# Analysis of Social Networking Privacy Policies

Leanne Wu, Maryam Majedi, Kambiz Ghazinour and Ken Barker
Department of Computer Science
University of Calgary
2500 University Dr. N.W.
Calgary, Alberta, Canada
{lewu, mmajedi, kghazino, kbarker}@ucalgary.ca

## ABSTRACT

As the use of social networks becomes more widespread and commonplace, users are beginning to question how their privacy is protected by social networks. In this paper, we review a privacy taxonomy for data storage polices and models and extend it to support social networking. We then apply the extended taxonomy to the privacy policies of six commonly used social networks, and present our findings with regards to how the published privacy policies of these social networks protect the privacy of users in reality.

## 1. INTRODUCTION

Social networking sites and services are an increasingly important part of how users experience the online world. They provide a means for people to share their thoughts, pictures, and other items they find interesting with their friends. But this same information may consist of items that they would be uncomfortable sharing with strangers.

Recent cases, such as Canada's challenge to Facebook's privacy policies, have shown a growing awareness on the part of the public with respect to how social networking sites and services treat data entrusted to them. It is important to understand to what extent the privacy of users of social networks is actually being protected, and also to understand how individual social networks compare to each other in terms of data privacy so consumers can make an informed choice about using their services. This paper presents a means to classify privacy policies of social networking sites and services, apply this technique to the privacy policies of six social networking sites and services, and present the results of such a classification process.

### 1.1 Social Networks & Related Work

Social networks refer to the informal connections individuals make amongst their friends and acquaintances. We observe that social networks are a phenomenon that are exploited by social networking sites, which strive to transform relationships between people and groups of people which already exist into an online network which can be traversed and exploited.

Boyd and Ellison[2] describe three fundamental characteristics of a social networking site. They observe that these sites provide a means for users to publish a profile which identifies themselves, for these users to identify other users with whom they share an acquaintance, and for the links between users to be used as a means to navigate through the user base.

Increasingly, social networking sites are now transforming into social networking services, and strive to bring users information about their social networks through any means available, not only the websites operated by the social networking services.

In this study, we examine the privacy practices and policies of six social networks: Facebook, LinkedIn, MySpace, Orkut, Twitter, and YouTube. Facebook and MySpace are selected for their widespread use, which are available at their respective websites. Orkut and YouTube are selected because they share a privacy policy in addition to their own privacy policies, LinkedIn is selected for its focus on workplace relationships and Twitter is selected because of its widespread use as a social networking service rather as primarily a social networking site.

This paper extends the existing privacy taxonomy [1] so that it can be used for social networks in addition to its intended use for data stores. We identify commonalities that exist for social networking services when placed in the extended privacy taxonomy. We analyze the privacy policies of six different social networks by identifying layers of data common to all, and derive areas of improvement for these policies.

As social networks become increasingly common, there has been increased interest in research on the privacy implications. Boyd and Ellison's [2] social networks survey provides a discussion of the characteristics common to social networks. One of the most comprehensive surveys of privacy in social networks by Preibusch *et al.* [5] deals with the data solicited by most social networks and divides the data into four categories which are based on the visibility of the data.

Studies on privacy in social networking services tend to concentrate on the prevalence of linkage (or inference) attacks. Since users of social networking services are explicitly connected together, such studies[5][3] believe such attacks are easy to execute. A study by Krishnamurthy and Wills [4] characterized privacy leakages within MySpace and Facebook, since these can lead to inference attacks. They


*PAIS 2010*, March 22–26, 2010, Lausanne, Switzerland.

describe the data gathered by the social network as "privacy bits," which are groups of data over which privacy controls are initiated by the same process. This is another manner in which to divide the data provided to various social networks into manageable units for analysis.

These studies highlight the importance of improving privacy in social networking sites and services. Yet such solutions will not be wholly technical, but require better privacy practices on the part of the social networking site or service and of individual data providers. To improve these practices, we must first understand areas of weakness in privacy policies published by various social networking services.

## 2. A TAXONOMY FOR PRIVACY

Barker *et al.*'s [1] privacy taxonomy is intended to help researchers classify and compare different privacy policies and models. Applied to social networking, this taxonomy allows us to identify what predicates are involved in privacy policies mentioned by social networking services.

We classify privacy policies and models by first determining what will be the *house*, or the party which is storing private data. For this study, the house will be each of the social networks under discussion. We also must determine the role of the social network user or member, about whom data is being stored. We define this party to be the *data provider* or *provider*, and it is the degree to which this party's privacy is being protected which we will attempt to determine.

The elements that form a privacy policy are purpose, visibility, granularity and retention that are relatively orthogonal to each other [1].

**Purpose:** Data providers have different motivations for providing data to other parties, and different expectations for how that data is used. Therefore, purpose is a fundamental property of data privacy. We define purpose as a specification of the legitimate reasons to access a specific piece of data or information. The purpose can be defined for a singular use, reused multiple times in different ways, or allow for unrestricted public access.

**Visibility:** Visibility defines who is allowed to access the provided data, for a legitimate purpose. This is crucial in a social network, where data is potentially available to parties the provider did not intend to provide access to. According to the taxonomy, data visibility can be restricted to data providers themselves, to the house, third parties and most universally, the whole world.

**Granularity:** The granularity predicate specifies the degree of precision that is revealed in response to a query for a given piece of data. When the exact value of the data (known as *microdata*) stored by a data store is returned, this is at a level of granularity defined as *specific*. Data is often aggregated in some manner in response to a query, which gives us a level of granularity of *partial*. The *existential* level offers even less precision, as the data store only reveals whether a given piece of data exists. Adjusting the granularity of the data provided in response to queries is a valuable technique in privacy protection, especially when dealing with sensitive values which exist in the data.

**Retention:** Retention specifies the time period during which access to data should be allowed. in general, the shorter the retention period, the less likelihood that the data in storage can be used in order to reveal information about the data provider. In the taxonomy, retention is expressed as a date (upon which the retention period of the data ends), or as $\infty$, which indicates that no retention period has been specified.

## 3. APPLYING THE TAXONOMY TO NETS

We study the privacy policies of six well-known online social networks by extracting parts of their privacy policies that are available online. The privacy policies of these social networking sites are often internally inconsistent, which we will selectively illustrate using the data privacy taxonomy. It should be mentioned we consider any inconsistencies between what is stated in these social network policies and actual privacy practice by these social networking sites to be out of the scope of this paper.

Six social networks were considered for placement on the privacy taxonomy and it becomes immediately clear all solicit data on the understanding that any of this data may be reused for any purpose, and that when accessed, the data will be accessed for its specific values. As a result, the placement of the privacy polices for all six social networks forms six nearly-identical lines in the same spot, with the lines depicting Facebook and Orkut being a bit shorter than the others, due to slightly more restrictions being placed on their data. We note that all of the privacy policies involved do not have clear statements with regards to the retention of the data provided.

Even though all of these privacy policies are similar when placed on the taxonomy, we observe that these sites and services are quite different in terms of how they handle user data, and furthermore, all of the privacy policies indicate that some data is handled differently than other data. We also observe that the use of "third-parties" in terms of visibility is not expressive enough in the context of social networks, since some third-parties will be treated differently than others.

### 3.1 Taxonomy extension to social networking

Social networking sites and services also introduce another group of third parties who can view a provider's data via the social network. While we can call the user's list of immediate contacts and friends "third parties," in general, the data provider will permit these users a greater degree of access than other types of third parties (which might include external companies given access to the data by the house). Thus, for the purposes of this study, we add the category "friends" to the visibility dimension, and place them between the house (the social network) and third parties.

We note that there is another group of users within a given social network who have access to a data provider's data. These are second-degree connections, the "friends-of-friends" who can access the data provider's data by navigating through a common contact or friend. This group of users may not have the same degree of accessibility to a given data provider as the contacts they share with the data provider, but this group typically has a different variety of access to the provider's data than a conventional third party. Thus, we also add the category "friends of friends" to the visibility dimension.

Another group who may enjoyed a greater degree of access to a data provider's data on a social network is the group of users who are not directly connected via common friends to the data provider, but who are members of that social network. In fact, the categories of "friends," "friends of friends," and the entire network can be defined between

**Table 1: Legend for Tables 2,3,4**

| Purpose | Visibility | Granularity |
|---|---|---|
| RSm=Reuse Same | H=House | S=Specific |
| RS=Reuse selected | F=Friends | P=Partial |
| RA=Reuse Any | FoF=Friends of Friends | |
| A=Any | N=Network | |
| | AW=All/World | |

the house and third parties on the visibility predicate of the social network privacy taxonomy, in the order listed.

## 3.2 Layers of Data

We have shown that social networking services share a great deal of similarities, to the point that when they are plotted on the privacy taxonomy, they are essentially identical. However, these services treat privacy differently, and thus, the question of how we can further analyze these services to identify these differences arises.

The privacy policies of most social networking services distinguish between four main layers of data, which mainly differ from each other in terms of purpose. We call these groups of data layers after the layered protocol used in networking, and because the data in each group often has very different privacy practices compared to the data in other groups. We note that each layer is not entirely discrete, since some pieces of data may be grouped as a part of adjacent layers, and some privacy policies may be vague as to the purpose of particular pieces of data.

**Registration:** This layer consists of the information required to identify the data provider uniquely among all the other users of the social network. Most of this data is personally identifiable, which is captured explicitly by the privacy policies of most social networks.

**Networking:** This layer consists of the information solicited by the social network to be released to its other users, in order to construct a network of contacts for the data provider. Note that with every contact added to a data provider's network additional information is released.

**Content:** This layer consists of the actual content with which the data provider actually participates in the social network. Discussion of this layer is often implicit in the privacy policies published by social networks, because individual users generally can choose who else can view this content. The actual privacy practices put into place to protect this layer can be complex, due to the preferences of individual users, and are not usually captured in the privacy policies of social networks.

**Activity:** This data consists of web server logs, information from cookies, as well as other means of gathering information about the data provider's activities on the social networking service. Data in this layer is often aggregated and provided to third parties for a variety of uses.

# 4. CLASSIFYING LAYERS OF DATA USING THE PRIVACY TAXONOMY

Now we consider each layer for each social network, and assign to each layer a position in the taxonomy. For the purpose of comparison, we consider each dimension separately, and list in Tables 2,3,4 each layer for each social network, and its corresponding value for each dimension of the taxonomy. We follow this analysis by considering each social network with its component data layers separately.

Table 1 shows the abbreviations used for each of Tables 2,3,4. Every entry for Tables 2,3,4 show what we discover when we place data in each layer (for each social network) on the relevant axis for each of purpose, visibility, and granularity. Retention is not included in this table, because it is not discussed in enough detail by any of the privacy policies. We use an arrow between two abbreviations to indicate that the privacy policy may cover the range between the two points, and a comma to indicate that the privacy policy contains discrete points on the axis.

## 4.1 Comparing purpose across data layers

We find that in general, social networks will collect data for any purpose, or for a single specific purpose, then later reused for any purpose. We find that because the data collected in the registration layer contains the most personally identifiable data, the policies of Facebook, MySpace, Orkut and YouTube aim to protect the data better by promising that this data will not be released as easily as some of the other data provided to them.

Data in the networking and content layers of all social networks is not restricted according to purpose, with the possibility that any of this data may be used for any purpose existing for both layers in all six social networks.

Because the data found in the activity layer can contain personally identifiable information, and because the raw data is of great value to the companies behind the social networks, LinkedIn and Twitter state in their privacy policies that information in their respective activity layers will only be released for very specific purposes.

## 4.2 Comparing visibility across data layers

The table for visibility shows the most diversity in terms of policies that are published for each data level of each social network, and illustrates that the privacy policies of social networking sites and services are most restrictive on this dimension.

The data in the registration layer of LinkedIn for is available only to the house, similar to the registration layers of the other five social networks. However, data in the networking layer may be visible only to third parties who are friends (if the data provider has chosen to protect their profile), or it may be visible to the world. The content layer in LinkedIn consists mostly of messages to other users, and of notifications that other users (who are friends) have acquired additional contacts. Therefore, much of this content is only visible to third parties who are friends. The privacy policy of LinkedIn makes explicit that the data in the activity layer is visible to the house, but that it may be made available to third parties such as potential advertisers.

YouTube, by contrast, makes the network and content layers visible to the the world (or the general public). This includes the user's profile and videos or comments which the user may have posted. This is a significantly different policy that what we have described for LinkedIn.

## 4.3 Comparing granularity across data layers

In general, the granularity of the data provided to social networking sites and services remains specific. The only layer in which common practice is to aggregate data is the activity layer, because of the sheer volume of data generated by the entire user base of a social networking site, in terms

**Table 2: Purpose across all four layers for each social network**

| Purpose | LinkedIn | Twitter | Orkut | Facebook | MySpace | YouTube |
|---|---|---|---|---|---|---|
| Registration | RA | RA | RSm | RA | RS | RA |
| Networking | A | A | A | RA→A | A | A |
| Content | A | A | A | RA→A | A | A |
| Activity | RS | RS | A | RA | A | A |

of web server logs, cookies, *etc.*. However, the aggregation may not be performed to protect user privacy, but for other reasons intended to benefit the social networking service or other consumers of the data. Therefore, this aggregation cannot be considered a privacy-preserving technique.

This content layer presents interesting findings when we examine social networking sites which are now in the process of becoming social networking services, such as Twitter, or Facebook. As data providers begin to use social networking services to aggregate their own data from various services in order to return a "feed" of their activities to their friends, the assumption that any content provided by the data provider be specific in granularity becomes increasingly spurious.

For instance, Twitter imposes a 140 character limit on all posts (or "tweets") by their users. This means that content posted to Twitter, such as pictures and video, is typically not uploaded directly to Twitter's own servers, but to those belonging to third party sites and services such as TwitPic[1], and what is actually posted is a URL (itself often shortened and therefore also aggregated by other third party services such as bit.ly[2].

## 4.4 Retention

Retention is generally overlooked by the privacy policies of the social networking sites and services. For example, the privacy policy of Facebook guarantees that data which had been visible on the social network to various third parties - friends, friends of friends, perhaps other Facebook users - can be made unavailable to these parties when a data provider chooses to deactivate their Facebook account, there is still no promise that Facebook will remove this data entirely. Only when a data provider is discovered to be underage (under the age of 13) does Facebook promise complete deactivation of that user's account and deletion of all of that user's data from their servers.

# 5. TAXONOMIZING THE NETS

Figures 1 and 2 show each data layer plotted for some social networks under consideration. It is clear that this permits the comparison of each data layer within a given social network.

We find that Facebook's privacy policy (not depicted) is consistent between its data layers, with the registration and activity layers describing a line such that purposes fall under "reuse any," the granularity is specific, and visibility varying between the owner and the house, with the option of being accessed by friends of the data provider. The content and network layers describe a similar point in space, except that purposes may vary between "reuse any" and "any," and the visibility of the data is opened also to friends of friends of the user, or even the membership of the social network at large.

Viewed in this manner, Facebook's privacy policy is quite different than that of MySpace or Orkut, both social networks which are viewed to be competitors of Facebook. Orkut differs from Facebook, with the data on the registration layer tightly controlled. In Orkut, we place the registration layer at the "reuse same" part of the purpose dimension, which indicates the use of this data is more restrictive. The data in the other layers, however, is placed with at the "any" point in the purpose dimension, which indicates the use of the data in the other layers is less restrictive than would be found in Facebook. Compared to Facebook, the data in the content and networking layer is much more exposed, since there is no option to restrict the purpose of the data.

LinkedIn, MySpace, and Twitter all provide to their members a choice between protecting selected parts of their user profiles and content, or completely exposing this data to the public. Breaking down their privacy policies to the corresponding data layers and plotting their results on the privacy taxonomy show that these policies are identical from the viewpoint of privacy taxonomy.

We see that the most exposed part of user data on these social networking sites and services belongs to the content layer. The networking layer is nearly as exposed, but the granularity of the data at this layer may be partial rather than specific (so data may be aggregated rather than being exposed as microdata). Data at the registration layer is more controlled, in terms of both visibility and purpose. The most tightly controlled data layer is the activity layer, which is aggregated and only used for specific purposes.

Despite the privacy polices of Orkut and YouTube both referring to the Google privacy policy, the privacy policies and practices of these two social networks are actually quite different. Everything on the networking, content and activity layers on YouTube is exposed to the public, whereas on Orkut this data can only be seen by friends. While the visibility of the registration layer is the same in both social networking sites, the purpose dimension for Orkut is more restrictive than that of YouTube.

We can use the following excerpts from the Orkut and Google privacy policies to illustrate how we plot the the social networks on the privacy taxonomy:

- "As an Orkut member, you can create a profile....This information may be accessed and viewed by other Orkut members, as determined by your privacy settings."
- "When you invite new members into your network or send messages through Orkut, we collect and maintain information associated with those messages, including email addresses and content."

The two excerpts above are associated with the networking and the content layers. Since no purpose is specified, we consider it as "any." The data is being used without any generalization, and therefore the granularity level is "specific." Finally, visibility level includes both the house and other members of Orkut.

Overall, each social network is generally quite good at restricting the visibility of the data, poor at controlling the

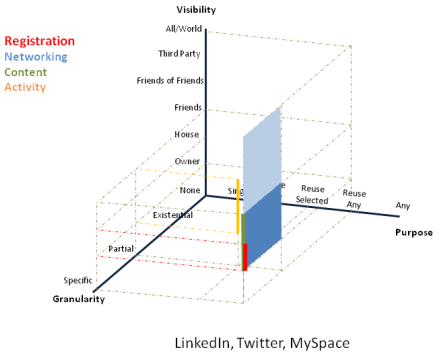
[1] `http://twitpic.com`
[2] `http://bit.ly`

**Table 3: Visibility across all four layers for each social network**

| Visibility | LinkedIn | Twitter | Orkut | Facebook | MySpace | YouTube |
|---|---|---|---|---|---|---|
| Registration | H | H | H | H | H | H |
| Networking | F→AW | F→AW | F→N | H→N | F→AW | AW |
| Content | F→AW | F→AW | F→N | H→N | F→AW | AW |
| Activity | H,TP | H,TP | H | H,TP | H,TP | H |

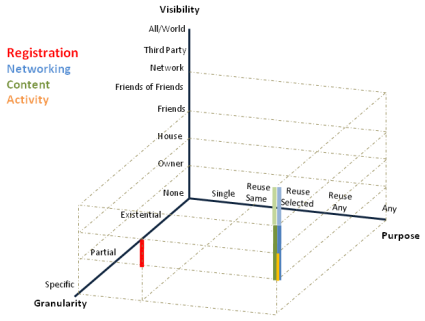
**Table 4: Granularity across all four layers for each social network**

| Granularity | LinkedIn | Twitter | Orkut | Facebook | MySpace | YouTube |
|---|---|---|---|---|---|---|
| Registration | S | S | S | S | S | S |
| Networking | S,P | S,P | S | S | S | S |
| Content | S | S,P | S | S | S | S |
| Activity | P | P | S | S | S | S |




LinkedIn, Twitter, MySpace

**Figure 1: The four data layers of LinkedIn, Twitter, and MySpace on the privacy taxonomy**




Orkut

**Figure 2: The four data layers of Orkut on the privacy taxonomy**

purpose and granularity of the data, with retention not being a consideration at all. We note that in most cases, the networking and content layers are placed on the privacy taxonomy as far away from the origin point as can be possible, which indicates that the data in this layer is at high risk of exposure. Data in the activity and registration layers are better protected and can be placed much closer to the origin point. As most personally identifiable data can be placed within these two layers, it seems that some effort is being made to protect the data considered the most sensitive to the users.

While social networking sites and services are intended to be venues in which members can share their data, and we do not anticipate a rigid privacy policy which makes this difficult, it is clear that the privacy policies of these sites and services may be improved significantly.

## 6. CONCLUSION & FUTURE WORK

By placing social networks on a well understood classification such as Barker et al.'s taxonomy, it becomes immediately clear which aspects of privacy are considered important by these sites. This arises because of the preceived value of the different data generally stored on each. However, the primary focus is clearly on visibility concerns so we call for more work that considers the aspects of purpose and retention. In the absence of such consideration it is clear that social network providers cannot provide the clarity required to help their users understand their policies in an unambiguous way.

## 7. REFERENCES


[1] K. Barker, M. Askari, M. Banerjee, K. Ghazinour, B. Mackas, M. Majedi, S. Pun, and A. Williams. A data privacy taxonomy. In A. P. Sexton, editor, *BNCOD*, volume 5588 of *LNCS*, pages 42–54. Springer, 2009.

[2] D. M. Boyd and N. B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 2008.

[3] J. He, W. W. Chu, and Z. V. Liu. Inferring privacy information from social networks. In *IEEE ICISI*, 2006.

[4] B. Krishnamurthy and C. E. Wills. Characterizing privacy in online social networks. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 37–42, New York, NY, USA, 2008. ACM.

[5] S. Preibusch, B. Hoser, S. Gürses, and B. Berendt. Ubiquitous social networks: opportunities and challenges for privacy-aware user modelling. In *Proceedings of the Workshop on Data*, 2007.