PASM: A Policy Aware Social Miner

by

Sharon Myrtle Paradesi

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2013

Author ....
Department of Electrical Engineering and Computer Science
January 23, 2013

Certified by ....
Lalana Kagal
Research Scientist, Computer Science and Artificial Intelligence Lab
Thesis Supervisor

Accepted by ....
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Theses

PASM: A Policy Aware Social Miner

by

Sharon Myrtle Paradesi

Submitted to the Department of Electrical Engineering and Computer Science
on January 23, 2013, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

The Policy Aware Social Miner (PASM) project focuses on creating awareness of how seemingly harmless social data might reveal sensitive information about a person, which could be potentially abused. It seeks to define good practices around social data mining. PASM allows people to create policies governing the use of their personal information on social networks. Using linked data, PASM semantically enhances the usage restrictions to ensure that potentially sensitive information is identified and appropriate policies are enforced. PASM also enables people to provide refutations for other information about them that is found on the Web. PASM encourages consumers of social information on the Web to use the mined data appropriately by enforcing data policies before returning the search results. PASM provides a solution to the following issue of privacy in social data mining - although people know that searches for data about them are possible, they have no way to either control the data that is put on the Web by others or indicate how they would like to restrict use of their own data. In a user study conducted to measure the performance of PASM in identifying sensitive posts as compared to the study participants, PASM obtained an F-Measure of 84% and an accuracy of 80%. Interestingly, PASM demonstrated a higher recall than precision, a property that was valued by the study participants as all but one participant indicated that they would prefer receiving false positives rather than false negatives.

Thesis Supervisor: Lalana Kagal
Title: Research Scientist, Computer Science and Artificial Intelligence Lab

Acknowledgments

I sincerely thank Prof. Hal Abelson for funding this research. I also thank him and Dr. Lalana Kagal for guiding my research in the right direction. Their technical expertise and support were invaluable. I especially thank Lalana for taking the time to proof-read this thesis. I also thank Dr. Joe Pato for providing input during the design of the user studies, guiding my interpretation of the results, and proof-reading this thesis. I thank Dr. Ilaria Liccardi for providing help with the design of the user study and for monitoring the first few interviews. I also thank Fuming Shih, Oshani Seneviratne, Ian Jacobi and other members of the Decentralized Information Group for their input over the course of this project.

Last, but not the least, I thank my father, Suresh Paradesi, mother, Dr. Mary Gollapalli, nephew, sister-in-law and brother, Matthias, Esther, and Martin Paradesi for their unconditional love, support and encouragement. Above all, I thank my God for His grace, mercy, and love towards me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the overarching problem and provides an overview of the thesis.

1.1 Problem Description

There is a large amount of personal data available on the Web about individuals that is generated by themselves and others. Furthermore, an increasing number of people on the Web are generating data about themselves and others that they consider to be sensitive. This sensitive data can be used negatively against them. For instance, a corporation or a person can use that data to decide a (potentially negative) course of action concerning that individual. For example, it is becoming fairly common for employers to use Web search tools (such as SocialIntelligence [13]) to gather information about a potential hire.

The severity of this phenomenon is highlighted by the Mosaic Theory [31], which is one of the primary motivators of this thesis. The Mosaic Theory states that seemingly harmless pieces of information could potentially reveal a damaging picture when pieced together. Individually, the pieces may be of no value because they do not reveal anything particularly significant or dangerous about the person. However, when the pieces are viewed together, the mosaic may present a remarkably different picture of the person. The process of forming a mosaic of someone is very easy to do using the Web because of the availability of various online services like search engines, social networks, etc. It is observed that people usually share information within a specific context or purpose. The problem is that the contextual integrity [28] of that information cannot be captured. People have no way to express these expectations. Although people know that searches for data about them are possible and that they can easily be profiled on the Web, they have no way to either control the data that is put on the Web by others or indicate how they would like to restrict usage of their own data.

This thesis attempts to provide a policy-based mechanism for privacy-aware mining of personal data on the Web. The system implementation, Policy Aware Social Miner (PASM), enables both creation and enforcement of policies that declare the intended use of data. This way, PASM enables people to control the "mosaic" that they present about themselves on the Web. The thesis also describes a user study that was designed to measure the performance of PASM in identifying sensitive posts as compared to the study participants.

1.2 Thesis Overview

The rest of this thesis is structured as follows.

- **Chapter 2** gives an overview of PASM by outlining the assumptions, motivating scenarios and potential solutions offered by PASM.
- **Chapter 3** demonstrates policies that one can create for the motivating scenarios, the system architecture and technical implementation of PASM, and the algorithm for linked data traversal that is used during the semantic enhancement process.
- **Chapter 4** describes the user studies conducted to measure the performance of PASM and the results obtained.
- **Chapter 5** describes research endeavors related to PASM and how PASM extends the current state of the art.
- **Chapter 6** concludes the thesis and discusses future work.

Chapter 2

Policy Aware Social Miner – Overview

This chapter provides an overview of Policy Aware Social Miner (PASM). It starts with the assumptions that were made when designing the system, two motivating scenarios, and a high-level discussion about the policies in PASM.

2.1 PASM Users

The users of PASM can be broadly categorized into the following two groups:

- *Data subjects:* These are people who generate data about themselves on the Web using social networks, blogs, and other services. Data about these people may also be generated by others on the Web.
- *Data consumers:* These are users who wish to conduct a Web search about a data subject with a specific intent (employment, insurance, etc.).

2.2 Features offered in PASM

Following are the features offered by PASM to the data subjects and data consumers.

- **Data Subjects:** PASM provides a platform for data subjects to annotate data about them on the Web in the following three ways:

1. If they own certain data, they can attach usage restrictions to specify their intended use for that data.
2. If they do not own certain data and wish to present a counter argument, they can write a statement refuting the implications of that data and link the statement to the data. Such a statement is referred to as a*refutation* in this research.
3. They can request and provide third-party refutations. A *third-party refutation* is a statement that someone other than the data subject provides for the data subject. A data subject can choose to verify or decline these statements.

- **Data Consumers:** PASM provides a platform for data consumers to search for data about Data Subjects in the following two ways:

  1. It enables them to participate in policy-aware searches about data subjects. After receiving the search terms and the intent of their search, PASM fetches data about the data subject and enforces the policies of the data subject before returning the results. PASM provides data consumers with a view of the complete story of an event or incident by highlighting the refutations created by the data subject.
  2. It gives them access to potentially more content about data subjects than what is publicly available about them on the Web in a policy-aware manner. It is possible that data subjects may make more data available through PASM because they would know that their usage restrictions and refutation links would be enforced during searches for their data.

2.3 Core Assumptions

In order to keep the design of PASM tractable, we made the following assumptions.

About Data Consumers: PASM makes the following assumptions about data consumers.

- They use PASM to search for data about a specific person with a specific intent. In other words, they do not use PASM for a casual Web search.

- They are willing to provide a truthful intent (purpose of search) that accurately reflects their reason for performing that search.
- If there are documents on the Web mentioning a data subject in a negative manner, they are interested in finding out both sides of the story by viewing the refutations created by the data subject.

Additional Assumptions: In addition to the assumptions mentioned above, the following assumptions were made about both types of the users (data consumers and data subjects) and the system (PASM) in general.

- Both the data consumers and the data subjects are willing to be identified using an authentication mechanism when using PASM. Further, they are not malicious users.
- PASM is a closed world system. The policies created in PASM would not benefit the data subjects if the data consumers choose to conduct a search for their information using other systems (such as Google, Bing, or other Web search engines).
- The searches conducted using PASM are logged for accountability.

2.4 Motivating Scenarios

In this section, two motivating scenarios are discussed, followed by a high-level discussion of the potential solutions provided by PASM. The protagonist, named *Alice Metzer* uses PASM as a data subject. Carol is Alice's mother and Danny is Alice's friend. Danny also happens to be use PASM as a data subject. Lastly, Bob is a medical insurance agent.

2.4.1 Data Originating from Alice

Carol recently had a heart attack and Alice's friends and relatives who are friends with Alice on Facebook post on her Facebook wall to inquire about her mother's health. They also inquire about the treatment process and recommend nearby cardiologists. Although it is her mother who has the medical condition, Alice fears that anyone who looks at these posts without knowing the context may mistakenly infer that Alice has heart problems. This would be especially problematic if that person happens to be a medical insurance agent, as this information could negatively affect her medical insurance.

2.4.2 Data Originating from Other Sources

Alice was involved in a protest that resulted in some arrests. However, she recently came across a news article that incorrectly stated that she was arrested though she was not. As she does not have permission to change the contents of that news article, she describes her side of the story on her blog. However, she is concerned that her blogpost may not be visible prominently in the search results of the major Web search engines.

2.5 Potential Solutions for the Scenarios

There are several ways to address the scenarios discussed in the previous section. For instance, an obvious solution for the problem highlighted in the first scenario would be to restrict access to the Facebook posts that Alice considers as being sensitive. However, not all users are aware of or inclined to set the necessary permissions on various social networking sites. It does not help that some social networking sites update their privacy controls on a regular basis. Even if a data subject is scrupulous about setting proper privacy controls on Facebook, it does not guarantee proper privacy controls on all data on Facebook about her. For instance, if a friend tags the data subject in a photo, then the privacy controls applied to this photo are different from what the data subject would normally expect. Therefore, there is a need to protect the data subjects from *misuse* rather than *access violations.*

The following policy-based solutions are provided to the data subjects: *usage restrictions* and *refutation links.* Usage restrictions declare the intended use of applicable data. These differ from security policies which are based on authentication and access controls. Refutation links declare personal refutations to content on the Web.

2.5.1 Usage Restrictions

In the first scenario involving Carol's heart attack, Alice can create a policy to protect posts on her Facebook account that contain the phrase "heart attack" . PASM would then filter matching posts from results during searches about her. PASM provides the following policy mechanism to attach usage restrictions to posts. The two main components of a usage restriction are: *keywords* and *intent*. Keywords are phrases (for example, "heart attack") that data subjects consider as sensitive for them. Intent denotes the purpose of a Web search that the data subjects wants to protect their data from. Posts that contain the keywords will not be made available to searches performed using specified intents. PASM uses the Respect My Privacy framework to provide the following five types of search intents: employment, commercial, financial, depiction and medical.

Semantic Enhancement: It is unreasonable to expect the data subject to provide a comprehensive list of sensitive keywords for protection. For instance, suppose that Alice has the following post "... is now using Esmolol." on her Facebook wall. Esmolol is a medicine that suppresses abnormal rhythms of the heart. This post, though related to heart attack, will not be filtered because none of the phrases in the post are explicitly mentioned in Alice's policy. As another example, there may be posts on Alice's Facebook wall related to heart attack but not explicitly mentioning those two words. It would be unreasonable to expect Alice to provide a list of terms related to heart attack, medications, and other associated concepts. To solve this problem, PASM semantically enhances Alice's policy to incorporate additional information about "heart attack". The *semantic enhancement* process finds additional information related to the keywords in the data subject's policy.

PASM uses the Linked Open Data [1] ontologies related to medicine to semantically enhance Alice's policies. More specifically, it uses ontologies such as (i) DrugBank [2], which lists drugs and their uses, (ii) SIDER [3], which lists the side effects of various drugs, and, (iii) Diseasome

[1] http://linkeddata.org/

[2] http://www4.wiwiss.fu-berlin.de/drugbank/

[3] http://www4.wiwiss.fu-berlin.de/sider/

[4], which provides information about various diseases. When deciding whether to filter a post based on a policy, PASM first divides the post into *noun phrases*, which are phrases with a noun as their head word. For example, in the following post – "recently suffered from a cardiac arrest", "cardiac arrest" is the noun phrase and PASM would compare it to Alice's semantically enhanced policy. The algorithm that is used to traverse the linked data ontologies is explained in detail in Section 3.3.

2.5.2 Refutation Links

In the second scenario involving Alice's alleged arrest, she can create a link between the news article and her blogpost so that people are properly informed about her side of the incident as well. According to Meriam-Webster dictionary, a *refutation* is the act of refuting, whereby one can prove an allegation wrong by argument or evidence. Usually, a refutation conveys an opposite connotation compared to the article it is refuting. It is important to note that PASM does not restrict or hide refuted documents from the search results. Instead, it aims to provide both sides of the story to a data consumer during a search by providing the data subject's refutation to a document along with that document in the search results. The goal is to prevent the data consumer from arriving at a misinformed conclusion due to a lack of complete knowledge of the incident under consideration.

Third-party Refutations: Suppose that Alice's friend Danny was present with Alice during the protest. Then Danny can potentially provide additional evidence that supports Alice's refutation. Using PASM, Alice can mention Danny as a trusted party in her refutation link. PASM would then inform Danny about Alice's refutation link and Danny can provide a supporting third-party refutation for Alice. PASM considers the trusted party to be a data subject as well. Thus, there are two types of data subjects with respect to third-party refutations: *requesters* and *supporters*. Requesters seek support for their refutation links and supporters provide supporting third-party refutations for requesters. The motivation for third-party refutations is to provide the data consumers with supporting evidences in addition to the data subject's refutation link.

[4] http://www4.wiwiss.fu-berlin.de/diseasome/

Chapter 3

Policy Aware Social Miner – Architecture

This chapter discusses the technical implementation details of PASM. First, the individual components of PASM are explained. Then, a sequence diagram describing the creation of third-party refutations is provided along with a screenshots of the system implementation.

3.1 Implementation Details

This section introduces the data sources accessed, the tools and APIs utilized by PASM, and, the implementation details of the PASM policies.

3.1.1 Data Sources

There are two main types of data sources used in PASM — one for the data that needs to be protected, and another for semantic enhancement. The first type of data source involves data on Facebook and the Web (using the Google Custom Search API [9]). The second type of data source involves the Linked Data ontologies described in section 2.5.1. To utilize these ontologies, the data subject is first asked to enter a resource describing the sensitive term using the DBpedia autosuggest tool when creating a usage restriction. PASM then follows the *sameAs* links from the DBpedia resource to external repositories. A mapping that lists the endpoints (external repositories) where resources of a particular type can be found is defined in PASM. A mapping is specified in N3 notation [11] in the form of triples – ⟨subject predicate object⟩. For example, to state "Boston is the capital of Massachusetts", one would create a triple as shown in Fig. 3-1.

```
dbpedia:Boston dbpedia:capitalOf dbpedia:Massachusetts.
```

Figure 3-1: *An N3 triple indicating that "Boston is the capital of Massachusetts". This snippet assumes that the dbpedia prefix has already been defined in the N3 document.*

An example mapping is shown in Fig. 3-2.

```
dpbedia-ontology:Disease pasm-mapping:correspondsToEndpoint <http://linkedlifedata.com/sparql>.
```

Figure 3-2: *Mapping from the type Disease to the corresponding endpoint. In the DBpedia ontology,* ***heart attack*** *is an instance of type Disease.*

As mentioned in section 2.5.1, the following three ontologies – DrugBank[4], Sider[5], and, Diseasome[3] – are used in PASM. These were made accessible through an endpoint in the linkedlifedata website [1]. Though the current deployment of PASM is restricted to medical domain, PASM can be extended to support other domains by using appropriate ontologies, endpoints, and, their corresponding mappings.

3.1.2 Tools and APIs

PASM makes use of the following five tools and APIs. First, it uses the Facebook API to fetch a data subject's data. Second, it uses Google Search API to search for information about the data subject on the Web. Third, it uses the DBpedia autosuggest tool to enable data subjects to enter "keywords" that are DBpedia resources when creating policies. Fourth, it uses Jena[2], a Java-based framework for building Semantic Web applications, to create, view, update, and, delete the PASM policies. Fifth, it uses NLTK to identify noun phrases present in the data subject's Facebook posts during a data consumer's search.

[1] http://linkedlifedata.com

[2] http://jena.apache.org/

3.1.3 PASM Policies

The policies (usage restrictions and refutation links) in PASM are also represented in N3 notation. In the first motivating scenario involving Carol's heart attack, Alice would create the following usage restriction shown in Fig. 3-3.

```
:alice
        pasm:hasUsageRestriction pasm:usage1;

pasm:usage1
        a pasm:UsageRestriction;
        pasm:hasKeyword
                <http://dbpedia.org/resource/Myocardial_infarction>;
        pasm:hasKeywordLabel
                "Heart Attack";
        pasm:hasRMPRestriction rmp:No-Medical;
        pasm:hasType
                <http://linkedlifedata.com/resource/semanticnetwork/id/T047> .
```

Figure 3-3: *A snippet of Alice's policy demonstrating the usage restriction for the first motivating scenario.*

In the second motivating scenario involving Alice's alleged arrest, Alice first writes a blogpost to refute the news article on *Wicked Local.* She then links this blogpost to that article using a refutation link. In order to highlight eye-witness accounts of the incident, she enlists the help of her friend Danny by providing his Facebook id as a *Trusted Party* when creating her refutation link. At the end of the sequence of actions as explained in section 3.2.1, Alice would have the refutation link shown in Fig. 3-4.

3.2 System Architecture

The architecture of PASM is illustrated in Fig. 3-5. The functional components of the architecture are explained from the perspectives of the interactions that a data subject and a data consumer would have with PASM.

3.2.1 Data Subject's Interaction with PASM

A data subject's interaction with PASM begins when the data subject installs PASM as a third-party plugin using Facebook and logs into the system using Facebook's OAuth

```
:alice
        pasm:hasRefutationLink pasm:refutation1;

pasm:refutation1
        a pasm:RefutationLink ;
        pasm:hasOffendingLink
                "http://www.wickedlocal.com/cambridge/news/x392617482/Cambridge-School..." ;
        pasm:hasRefutingLink
                "http://alicemetzger.com/blog?p=12" ;
        pasm: hasThirdPartyRefutation pasm:tpr1 ;
        pasm:hasTrustedParty "100001504048992" .

pasm:tpr1
        a pasm:ThirdPartyRefutation ;
        pasm:hasContent
                "I met with Alice two days after and can attest that she was, in fact, not arrested." ;
        pasm:hasSupporter "100001504048992" ;
        pasm:hasRequester "815853080" ;
        pasm:hasStatus "Verified" ;
        pasm:hasURI "" .
```

Figure 3-4: *A snippet of Alice's policy demonstrating the refutation link and third party refutation for the second motivating scenario.*

mechanism [8]. After successful authentication using the *Authentication* component, the data subject can use the data subject interface to create, view, update, or, delete his or her policies.

The data subject can create usage restrictions and refutation links using the *Policy Creator.* The usage restrictions are then sent to the *Semantic Enhancer* to gather additional data related to the mentioned keywords. Finally, the usage restrictions and refutation links are stored in a centralized *Policy Repository.*

Although the current implementation is tied to Facebook, one can extend PASM to operate with other social networks such as, Twitter, Flickr, Google+, or LinkedIn. Since data subjects only specifies keyword(s) and intents in the their usage restrictions, these restrictions can be applied to data across multiple social networks. Further, adding authentication mechanisms used by other social networks is non-trivial, but straightforward.

Third-party Refutations: Data subjects (requesters and supporters) interact with PASM as shown in Fig. 3-6 to create third-party refutations.

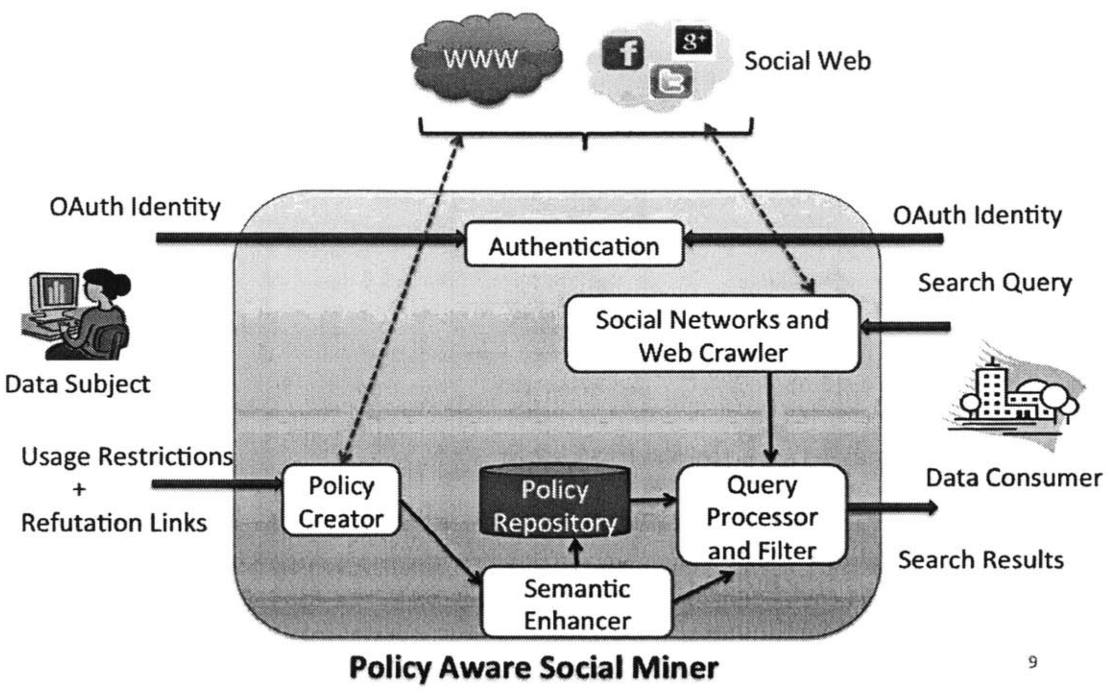


Figure 3-5: *High-level overview of the system. This figure shows how the data subject and consumer interact with PASM and the data sources used in their interactions. The thick arrows show external interaction (of PASM) with the data subject and consumer, while the narrower ones depict the internal flow of control and data. The red, dashed arrows indicate the external (Web and social networks) search performed by PASM.*

First, the requester (Alice) creates a refutation link by providing PASM with the offending link, refuting link and a list of trusted parties. The list of trusted parties is basically a list of Facebook ids of Alice's friends on Facebook who also use PASM as data subjects. Alice trusts these people to provide supporting third-party refutations for her. Second, PASM alerts the supporter (Danny) that Alice mentioned him as a trusted party for the following refutation link. Danny can then provide a third-party refutation in the form of a comment using PASM or write a post elsewhere on the Web and provide that link to PASM. The system (PASM) then takes that comment or URL and stores it (along with the actual refutation) in Alice's policy. It then notifies Alice about Danny's third-party refutation. Alice can choose to either verify or decline Danny's third-party refutation.

When the data consumer (Bob) performs a search on Alice, he would see search results

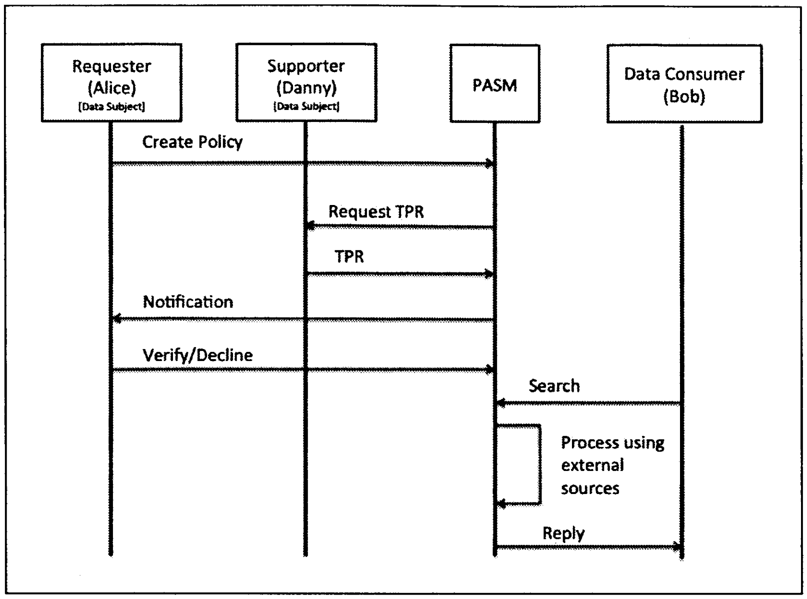


Figure 3-6: *Sequence Diagram showing the sequence of actions performed during the creation of a third-party refutation.*

about Alice returned by the Google Custom Search Engine. However, he would also see Alice's refutation link and Danny's supportive third-party refutation (verified by Alice) as well. Unverified third-party refutations are suppressed.

3.2.2 Data Consumer's Interaction with PASM

A data consumer's interaction with PASM begins when the data consumer installs PASM as a third-party plugin using Facebook and logs into the system using Facebook's OAuth mechanism [8]. After successful authentication using the *Authentication* component, the data consumer can use the data consumer interface to conduct a search using the data subject's Facebook id, search keyword(s) guiding the search, and, the intent of the search. The identity of the data consumer, along with the search keywords and intent are logged for accountability [39].

PASM then searches Alice's Facebook account and the Web using the *Social Networks and Web Crawler* component and looks for data that mention Alice in addition to the keyword(s) entered during the search. It then determines, using Alice's semantically enhanced usage restrictions, whether each result from Facebook should be filtered or not. It also determines, using Alice's refutation links, whether any results returned by the Google Custom Search Engine were refuted by Alice. If they were, the corresponding refutation links and verified third-party refutations (if any) are returned with the results. The *Query Processor and Filter* component is responsible for processing the query and for applying Alice's policy on the search results before returning them to the data consumer.

3.3 Semantic Enhancement of Usage Restrictions

PASM uses the following procedure to understand the keywords provided during the creation of usage restrictions. First, data subjects use the DBpedia autosuggest tool to identify resources that describe keywords that they consider to be sensitive. PASM then uses DBpedia to identify other related concepts. Using Linked Data and mapped endpoints, PASM attempts to identify types of related resources. This is an iterative procedure, which depends on the data subject's input as well as on the endpoint's availability. During each iteration of this procedure, PASM identifies the types of the resources related to the resources already identified as sensitive. It then displays this list of types to the data subjects, who can choose to select the types that they find relevant.

The pseudocode in Fig. 3-7 shows the semantic enhancement procedure, which begins with a list of initial resources that the data subject enters using the DBpedia autosuggest tool. For each resource in this list, PASM first identifies the resources that are similar to it using DBpedia. Once it has a list of similar resources, it then fetches the types of those similar resources and displays this list of types to the data subject. Depending on the data subject's selection of types, it fetches additional resources related to resources of those selected types. This process terminates when the user does not select any additional types, or when there are no new types of resources found at the endpoint. PASM currently makes use of the linkedlifedata.com [10] endpoint for ontologies related to the medical domain. PASM also uses the DBpedia endpoint [6] to find the "sameAs" resources.

Algorithm 3.3.1: USERGUIDEDTRAVERSAL(R)

```
for each $r \in R$
  do {$S = (SAMEAS(r, “http://dbpedia.org/sparql''), MAPPING(TYPE(r)))$
for each $r \in R$
  do { for each $(s, e) \in S$
         do {$T = T \cup TYPE(s, e)$
       { //till no more selections
         show T to user
         for each $t \in T$ selected by user
           do { $\{s1\} = SELECT(s, e)$
                for each $s' \in \{s1\}$
                  do { if $s' \notin R$
                         then { $R = R \cup s'$
                                $T = T \cup TYPE(s', e)$
return (T)
```

Figure 3-7: *UserGuidedTraversal(R): traverses linked data at the endpoint using the list of initial resources (R). This procedure returns a list of types (T) that the data subject selects.*

Algorithm 3.3.2: DIRECTCOMPARISON(P, R, e)

```
for each $p \in P$
  do { for each $r \in R$
         do { for each $n \in p$ //n = noun phrase
                do { if $SELECT(n, r, e)$
                       then { remove $p$ from $P$
                              break //next post
return (P)
```

Figure 3-8: *DirectComparison(P, R, e): filters a post if the noun phrases in the post match content of the resources identified by semantic enhancement. This procedure returns a list of residual, unfiltered posts.*

The pseudocode in Fig. 3-8 shows the comparison procedure that is performed when given (i) the list of Facebook posts of the data subject, (ii) the list of keywords mentioned in the usage restrictions, and, (iii) the endpoints where the similar resources can be obtained. PASM identifies the noun phrases in the posts using NLTK and compares those phrases to the content of linked data resources. If a match occurs, the corresponding post is removed from the list. Finally, the residual list of unfiltered posts is returned to the data consumer.

The individual SPARQL [14] queries mentioned in capital letters in the pseudocode in Figs. 3-7 and 3-8 are explained below.

The SPARQL query in Fig. 3-9 returns a list of endpoints that contain resources corresponding to the types provided in the input.

```
SELECT ?e
WHERE {
        <t> pasm-endpoint:correspondsToEndpoint ?e.
}
```

Figure 3-9: *MAPPING(t): returns the endpoints where resources of type <t> can be found.*

The SPARQL query in Fig. 3-10 returns the objects of the triples from the specified endpoint where (i) the corresponding subject matches the URI of the keyword provided in the input, and, (ii) the content of the object contains the noun phrases provided in the input.

```
SELECT ?o
WHERE
SERVICE <e>{
        <u> ?p ?o.
        FILTER(REGEX(?o, n, "i"))
}
```

Figure 3-10: *SELECT(n, u, e): returns all resources that appear as the object in the triples containing <u> as the subject.*

The SPARQL query in Fig. 3-11 returns the subjects (or, objects) of the triples from the specified endpoint where the corresponding object (or, subject) is the URI of the resource provided in the input.

```
SELECT ?o
WHERE
SERVICE <e>{
        <s> ?p ?o.
}
}
UNION
{
SELECT ?s
WHERE
SERVICE <e>{
        ?s> ?p <s>.
}
```

Figure 3-11: *SELECT(s, e): returns all resources that appear either as the subject or as the object in the triples containing <s> as the object or as the subject respectively.*

The SPARQL query in Fig. 3-12 returns the resources present at the endpoint that are similar (linked by the property "sameAs") to the resources mentioned by the data subject.

```
SELECT ?o
WHERE
SERVICE <e>{
        <u> owl:sameAs ?o.
}
```

Figure 3-12: *SAMEAS(u, e): returns resources that are linked to resource <**u**> via the **owl:sameAs** links.*

The SPARQL query in Fig. 3-13 constructs triples of types of similar ("sameAs") resources at a particular endpoint, along with their labels and available definitions.

```
SELECT ?type1 ?label1 ?definition1
WHERE
SERVICE <e>{
        ?s ?p <sa>.
        ?s rdf:type ?type1.
        ?type1 rdfs:label ?label1.
        OPTIONAL{?type1 skos:definition ?definition1}.
        FILTER((LANG(?label1) = ‘‘’’ || LANG(?label1) = ‘‘en’’))
}
}
UNION
{
SELECT ?type2 ?label2 ?definition2
WHERE
SERVICE <e>{
        <sa> ?p ?o.
        ?s rdf:type ?type2.
        ?type2 rdfs:label ?label2.
        OPTIONAL{?type2 skos:definition ?definition2}.
        FILTER((LANG(?label2) = ‘‘’’ || LANG(?label2) = ‘‘en’’))
}
```

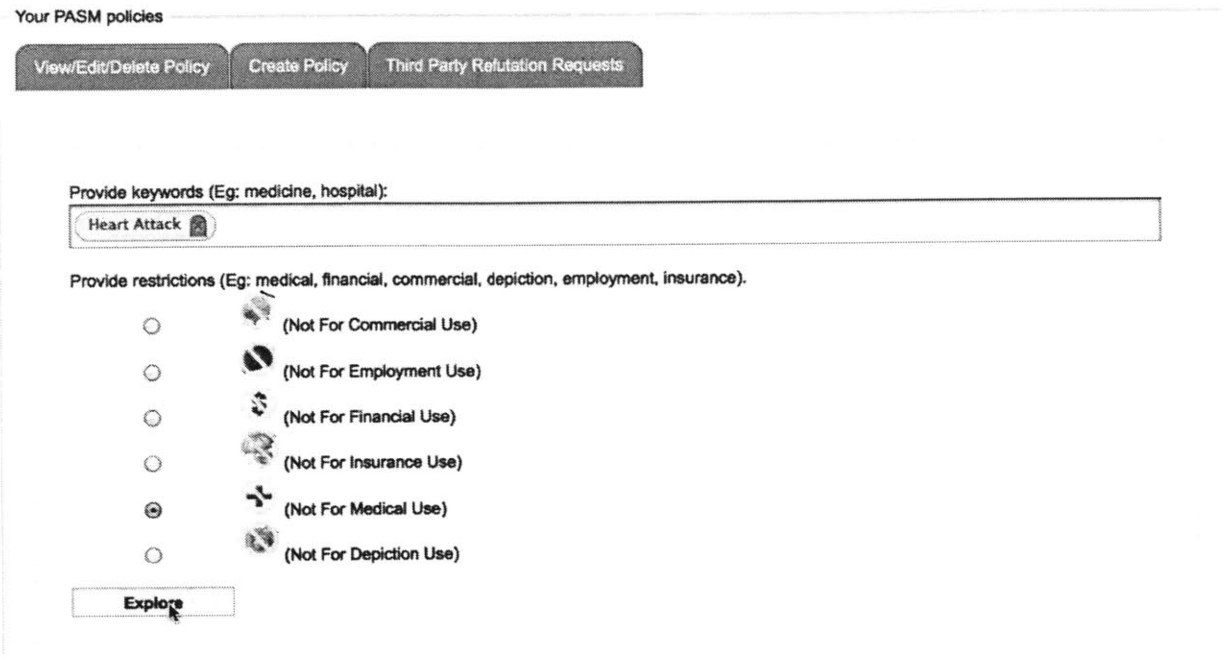
Figure 3-13: *TYPE(sa, e): first identifies the types of resources that are linked to the* <**sa**> *resource either as a subject or as an object. This procedure then returns the list of types along with the corresponding labels and definitions.*

3.4 Implementation Screenshots

A working implementation of PASM is currently available [3] and has interfaces for both data consumers and data subjects.

Data Subject Interface: Alice uses the interface shown in Fig. 3-14 to create a usage restriction for the first motivating scenario. Internally, PASM keeps track of the URI corresponding to the resource labeled “Heart Attack”.

[3] http://musigma.csail.mit.edu:2020/pasm.html



Your PASM policies

Provide keywords (Eg: medicine, hospital):

Heart Attack

Provide restrictions (Eg: medical, financial, commercial, depiction, employment, insurance).

○ (Not For Commercial Use)

○ (Not For Employment Use)

○ (Not For Financial Use)

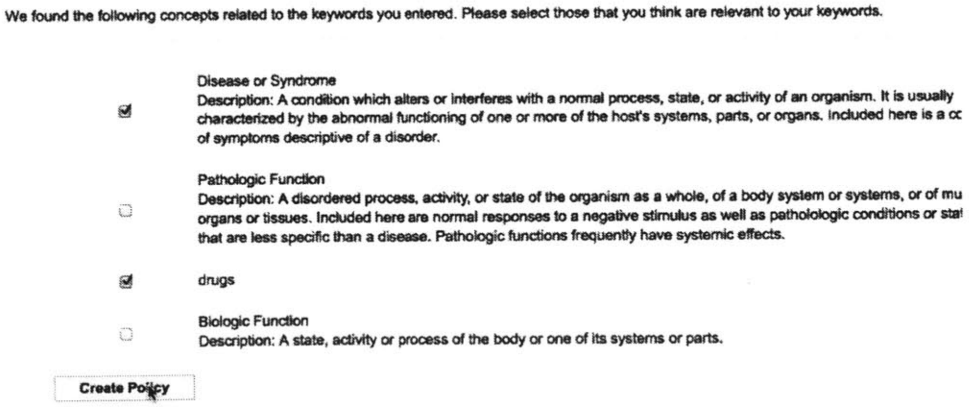
○ (Not For Insurance Use)

● (Not For Medical Use)

○ (Not For Depiction Use)

Figure 3-14: *Creating Usage Restriction*

After clicking on "Explore" button, Alice sees a list of types that correspond to the resources that the Semantic Enhancer identifies as relevant to "heart attack". She then selects the first and third types as shown in Fig. 3-15. These selected types are then stored in the usage restriction along with the keyword "Heart Attack" and the intent "Medical". A complete usage restriction definition in N3 representation is provided in Appendix 7.3.



We found the following concepts related to the keywords you entered. Please select those that you think are relevant to your keywords.

☑ Disease or Syndrome
Description: A condition which alters or interferes with a normal process, state, or activity of an organism. It is usually characterized by the abnormal functioning of one or more of the host's systems, parts, or organs. Included here is a cc of symptoms descriptive of a disorder.

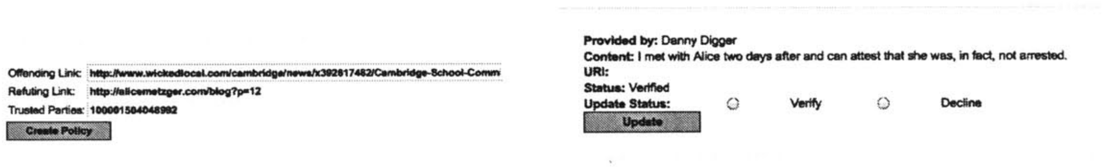
□ Pathologic Function
Description: A disordered process, activity, or state of the organism as a whole, of a body system or systems, or of mu organs or tissues. Included here are normal responses to a negative stimulus as well as patholologic conditions or sta that are less specific than a disease. Pathologic functions frequently have systemic effects.

☑ drugs

□ Biologic Function
Description: A state, activity or process of the body or one of its systems or parts.

Figure 3-15: *Semantic Enhancement.*

Alice uses the interface shown in Fig. 3-16 (a) to create a refutation link for the news article by linking the URL of the news article to her blogpost. After naming Danny as her trusted party (using Danny's Facebook id) and receiving his supporting refutation, she can choose to verify or decline his third-party refutation as shown in Fig. 3-16 (b).



Offending Link: http://www.wickedlocal.com/cambridge/news/x392817482/Cambridge-School-Comm
Refuting Link: http://alicemetzger.com/blog?p=12
Trusted Parties: 100001504048992

(a) Creating a refutation link

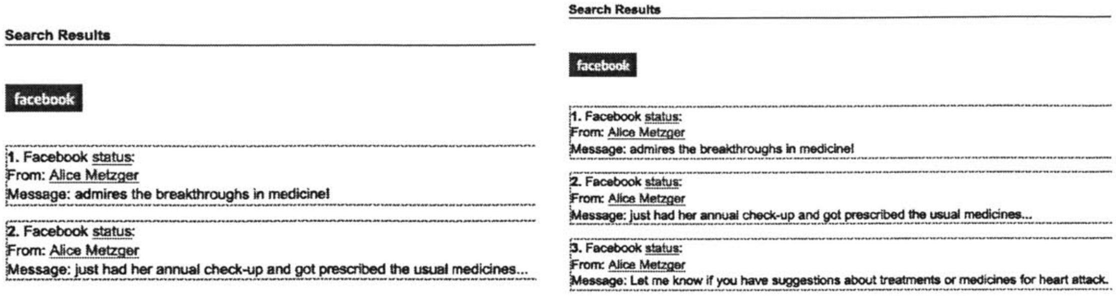
Provided by: Danny Digger
Content: I met with Alice two days after and can attest that she was, in fact, not arrested.
URI:
Status: Verified
Update Status: ○ Verify ○ Decline

(b) Verifying a third-party refutation

Figure 3-16: *Refutation Link and Third-party Refutation*

Data Consumer Interface: Bob, the data consumer, enters Alice's Facebook id, search keyword(s), and, the intent of the search into PASM's data consumer interface. Suppose that Bob is searching for data about Alice containing the keyword "medicine" and truthfully declares that the intent of the search is about medical insurance.

Upon submitting the form, Bob would see Fig. 3-17(a) which shows data from Alice's Facebook profile that pass the filters created by her semantically enhanced policies. To illustrate the application of usage restrictions, Fig. 3-17(b) shows the results returned by PASM when Bob searches for the same keyword ("medicine") but with a different intent (employment). As the new intent (employment) does not match the filters created by Alice's usage restrictions for posts containing heart attack, Bob is able to view additional posts from Alice's Facebook profile that mention heart attack. The additional posts shown in Fig. 3-17(b) are ambiguous (as it is not clear whether it is *Alice* or *someone else* that has heart problems) and thus highlight the dangers of the Mosaic Theory mentioned in Chapter 1.



Search Results

1. Facebook status:
From: Alice Metzger
Message: admires the breakthroughs in medicine!

2. Facebook status:
From: Alice Metzger
Message: just had her annual check-up and got prescribed the usual medicines...

(a) intent: medical

Search Results

1. Facebook status:
From: Alice Metzger
Message: admires the breakthroughs in medicine!

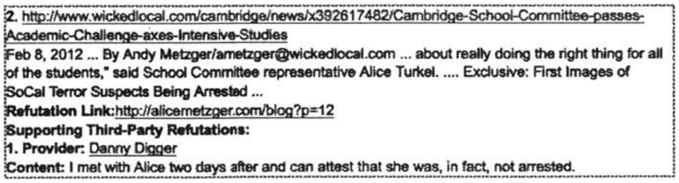
2. Facebook status:
From: Alice Metzger
Message: just had her annual check-up and got prescribed the usual medicines...

3. Facebook status:
From: Alice Metzger
Message: Let me know if you have suggestions about treatments or medicines for heart attack.

(b) intent: employment

Figure 3-17: *Search results for two searches using the keyword "medicine" ((a) was performed with a medical intent, while (b) was performed with an employment intent). Both searches return data from Facebook. However, note the additional results in (b).*

The Web search performed on Alice using her name and the keyword "medicine" returns results from the Google's Custom Search Engine. This is analogous to a normal Web search, where the presence of a subset of keywords in a document would trigger the insertion of that document into the search results. Since Alice does not own this content, PASM searches her refutation links to see whether she created refutations for any of the returned results. It then finds the refutation link for the *Wicked Local* news article in Alice's policy. PASM then prominently displays the following below the URL of the news article: (i) Alice's refutation, and, (ii) Danny's supporting third-party refutation that Alice verified. This annotation of the custom search results is shown in Fig. 3-18. This makes Bob aware of Alice's counter-argument to the document that stated incorrect "facts" about her arrest and thus informs Bob about Alice's side of the story as well.



2. http://www.wickedlocal.com/cambridge/news/x392617482/Cambridge-School-Committee-passes-Academic-Challenge-axes-Intensive-Studies
Feb 8, 2012 ... By Andy Metzger/ametzger@wickedlocal.com ... about really doing the right thing for all of the students," said School Committee representative Alice Turkel. .... Exclusive: First Images of SoCal Terror Suspects Being Arrested ...
Refutation Link:http://alicemetzger.com/blog?p=12
Supporting Third-Party Refutations:
1. Provider: Danny Digger
Content: I met with Alice two days after and can attest that she was, in fact, not arrested.

Figure 3-18: *Search Results from Google custom search engine*

Chapter 4

User Studies

4.1 Introduction

This chapter discusses the user studies performed during the course of this research. It first describes the pre-study conducted to collect a list of commonly-identified sensitive terms and then describes the main user study to measure the performance of PASM.

4.2 Pre-Study – Sensitive/Negative Phrases

In order to gauge what terms may be considered sensitive, a pre-study was conducted to compile a list of such terms by crowd-sourcing the task.

4.2.1 Methodology

This section describes the goal and procedure of the pre-study.

Goal: The goal of this pre-study was to compile a list of terms that may be considered "sensitive" by the participants and thus harmful to their reputation.

Procedure: The pre-study was conducted on Amazon's Mechanical Turk [1]. We recruited

[1] https://www.mturk.com/mturk/welcome

100 participants and divided them into two groups of 50 each. We selected the following ten categories that could potentially contain sensitive keywords: Action, Disease, Drink, Object, Role, Behavior, Event, Food, Location, and, Medical. Five of the ten categories were assigned to each group of participants. We then asked the participants to list three words or phrases that they consider to be sensitive for each of those five categories.

4.2.2 Results

At the end of the study, we had a list of 1500 "sensitive" words. However, most of them were proper nouns and could not be considered sensitive in general without a proper context. For example, we received names of places, soft drinks, and food which were not sensitive on their own, but may have been sensitive to certain participants under a particular context. In order to identify commonly-perceived sensitive words, we grouped similar terms together and counted the number of times these terms appeared in the list. Interestingly, the two categories *Disease* and *Medical* contained terms with the largest frequency of occurrences. The list of top-ten terms (groups of similar responses) provided by participants for the categories Disease and Medical is shown in Table 4.1.

Keywords	Number of users
cancer	30
drug(s)/medicine(s)/tablet(s)	16
disease/illness/ailment/sick(ness)	15
needles/injection(s)/syringe(s)/shots/stitches	14
operate/operation	12
stroke/heart attack/cardiac arrest	10
pain(ful)/suffering/weak(ness)/ache	9
diabetes/sugar	8
doctor	8
HIV/AIDS/HIV +ve	7

Table 4.1: *Top ten crowd-sourced sensitive terms related to health.*

The term Cancer ranks first in the list shown in Table 4.1. However, there are different types of cancer and some of them are gender-based. Thus, we decided to use the next ranked diseases – Heart Attack and Diabetes – as topics of posts in the study described in section 4.3. Further, we used Football as third (control) topic in the study as it did not occur in this list.

4.3 User Study – Performance of PASM

This study was conducted to understand how well PASM recognizes sensitive posts as compared to the study participants.

4.3.1 Methodology

This section discusses the research question, hypothesis of the study, recruitment of participants, and, the study procedure.

Research Question: How well can PASM identify that a particular post is sensitive?

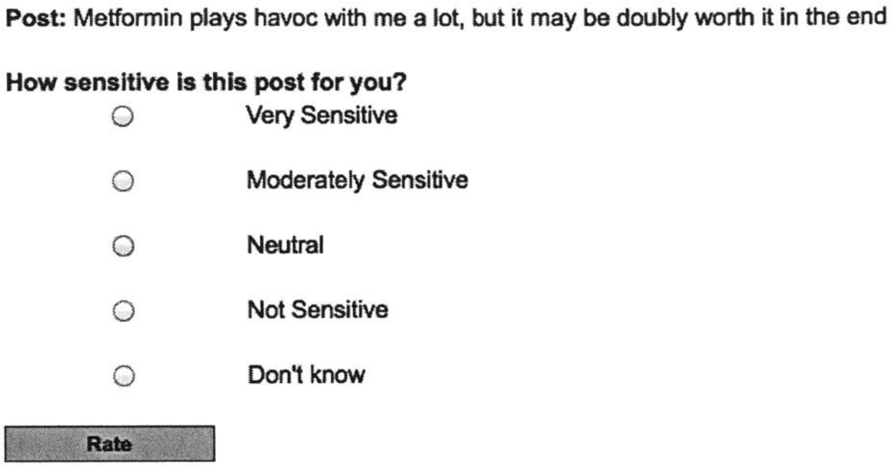
Hypothesis: Given keywords, PASM can identify posts that may be sensitive to the participant with a high F-Measure.

Participants: For this study, we recruited participants mostly from the CSAIL lab at MIT. A more detailed explanation about the participant population characteristics is given in section 4.3.2. Not all of them were researchers in the field of privacy or knew about this research. In order to familiarize everyone with this research, we informed the participants that the study was about privacy in social networking sites. We then briefed them about the task prior to the start of the study. During the study, we asked the participants to respond to an optional demographic survey about their gender, age, qualification and Facebook privacy usage (private or public). The survey was followed by the task of rating the sensitivity of 15 posts and finally we asked the participants to provide subjective feedback.

Procedure: The following task was shown to the participants at the start of the study: *"Imagine that you are using a tool that works with your data and displays your Facebook* *personal status updates and wall messages. You sometimes post some health related issues that might affect you, your family/close friends or just general health issues important to you. You also post more general information about sports, music etc."*

The participants were then asked to rate 15 posts using a 5-point Likert scale (1 being very sensitive, 4 being not sensitive, and, 5 indicating that they are not sure). The posts were a mix of public Facebook status updates and wall messages, some of which are personal. We chose to use synthetic data for this study because of the following two reasons. First, the frequency of sensitive posts that a person has on Facebook is unknown. Average users of social networking sites typically do not post a lot of sensitive health-related posts on their social networks. Using synthetic data allowed us to know the density of sensitive posts used in the study. Second, using synthetic data allowed us to calibrate sensitivity across different individuals. We had an "objective" external determination of the sensitivity of a post, and then got the per-user scoring. In order to prepare this data, we used the Facebook API to retrieve posts matching the following terms – "diabetes", "heart attack", and, "football". We selected 15 posts that (i) described personal experiences with diabetes and heart-attack, (ii) were not explicitly about diabetes and heart attack, but contained related terms, and, (iii) were about football. As an example for (ii), we collected posts that mentioned 'Metformin', a drug for diabetes. In this way, we exploited the semantics of the keywords so that the semantic enhancement procedure could be tested. We had a larger proportion of posts about diabetes and heart attack than about football.

In order to overcome ordering effects that could be created by showing the posts in a particular order for all the participants (for example, showing those related to diabetes first), we randomized the order of the posts shown to the participants. Further, to avoid cross-influencing the decisions of the participants' rating of one post on the other posts, we displayed each post on a separate page along with the corresponding 5-point Likert scale. A screenshot of the interface shown to the participants is provided in Fig. 4-1.



Post: Metformin plays havoc with me a lot, but it may be doubly worth it in the end

How sensitive is this post for you?

○ Very Sensitive

○ Moderately Sensitive

○ Neutral

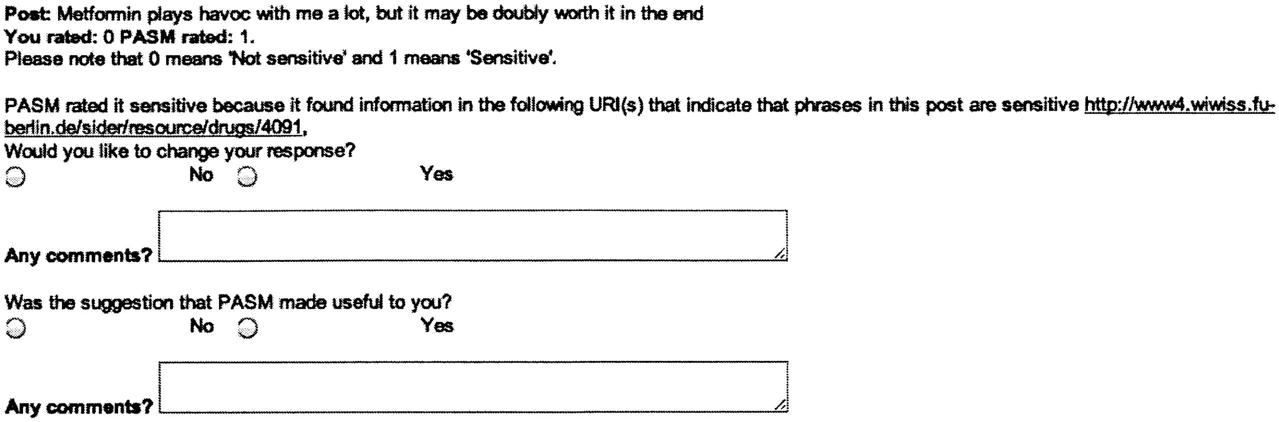
○ Not Sensitive

○ Don't know

Figure 4-1: *User study interface with a post and the corresponding 5-point Likert Scale.*

PASM also rated these 15 posts using the Semantic Enhancer component. However, since PASM needs to decide whether to filter a post from the results to be shown to a data consumer, it uses a binary scale – sensitive/not sensitive. Thus, the participants' ratings on the 5-point Likert scale were scaled to a binary scale. A rating of 1 or 2 on the Likert scale was considered to be sensitive, while that between 3 and 5 was considered to be not sensitive. Discrepancies that occurred between the participant's answer and PASM's rating (both on binary scale) for a post were flagged at the end of the study. The participants were then asked questions regarding the discrepancy as shown in Fig. 4-2. If the participants gave a scaled rating of "not sensitive" and PASM classified that post as "sensitive", they were given a justification of why PASM made this decision. The justifications highlighted the linked data resources that PASM found during the semantic enhancement process. The participants were then asked if they would like to change their answer and provide subjective feedback. They were also asked whether the suggestion made by the PASM's rating (sensitive or not sensitive) was useful to them.

Finally, we explained about PASM and asked the participants to provide subjective feedback about their opinion concerning such a system, the justifications displayed in the study, and, whether they preferred to receive false positive or false negatives by a tool that attempted to identify sensitive posts.



Post: Metformin plays havoc with me a lot, but it may be doubly worth it in the end
You rated: 0 **PASM rated:** 1.
Please note that 0 means 'Not sensitive' and 1 means 'Sensitive'.

PASM rated it sensitive because it found information in the following URI(s) that indicate that phrases in this post are sensitive http://www4.wiwiss.fu-berlin.de/sider/resource/drugs/4091,
Would you like to change your response?
○ No ○ Yes

Any comments?

Was the suggestion that PASM made useful to you?
○ No ○ Yes

Any comments?

Figure 4-2: *Justification and follow-up questions.*

4.3.2 Results

This section presents the participant population characteristics, summarizes the objective results about the ratings, and, discusses the subjective responses provided during the study.

Participant Population Characteristics: We recruited 23 participants (ten female and thirteen male). Most of them were part of the CSAIL lab at MIT. However, we had three participants who were not affiliated with the lab. Twelve users were between 20 and 25 years of age, six were between 26 and 30, three were between 31 and 35, and, two were between 50 and 70. Eighteen participants were graduate students, one was a software consultant, one was a stay-at-home parent, one was a professor, one was an undergrad and one was a research scientist. One participant declared that they did not have any social networking account and another declared that they did not actively use social networking websites.

Post Ratings: Table 4.2 shows (i) the aggregate post ratings, (ii) number of the participants (per post) who expressed a change in attitude or awareness based on the information presented by PASM, and, (iii) number of the participants (per post) who found PASM's justifications to be useful.

Post	True	True	False	False			Don't		
ID	Positives	Negatives	Positives	Negatives	Errors	Neutral	Know	Change	Useful
$1^{(*)}$	0	16	0	7	7	6	0	-	0
$2^{\dagger(*)}$	22	0	1	0	1	1	0	0	0
$3^{\dagger(*)}$	15	0	8	0	8	5	3	6	6
$4^{\dagger(*)}$	22	0	1	0	1	1	0	1	1
$5^{\dagger(*)}$	23	0	0	0	0	0	0	0	0
$6^{\dagger(*)}$	14	0	9	0	9	6	0	4	5
$7^{\dagger(*)}$	22	0	1	0	1	1	0	1	1
$8^{(**)}$	0	23	0	0	0	6	0	0	0
$9^{(**)}$	0	23	0	0	0	5	0	0	0
$10^{(**)}$	0	20	0	3	3	6	0	-	2
$11^{\dagger(***)}$	20	0	3	0	3	3	0	0	2
$12^{(***)}$	0	12	0	11	11	8	0	-	3
$13^{\dagger(***)}$	23	0	0	0	0	0	0	0	0
$14^{\dagger(***)}$	19	0	4	0	4	4	0	3	3
$15^{\dagger(***)}$	2	0	21	0	21	5	6	14	17
Total	182	94	48	21					

Table 4.2: *Aggregate ratings of posts in the user study. Posts marked with † were rated as sensitive by PASM. Posts marked with* $^{(*)}$ *are related to diabetes, those marked with* $^{(**)}$ *are related to football and those marked with* $^{(***)}$ *are related to heart attack. Posts 5 and 13 were consistently rated as sensitive by all the users while posts 8 and 9 were consistently rated as not sensitive.*

We can group these ratings into five categories based on the agreement between PASM's and the participants' ratings. Note that the agreements are indicated in the following columns – *True Positives* (for posts that are sensitive) and *True Negatives* (for posts that are not sensitive). The opposite notion of agreement is a discrepancy and these are indicated in the following columns – False Positives (PASM rated the post as sensitive while participant rated it as not sensitive) and False Negatives (PASM rated the post as not sensitive while participant rated it as sensitive). The five categories based on the following levels of agreement – *maximum, high, medium, low, minimum* – are explained as follows.

1. *Agreement is maximum:* Posts 5, 8, 9, and, 13 come under this category. Posts 5 and 13 were about diabetes and heart attack respectively while posts 8 and 9 were about football. All participants provided a rating consistent with that of PASM.

2. *Agreement is high:* Posts 2, 4, and, 7 come under this category. These were about diabetes and only one of the twenty three participants gave a rating different from that of PASM for each of these posts. One of the participants found PASM's justification to be helpful for these posts and mentioned *"I didn't know metformin is a drug. I thought it might have been some chemical material in food."*.

3. *Agreement is medium:* Posts 10, 11, and, 14 come under this category. These were about football and heart attack. Post 10 talks about football, but three participants rated it as being sensitive because they thought it revealed private information about their work schedule (the post mentioned about working on the weekend). One participant commented *"Indicates that author of the post will not be home during a certain time - possible potential for robbery?"*. Although Posts 11 and 14 were about heart attack, they were not considered to be sensitive. A participant explained their reasoning by stating *"I thought I had a heart attack but it was not. Unless I was medically inclined, I would think it unlikely an average person would conclude I necessarily had heart trouble [...]"*. PASM's justifications were generally not considered useful as mentioned in the following comment *"I'm already familiar with cholesterol from everyday interactions with media and people, and the definitions were a little too verbose for me to quickly discern and decide whether they were relevant to me"*

4. *Agreement is low:* Posts 1, 3, 6, and, 12 come under this category. These posts were related to diabetes and heart attack. Posts 1 and 12 were rated by PASM to be not sensitive, while some participants rated them otherwise. Interestingly, the reasons given by the participants indicated that they were concerned about the privacy of others (post 1) or the fact that their employment information was described in the post (post 12). For post 1, two participants commented as follows. *"Thought the post was sensitive because it revealed low-level health information about a minor (4th grader named Drew)."*, and, *"1. It*

did not take into account that I will be revealing my calendar/schedule for a future date. 2. It suggests that revealing information on another person other than me is not sensitive, when in fact it is.". Posts 3 and 6 were rated by PASM as being sensitive, while some participants rated otherwise. The main reason for this was indicated to be the lack of specificity of the post as mentioned by the following comment *"Neutral [because] it does not give specific information about my health status, it's just a general observation."*. Regarding PASM's justifications, some participants found them to be helpful and mentioned that *"I didn't know metformin is a drug [...]"*. Others who did not find PASM's justification helpful cited the lack of a clear explanation. For example, one participant mentioned, *"I don't know why people take metformin. the page didn't tell me that."*.

5. *Agreement is minimum:* Post 15 (*Learned about the funny noun Esmolol today and its importance in my life.*) falls under this category. This post mentioned a drug for heart disease and was rated by PASM to be sensitive. However, a majority of the users (21 out of 23) rated it as not sensitive. Interestingly, 11 of those 21 users selected options 3 (neutral) or 5 (not sure). In other words, these 11 users did not explicitly rate the post as being not sensitive. The most common causes for this discrepancy was due to lack of knowledge of what the terms meant or the impersonal nature of the post. Lack of knowledge of what the terms meant was one of the problems with using synthetic data. Presumably, the subjects would be more aware of the meaning of their own posts. Conversely, an individual may be unaware of additional meanings conveyed by a post and PASM can be helpful when it points out these meanings. The following were some of the comments received from the participants, *"the post is an impersonal expression."*, *"It could be simply that I learned what a new word means, and sharing the information to express happiness. If I were a biology student, I learned it from class today."*, and, *"I had never run into the term Esmolol before; knowing that it is a drug that deals with some kind of heart trouble makes me want to keep this info more private - restricted to family/close friends, rather than open to whoever wanders by online."*. Majority of these participants (17 out of 21) found PASM's justifications to be useful and one commented *"PASM gave me the info that Esmolol is a heart-problem drug, which I did not know."*.

Precision and Recall Metrics: From the aggregated ratings listed in Table 4.2, we can compute the performance metrics of PASM as follows.

Precision: $\frac{\text{True Positives}}{\text{True Positives+False Positives}} = \frac{182}{182+48} = 0.791304$

Recall: $\frac{\text{True Positives}}{\text{True Positives+False Negatives}} = \frac{182}{182+21} = 0.896552$

F-Measure: $\frac{\text{2 * Precision * Recall}}{\text{Precision + Recall}} = \frac{2*0.791304*0.896552}{0.791304+0.896552} = 0.840647$

Accuracy: $\frac{\text{True Positives + True Negatives}}{\text{True Positives+True Negatives+False Positives+False Negatives}} = \frac{182+94}{182+94+48+21} = 0.8$

These values indicate that PASM's recall, in this user study, is slightly better than its precision. Thus, PASM is better at classifying more of the sensitive posts as actually being sensitive, compared to correctly classifying posts as being sensitive or not. In other words, PASM is better at identifying truly sensitive posts. It is vital to note, however, that this is a very small study - both in the number of subjects and in the number of posts classified.

4.3.3 Discussion

This section describes the subjective responses provided by the participants regarding PASM, the justifications provided during the study, and, their opinion about false negatives versus false positives.

About PASM: All 23 participants found PASM to be an interesting system and said that they would be interested in using it. Most participants thought that PASM could serve as a cautionary tool that could help social network users think twice before posting data. Participants stressed that this is important mainly because people do not generally perceive the ramifications that their data on the Web can have on them. In other words, they do not see how their own data can be used against them. Something as innocuous as eating in a particular place can be used for targeted advertising. One participant mentioned that such a tool would be especially good for younger children who are legally allowed to join social networks but may not necessarily know what is appropriate or inappropriate to post on the Web. Another participant mentioned that a tool like PASM would be helpful by providing a third-party, independent perspective on why a particular post is sensitive.

Some of the participants compared PASM to features present on commercially-available tools. One example provided was the feature called "Email Goggles" on Google's GMail. This feature asks users to solve simple math problems before sending an email on Friday nights. Another example given was the GMail's pop-up reminder asking if a user forgot to add an attachment when the user mentions the word "attachment" in the content of the email but does not attach one. A third example was Google search's "This site may harm your computer message". One user suggested that PASM could be designed similar to such a feature by advising users that to change their privàcy setting due to the sensitivity of a post.

Five participants mentioned the importance of context when considering sensitivity of a post. For example, one participant mentioned being comfortable with sharing certain content of their Facebook profile only with their family members. However, that participant did not want to share that data with friends, let alone strangers. Generally speaking, these participants correlated the context of a post with the recipients of that post. One participant mentioned about being very careful when posting and if the post was misconstrued, they would respond with an appropriate explanation. After reminding them about the data consumers who may actually view the data and then make decisions about them (probably without asking for clarifications from them), the participant agreed that sensitivity of a post from the reader's perspective is actually more important than the sensitivity of that post from the poster's perspective. In other words, the participants agreed that people who post on their own Facebook profile should consider whether a post would be considered sensitive by others reading that post (especially those who have power over them). One participant mentioned that the notion of sensitivity varies with culture, while another pointed that some posts could be sensitive for the person who posts information, while others could be sensitive for those referred to in the post (citing post 1 as an example).

Another important issue was about autonomy - whether users should have the final say on determining whether the post is sensitive or not. This issue was raised in the cases where PASM rated some posts as being not sensitive while the participants considered them to be sensitive. The participants argued that they would be able to determine whether a post was not sensitive themselves and because identifying a post as not sensitive usually had no dire consequences, they were not concerned about those posts. Another side of this argument concerns users who really want to share such information, either because they are interested in receiving feedback or assurance about the concerned matter or because the issue is a well-known fact about that participant. In such cases, they wanted the ability to override PASM's suggestions to minimize annoyances. Interestingly, because sensitivity is a subjective issue, some participants mentioned that they would probably disregard objective metrics (a sensitivity score, for example), but would carefully consider any reference information pointed out to them. A few participants suggested that having categories of sensitive words (not only about health, but also about politics, politically incorrect information or content that may offend other people) would be helpful.

Finally, participants suggested having different categories of sensitivity that would indicate different *"threat levels"* for their privacy or reputation. Frequently suggested categories included politics, health, family, location, employment, and financial. Having such categories and settings for them would enable participants to either turn these warnings off or select the ones that are applicable to them personally. That way PASM could focus on classifying posts belonging to the selected categories only.

Justifications: In general, most of the participants found the justifications to be useful. Participants described that they could be viewed as a "wake-up call" or a "moral compass". Some also pointed out that a tool that provides feedback and trains the user for how to behave in other contexts has the potential to educate users on how to speak in public and not offend people. A participant mentioned about not knowing that a particular term was a medication. Once that participant was made aware of this fact, they wanted to change the rating that they entered during the study. Another useful aspect of the justifications (as mentioned by the participants) was that it showed accountability on the part of PASM. One participant mentioned that they would not want a system classifying posts as sensitive without providing an explanation.

Among the Linked Data resource descriptions shown to the participants during the study, the participants found most of them to be helpful, but some felt that one resource description was incomprehensible. Some participants commonly mentioned that the resource desciptions were too technical, hard to read, not very expressive, not in layman's terms, not visually sufficient, and, contained too much information (some of which was either unnecessary or in a different language). These factors made the resource descriptions difficult to focus on or understand their summary. Participants suggested condensing the content, showing the main topic/summary clearly (preferably highlighted by making it bold or increasing its font size), and, explaining the content assuming that a lay person was reading it.

Interestingly, there were many innovative suggestions for implementing the justifications. Some users suggested showing Wikipedia content instead of linked data format (DBpedia). Some participants requested showing a few links from various websites (and perhaps a summary of the relevant content from those websites) so that they could select the ones they would like to view. Others suggested showing case studies or news articles that would indicate why information about this keyword is considered to be sensitive. For example, if the post is health-related, a link to an article that gives an anecdotal reason why one would not want to post this health information for public consumption. If someone actually posted something similar and got into trouble for doing so, displaying an anecdotal report would better equip the user to decide that a particular post is indeed sensitive. A participant commented that though the information was helpful, it was *"not scary enough"*. Another unique suggestion by a participant was to have a cheatsheet of what not to post on social networks (more as a reference and not necessarily interactive).

These suggestions indicate that PASM did not adequately meet the needs of users who wanted to know why it rated a particular post as sensitive. Upon enquiry, some participants mentioned that they would prefer finding out about the audience to whom such posts may be sensitive in addition to a comprehensible summary of why that post is sensitive. One participant mentioned that a prompt like "if an employer or professor could view that post, it could damage your reputation" would be a helpful hint. Another comment along the same lines was to have PASM suggest who to share that post with or what privacy setting to choose for this post. Instead of just declaring that the post is sensitive, participants mentioned that it would be helpful to have PASM give explicit prompts such as "Are you sure you want to post this?" or "Some of the words in this post are rated as sensitive. Are you sure you want to post it?". Lastly, a participant mentioned that it would be helpful to not only explain the meaning of a term, but also the consequences of using it. For example, "if you share this post, people may think you have diabetes."

There were a few comments about the technical implementation and user interface design of the justifications. Some participants suggested having such a feature integrated in social networks so that when one types, the system can automatically process it and then show a warning without the need for pop-ups, button clicks or other obstructive implementation mechanisms. One participant suggested implementing this as a browser extension with the data (sensitive words detected) stored locally and not shared with a commercial entity. Other usability concerns were about the presentation. Some participants preferred to have visual indications of the sensitivity of the post either through the use of color (red being very sensitive) or icons. One participant mentioned that in such a system, they would give a higher preference to usability, rather than correctness. Another aspect of usability involved control. Participants preferred systems that gave them the final say about whether to classify the post as being sensitive or not.

False Positives versus False Negatives: Out of the 23 participants, 22 indicated that they would prefer having false positives than false negatives. Most of the participants in this group stated that being cautious is good for privacy. Having the system flag a post as sensitive would cause them to think more about potential privacy concerns before publishing such posts on social networks. Some participants mentioned that they were very cautious about their privacy on social networks anyway and would therefore prefer an overcautious system as a "sensitivity filter". Some participants pointed out that "it is better to be safe than sorry" and once damage is done by to one's reputation, the damage may be irreversible. Participants also pointed out that if a system had more false negatives than false positive, they would suspect that something was wrong with the system. The only participant who preferred false negatives to false positives stated that they prefer receiving less intrusions, and if the system became annoying, they would actually stop using it.

Annoyance is an important factor to consider when a system may produce more false positives than false negatives. Many participants, who initially preferred having false positives added a caveat that they would prefer a system that did not annoy them too much. Most of them preferred a trainable system that would initially be over cautious and then tone down based on the user's responses, by learning what the user wants. As an example, a participant suggested using the current posts in the user's Facebook account to train the learning algorithm. Posts that have modified privacy settings may potentially have content that the user thinks is sensitive. By analyzing them, PASM could get a general sense of which categories or content that user would personally want to classify as being sensitive. Another learning opportunity would be when the user explicitly overrides the system's rating. Through this interaction, PASM should be able to identify that certain topics or meanings of phrases are not sensitive for this particular user. Another dimension of interactive learning is to see the posts where the user agreed with PASM's rating and then automatically classify posts containing similar meaning/phrases in the future.

One participant mentioned that they would be able to tolerate annoyances as long as their privacy could be protected. Others mentioned that they could tolerate intrusive nagging up to a limit (for example, one out of ten ratings being wrong is okay, but not one out of every two ratings). The most common response about the intrusive nagging was that they would start ignoring them after a while. Some users provided constructive feedback by suggesting creating a settings page where users can be given controls to turn the feature off completely or personalize the settings based on their individual need. Most participants mentioned that the ability to override the default settings would be beneficial.

Chapter 5

Background and Related Work

This chapter describes related research in the areas of privacy in social networks, policy-based privacy research in social networks and linked data traversal.

5.1 Privacy and Social Networks

This section describes the state of the art for ensuring privacy in social networks.

A recent PEW study [26] states that there is an increasing number of people who are setting more private access controls compared to those in 2009. However, half of the people surveyed reported having difficulty in managing their privacy controls and over 11% reported that they regretted posting something online. The risks to privacy invasions were attributed more to others than to oneself which may be due to a third-person effect. Staddon *et. al.* [37] investigated whether privacy concerns were a turn-off when using social networks. They noted that people with privacy concerns report to be less engaged when using social networks, compared to those without such concerns. Furthermore, [17] describes the following paradox between the control and publication of data – less control actually made people more aware of the need for privacy.

Despite such a lack of confidence in privacy controls, two types of mechanisms have been introduced to help educate users and ensure their privacy. First, the social networking sites have mechanisms in place to restrict access to data based on the user's privacy settings. Examples of these mechanisms include Facebooks View As (Audience View) feature and friend lists, Google's circles on Google+ and Twitter's mechanism of direct and private tweets. Second, third-party services (like the social networking privacy experiment *We know what you're doing* [15]) aim at educating users about how some content may be inappropriate and are still being shared on Facebook and Twitter.

One way to ensure privacy would be to use anonymization and decentralized approaches for social networks. However, Narayanan and Shmatikov [27] identified that anonymization is not a particularly effective technique to preserve privacy given the fact that there are various social networks and one may be able to recognize the identity of a person on the Web. For example, they found out, using a re-identification technique based on the network graphs, that 24% of the accounts on Twitter are also present on Flickr, while 5% of Flickr accounts correspond to that on Twitter. Regarding decentralized approaches, [24] identified that protecting metadata is not a trivial task in a system employing peer-to-peer architecture.

Privacy is not clearly defined in terms of ownership because data belonging to a particular user's account could actually be about another user. For example, a friend of a user may tag the user in photos that belong to the friend. In such a situation, the privacy of the user (and not the user's friend) is at stake. Some approaches to solving this issue involve collaboratively setting privacy policies [34]. The "Same-As" privacy management [35] takes advantage of the user's and their friends' opinions in collaboratively creating privacy policies.

Most social networking sites allow users to restrict access to data that are uploaded by the users, but not by others. unFriendly [38] attempts to resolve such conflicting privacy policies as the sites need to honor the privacy settings of the people involved in the content who may not necessarily be the owners. They create a framework for multi-party conflicts by formalizing and creating inference-based techniques to classify the conflicts. Another technique that can be helpful in this situation is a system that predicts privacy policies or ways to protect one's data. Song *et. al.* [33] created a tool that can automatically suggest sensitive labels which can further be incorporated with privacy policies for protection.

Enabling and enforcing refutation links is necessary when people want to dispute content on the Web over which they do not have control over. Ennals *et. al.* [19, 20] discuss how to manage disputes on the Web and provide users with notifications when certain content on the Web is disputed by other online sources or people. Users can also actively participate by highlighting pieces of text with which they disagree and provide a summary about their point of view. The system, DisputeFinder, works as a browser extension that maintains a centralized database with the collection of disputed claims. It then shows these disputed claims to users when they browse the pages about the disputed topic.

5.2 Policy-based Privacy in Social Networks

This section describes related research directions involving policies and privacy mechanisms for social networks.

Due to a lack of user awareness and proper privacy protection tools, a lot of personal and sensitive information is being made accessible to authorities, strangers and recruiters/employers. Over one billion active users [7] share information on Facebook. It may be difficult for some users to create fine-tuned privacy policies, but there is a greater inherent difficulty in selecting recipients of social media content. For instance, deciding which people or categories of people should have access to users' information is often cumbersome since it requires users to constantly manage their privacy settings or friend lists.

Access controls are usually a-priori techniques, allowing users to construct barriers around their data. However, even if access controls for certain data are rigorously implemented, the people who are valid recipients could potentially misuse that data. In order to addresses undesired use by authorized parties, "Respect My Privacy" (RMP) [25] was created. RMP is a policy specification language that helps users create usage control policies that govern how their social network data can be used. It offers a pre-defined set of usage policies that are similar in nature to a Creative Commons [1] license. The tool described in [25] allows users to generate usage control policies and visualize them using Tabulator (a semantic web-friendly browser plugin).

On a similar note, in order to increase privacy awareness and help users maintain better control over their data, [16] outlines an XML-based policy framework called UPP (User Privacy Policy) which incorporates access rights, reputation and entities that can view data. Clifton *et. al.* in [18] propose a privacy framework for data sharing and integration by predicting the matches without revealing sensitive data and enabling querying across sources using semantic correspondences. Another ontology defined for this same purpose is called *Privacy Preference Ontology (PPO)*, which is a fine-grained access control specification for linked data (RDF documents) using the Web Access Control (WAC) vocabulary. The authors posit that this has a direct application in social networking applications where users can specify their policies using their Privacy Preference Manager, which can also enable the resolution of conflicting privacy preferences.

CoPE (Collaborative Privacy Management) [36] is an application on Facebook that can be used as a collaborative privacy mechanism. It defines an extended notion of a "content stakeholder" by permitting users to indicate who else might be a stakeholder for a particular data object (the primary focus of this research is the photos on Facebook). Once the stakeholders are defined, CoPE enables the identified stakeholders to have their privacy policies applied on these data objects as well.

5.3 Linked Data Traversal

This section describes related approaches on traversing Linked Data ontologies. As described in Section 3.3, PASM uses Linked Data to semantically enhance the usage restrictions.

Traversal along linked data graphs is especially useful for query executions where one cannot assume to know about all the data sources prior to the execution of the queries. Further, by dynamically identifying the data sources, one can tap into the full potential of the Web. In general, there are three kinds of approaches for query execution using SPARQL. The first approach is a top-down approach where the system assumes to have access to all the data sources that it would need. This approach executes the query on those identified sources and obtains the desired result. The second approach is a bottom-up approach where the system searches for additional sources during the execution of the query. In this approach, data sources are updated with additional data that are obtained dynamically. The third approach employs a mixed combination of the previous two strategies.

Hartig et. al. [21] describe an approach to identify reachability of linked data resources for the purpose of ensuring that the entire linked data graph is covered during the query execution. They attempt to answer queries without crawling large portions of the Linked Data graph that may not be relevant to the user's query. They first evaluate parts of the query and use the URIs obtained from the results of the partial evaluations for the rest of the query. An issue with using partial evaluations of queries and dynamically finding new sources of data is that such pipelines have a higher tendency to block. For instance, one portion of the query may need data from another part of the query and thus has to wait for the other one to complete execution. In other words, the iterators are blocked because a URI fetched previously has to be used in the specified time interval; otherwise, the URI is no longer applicable or useful. The authors in [21] prevent this and implement a non-blocking iterator-based approach by introducing a reject function that decides whether to keep the resulting URIs or discard them, thereby allowing these iterators to execute in parallel. Another approach outlined in [23] describes a system called LD-Spider that attempts to get the data locally in order to speed up the query execution process.

Passant and Mendes [29] describe a push-based service called PubSubHubbub, in which updates to linked data are proactively pushed to users (source or requester). In such a service, the system pushes information to the data consumer even though he or she did not search for that information.

5.4 PASM and Related Research

This section situates PASM among the related research described so far.

The development of PASM was motivated by the following three works. First is the *Mosaic theory* [31], which states that apparently harmless individual pieces of information could potentially reveal a damaging picture about an individual when pieced together. Second is *contextual integrity* [28], where contexts are defined to govern significant aspects like roles, expectations, and, behaviors. Data intended to be viewed in one context may have a completely different meaning when put in a different context. Third is *Information Accountability* [39], which makes the case for decentralized accountable systems. Such systems would make information usage actions (such as searches) more transparent by using logged transactions, a robust policy-language framework and policy-reasoning tools.

These concerns, although primarily about surveillance data, also apply to social networks data and other data on the Web. As [32] articulates, Under Section 7 of the National Labor Relations Act (NLRA), employees have the right to self-organize for the purpose of collective bargaining for mutual aid or protection. Section 8 of the same act prohibits employers from interfering with such activities. In the context of social media, Russell articulates that an employee's use of social media must be concerted and made for purpose of mutual aid or protection in order to be protected under sections 7 and 8 of NLRA. In other words, he acknowledges that online rants against an employer are unlikely to qualify for protection from firing based those two sections of NLRA. As of now, the best solution is to update the privacy settings on social media like Facebook and Twitter to limit conversations with coworkers as much as possible.

Although social networks have in-built mechanisms for privacy of their users, they are inadequate to completely express the context that the data subjects want their data to be viewed in. Further, there are tools like SocialIntelligence [13] (an FTC-approved commercial system) can be used by employers to search for information about an employee. If any data of the employee matches the employers search criteria, SocialIntelligence notifies the employer of the same. When compared to PASM, SocialIntelligence would correspond to the data consumer's interface. It provides the same data gathering function for employers, but it does not have any of the data subject privacy policy features. An analogous commercial product for the data subject's interface is Reputation.com [12] which is a service aimed to replace malicious reviews with truthful, positive feedback.

PASM extends RMP by both incorporating linked data during the creation of the policies and by specifying refutation links through which the data subject could create policies over data that they do not personally control or own. Pato *et. al.* [30] demonstrated the prototype of an accountable mechanism that logs provenance and inference results. By using a tool called "Aintno", data subjects would be able to analyze whether, and how, their personal data on social networking sites was used when a particular decision was made about them. PASM attempts to preemptively help the data subjects by enabling them to declare the intended usage of their data using policies.

PASM differs from the approach described in [22] by not using the non-blocking iterator-based approach when traversing linked data. Instead, PASM employs a *user-directed traversal* where data subjects can specify the types of related concepts that they would like PASM to consider when filtering the search results. In the first motivating scenario, PASM would obtain the following concepts related to "heart attack" when creating Alice's policy – *Disease or Syndrome*, *Pathologic Function*, *Drugs*, and, *Biologic Function*. Alice then includes the first and third types in her policy, as shown in Fig. 3-15.

PASM differs from PubSubHubub [29] because it does not actively push linked data updates and re-run the search. This is done so that as data consumers do not receive additional information about data subjects unless they perform another explicit search. This practice would also be conducive for information accountability because explicit logging of the search queries can be performed. Lastly, by not storing the related resources directly in the policy, PASM performs a new search using the types specified in the data subject's policy each time a data consumer performs a search about that data subject. Thus, if a new resource has been added to the Linked Data ontology since the time the policy was created, PASM would be able to incorporate that resource during its semantic enhancement procedure.

Chapter 6

Conclusion

This thesis demonstrated a need for allowing users to control their privacy on the Web and described a system that accomplishes this task – Policy Aware Social Miner (PASM). With the help of PASM, data subjects are given the ability to create usage restrictions over the data they control and refutation links for data that they do not control. Trusted third parties can provide refutations for data subjects who can then choose to verify or decline those third-party refutations. PASM also provides a way for the data subjects to semantically enhance their usage restrictions by traversing linked data. Finally, PASM provides a way for the data consumers to execute policy-aware searches of data subjects.

The significant contributions of this thesis are: a framework called PASM for policy-aware search on the Web, an algorithm for data subject-based linked data traversal for *Semantic Enhancement* of the data subject's usage restrictions, and, a user study that measured the performance of PASM in identifying sensitive posts as compared to the study participants.

Three of the interesting findings from the user study are as follows. Some of the study participants considered the privacy concerns of people mentioned in the posts when making decisions about sensitivity of those posts. Some other participants consider privacy as multi-dimensional (posts that mentioned work schedule, location information, etc. were considered sensitive by them). We also found out that the *Semantic Enhancer* component in PASM is a useful tool to educate users about terms that could potentially be considered sensitive.

6.1 Future Work

Looking forward, the following improvements will be beneficial to PASM:

- PASM gives a lot of control to the data subjects who can potentially abuse the system by specifying very broad usage restrictions or by providing misleading refutations. Further, the data consumer may have legitimate reasons and rights to view certain content, but PASM may not allow that. One avenue for further investigation is to analyze the needs of the data consumers and provide stronger incentives for them to participate in policy-aware searches.

- The linked data traversal fetches related concepts based on the users' input but does not make use of semantic input from the user. For example, understanding the context in which certain terms are sensitive for the user would help PASM understand and apply these usage restrictions in a more robust manner.

- The data subjects may or may not necessarily know what related concepts are collected and whether they are sensitive or not. It would be worthwhile to investigate whether having the data subjects select the endpoints they are interested in or the categories of those endpoints would be beneficial.

- Some of the study participants mentioned that it would be helpful to give them an option to select the topics they consider to be sensitive and specify the amount of notifications they are comfortable with receiving.

- For better security, data subjects could adopt a *decentralized approach* and host their policies on their own servers. A possible approach is to use data.fm [2], which is a read/write linked data service.

- It would be good to extend PASM to support semantic enhancement over other domains besides the medical domain.

Chapter 7

Appendix

7.1 Posts

The following 15 posts were used in the user study.

1. Getting pumped for the big walk tomorrow to help raise money for Juvenile Diabetes research. So thrilled to be part of Drews Crew and honored to show our support for one VERY amazing 4th grader and his loving, dedicated family

2. Had a scary day today. My blood sugar went over 400 and no matter how much Insulin I gave myself, it would not come down. Finally got it down to 101 about 30 minutes ago. I am relieved but I still feel kind of crappy. Stupid Diabetes. If anyone has an extra pancreas laying around, let me know...mine is broken.

3. Metformin plays havoc with me a lot, but it may be doubly worth it in the end

4. I ate a burrito,had a cupcake and 2 cups of coffee this morning. Just checked my sugar, 88. Whats up with that? Could it be my 1000 mgs of Metformin??? wow, thank you God. He must be watching over ME. I wonder what my test will say in a couple of weeks.

5. Had CT Scan this afternoon. Everything went OK. Had to take Prednisone (steroid) for a day and Benadryl because I am allergic to iodine. Also had to be off of Metformin???

(my diabetes pill) Tuesday night and can not start taking it until Sat night. Now to wait for the results. one down....two to go. My brother, Alan, had to take me because did not know how I would react to benadryl.

6. I learned a good lesson today do not drink milk with metformin OMG! I wanna go home n lay with my body pillow.', 'I, a good lesson, a good lesson today, home, home n, metformin, metformin OMG, milk, my body, my body pillow, not drink milk, today

7. Finally got back from the doctor office. I finally got put on metformin, so hopefully all that stuff gets back in swing. And I finally got my knee looked at. If the stars are aligned correctly, I just have to do physical therapy. However, worst case scenario, it will need surgery. Here is to hoping for physical therapy

8. Weekend football wrap up. Wins for ASU and the Cards. Yay! Another loss for Army and T-bird had a 3 game win streak snapped.

9. Who loves Georgia Football? Go Dawgs

10. Oh yeah I will be working but but I am sure someone will keep me posted —Monday night football —Go Dallas Cowboys

11. Home from the hospital and all thing are ok thought i was having a heart attack sunday morning went to the hospital had all kinds of tests sunday and this morning not a thing was found wrong and i am very happy how you all have to put with for a while longer.

12. I never sweat or lose my cool when I audit multi million dollar businesses But I have a mini heart attack when I think about my daughter going to highschool

13. My cholesterol is 251 again. At least Kaiser is fast. They emailed me the results 4 hours after I got my blood drawn. As far as I am concerned, Earth Balance to the rescue

14. The cholesterol is back up again, no wonder my head feels like its swimming again

15. Learned about the funny noun Esmolol today and its importance in my life.

7.2 Questions asked in the Study

The user study was divided into the following three parts – a demographic survey, rating the posts, and, an exit survey.

The following questions were asked during the demographic survey. The participants had the option to leave these questions blank, or select from one of the options shown below.

- What is your gender? Options: Male, Female
- What is your age? Options: 18-20, 21-25, 26-30, 31-40, 41-50, 51-60, 61 or greater
- What is your highest education level? Options: High School, College, Graduate
- Are you concerned about your privacy on social networking sites? Options: Yes, No
- How often do you check/change the privacy settings of your social networking profile? Options: Weekly, Monthly, Yearly, Rarely, Never
- Do you use social networking sites? Options: Yes, No
- If yes, please check all that apply
  - Facebook:
    - What I do. Options: Private, Public, Other
    - Where I go. Options: Private, Public, Other
    - Pictures (myself). Options: Private, Public, Other
    - Pictures (myself and family). Options: Private, Public, Other
    - Pictures (myself and friends). Options: Private, Public, Other
    - Health-related. Options: Private, Public, Other
    - Private Information. Options: Private, Public, Other
    - Other. Options: Private, Public, Other
  - Twitter:

    * What I do. Options: Private, Public, Other
    * Where I go. Options: Private, Public, Other
    * Pictures (myself). Options: Private, Public, Other
    * Pictures (myself and family). Options: Private, Public, Other
    * Pictures (myself and friends). Options: Private, Public, Other
    * Health-related. Options: Private, Public, Other
    * Private Information. Options: Private, Public, Other
    * Other. Options: Private, Public, Other

- Google+:
    * What I do. Options: Circles, Public, Other
    * Where I go. Options: Circles, Public, Other
    * Pictures (myself). Options: Circles, Public, Other
    * Pictures (myself and family). Options: Circles, Public, Other
    * Pictures (myself and friends). Options: Circles, Public, Other
    * Health-related. Options: Circles, Public, Other
    * Private Information. Options: Circles, Public, Other
    * Other. Options: Circles, Public, Other

Rating of the posts during the study is described in section 4.3. Finally, the following questions were asked in the exit survey.

- What do you think of a system like PASM that tries to identify if your posts are sensitive?
- What do you think about the justifications that PASM showed you. What about the URIs and their content?
- Would you rather prefer false negatives or false positives?

7.3 Alice's Policy

Alice's policy based on the scenarios described in chapter 2 is shown below.

```
<http://musigma.csail.mit.edu:2020/profiles/pasm#refutation1>
      a       <http://musigma.csail.mit.edu:2020/profiles/pasm#RefutationLink> ;
      <http://musigma.csail.mit.edu:2020/profiles/pasm#hasOffendingLink>
              "http://www.wickedlocal.com/cambridge/news/x392617482/Cambridge-School..." ;
      <http://musigma.csail.mit.edu:2020/profiles/pasm#hasRefutingLink>
              "http://alicemetzger.com/blog?p=12" ;
      <http://musigma.csail.mit.edu:2020/profiles/pasm#hasThirdPartyRefutation>
              <http://musigma.csail.mit.edu:2020/profiles/pasm#tpr1> ;
      <http://musigma.csail.mit.edu:2020/profiles/pasm#hasTrustedParty>
              "100001504048992" .

<http://musigma.csail.mit.edu:2020/profiles/pasm#tpr1>
      a       <http://musigma.csail.mit.edu:2020/profiles/pasm#ThirdPartyRefutation> ;
      <http://musigma.csail.mit.edu:2020/profiles/pasm#hasContent>
              "I met with Alice two days after and can attest that she was, in fact, not arrested." ;
      <http://musigma.csail.mit.edu:2020/profiles/pasm#hasSupporter>
              "100001504048992" ;
      <http://musigma.csail.mit.edu:2020/profiles/pasm#hasRequester>
              "815853080" ;
      <http://musigma.csail.mit.edu:2020/profiles/pasm#hasStatus>
              "Verified" ;
      <http://musigma.csail.mit.edu:2020/profiles/pasm#hasURI>
              "" .

<http://musigma.csail.mit.edu:2020/profiles/pasm#usage1>
      a       <http://musigma.csail.mit.edu:2020/profiles/pasm#UsageRestriction> ;
      <http://musigma.csail.mit.edu:2020/profiles/pasm#hasKeyword>
              <http://dbpedia.org/resource/Myocardial_infarction> ;
      <http://musigma.csail.mit.edu:2020/profiles/pasm#hasKeywordLabel>
              "Heart Attack" ;
      <http://musigma.csail.mit.edu:2020/profiles/pasm#hasRMPRestriction>
              <http://dig.csail.mit.edu/2008/02/rmp/rmp-schema#No-Medical> ;
      <http://musigma.csail.mit.edu:2020/profiles/pasm#hasType>
              <http://linkedlifedata.com/resource/semanticnetwork/id/T047> .

<http://musigma.csail.mit.edu:2020/data815853080#me>
      <http://musigma.csail.mit.edu:2020/profiles/pasm#hasRefutationLink>
              <http://musigma.csail.mit.edu:2020/profiles/pasm#refutation1> ;
      <http://musigma.csail.mit.edu:2020/profiles/pasm#hasUsageRestriction>
              <ttp://musigma.csail.mit.edu:2020/profiles/pasm#usage1> .
```

Figure 7-1: *Alice's policy*

Bibliography

[1] Creative Commons. http://creativecommons.org/.

[2] Data.fm. http://data.fm.

[3] Dataset – diseasome. http://www4.wiwiss.fu-berlin.de/diseasome.

[4] Dataset – drugbank. http://www4.wiwiss.fu-berlin.de/drugbank.

[5] Dataset – sider. http://www4.wiwiss.fu-berlin.de/sider.

[6] Dbpedia sparql endpoint. http://dbpedia.org/sparql.

[7] Facebook newsroom. http://newsroom.fb.com/key-facts.

[8] Facebook OAuth. http://developers.facebook.com/docs/reference/dialogs/oauth/.

[9] Google Custom Search. http://www.google.com/cse.

[10] LinkedLifeData SPARQL endpoint. http://linkedlifedata.com/sparql.

[11] N3. http://www.w3.org/teamsubmission/n3/.

[12] Reputation.com. http://www.socialintel.com/.

[13] SocialIntelligence. http://www.socialintel.com/.

[14] SPARQL. http://www.w3.org/tr/rdf-sparql-query/.

[15] We know what you're doing. http://www.weknowwhatyouredoing.com/.

[16] E. Aimeur, S. Gambs, and A. Ho. UPP: User privacy policy for social networking sites. In *Internet and Web Applications and Services, 2009. ICIW'09. Fourth International Conference on*, pages 267–272. IEEE, 2009.

[17] L. Brandimarte, A. Acquisti, and G. Loewenstein. Misplaced confidences: Privacy and the control paradox. *Social Psychological and Personality Science*, 2012.

[18] C Clifton, M Kantarcioğlu, A Doan, G Schadow, J Vaidya, A Elmagarmid, and D Suciu. Privacy-preserving data integration and sharing. In *Proceedings of DMKD 2004.*, New York, NY, USA, 2004. ACM.

[19] R Ennals, D Byler, J Agosta, and B Rosario. What is disputed on the web? *Proceedings of the 4th workshop on Information credibility WICOW 10*, 2010.

[20] R Ennals, B Trushkowsky, and J Agosta. Highlighting disputed claims on the web. *Proceedings of the 19th international conference on World wide web WWW 10*, (Figure 3):341, 2010.

[21] O. Hartig and J. Freytag. Foundations of Traversal Based Query Execution over Linked Data (Extended Version). *Arxiv preprint arXiv11086328*, page 15, 2011.

[22] O. Hartig and J.C. Freytag. Foundations of traversal based query execution over linked data. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 43–52. ACM, 2012.

[23] R. Isele, A. Harth, J. Umbrich, and C. Bizer. Ldspider: An open-source crawling framework for the web of linked data. In *Poster, International Semantic Web Conference*, 2010.

[24] S. Jahid, S. Nilizadeh, P. Mittal, N. Borisov, and A. Kapadia. Decent: A decentralized architecture for enforcing privacy in online social networks. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, pages 326–332. IEEE, 2012.

[25] T. Kang and L. Kagal. Enabling Privacy-Awareness in Social Networks. In *Intelligent Information Privacy Management*, volume 2010, pages 98–103. AAAI, 2010.

[26] M. Madden. Privacy Management on Social Media Sites. *Pew Internet Report*, 2012.

[27] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *Security and Privacy, 2009 30th IEEE Symposium on*, pages 173–187. IEEE, 2009.

[28] H. Nissenbaum. Privacy as Contextual Integrity. *Washington Law Review*, 79, 2004.

[29] A Passant and P Mendes. sparqlpush: Proactive notification of data updates in rdf stores using pubsubhubbub, 2010.

[30] J. Pato, S. Paradesi, I. Jacobi, F. Shih, and S. Wang. Aintno: Demonstration of Information Accountability on the Web. In *Privacy, security, risk and trust (PASSAT), 2011. IEEE.*, pages 1072–1080. IEEE, 2011.

[31] D. Pozen. The Mosaic Theory, National Security, and the Freedom of Information Act. *Yale Law Journal*, 2005.

[32] B. Russell. Facebook firings and twitter terminations: The national labor relations act as a limit on retaliatory discharge. *Wash. JL Tech. & Arts*, 8:29–61, 2012.

[33] Y. Song, P. Karras, Q. Xiao, and S. Bressan. Sensitive label privacy protection on social network data. In *Scientific and Statistical Database Management*, pages 562–571. Springer, 2012.

[34] A.C. Squicciarini, M. Shehab, and F. Paci. Collective privacy management in social networks. In *Proceedings of the 18th international conference on World wide web*, pages 521–530. ACM, 2009.

[35] A.C. Squicciarini, M. Shehab, and J. Wede. Privacy policies for shared content in social network sites. *The VLDB JournalThe International Journal on Very Large Data Bases*, 19(6):777–796, 2010.

[36] A.C. Squicciarini, H. Xu, and X.L. Zhang. Cope: Enabling collaborative privacy management in online social networks. *Journal of the American Society for Information Science and Technology*, 62(3):521–534, 2011.

[37] J. Staddon, D. Huffaker, L. Brown, and A. Sedley. Are privacy concerns a turn-off?: engagement and privacy in social networks. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, page 10. ACM, 2012.

[38] K. Thomas, C. Grier, and D. Nicol. unfriendly: Multi-party privacy risks in social networks. In *Privacy Enhancing Technologies*, pages 236–252. Springer, 2010.

[39] D. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G. Sussman. Information Accountability. pages 82–87, June 2008.