What is Information Ethics?", *Computers and Society Magazine*, Volume 32 - Issue 8, June-2004, [http://www.computersandsociety.org/sigcas\\_ofthefuture2/sigcas/subpage/sub\\_page.cfm?article=909&page\\_number\\_nb=1](http://www.computersandsociety.org/sigcas_ofthefuture2/sigcas/subpage/sub_page.cfm?article=909&page_number_nb=1)