Intelligent Information Management, 2012, 4, 123-133
http://dx.doi.org/10.4236/iim.2012.44019 Published Online July 2012 (http://www.SciRP.org/journal/iim)

Experimentation with Personal Identifiable Information

Sabah Al-Fedaghi, Abdul Aziz Rashid Al-Azmi
College of Engineering and Petroleum, Kuwait University, Kuwait City, Kuwait
Email: sabah@alfedaghi.com, fortinbras.222@kuniv.edu.com

Received May 22, 2012; revised June 29, 2012; accepted July 17, 2012

ABSTRACT

In this paper, actual personal identifiable information (PII) texts are analyzed to capture different types of PII sensitivities. The sensitivity of PII is one of the most important factors in determining an individual's perception of privacy. A "gradation" of sensitivity of PII can be used in many applications, such as deciding the security level that controls access to data and developing a measure of trust when self-disclosing PII. This paper experiments with a theoretical analysis of PII sensitivity, defines its scope, and puts forward possible methodologies of gradation. A technique is proposed that can be used to develop a classification scheme of personal information depending on types of PII. Some PII expresses relationships among persons, some specifies aspects and features of a person, and some describes relationships with nonhuman objects. Results suggest that decomposing PII into privacy-based portions helps in factoring out non-PII information and focusing on a proprietor's related information. The results also produce a visual map of the privacy sphere that can be used in approximating the sensitivity of different territories of privacy-related text. Such a map uncovers aspects of the proprietor, the proprietor's relationship to social and physical entities, and the relationships he or she has with others.

Keywords: Personal Identifiable Information; Public Policy Issues; Privacy; Sensitivity

1. Introduction

Personal identifiable information (PII) is vital in today's privacy legislation, according to Schwartz and Solove [1]:

> *Personally identifiable information (PII) is one of the most central concepts in information privacy regulation. The scope of privacy laws typically turns on whether PII is involved. The basic assumption behind the applicable laws is that if PII is not involved, then there can be no privacy harm.*

The Department of Homeland Security (DHS) defines PII as "*Any information that permits the identity of an individual to be directly or indirectly inferred, including any information which is linked or linkable to that individual*" [2]. McMeekin [3] notes that "PII is generally defined as information *about* or *associated with* an individual".

Privacy laws in their various forms typically prohibit unconstrained handling of PII, though they do not recognize sensitive versus non-sensitive PII [4]. In Ohm's view, laws related to PII have drastically failed to protect individuals' privacy, and the notion of PII should be abandoned [5]. Other scholars dispute this view [6]. Ohm's view evolved because of some results in de-anonymization that adopted a definition of PII that does not generalize it to any collection of secondary attributes that uniquely identify a person. Simply, de-anonymization succeeds only because of failure of an anonymization technique to remove all identities embedded in the records.

Not all PII is sensitive information (see **Figure 1**). There is a point that must be exceeded to begin to consider PII sensitive. Social networks depend on the fact that individuals willingly publish their own PII, causing more dissemination of sensitive PII that compromises individuals' information privacy. This may indicate that PII sensitivity is an evolving notion that needs continuous evaluation. On the other hand, many Privacy-Enhancing Technologies (PETs) are being devised to help individuals protect their privacy [7], indicating the need for this notion.

The sensitivity of PII is one of the most important factors in determining an individual's perception of privacy [8]. In data protection law, the principle of sensitivity holds that the processing of certain types of data should be subject to more stringent controls than other personal data [9]. In general, the notion of sensitivity is a particularly difficult concept. In many situations, sensitivity seems to depend on the context, and this cannot always be captured in a mere linguistic analysis; however, this does not exclude the possibility of "context-free" sensitivity, as proposed in this paper.

Additionally, creating context-free sensitivity can provide an initial classification of information that can be

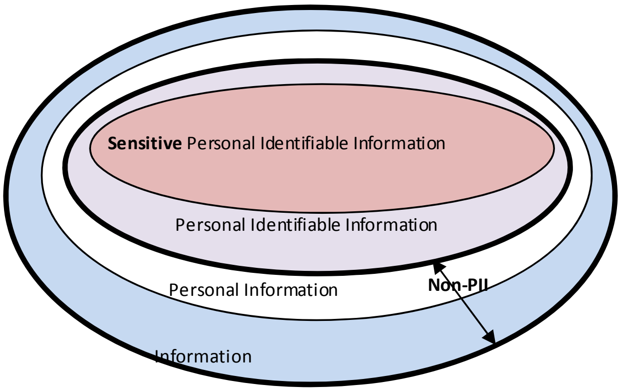
Sensitive Personal Identifiable Information
Personal Identifiable Information
Personal Information
Non-PII
Information

Figure 1. Information and PII.

further refined through either a knowledge-based system, a manual system, or both. A report about an identified person has some significance for that person regardless of the context (e.g., *they were talking about you*). The sensitivity of the matter is always there regardless of context when mention of the identified person is associated with a behavior (e.g., *they were talking about your visit* vs *about your outrage*).

In practice, it has been rather difficult to identify categories of sensitive data, especially personal identifiable information (PII). Bing [10] proposed assigning a level of PII sensitivity, from the most sensitive to the least, as follows: 1) inherently sensitive, intimate (e.g., medical or sexual) information, 2) judgmental data that could lead to harm for the data subject, and 3) biographical data that provides access to more sensitive data.

In this paper, we establish a semi-automated methodology for measuring PII sensitivity starting from initial values that can be refined manually and by self-learning from previous evaluations. The methodology is built on anatomizing PII in pieces (e.g., the identified person vs the action) and kinds (e.g., a single person vs relationships among persons). We aim to build sensitivity measurements upon linguistic units to provide a syntactical base for dealing with the question: why is some PII more sensitive than other PII?

The approach involves a linguistic inquiry to discover the "tendencies" of different types of PII that ignite different levels of sensitivity. There are many advantages to developing such a methodology, including 1) providing privacy management for the complexity of many inquiries for PII, and 2) managing the act of consenting to disclose/withhold PII such that it is configured *a priori*.

2. Related Works

Not all pieces of PII are equal in privacy significance. To our knowledge, there are no works in this area besides the usual objective statistical measures. In authentication systems, PII is generally limited to identification/contact information. The Platform for Privacy Preferences Project (P3P) does not classify categories of data according to their sensitivity. Microsoft's IE6 restricts the use of cookies under certain scenarios, but only if "sensitive" categories (e.g., GOV) are included. Some works propose a vocabulary for composing policies that allow or deny access to PII and facilitate the ways in which semantic web processors reason through policies [11].

PII that embeds sensitive information is likely to be targeted for exploitation. The unauthorized exposure of such PII does definite damage to the privacy rights of the proprietor (the person about whom the PII informs). Today we hear a lot of news about some company or government losing vast amounts of customer or client information, which is in many ways PII. The Privacy Rights Clearinghouse has reported that hundreds of millions of personal records have been improperly exposed since 2005 [12].

Many efforts are under way to specify rules and regulations for handling PII. For example, the US Department of Homeland Security [2] specifies steps that must be taken to safeguard PII. It clearly explains how to identify and protect PII, and what to do in case PII is compromised. Nevertheless, its category system is static and restricted.

The closest work to our problem is a study by Fule and Roddick [13] to detect privacy and ethical sensitivity in data-mining results. Fule and Roddick studied the potential sensitivity of information extracted from a database. They observed that evaluating such sensitivity is "context dependent and thus global measures of sensitivity cannot be adopted" [13]. They propose a system to address "the subjective nature of ethics and privacy" by automatically rating all generated results of a query using user-defined sensitivity values. They accomplish this goal by storing a set of privacy and ethical sensitivity values in the range 0 to 10 for each attribute or attribute value. The system evaluates a combination function to form a sensitivity rating for a rule in the mining process. In addition to data-item sensitivity, the function takes into consideration its position (whether antecedent or consequent) and structural aspects, such as non-leaf values within a hierarchy.

3. Defining Personal Information

The notion of an identifiable person divides entities into two fundamental types: natural persons and others. This approach has been described in [14]. PII is any information that has *referent*(*s*) of type natural persons. There are two types of personal information:

1) Atomic PII (APII) is information that has a single human referent.

2) Compound PII (CPII) is information that has more than one human referent.

"Atomic" in this definition refers to the "subject" of the statement and not to the composition of the statement expressing that fact. Thus, *John is tall and handsome*, *John is tall*, and *John is handsome* are all APII, even though the first contains the second and third statements. A single referent does not necessarily imply a single occurrence of a referent. Thus, "*John* wounded *himself*" has one referent.

According to the definition of atomic PII, or APII, every assertion about an identified individual is his or her atomic PII. While identifiability is a strict measure of what is PII, "sensitivity" is a notion that is hard to pin down. In this paper, we pursue an approach involving a linguistic inquiry to discover the "tendencies" of PII to ignite different levels of sensitivity.

The relationship between individuals and their own APII is called *proprietorship* [14]. If *p* is a piece of atomic PII of a person, then *p* is proprietary PII of its proprietor. Compound PII or CPII is proprietary information of its proprietors.

Any CPII is privacy-reducible to a set of APII. 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*; however, it is obvious that the privacy-reducibility of a CPII causes a loss of "semantic equivalence" since the identities of the proprietors in the original PII are separated. Semantic equivalency here means preserving the totality of information: the atomic PIIs and their link. This topic is not a main concern of this paper.

Defining PII as "information identifiable to the individual" does not mean that the information is "especially sensitive, private, or embarrassing. Rather, it describes a relationship between the information and a person, namely that the information—whether sensitive or trivial—is somehow identifiable to an individual" [15]. The *significance* of PII derives from its privacy value to a human being.

From an informational point of view, an individual is a bundle of his or her PII. PII comes into being not as an independent piece of information, but rather as a constitutive part of a particular human being [16]. PII ethics is concerned with the "moral consideration" of PII because PII's "well-being" is a manifestation of the proprietor's welfare [17-19].

4. PII Sensitivity

We are interested in evaluating the sensitivity of a piece of atomic PII according to the construct of sentences. The technique lends itself to automation and can be complemented with other human-based or context-based methods.

In Fule-Roddick's system, "sensitivities associated with fields can be created either by someone with expert knowledge of what is socially acceptable or through the gathering of societal perceptions using other means such as surveys" [13]. We will uncover sensitivity by analyzing atomic PII. Our plan for discerning APII is as follows:

1) Removing non-PII embedded in APII, e.g., *John's house is big* embeds the non-PII *the house is big*. The justification for this is that we want to focus directly on things that describe or relate to the proprietor.

2) Separating APII into two types: a) that portion of APII where the proprietor is the only object of reference (e.g. *John is nice*) and b) portions where there are several objects of reference, e.g., *John has a horse*. The justification for this is that we want to separate a proprietor's features from his or her relationships with non-human objects in the world.

3) Simplifying the resultant assertions, e.g., *John is tall and dark* is simplified to *John is tall* and *John is dark*.

4) In addition to the proprietor, identifying the source of sensitivity in the verb (action), the rest of PII, or both.

5) Accordingly, comparing the result with a sensitivity list of words or phrases.

A piece of PII is information about:

a) Aspects of its proprietor (e.g., short, tall, funny).

b) His or her association with non-person "things" (e.g., house, dog, organization).

c) His or her relationships with other natural identifiable persons (e.g. wife, friend, employee).

Accordingly, we first try to isolate language structures that assist in recognizing these types of information. We call PII that "focuses" on the proprietor self-PII. Then, we isolate types of PII that describe aspects of the proprietor "v" (called singleton self-PII), from those that contain one or more referents to non-human objects (called multitude self-PII).

Definition: APII is said to be *self-APII* (SPII) if its *subject* is its proprietor and only its proprietor.

For example, *John's house is burning* is not self-assertion because it expresses two pieces of information: a) *John has a house* and b) *the house is burning.* The statement *John has a house* is self-APII, or SPII, because its "subject" is its proprietor. *The house is burning* is non-PII because its "subject" is not a person but a house. The term "subject" here means the entity about which the information is communicated. In many cases, this means that the individual affects/is affected by the verb(s) of the sentence.

Proposition: Every APII is reducible to a set of SPII and non-personal information.

Discussion: The reduction process from a single piece of APII to SPII involves the following:

a) Recognizing *entities* (referents) in APII.

b) Separating the proprietor from other referents.

c) Identifying the relationship (e.g., has) between the proprietor and other entities.

d) Constructing two types of information: 1) information in which the proprietor is the subject; and 2) information in which other entities are the subjects.

The proposition is intuitively reasonable. It reflects the commonsense notion that a piece of information is about entities in reality that can be classified into different categories. This level of description is not unreasonable, since it is a first attempt at pursuing such a difficult semantic notion as sensitivity. The method of discerning "meaning" within APII is a development in the right direction, even if the formal methods of PII analysis and practical applications of the analysis are not readily apparent.

Note that this reduction process aims to identify what makes PII sensitive information. Eliminating non-PII focuses the attention on the proprietor's role in the information.

We justify our descriptive approach on the grounds that the whole concept of digging up the semantic roots of privacy in linguistic texts seems to have no background on which we can build. Even standard semantic analysis appears to depart in different directions, as discussed next.

In language studies, the process of determining different entities in an assertion starts with identifying "nouns (determiners)". In **Figure 2**, we give an example of a standard semantic analysis of a typical sentence. The two circles denote our reduction to self-PII and non-PII.

Standard semantic analysis does not pay attention to privacy-based units and uses the Verb Phrase and the Prep Phrase to form a Verb Phrase, then to form a sentence (dotted lines in the figure). Consequently, our approach is to draw a descriptive road map in order to see the total picture by pursuing privacy-based sensitivity semantics.

In pursuit of further privacy-based syntactical units, we notice that self-PII can also embed special types of information.

Definition: The PII q is a *singleton SPII* (SSPII) if it is an SPII such that its proprietor is the only entity in q; otherwise, SPII is called a *multitude SPII* (MSPII). A singleton *SPII* has the general structure (subject attribute object) where the object is an *aspect* (*property*, *character*, etc.) of the proprietor. A multitude SPII has the general structure (subject predicate object) where the object is not an *aspect*.

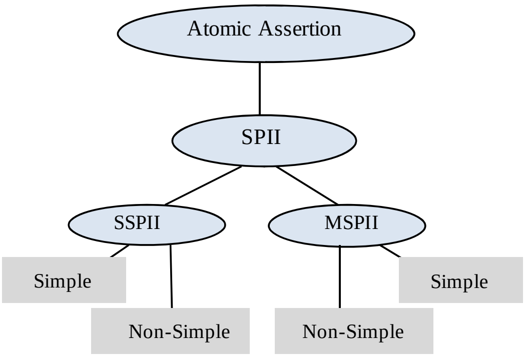

Figure 2. A categorization of different types of atomic PII.

Proposition: Every self-APII is reducible to a set of singleton SSPII and multitude MSPII.

Discussion: The implication in this proposition is that we try to deduce from any given self-APII as many embedded singleton SPIIs as possible. Any SPII is, by default, a proprietor with possible other non-person objects; thus, the proposition is reasonable. Note that a singleton SPII simply associates some aspect with the proprietor (e.g., tall, short). Thus, one part of our goal of spotting sensitive meaning is made easy as the sensitivity question is reduced to: what features of a proprietor are more sensitive than other features (e.g., *John is a quiet person* vs *John is a nervous person*)?

We can further explore the structure of self-APII to identify more primitive types of linguistic structures.

Definition: A simple SPII is information that includes a single SPII. In other words, a simple SPII is a simple proposition, where a proposition is a claim about a subject expressed as an assertion.

Figure 2 shows a classification of different types of PII discussed so far.

Proposition: Every singleton or multitude SPII is reducible to a set of simple PII.

Discussion: The implication in this proposition is that we try to simplify the set of singleton and multitude SPII as much as possible. The process of reduction revolves around the token (e.g., noun) that identifies the proprietor of the SPII and such connections as "and", "or", and so on. Additionally, the process of producing PII from another PII preserves identity. Thus, we reach simple PII when there is a single predicate associated with this identity. Reducing the original piece of PII to a set of simple PII refers to isolating the privacy aspects of the original PII. Whether the resultant set of simple PII is semanticcally equivalent to the original one is an interesting problem, but it is not our concern here. The reduction process of the original PII is rather utilized to identify "privacy centers" and use these "centers" to measure the sensitiveity of PII.

SPII can be a complicated expression. For example, in *Farmer John's house is burning*, we have the *simple* SPIIs *John is a farmer*, and *John has a house*. Moreover, an expression such as *John's business dealings with religious charities raise uncomfortable questions* is considered SPII because John's dealings are his own activity and not something ontologically independent from him. Atomic SPII is "pure" atomic PII that expresses actions or attributes of the proprietor.

Sensitivity in a simple PII can be traced to two sensitivity locations: its verb part or its rest-of-the-PII part, or to both parts. Identifying the source of sensitivity in these simple linguistic units can be complemented with semantic ranking (e.g., verbs: *He* "*taught*" vs "*molested*" *juveniles* and non-verbs: *He engaged in* "*discussion*" vs "*sex*").

Accordingly, we can build a ranking system of information of the type "Someone [stranded, murdered, gave charity to, belongs to, …] some people [14]". The pieces of atomic PII can be value-ranked according to the sensitivity of the verb. We concentrate here on sensitive personal identifiable "information" or "facts" and ignore the issues related to metaphors, clichés, etc.

It is still difficult to provide an automated understanding of unrestricted natural language because of its involved theoretical complexities. For textual information, our methodology introduces the basic principle of the approach. Of course, a number of improvements and refinements can be introduced in order to develop a more sophisticated mechanism. A great deal of research is needed at the linguistic level to develop a "PII Analyzer". Reading text to identify PII is a tedious process since it may be scattered throughout the text. A "PII Analyzer" will assist people in locating and analyzing PII in documents. It will perform various tasks, such as locating all the occurrences of PII in a text, ranking pieces of PII according to their sensitivity, suggesting possible replacements to reduce the level of sensitivity, and so forth. It may be used alone, or it may be connected to a knowledge system.

5. Sensitivity Analysis

Next we investigate the technique of measuring the sensitivity of a given APII. This problem and the approach we describe have been described in [20]; here we apply it to English texts, instead of to Arabic as in that paper [20].

Simple SSPII and MSPII are linguistic constructs that express an aspect or a relationship about or involving a proprietor. We view each of these constructs as a three-part structure: 1) the subject: represents the proprietor, 2) the predicate (verb): denotes the action, and 3) the extension (remainder): denotes other parts beyond the subject and predicate. Such a categorization parallels the traditional grammatical elements: noun, verb, and particle. For example, in the MSPII: *John lives in a house* we have: subject: *John*, predicate: *lives*, and extension: *a house*. Other examples are the SSPII: *John thinks*, *John is beaten*, *John works in the city*, *the bat hit John*, etc. Note that the extension (the rest-of-PII) is optional. The subject (the agent in the semantic/linguistic—not grammatical—sense) of each of these SSPIIs is John; each describes his physical, mental, or emotional existence and welfare. The predicate is always a "sign" of something said of the (human/individual) subject (proprietor).

Note that the verb is a loaded concept and its sensitivity depends on the context (other parts of the sentence). This may require a second-order evaluation of the sensitivity of the word in the presence of other words. For example, the verb *place* may indicate *to put in or set*, *to assign a position in a series*, *to give* (*an order*) *to a supplier*, … At this stage of our research we will manually assign the initial value of sensitivity of a verb, so it is possible that *to place* may have two different values in the same text. In this case there is a merging of the verb and that part of the context that affects the verb. For example, *placed in a dangerous and vulnerable position of harm* will be assigned a sensitivity value of *endangered*, *might be harmed*.

Accordingly, sensitivity can be traced to these three substructures: a) proprietor, d, b) predicate, c, and c) remainder, r. Thus, given a simple SPII x, we use the equation:

$$S(x) = \alpha S(d) + \beta S(c) + \gamma S(r) \qquad (1)$$

where S(d), S(c), and S(r) are values in the interval [0,10] of the sensitivity of d, c, and r, respectively. The factors α, β, and γ in (1) represent different weights associated with the terms since it is possible that the three parts do not contribute equally to S(x). We ignore the possibility of nonlinearity. Consider *John likes pornography*; we expect *pornography* to have a high sensitivity value close to 10. In *John lies about his age*, we expect *lies* to have a higher sensitivity value, and in *Ayman al-Zawahiri* (*al Qaeda leader*) *is hiding in Afghanistan*, we expect *Ayman al-Zawahiri* to be the sensitive part of the sentence. Our strategy is to develop a self-learning system with given initial sensitivity values that are tuned according to previous use; thus, as a first approximation of the sensitivity of S(x), we assume that the three parts of the simple self-APII contribute equally to S(x); *i.e.*, $\alpha = \beta = \gamma$. In general this may not be exactly true.

Alternatively, S(x) can be calculated as:

$$S(x) = \alpha S(d) \cdot \beta S(c) \cdot \gamma S(r) \qquad (2)$$

Several factors may give justification to Equation (2) including inter-parts sensitivity. For example, the act of a *celebrity* (sensitive proprietor) who *steals* (sensitive action) is multiply sensitive in comparison with the same act performed by an ordinary person. We plan to experiment with Equation (2) to compare the results with Equation (1) in implementation and testing.

The basic steps performed by the system are as follows:

Input: Simple SPII

1) Break the sentence down into three parts: proprietor, verb, and remainder-of-PII.

2) Calculate the sensitivity value of each part.

3) Calculate total sensitivity.

Figure 3 shows a general view of the proposed mechanism that calculates the sensitivity value of a given PII, x. In the figure, *f* represents a function associated with the three substructures used to calculate the sensitivity value of that part.

The function *f*(c) for a given verb can be realized by use of a look-up table. Similar tables are utilized for *f*(d) and *f*(r). The values in the look-up table can be selected according to commonsense criteria such as effects, culture, and celebrity. Associations among words may be a significant factor. For example, *Alice felt naked* is less sensitive than *Alice walked naked*. This is a second level of sensitivity calculation that will be pursued in future study.

6. Experimentations

In this section we experiment with three textual documents for the purpose of identifying pieces of PII and locating sensitivity centers. We will use both Equations (1) and (2) in our analysis. The sensitivity is presented in terms of a graphical map showing proprietors, their aspects, and relationships to other individuals and non-individuals.

6.1. Psychiatric Report

The following medical record belongs to a psychiatric' patient named Lucy [21]. Pieces of PII are numbered in the text and underlined words will be examined for levels of sensitivity.

Since receiving the diagnosis of neural tumor (1), "*Lucy*" has felt depressed and anxious about her health (2). Lucy has experienced two nights of restless sleep (3). She has lost enthusiasm for her usual activities, such as going shopping (4) and taking care of her son "Tim" (5). She reports having no energy for maintaining her work or social life (6). She has also become more irritable and aggressive (7), which is putting additional pressure on her family (8). She admits to being preoccupied with thinking about her illness (9) and is having trouble concentrating on daily activities (10). She reports feeling tired (11) but too scared to sleep for fear that she will not wake in the morning (12). In conjunction with her depressive symptoms, Lucy is also experiencing excessive anxiety (13).

Her anxiety is associated with restlessness, tiredness, irritability, insomnia, and difficulty in concentrating (14). Other symptoms include palpitations, tachycardia and flushing (15). Lucy expresses concern over the impending biopsy report (16), due sometime in the next two days, asking "Am I going to die? Does the tumor mean cancer?" (17). Lucy also expresses concern over her son's welfare while she is hospitalized (18). In the last month, her fiancé James and her mother Hermione have been looking after both her and her son (19).

In the following discussion, we assume persons are not celebrities or public figures, so we ignore the proprietor's sensitivity and leave it with the value of 1. Also, the weight factors α, β, and γ are assigned a value of 1 in this first experiment. **Table 1** quantifies the level of sensitivity of the verb (in the range of 0 - 10) and the rest-of-PII portion. These sensitivity levels are assigned manually. In further research work, we plan to have such assignment be initial values for a self-learning knowledge system that stores sensitivity values for words and phrases.

In **Table 1**, we have estimated the level of sensitivity by using Equations (1) and (2) to evaluate verbs, rest-of-PII, and overall sensitivity levels. From such a table we draw **Figure 4**, a conceptual map of different types of information. We call these maps *sensitivity spheres*.

Figure 4 depicts the PII sphere of Lucy from different sides: CPIIs, and (simple) SSPIIs and MSPIIs. One objective of our research is to produce such diagram for, say, privacy officers to be used as a visual aid in deciding various sensitivities (hence security level) of PIIs. For example, it may be decided that the MSPII has the highest sensitivity since it includes {a report about *Lucy has a tumor*}.

The sensitivity of descriptive aspects of Lucy as described by the set of SSPII may be sensitivity ranked as shown by the large darkly graded bracket enclosing *hospitalized* to *concentrating*. The information {hospitallized, aggressive, depressed} may be restricted (e.g.,

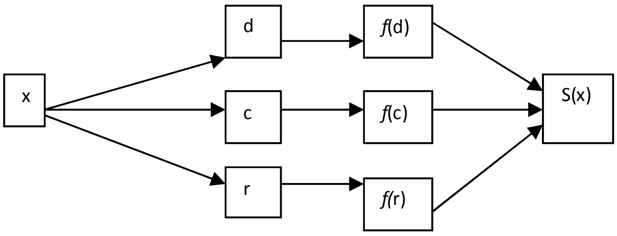

Figure 3. Calculated sensitivity mechanism.

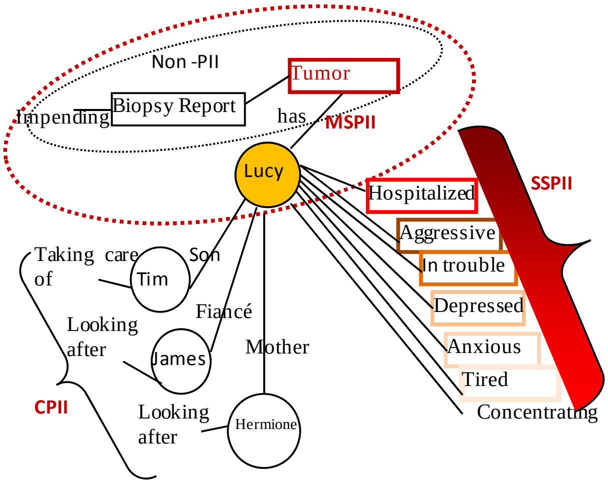

Figure 4. Lucy's PII sphere.

Table 1. Levels of sensitivity of the psychiatric report.

PII #	Verbs	Verbs Sensitivity	Rest-of-PII	Rest Sensitivity	Combined Sensitivity Equation (1)	Combined Sensitivity Equation (2)
1	receive	3	diagnosis of a tumor	9	12	27
2	feel	5	depression, anxiety	5	10	25
3	experience	1	restless sleep	10	11	10
4	lose	6	enthusiasm	4	10	24
5	take care	1	caring for her son	2	3	2
6	having	1	maintain her life	5	6	5
7	becomes	3	more aggressive	8	11	24
8	put	1	pressure on family	10	11	10
9	admit	6	about illness	6	12	36
10	having, concentrating	4	effect on daily activities	2	6	8
11	report	2	tired	8	10	16
12	scared, sleep	6	dying during sleep	5	11	30
13	experience	2	excessive anxiety	9	11	18
14	associate	1	symptoms on her behavior	10	11	10
15	include	2	physical symptoms	10	12	20
16	express	5	impeding report	1	6	5
17	ask	3	tumor and cancer	10	11	30
18	express	7	son's welfare	4	11	28
19	look after	2	family members	7	9	14

access) more than the rest of Lucy's aspects.

Further examination of **Table 1** shows that Equation (2) is more suitable for "contrasting" the levels of sensitivity. In row 9, "admit" and "about illness" raise the sensitivity to 36, greater than other PIIs. Equation (1) gives it 12, high sensitivity, but not far from other PIIs. In row 17, using Equation (2) with "tumor and cancer" yields a high sensitivity of 30, greater than most of the other PIIs.

Equation (1) results in a high sensitivity level of 11, a value shared with many other PIIs. So, it seems that Equation (2) is more suitable for sensitivity analysis because it magnifies extensive sensitivity.

6.2. Medical Examination

This example is a medical report for a patient suffering from stiff muscles [22]. The medical report is as follows.

Musculoskeletal Exam: A 36-year-old patient "Amy" presents with a stiff neck (1) that has affected her for two days (2).

History of symptoms:

- **Chief complaint**—Neck pain, radiating from the left shoulder, often causing a headache (3).
- **History of present illness**—Patient woke yesterday morning with a stiff neck; it grew more painful throughout the day (4). She woke today with a severe headache (5) and her shoulder was affected (6).
- **Review of systems**—She had normal blood pressure; no pain in arms; allergic to penicillin (7).
- **Constitutional:** She had a temperature of 99; blood pressure 120/80; weight 140 (8).

Musculoskeletal Examination Details:

1) Examination of the patient reveals a limited range of motion and neck is tight on the left side (9).

2) She remarks on severe pain on neck palpation (10).

3) Patient cannot raise her left arm above her head (11).

4) Patient's spine appears properly aligned (12).

5) Patient's lower extremities are not affected (13).

Neurological: Patient seems agitated and stressed (14), and states that her lack of sleep the prior evening has triggered a "blue" mood (15).

Medical Decision: Patient's neck and shoulder pain most likely caused by a pulled muscle (16). No sign of nerve or spinal irregularities. Patient should return in three days for more testing (17) if the pain and/or headaches do not subside. The physiatrist "Dr. Jones" prescribes a painkiller for the patient (18) and asks her to call in three days to report her progress (19).

Table 2 shows the statistics for this medical report. Assume that Amy is a public figure leading to S(d) = 5, and assume that the weight factors α, β, and γ are set to 1. **Figure 5** depicts Amy's PII sphere with all other aspects, objects, and other individuals.

Because in this example Amy is a celebrity, we see a

Table 2. Levels of sensitivity of the medical examination report.

PII #	Verbs	Verbs Sensitivity	Rest-of-PII	Rest Sensitivity	Combined Sensitivity Equation (1) $\alpha, \beta,$ and $\gamma = 1$, S(d) = 5	Combined Sensitivity Equation (2) $\alpha, \beta,$ and $\gamma = 1$, S(d) = 5
1	present, with	2	stiff neck, 36 old	4	11	40
2	affect	5	lasting for two days	3	13	75
3	radiating , cause	5	neck pain from shoulder, headache	5	15	125
4	wake, grow	2	yesterday, stiff neck through the day	3	10	30
5	wake	3	severe headache	5	13	75
6	affect	5	her shoulder	2	12	50
7	have	1	blood pressure, allergic	6	12	30
8	have	1	vital signs and status	8	14	40
9	have, is	1	limited neck movement	6	12	30
10	remark	2	severe pain in neck	6	13	60
11	cannot	2	limited arm movement	5	12	50
12	appear	3	proper and aligned	3	11	45
13	aren’t	2	lower extremities	4	11	40
14	seem	1	agitated and stressed	7	13	35
15	has, trigger	2	mood changes	4	11	40
16	cause	2	pain source “pulled muscle”	5	12	50
17	shall, testing	2	return for tests	3	10	30
18	prescribe	2	medication	5	14	70
19	ask, call, report	3	patients’ status	4	12	60

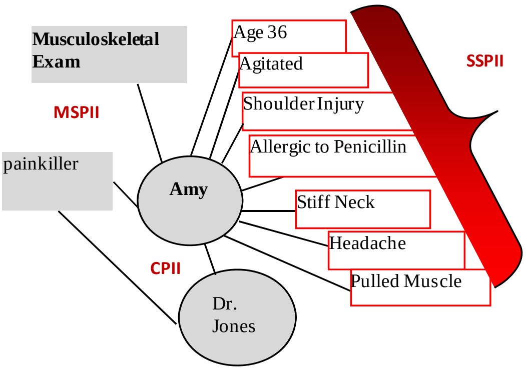

Figure 5. Amy’s PII sphere.

large difference in sensitivity levels, especially using Equation (2). This is a clear indication that when we combine the sensitivity of the verbs and rest-of-PII, and the proprietor is a public figure, we get more magnified differences.

The new features in this experiment and in **Figure 5** are as follows:

- Suppose that celebrity identification is 10 times more sensitive than identification of an ordinary person. Then, again, Equation (2) would greatly magnify the sensitivity of PII. For example, in row 17, news of the celebrity Amy admitted for medical testing would be 5:10 as sensitive as that of an ordinary person. Equation (2) produces 30:6.
- The CPII embeds multi-sensitivities: the relationship of Dr. Jones with a celebrity, the SSPII *Amy takes painkiller*; the SSPII *Dr. Jones prescribes painkiller*, and the total sensitivity of associating Amy, Dr. Jones, and the painkiller.

6.3. Lawsuit Report

This example is part of a legal text [23] about a case of murder of Maria Teresa, who was killed by her estranged husband as a direct result of neglect by sheriffs’ deputies. The lawsuit is as follows:

For more than a year prior to her murder on April 15, 1996 (1), Maria Teresa was repeatedly dismissed, ignored, and even ridiculed by employees and supervisors of the Sheriff’s Department (2) and as a direct consequence, was placed in an increasingly dangerous and vulnerable position of harm from her estranged husband (3). Specifically, in just the last three months of her life, between January 15, 1996, and April 15, 1996, Maria Teresa made at least twenty different and distinct reports and pleas for help and protection to the Sheriff’s Department (2).

Many of these reports were witnessed by others (3). Some of these reports were supplemented by witnesses who independently described Avelino’s conduct, including his threats to kill (4). These reports included descriptions of Avelino's continuous stalking, which is a felony when a restraining order is in effect or when the stalking is repeated (5). Often, Defendant deputies responded to Maria Teresa's home, and were shown the restraining order with its narrative of physical and sexual abuse, spoke with her in person at the Defendant's substation, or spoke on the phone with her (6). Despite the repeated proofs and warnings, the Defendants reacted with dismissiveness, disdain, and obstruction (7).

We have numbered PIIs and underlined the verbs in the text. The results of our examination of the PIIs are shown in **Table 3** and **Figure 6**. In this experiment we assume that:

- Maria Teresa is not a public figure, making S(d) = 1.
- Since this is an investigative report regarding a murder (an act), the weighting factors are set as $\alpha = 1$ and $\beta = \gamma = 1.5$ to emphasize the verb and its description.

As mentioned, we plan such a manual initialization to be a starting point for a semi-automated learning system that will revise the initial values. This system may be adjusted to match actual and perceived privacy concerns that may arise. The system's ultimate goal is to give a relatively real sense of the level of sensitivity.

Figure 6 shows Maria's PII sphere according to this text. The sphere has rich CPIIs between Maria and Avelino. CPII sensitivity is a function of both proprietors' sensitivity. Since the verb and the non-proprietor part of the CPII are also highly sensitive matters in this case, the system ought to give this area of the sphere high sensitivity marks. If this is not public information, then the consent of both proprietors (and their families/attorneys)

Table 3. Levels of sensitivity of the legal report.

PII #	Verbs	Verbs Sensitivity	Rest-of-PII	Rest Sensitivity	Combined Sensitivity Equation (1) $\alpha = 1$, β and $\gamma = 1.5$, S(d) = 1	Combined Sensitivity Equation (2) $\alpha = 1$, β and $\gamma = 1.5$, S(d) = 1
1	murder(ed)	10	April 15, 1996	5	18.5	75
2	dismissed, ignored, and even ridiculed	6	repeatedly, by employees and supervisors of the Sheriff's Department	5	14.5	45
3	placed (in dangerous and vulnerable position of harm)	8	from her estranged husband	6	18	72
4	made at least twenty different and distinct reports and pleas	2	reports	4	9	12
5	were, describe	2	Avelino's conduct	6	15	18
6	include, repeat	3	Avelino's stalking	8	16	36
7	respond, were, speak	3	defendant's meeting Maria Teresa	5	11.5	22.5
8	repeat, respond	3	proofs and warnings	9	17.5	40.5

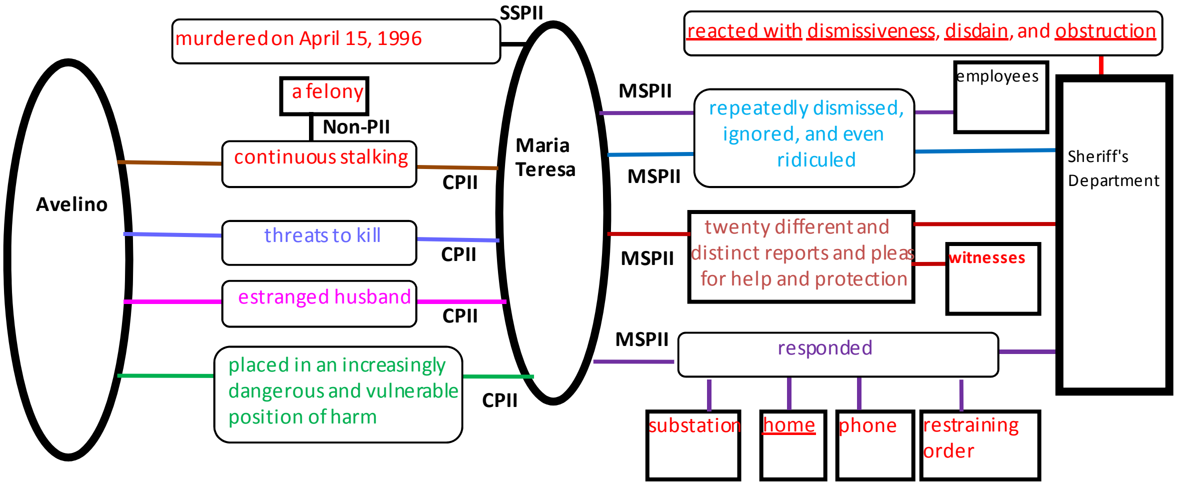
SSPII
murdered on April 15, 1996
a felony
Non-PII
continuous stalking
CPII
Maria Teresa
Avelino
threats to kill
CPII
estranged husband
CPII
placed in an increasingly dangerous and vulnerable position of harm
CPII
reacted with dismissiveness, disdain, and obstruction
MSPII
repeatedly dismissed, ignored, and even ridiculed
employees
MSPII
Sheriff's Department
twenty different and distinct reports and pleas for help and protection
MSPII
witnesses
MSPII
responded
substation
home
phone
restraining order

Figure 6. Maria's PII sphere.

may be required before releasing this information.

Note that identifying APIIs and CPIIs draws the boundaries of the privacy rights of proprietors. The CPII is a shared privacy territory that is “owned” by its proprietors. Both proprietors have equal rights of ownership of their CPII, and our reduction does not reduce this important aspect of the dual nature of the ownership.

Note that a very sensitive part of the diagram is the non-PII accusations that the sheriff’s department reacted with dismissiveness, disdain, and obstruction of complain. But this is not a privacy-based sensitivity. This information is definitely within the boundary of free speech (press). Other parts of the sphere may need anonymization before handling.

The difference between SSPII and MSPII is of some importance. The MSPII may refer to a group of people (e.g., employees, witnesses) and institutions (substation) that may object to engaging their identification in any handling of such PII. On the other hand, the SSPII is the sole “property” of its proprietor, and no other individual may claim it in any circumstance.

7. Conclusions

This paper proposes a theoretical approach to the notion of PII sensitivity. The methodology is a road map for developing a classification of PII sensitivity, starting from atomic PII, in order to identify pivots of sensitivity that reflect the significance of PII. Some PII expresses a relationship among persons, some specifies aspects and features of a person, and some describes a relationship with non-human objects.

The paper also experiments with actual PII texts that are analyzed to capture different types of PII sensitivities. Results point to a possible semi-automatic system that evaluates levels of sensitivity in these texts.

From the experiments, we can conclude that a portion of PII sensitivity can be calculated based on context-free analysis. Decomposing PII into privacy-based portions helps in factoring out non-PII information and focusing on a proprietor’s related information.

Building a visual map of the privacy sphere can be used to approximate the sensitivity of different territories of the privacy-related text. The sensitivity levels can be adjusted to match certain criteria and context of the given text document.

REFERENCES

[1] P. M. Schwartz and D. J. Solove, “The PII Problem: Privacy and a New Concept of Personal Identifiable Information,” *New York University Law Review*, Vol. 86, 2011, p. 1814.

[2] DHS, “Handbook for Safeguarding Sensitive Personally Identifiable Information,” The Privacy Office, US Department of Homeland Security, Washington DC. www.dhs.gov/privacy_

[3] M. McMeekin, “Personally Identifable Information (PII) Incident Security Guidelines,” Northern Rockies Coordinating Group NRCG, 2010

[4] J. Yakowtiz, “Tragedy of the Data Commons,” 2011. www.papers.ssrn.com/sol3/papers.cfm?abstract_id=178979

[5] P. Ohm, “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization,” *UCLA Law Review*, Vol. 57, 2010, p. 1701. www.epic.org/privacy/reidentification/ohm_article.pdf

[6] J. Yakowitz, “Space Invaders: Intrusion in the Digital Age,” 2012. http://ssrn.com/abstract=2019079

[7] “W3C Platform for Privacy Preferences.” www.w3.org/p3p

[8] S. Lederer, C. Beckmann, A. Dey and J. Mankoff, “Managing Personal Information Disclosure in Ubiquitous Computing Environments,” 2003. http://www.intel-esearch.net/Publications/Berkeley/070920030922_139.pdf

[9] L. Bygrave, “Data Protection Law,” 2002. http://www.michaelkirby.com.au

[10] J. Bing, “Classification of Personal Information, with Respect to the Sensitivity Aspect,” *Proceedings of the First International Oslo Symposium on Data Banks and Societies*, Oslo, 1972, pp. 98-150.

[11] R. Lee, “Personal Data Protection in the Semantic Web,” 2002. http://www.w3.org/2002/01/pedal/thesis.html

[12] Clearing House Privacy Rights Clearinghouse, “Chronology of Data Breaches Security Breaches 2005-Present,” 2005. http://www.privacyrights.org/ar/ChronDataBreaches.htm

[13] P. Fule and J. Roddick, “Detecting Privacy and Ethical Sensitivity in Data Mining Results,” *Twenty-Seventh Australasian Computer Science Conference* (*ACSC* 2004), Dunedin, 2004. http://crpit.com/confpapers/CRPITV26Fule.pdf

[14] S. Al-Fedaghi, “How Sensitive Is Your Personal Information?” *The* 22*nd ACM Symposium on Applied Computing (ACM SAC* 2007), Seoul, 11-15 March 2007, pp. 165-169.

[15] J. Kang, “Information Privacy in Cyberspace Transactions,” *Stanford Law Review*, Vol. 50, No. 4, 1998, pp. 1193-1294. doi:10.2307/1229286

[16] L. Floridi, “Information Ethics: On the Philosophical Foundation of Computer Ethics,” *ETHICOMP*98 *Fourth International Conference on Ethical Issues of Information Technology*, 1998. http://www.wolfson.ox.ac.uk/~floridi/ie.htm

[17] S. Al-Fedaghi, “Crossing Privacy, Information, and Ethics,” 17*th International Conference Information Resources Management Association* (*IRMA* 2006), Washington DC, 21-24 May 2006. www.irma-international.org/viewtitle/32702/

[18] S. Al-Fedaghi, “The Ethics of Information: What Is Valued Most,” *The Open Ethics Journal*, Vol. 3, No. 1, 2009,

pp. 118-126. doi:10.2174/1874761200903010118

[19] S. Al-Fedaghi, "Personal Information Ethics, Encyclopedia of Information Ethics and Security," M. Quigley, Ed., Information Science Publishing, Hershey, 2007, pp. 513-519. doi:10.4018/978-1-59140-987-8.ch076

[20] S. Al-Fedaghi and Fl. Al-Haqan, "Privacy Sensitivity: Application in Arabic," *International Conference on Asian Language Processing* (*IALP* 2009), Singapore, 7-9 December 2009.

[21] National Institute of Health, "A Sample Health Record." http://www.nlm.nih.gov/medlineplus/magazine/issues/summer09/articles/summer09pg17.html_

[22] Supercoder.com, "Avoid under Coding: Billing for High-Level, Single-System E/M Services." http://www.supercoder.com/articles/articles-alerts/pmc/avoid-undercoding-billing-for-high-level-single-system-em-services/_

[23] Women's Justice Center, "The Maria Teresa Macias Case." http://www.justicewomen.com/macias_case_as_filed.html