Motivation

Online Vs. Offline Networks

Online Social Networks - Privacy Implications

Analysis: The Facebook.com
  - Patterns of information revelation and their privacy implications

Conclusions
Motivation

Why study privacy in online social networks?

- Two main reasons:
  1. Mass adoption of online social networks
  2. Information revelation behavior of participants
1. Mass adoption

- Number of online social networking sites has increased
- Dramatic increase of online network participants each year
- Important to note:
  - Users may have the same information on different sites
  - Users may be anonymous on some sites and identified on other sites
Motivation
Motivation

2. Information revelation behavior of participants

- Based on observation, there is an apparent openness for individuals to reveal personal information to networks of loosely defined acquaintances and in some cases, complete strangers.

- Why?
Online Vs. Offline Networks

- Social network theory (offline networks) has been used to discuss online incarnations of social networks.

- The specific use of “offline” social network theory to study information revelation (and implicitly, privacy choices) in online social networks highlights significant differences between the offline and online scenarios.
Online Vs. Offline Networks

- Offline social networks contain diverse relations.
  - Examples – Family, Friend, Co-Worker, Roommate, Acquaintance, Classmate, Teammate, Enemy, etc.

- Online social networks simplify relations to simplistic binary relations such as “Friend or not”.
  - How does someone qualify as a “Friend or not”? What is the measurement?
  - Most users tend to list anyone (as a Friend) who they know and do not actively dislike. This often means that people are indicated as Friends even though the user does not particularly know or trust the person.
Online Vs. Offline Networks

- A person’s strong ties may not be significantly increased by online networking technology.

- Weak ties could increase substantially, because the type of communication that can be done cheaper and easier with new technology is more conducive to weak ties.
Online Vs. Offline Networks

- An offline social network may include up to a dozen intimate or significant ties and 1000 to 1700 “acquaintances” or “interactions”.

- Online social networks can list hundreds of direct “friends” and include hundreds of thousands of additional “friends” within just three degrees of separation from a subject.
Online Vs. Offline Networks

- In an online network, thousands of users may be classified as friends of friends of an individual and become able to access her personal information, while, at the same time, the threshold to qualify as a friend is low.

- Hence trust in and within online social networks may be assigned differently and have a different form of meaning than in their offline counterparts.
Privacy implications depend on the information provided to the site.

Specifically:
1. The level of identifiability of the information
2. The possible recipients of the information
3. The possible uses of the information
Online Social Networks - Privacy Implications

1. Level of identifiability
   - Sites that don’t expose user identity may provide enough information to identify the profile’s owner
   - Examples:
     - Face re-identification through photos used across different sites
     - Demographic data
     - Category-based representations of interests that reveal unique or rare overlaps of hobbies or tastes
   - Information Revelation (Two possibilities)
     - Identify “anonymous” profile through previous knowledge of profile owner’s characteristics or traits.
     - Allowing a party to infer previously unknown characteristics or traits about an identified profile.
2. Possible Recipients – Who has access to the profile information?

- Hosting site / Company
- The site’s social network (in some cases site visitors)
- Hackers
- Government Agencies
Online Social Networks - Privacy Implications

3. Possible uses – how can social network profile information be used?

- Dependant upon information provided (may be extensive and intimate in some cases)

- Possible uses (risks)
  - Identity theft
  - Online/physical stalking
  - Embarrassment
  - Blackmail
Online Social Networks - Privacy Implications

- Regardless of implications, information is willingly provided. Why?
- Different factors are likely to drive information revelation.
  - Benefit of selectively revealing data to strangers may appear larger than the perceived costs of possible privacy invasions.
  - Peer pressure or herding behavior.
  - Relaxed attitudes (or lack of interest in) personal privacy.
  - Incomplete information about possible privacy implications.
  - Faith in networking service or trust in its members.
  - Service’s user interface may drive unchallenged acceptance of default privacy settings.
Gross and Acquisti investigate information revelation behavior in online networking using actual field data about the usage and the inferred privacy preferences of more than 4,000 Carnegie Mellon University (CMU) students on Facebook.com.
Analysis - The Facebook.com

- Facebook.com (Circa 2005)

The Facebook is an online directory that connects people through social networks at schools. Now there are two Facebooks: one for people in college and one for people in high school. The site is open to a lot of schools, but not everywhere yet. We're working on it.

You can use Facebook to:
- Look up people at your school.
- See how people know each other.
- Find people in your classes and groups.
Analysis - The Facebook.com

- In 2005 Facebook.com was a college-oriented social network site.
- Intriguing candidate for study. Sense of trust and intimacy may be larger due to the following.
  - Validity expectations may increase due to the requirement of a college e-mail account.
  - Apparent sharing of a physical environment with other members of the network – a college campus.
- Privacy expectations may not be matched by privacy reality.
  - Members can’t control the expansion of their own network.
  - Networks can be easily accessible by outsiders.
In June 2005, the authors searched for all “female” and all “male” profiles for CMU Facebook members using Facebook’s advanced search feature and extracted their profile IDs.

Using the extracted IDs, they downloaded a total of 4540 profiles – virtually the entire CMU Facebook population at the time of the study.
Figure 1: Age distribution of Facebook profiles at CMU. The majority of users (95.6%) falls into the 18-24 age bracket.
Table 1: Distribution of CMU Facebook profiles for different user categories. The majority of users are undergraduate students. The table lists the percentage of the CMU population (for each category) that are users of the Facebook (if available).

<table>
<thead>
<tr>
<th>Category</th>
<th># Profiles</th>
<th>% of Facebook Profiles</th>
<th>% of CMU Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undergraduate Students</td>
<td>3345</td>
<td>74.6</td>
<td>62.1</td>
</tr>
<tr>
<td>Alumni</td>
<td>853</td>
<td>18.8</td>
<td>-</td>
</tr>
<tr>
<td>Graduate Students</td>
<td>270</td>
<td>5.9</td>
<td>6.3</td>
</tr>
<tr>
<td>Staff</td>
<td>35</td>
<td>0.8</td>
<td>1.3</td>
</tr>
<tr>
<td>Faculty</td>
<td>17</td>
<td>0.4</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Table 2: Gender distribution for different user categories.

<table>
<thead>
<tr>
<th>Category</th>
<th>Male</th>
<th>Female</th>
<th>% of Category</th>
<th>% of CMU Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>2742</td>
<td>1781</td>
<td>60.4</td>
<td>-</td>
</tr>
<tr>
<td>Undergraduate Students</td>
<td>2025</td>
<td>1320</td>
<td>60.5</td>
<td>62.0</td>
</tr>
<tr>
<td>Alumni</td>
<td>484</td>
<td>369</td>
<td>56.7</td>
<td>-</td>
</tr>
<tr>
<td>Graduate Students</td>
<td>191</td>
<td>79</td>
<td>70.7</td>
<td>6.3</td>
</tr>
<tr>
<td>Staff</td>
<td>23</td>
<td>12</td>
<td>65.7</td>
<td>-</td>
</tr>
<tr>
<td>Faculty</td>
<td>17</td>
<td>0</td>
<td>100</td>
<td>3.4</td>
</tr>
</tbody>
</table>


In general, CMU Facebook members provided large amounts of information.

- 90.8% of profiles contained an image.
- 87.8% revealed their birth date.
- 39.9% listed a phone number
- 50.8% listed their current residence.
- 62.9% listed their relationship status.

Across most categories, the amount of information revealed by female and male users was similar. A notable exception was the phone number, disclosed by substantially more male than female users (47.1% vs. 28.9%).
Figure 2: Percentages of CMU profiles revealing various types of personal information.
In addition to types of information disclosed, Facebook profiles tend to be fully identified with each participant’s real first and last names.

Easy to connect the real first and last name of a person to the information provided – which may include residence.
How valid is the information?

Determining the accuracy of information is nontrivial for most cases.

Validity evaluation is restricted to the measurement of the manually determined perceived accuracy of information on a randomly selected subset of 100 profiles.
Names were manually categorized as being one of the following.

- **Real Name** – Name appears to be real (example – can be matched to the visible CMU e-mail address provided at login).
- **Partial Name** – Only a first name is given.
- **Fake Name** – Obviously fake name.

**Table 3: Categorization of name quality of a random subset of 100 profile names from the Facebook.** The vast majority of names appear to be real names with only a very small percentage of partial or obviously fake names.

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage Facebook Profiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Name</td>
<td>89%</td>
</tr>
<tr>
<td>Partial Name</td>
<td>3%</td>
</tr>
<tr>
<td>Fake Name</td>
<td>8%</td>
</tr>
</tbody>
</table>
• Vast majority of profiles contained an image (90.8%).

• To assess the quality of the images provided the authors manually labeled them into one of four categories.
  • Identifiable – Image quality is good enough to enable person recognition.
  • Semi-Identifiable – Person is not directly identifiable. Other aspects (hair color, body shape, etc) are visible.
  • Group Image
  • Joke Image
The same evaluation was repeated for Friendster, where the profile name is only the first name of the member (which makes Friendster profiles not as identifiable as Facebook profiles).

Table 4: Categorization of user identifiability based on manual evaluation of a randomly selected subset of 100 images from both Facebook and Friendster profiles. Images provided on Facebook profiles are in the majority of cases suitable for direct identification (61%). The percentage of images obviously unrelated to a person ("joke image") is much lower for Facebook images in comparison to images on Friendster profiles (12% vs. 23%).

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage Facebook Profiles</th>
<th>Percentage Friendster Profiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifiable</td>
<td>61%</td>
<td>55%</td>
</tr>
<tr>
<td>Semi-Identifiable</td>
<td>19%</td>
<td>15%</td>
</tr>
<tr>
<td>Group Image</td>
<td>8%</td>
<td>6%</td>
</tr>
<tr>
<td>Joke Image</td>
<td>12%</td>
<td>23%</td>
</tr>
</tbody>
</table>
Friends networks can also contribute to data validity and identifiability since adding a friend requires explicit confirmation.

Facebook users typically maintain a very large network of friends.

On average, CMU Facebook members list 78.2 friends at CMU and 54.9 friends at other schools.

Figure 3: Histogram of the size of networks for both CMU friends (a) and non-CMU friends (b). Users maintain large networks of friends with the average user having 78.2 friends at CMU and 54.9 friends elsewhere.
Default Settings

- Facebook provides a sophisticated interface to control profile searchability and visibility.
- By default, everyone on Facebook appears in searches of everyone else, independent of the searcher’s institutional affiliation. Search results contain the users’ full names along with the profile image, the academic institution that the user is attending, and the users’ status there.
- Facebook reinforces these default search settings by labeling it “recommended” on the privacy preference page.
- Also by default, the full profile (including contact information) is visible to everyone else at the same institution.
Default Settings

- To test how CMU Facebook members selected their own privacy settings, accounts were created at different institutions.

- **Profile Searchability**
  - Measured the percentage of users that changed the search default setting
    - from being searchable to everyone on Facebook
    - to only being searchable to CMU users.
  - A list of profile IDs currently in use at CMU was created and compared to a list of profile IDs visible from a different academic institution.

Only 1.2% of the users (18 female, 45 male) made use of this privacy setting.
Default Settings

- **Profile Visibility**
  - Evaluated the number of CMU users that changed profile visibility by restricting access to CMU users.
  - The list of profile IDs currently in use at CMU was used to evaluate which percentage of profiles were fully accessible to an unconnected user (not friend or friend of a friend of any profile).

Only 3 profiles (0.06%) were restricted to CMU users only.
The Facebook.com Privacy Implications

- It appears that the population of Facebook users studied is oblivious, unconcerned, or pragmatic about their personal privacy.
- Users may put themselves at risk for a variety of attacks on their physical or online persona.
  - Personal data is generously provided and limiting privacy preferences are sparingly used.
  - Profiles disclose a variety of personal information.
  - Public linkage to real identity.
Stalking

Potential adversary (with an account at the same institution) can determine the likely physical location of the user for large portions of the day based on profile information about

- residence location
- class schedule
- location of last login.
Re-identification

• **Demographics**
  - 45.8% list birthday, gender, and current residence. An adversary with access to the CMU section could link users to outside, de-identified data sources such as hospital discharge data.

• **Face Re-Identification**
  - Using a commercial face recognizer, it was possible to correctly link facial images from Friendster profiles without explicit identifiers with images obtained from fully identified CMU web pages.
Re-identification

- **Social Security Numbers**
  - Hometown and birth-date can be used to estimate the first three and middle two digits of a social security number.
  - Possible to obtain last four digits (often used in unprotected logins and passwords) through social engineering.

- **Identify Theft**
  - Majority of profiles contain current phone number and residence which are often used for verification by financial institutions.
Privacy implications of revealing personal information may extend beyond their immediate impact, which can be limited.

With low and decreasing costs of storing digital information, it is possible to monitor and record the evolution of the network and its users’ profiles, thereby building a digital dossier for its participants.

Users may not be concerned about the visibility of personal information now, but may be later when the data could still be available.
Fragile Privacy Protection

• Mechanisms protecting Facebook’s network can be circumvented.

  • **Fake E-mail Address** – An adversary can receive a confirmation e-mail from Facebook by attempting to remotely access a hacked or virus-infected machine or physically accessing a networked machine.

  • **Manipulating Users** – Social engineering can be used to become a user’s friend to access profile information. According to a cited paper, there is an instance where a Facebook users used an automatic script to contact 250,000 Facebook users across the country and asked to be added as their friend. 75,000 of the 250,000 recipients accepted.

  • **Advanced Search Features** – Facebook makes the advanced search page of any college available to anyone in the network. Various profile information can be searched and used to reconstruct previously inaccessible information by keeping track of returned profile IDs.
Table 5: Overview of the privacy risks and number of CMU profiles susceptible to it.

<table>
<thead>
<tr>
<th>Risk</th>
<th># CMU Facebook Profiles</th>
<th>% CMU Facebook Profiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real-World Stalking</td>
<td>280 (Female) 580 (Male)</td>
<td>15.7 (Female) 21.2 (Male)</td>
</tr>
<tr>
<td>Online Stalking</td>
<td>3528</td>
<td>77.7</td>
</tr>
<tr>
<td>Demographics Re-Identification</td>
<td>1676</td>
<td>44.3</td>
</tr>
<tr>
<td>Face Re-Identification</td>
<td>2515 (estimated)</td>
<td>55.4</td>
</tr>
</tbody>
</table>
Conclusions

• Online social networks are both vaster and looser than their offline counterparts.
  • Possible for a profile to be connected to thousands of other profiles through the network’s ties.

• In the study of CMU users of Facebook
  • Quantified individuals’ willingness to provide large amounts of personal information has been.
  • Shown how unconcerned its’ users appear to privacy risks based on how personal data is generously provided and limiting privacy preferences are hardly used.
  • Based on the information they provide online, users expose themselves to various physical and cyber risks.