Understanding and Specifying Social Access Control Lists

Mainack Mondal
MPI-SWS
mainack@mpi-sws.org

Yabing Liu
Northeastern University
ybliu@ccs.neu.edu

Bimal Viswanath
MPI-SWS
bviswana@mpi-sws.org

Krishna P. Gummadi
MPI-SWS
gummadi@mpi-sws.org

Alan Mislove
Northeastern University
amislove@ccs.neu.edu

ABSTRACT

Online social network (OSN) users upload millions of pieces of content to share with others every day. While a significant portion of this content is benign (and is typically shared with all friends or all OSN users), there are certain pieces of content that are highly privacy sensitive. Sharing such sensitive content raises significant privacy concerns for users, and it becomes important for the user to protect this content from being exposed to the wrong audience. Today, most OSN services provide fine-grained mechanisms for specifying social access control lists (social ACLs, or SACLs), allowing users to restrict their sensitive content to a select subset of their friends. However, it remains unclear how these SACL mechanisms are used today. To design better privacy management tools for users, we need to first understand the usage and complexity of SACLs specified by users.

In this paper, we present the first large-scale study of fine-grained privacy preferences of over 1,000 users on Facebook, providing us with the first ground-truth information on how users specify SACLs on a social networking service. Overall, we find that a surprisingly large fraction (17.6%) of content is shared with SACLs. However, we also find that the SACL membership shows little correlation with either profile information or social network links; as a result, it is difficult to predict the subset of a user's friends likely to appear in a SACL. On the flip side, we find that SACLs are often reused, suggesting that simply making recent SACLs available to users is likely to significantly reduce the burden of privacy management on users.

1. INTRODUCTION

Online social networks (OSNs) are now a popular way for individuals to connect, communicate, and share content; many of them now serve as the de-facto Internet portal for millions of users. On these sites, users are encouraged to establish friendships and upload content, providing an incentive for users to return. As a result, social network users today have hundreds of friends and many thousands of pieces of content. These same users are also expected to *manage* their privacy—i.e., select the appropriate privacy setting for each piece of content—a task that is both time-consuming and complex [34].

While much OSN content is shared with default settings (e.g., visible to all of a user's friends), certain sensitive content is often shared with subsets of friends. For example, on Facebook, users may explicitly enumerate friends to allow or deny the ability to view a photo, or create friendlists for the same purpose. We refer to these resulting sets of users who are able to access content as *social access control lists* (social ACLs, or SACLs); by definition, a SACL is a proper subset of a user's friends who are selected by the user to access a piece of content. Due to the privacy-sensitive nature of the content the SACLs protect, one of the hardest parts of using today's OSN privacy management tools is defining appropriate SACLs for different pieces of content.

Many prior studies have examined the privacy concerns that arise when users share content on Facebook, such as the problem of "over-sharing" content with default settings that make the content visible to everyone in the network [34]. As a solution, researchers have proposed grouping friends into subgroups based on their relationship type (e.g., high school friends, work colleagues, family) or community structure in the one-hop network of the user, and sharing content with specific subgroups [16]. However, most of these approaches rely on small scale user studies where they conduct a survey to understand the privacy preferences of users to evaluate their technique. None of these approaches have been evaluated on how well they capture real privacy preferences specified using SACLs. Given that content shared with SACLs is likely to be the most privacy sensitive (and therefore, likely the most important), having an understanding of the SACLs in-use is crucial to designing improved privacy mechanisms for OSN users.

In this paper, we make three contributions: *First*, we conduct the first large-scale measurement study of use of SACLs in OSNs. Using a popular Facebook application installed by over 1,000 users, we collect a total of 7,602 unique SACLs specified by users.[1] We find that over 67% of users are sharing at least some of their uploaded content using SACLs, and that 17.6% of all content is shared with a SACL; these observations underscore the important and unstudied role that SACLs play in users' privacy management.

Second, we focus on understanding the membership of SACLs (i.e., how are the friends who are allowed to view

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

[1]Our study was conducted under Northeastern University Institutional Review Board protocol #14-01-09.

a piece of content similar to each other, but different from other friends?). Examining the in-use SACLs that we collected, we find that for less than 10% of SACLs all the members of the SACL share a common profile attribute. Moreover, we find that only 20% of SACLs show strong community structure in the links between their members. Taken together, these results suggest that SACLs are likely to be difficult to detect automatically. This result is surprising given the existing work on automatically grouping friends based on network structure or attributes for better privacy management [16, 39]; We suspect that this difference occurs because these prior studies did not evaluate their techniques against ground-truth data about fine-grained content sharing in OSNs.

Third, we explore the difficulty faced by users in specifying SACLs today. Overall, we find that the complexity of SACLs (as defined by the number of terms[2] a user must select when creating a SACL) is quite high for a non negligible fraction of our users: over 18% of users specify more than 5 terms per SACL on average. We observe that there is significant room for improvement in reducing the burden of specifying SACLs, and we find that simply allowing users to re-use previously used SACLs reduces much of the user overhead: for the vast majority ($> 80\%$) of users, 90% of their content is shared with fewer than 5 unique SACLs.

The remainder of the paper is organized as follows. Section 2 presents background and related work on SACLs and OSN privacy. Section 3 describes how we obtained our SACL data set, and Section 4 provides some high-level statistics on SACLs. Section 5 explores the relationship between SACL members, while Section 6 investigates the user overhead in specifying SACLs today. Finally, Section 7 concludes.

2. BACKGROUND AND RELATED WORK

In this section, we first provide some background on how social networking sites have evolved in helping users to manage their data privacy today. Our focus is on the fine-grained privacy management tools that enable the sharing of privacy sensitive content on OSNs.

In this paper, we focus on the largest social networking site as of March 2014 — Facebook. Up until 2005, Facebook split users on the site into different regional networks (based on geography, workplace or educational institution). By default, each user would share all of her content with everyone in the regional network and the service lacked any concrete privacy controls for sensitive data. By 2009, Facebook had 300 million [15] users and some regional networks grew too large (e.g., in India and China) to be used for privacy settings. There were widespread demands for better privacy management mechanisms for users [7], and by the end of 2009, Facebook rolled out more fine-grained privacy controls.

2.1 Mechanisms for privacy management

In December 2009, Facebook made an important change which allowed users to set access control policies for content they publish on a per-post basis [23]. For example, a user can share a particular photo with only family members and close friends. This change allowed users to customize their privacy settings on a per-content basis, instead of simply adopting the default privacy setting offered by Facebook, which allows access to "everyone" (all users on Facebook) [19].

Facebook introduced an additional mechanism called *friend lists* [26] to complement their existing fine-grained privacy controls. Users can create friend lists and add a subset of their friends to each of these lists. For example, a user can create a list called "co-workers" and manually add all of her friends who are co-workers into that list. This allows the user to group her friends into different lists that might be meaningful to her in terms of sharing content. Now, instead of handpicking individual friends for specifying a privacy setting for each content, users can use their pre-created friend lists for specifying privacy settings (e.g., share this photo with "soccer buddies" list). Friend lists are private to the user who creates them.

By October 2010, Facebook observed that only a small percentage (5%) of Facebook users had ever created friend lists [27]. This could be due to the manual effort required of the users to create and maintain friend lists. To help users further, Facebook started automatically creating friend lists for the user and populated the lists with a specific subset of the user's friends [14]; these lists are called *smart lists*. This automation is done by leveraging the profile attributes of the user and the user's friends, e.g., employer, location, family and education information provided by users. An example would be a list called "Family" that automatically groups all the friends of the user who have marked the user as a family member. In addition, Facebook also creates two empty smart lists for the user, "Close Friends" and "Acquaintances". However, instead of auto-populating these two lists, Facebook only shows friend recommendations to the users based on the interaction between the user and her friends. In this paper, we will refer to all of these Facebook-created smart lists as *Facebook lists*. Moreover, when using the term *lists* we are referring to both the user-created friend lists as well as Facebook lists.

So far, we observed that there are different ways in which a user can specify which friends have access to a piece of content on Facebook today. In the rest of the paper, we will use the term *social access control lists* (social ACLs or SACLs) to refer to such privacy policy specifications. A more precise definition is below.

Social access control lists (SACLs): A SACL is a privacy policy specification attached to a piece of content containing a proper subset of the user's friends; friends specified in the SACL have access to view and perform other actions on the content (e.g, liking or commenting). SACLs can be specified using different mechanisms provided by Facebook: allowing or denying access to individual friends one by one, specifying friend lists, using Facebook lists, or using a combination of handpicked friends and lists.[3]

It is important to note that SACLs *only* encompass custom settings by users and do not include the Facebook predefined access permissions: "everyone" or "public" (share with all Facebook users), "regional network" (share with

[2]When creating a SACL, a user can specify either individual friends or pre-created lists of friends; we refer to both of these as *terms*.

[3]Facebook also allows users who are *tagged* in a specific post to see the content [22, 24]. However, since users did not specify tagged friends explicitly through the privacy management interface, we do not consider them to be the part of SACLs. We leave exploring privacy expressed through tags to future work.

everyone in a regional network, deprecated in 2009), "all friends-of-friends" (share with all friends-of-friends), "all friends" (share with all friends), and "only me" (only visible to the user who uploaded the content).

2.2 Related work

Now that we understand the background of this paper, we discuss related work along three directions.

Understanding privacy awareness of users Researchers have studied the privacy awareness of social networking users [11, 30, 46]. These studies examine the profile information sharing behavior of users over a long period of time (e.g., 7 years) to understand if users' attitude towards their data privacy changes over time. Dey et al. [11] and Stutzman et al. [46] have shown quantitative evidence that Facebook users are sharing fewer profile attributes (such as hometown, birthdate and contact information) publicly over time. Social media discussions about Facebook privacy [8,13,44,45,47] and Facebook regularly rolling out new fine grained privacy management features [21] for the last few years have caught users' attention and potentially increased their concerns about available privacy controls.

How effective are users in managing their privacy settings? Recent studies have explored how effective users are in managing their privacy settings. Studies have shown that there exists a mismatch between desired and actual privacy settings when users share content on Facebook [4,5,29,34,36]. Liu et al. [34] conducted a user survey about privacy preferences for photos uploaded by users on Facebook. They found that privacy settings match users' expectations only 37% of the time, and when wrong, users are exposing their content to a much wider audience (e.g., all friends, friends-of-friends or even everyone on Facebook) than they intended. While the exact reason for incorrect privacy settings is hard to infer, it could be due to poor privacy management user interfaces or the significant cognitive burden required to manage privacy of their sensitive content.

Techniques for better privacy management Several techniques have been proposed to reduce the burden on users when managing their privacy settings. We can organize work in this space into two high level categories: (1) The first approach is to assist in automatically pre-defining grouping of friends that might be meaningful to the user for sharing sensitive content later. Facebook allows the user to pre-define such friend groupings using the friend list feature. But friend lists on Facebook have to be manually specified by the user today and this user overhead could be reduced by these approaches. (2) The second approach is to help the user on the fly to specify SACLs while sharing content. They predict SACL specifications with some input from the user. For example, if the user gives the name of a few friends that he wants to share content with, these approaches can automatically predict the remaining members of the SACL or provide recommendations of other possible SACL members. Now we will explain different proposals that fall in the above two categories.

PViz [39] is a proposal from first category that can automatically detect friend lists for a Facebook user and use it for better privacy policy visualization for the user. It leverages the network structure of the subgraph induced by the user's friends (i.e., the user's "one-hop subgraph") to detect friend lists using a modularity-based community detection algorithm. Using information extracted from friends' profile, it can also automatically assign a label to each detected list. This helps the user to understand the composition of a list. Based on these predefined lists, PViz points out to users which of her friends from a list can view a particular content. PViz presents a user study based on 20 users, who find PViz useful for understanding their existing privacy settings better.

However, many previous works [1, 16, 48] fall in the second category. They focus on recommending friends on the fly to the users as the user starts sharing a piece of content and selects a few intended friends. Privacy Wizard [16] is one example of such a tool. Privacy Wizard leverages network structure and profile attributes (like gender, age, education, work, etc) to recommend friends for inclusion in a privacy setting. The process starts as the user tags a few of her friends as "allowed" or "denied" for a content. It then uses a machine learning algorithm to classify the remaining untagged friends into an allow or denied category. The authors designed a survey experiment with 45 Facebook users and 64 profile data items to evaluate the accuracy of their tool. They observed that on average if a user tags 25 of her friends, the wizard configures her privacy setting with high accuracy. However in their experiments they did not look into the ground truth data on how a user actually specifies SACLs while sharing sensitive content.

We conduct a large scale study comprising of 1,165 users and all their uploaded content to focus on the privacy settings used by users. Most prior work tries to approximate ground truth privacy preference data by asking user privacy preferences explicitly, most of the time via surveys or via a combination of surveys and profile data collection using apps [30]. However, none of these studies looked into the "ground truth" data on SACLs (i.e., how a normal user would share their content using SACLs without any external intervention). To the best of our knowledge, we are the first ones to look into ground-truth data on users' usage of SACLs to propose insights on how to assist users to reduce the overhead of specifying SACLs.

2.3 Key questions

While fine-grained privacy controls put users in better control of their data privacy, it is not clear how users are using these privacy mechanisms. In this work, we take a first look at SACLs specified by 790 Facebook users (users who created at least one SACL) for 212,753 pieces of uploaded content. Our analysis focuses on the following key questions:

- *How are users using SACLs today?* We analyze how often SACLs are specified by users and how different types of content are shared using SACLs.
- *Is SACL membership predictable?* We analyze characteristics of SACL members to understand if they have something in common that other friends of the user do not. Our analysis explores whether members have similar profile attributes, exhibit strong social network connectivity with each other, or share similar activity levels. If we are able to separate out SACL members from among all friends, we may be able to automatically create SACLs for users.
- *What is the user overhead in specifying SACLs?* We

examine the overhead that users spend specifying SACLs today. The intuition is that, the more work required to create SACLs, the less usable the privacy mechanisms are.

- *What is the potential for reducing user overhead?* Based on our insights gained from analyzing SACL membership, we quantify the potential for further reducing the complexity of SACL specifications. Our findings serve as guidelines for designing better privacy management tools in the future.

3. DATASET

Now we describe the dataset we have collected about in-use SACLs on OSNs today.

3.1 Collecting data about SACLs in Facebook

Obtaining data at scale about user privacy specifications is quite challenging. In Section 2.2, we observed that most previous work used small-scale data about user privacy preferences. There are two main challenges in collecting large-scale data: First, we need permission to view the user SACLs. This is challenging as private settings on Facebook are private to the user; they cannot be obtained by crawling publicly visible user profiles. To address this challenge, we use the Facebook API [17], which offers methods to collect data about in-use privacy settings (provided the user gives us permission to do so). We therefore developed a Facebook application that helps users to better manage their privacy settings, and recruited users for the application. The Facebook application requests consent from the user to collect data about their SACL specifications for our research study. The data collection was performed under an approved Northeastern University Institutional Review Board protocol.

The second challenge is recruiting large number of users for the study who can provide consent to access their private SACL information. The traditional approach for recruiting users rely on personal communication (e.g., via email) or through an open call posted on a public bulletin board at a university or research lab. In such cases, the number of users that could potentially be recruited is usually limited to a few tens or hundreds. Another approach is to use online crowdsourcing platforms like Amazon Mechanical Turk (AMT) [38]. However, using a Facebook application violates the AMT policy that workers should not be required to download and install an application [2].[4] Instead, to recruit a variety of users at scale, we leverage the Facebook advertising platform.

3.1.1 Facebook Application: Friendlist Manager

We developed the Facebook app "Friendlist Manager" (or FLM) [25,35] that helps the user to automatically create and update friend lists that could be used for specifying SACLs. This application reduces the user burden of manually creating friend lists. FLM automates list creation by leveraging the network structure in the "one-hop subgraph" of the user. It uses network community detection algorithms [6, 40] to find overlapping communities in the one-hop subgraph. We found that users found FLM to be helpful; 480 (41%) of our users allowed FLM to update at least one of their lists. It is important to note that for the analysis in this paper, we only consider the content users shared and members of friend lists users had *before* installing FLM; this ensures that usage of our app does not impact the results.

When installing FLM, we request permission to access the following types of user data: basic user profile details including workplace, education, current city and family; privacy settings (including SACLs) used for all uploaded content (photos, videos, statuses, notes, music, questions, Shockwave Flash Player (SWF) movies, and checkins); and the friends, friend lists, and Facebook lists of the user. Should the user choose to not grant us access to their content, they are still allowed to use the application.

3.1.2 Recruiting users

To recruit users, we set up an advertisement for FLM on the Facebook advertising platform. The Facebook advertising platform allows us to reach out to the large Facebook population and target users with specific demographics. Our ad included the following text:

> *Need help to better organize your list of friends? Give FLM a try!*

Starting from June 20th 2012, we ran five ad campaigns for 10 days targeting 10 countries where English is an official language: USA, UK, Australia, New Zealand, Canada, India, Pakistan, Singapore, South Africa, and Philippines. In total, 232 users installed our app during this period. After this initial push, our app received a steady stream of new users through August 2013; in total, we observed 1,007 additional users after the advertising campaign ended. We believe FLM also spread "virally", with users "liking" or recommending the app to their friends. While it is hard to trace the source of these new 1,007 users, we found that 59 of them were friends of users who installed FLM through ads. The remaining users likely found FLM through search tools (e.g., Google Search) or through word-of-mouth based propagation.

Overall, a total of 1,239 users installed our application. For the purpose of this study, we only focus on the 1,165 users (94%) who gave us permission to access all the data we required for our research study.

3.1.3 User bias

One potential issue with user studies is a bias in the user population. In our case, it is challenging to obtain a random set of Facebook users. This is a fundamental issue with most user studies, and the common methodology is to carefully characterize the users under study to understand how diverse the users are in terms of demographics. Our user population is by no means random, and we report the demographics and behavior of users below. We believe that the users who installed FLM are those who are interested in creating friend lists to better manage their privacy settings. It is known that Facebook has been promoting friend lists as a way to more efficiently specify privacy preferences [14]. Thus, our user sample is most likely biased towards privacy-aware Facebook users. Additionally, ads can only be shown to users when they are logged in to Facebook; we are therefore likely to get users who are active on Facebook.

3.1.4 Ethical Considerations

The data we collected in FLM is highly privacy sensitive, and we took great care to respect users' privacy. First, we

[4] Our data collection methodology requires users to install a Facebook application.

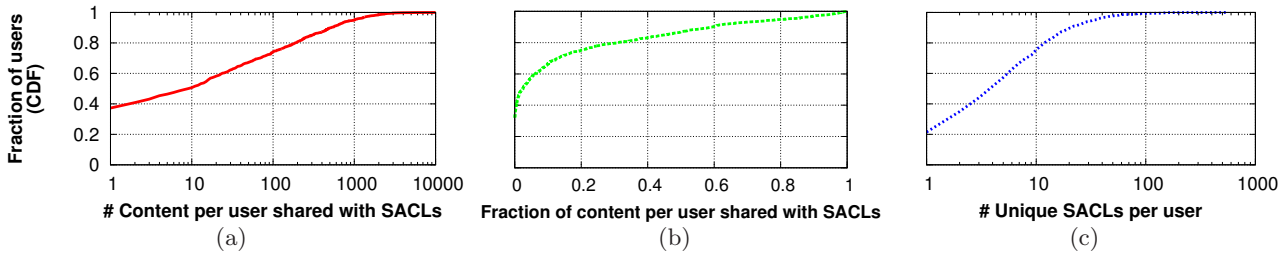

Figure 1: (a) Cumulative distribution of number of pieces of content uploaded by users using SACLs. A significant fraction (67.8%) of users in our dataset upload at least one content using SACLs. (b) Cumulative distribution of percentage of content per user shared using SACLs. More than 200 users in our dataset uploaded more than 30% of their content using SACLs. (c) Cumulative distribution of number of unique SACLs specified by each user who upload at least one content using SACLs.

conducted our study under Institutional Review Board approval. Second, we only report aggregate statistics here, and in any future papers. Third, we will never release any non-aggregated data to third parties. All of these steps were also included as part of FLM's Privacy Policy (provided to users when installing FLM).

3.2 FLM user demographics

We now examine the demographics of users who installed FLM and allowed us to collect their data. Users usually self-report their location, age, gender and education on their profile. We examine the "current location" attribute to estimate the location of users at a country level and find that 952 (82%) users have provided location information. According to this information, users are from 75 countries covering six continents. There were 19.2% users from North America, 18.1% from Europe and 35.5% from Asia. The top five countries were United States (20%), Pakistan (14%), India (7%), Brazil (7%), and Philippines (6%). Thus, we have users from a diverse set of geographic locations.

Next, we examine the age of users. Our users ranged between 18 and 65 and older, with the median age being 29; this distribution is in-line with the overall U.S. Facebook population [41]. Only 1.2% of users did not specify their gender and for the rest, we observed a strong male bias with 76% of our users being male; this differs from the overall U.S. Facebook breakdown of 47% male [41]. Finally, for the education level (reported by 73.8% of users), 67.8% of users have been to college, while 5.9% of them have been to graduate school. All these statistics demonstrate that we have a diverse set of users in our dataset.

As our users are recruited from a social network, one additional concern is that the users might be a "close-knit" group of friends (in terms of friendship), and not a more general sample of the user population. To evaluate whether this is the case, we check how closely related our users are by examining the number of users who are friends on Facebook, and the number of user pairs with at least one common friend. Out of the 678,030 possible pairs of users $[\binom{1,165}{2}]$, 44 (0.01%) were direct friends and 1,266 (0.19%) were not direct friends but had at least one friend in common. Thus, while our population does show some correlation with the social network (unsurprising, given the viral spreading we observed before), the user population is not strongly biased towards one small region of the entire Facebook social network.

Finally, we examine the activity of users in terms of uploaded content. Overall, our 1,165 users have an average of 518 friends (median 332), and uploaded on average 1,040 pieces of content (median 506). 1,003 (84%) users have uploaded more than 100 pieces of content. Only 39 (3.3%) users have uploaded fewer than 10 pieces of content while 3 (0.2%) of them have uploaded none. When we look at activity of users over time, we observe that the activities of our users spanned over 8 years from 2005 to 2013. We further find that 90% of users have been active for more than 20% of weeks since they joined Facebook. Our analysis of users suggests that we have a fairly diverse population most of whom are actively uploading content on Facebook.

4. SACL USAGE

We begin by examining the usage of SACLs by OSN users. Specifically, we investigate how often and for what types of content users specify SACLs.

1. How widely are SACLs used? We first examine how often different users share content with SACLs, using the FLM user set described in the previous section. Figure 1(a) presents the cumulative distribution (CDF) of the number of content shared using SACLs per user. We observe that a majority of our users are using SACLs for content sharing: 790 (67.8%) users out of 1,165 shared at least one piece of content using a SACL. In total, these 790 users uploaded 212,753 pieces of content using SACLs; this content accounts for 17.6% of all content uploaded by all 1,165 users. In the remainder of the paper, we focus only on these 790 users and the content they uploaded using SACLs. We note that the fraction of users using SACLs in our dataset is comparable to that reported for Google+ [31], where 74.8% of the users used SACLs. However, these Google+ users shared significantly more (67.8%) of their content with SACLs. This difference in the percentage of shared contents in Facebook and Google+ is likely due to the differences in user interface between the platforms. We leave a full exploration of the comparative use of SACLs across online social networks to future work.

Next, we observe that users use SACLs to different extents. In particular, we examine the percentage of content that each user shares with SACLs in Figure 1(b) (i.e., for each user, what fraction of their content is shared with a SACL?). We observe a biased distribution across users, but a significant fraction of users select SACLs for much of their

Content type	Total content items	Number shared with SACLs	Percent shared with SACLs
Status	786,800	139,112	17.7%
Photo	264,714	45,308	17.1%
Video	111,676	20,880	18.7%
Album	26,527	4,415	16.6%
SWF	9,794	1,554	15.9%
Note	8,500	883	10.4%
Checkin	3,224	548	17.0%
Question	374	25	6.7%
Music	355	27	7.6%
Offer	9	1	11.1%
Total	1,211,973	212,753	17.6%

Table 1: Distribution of the number and percentage of content shared with SACLs across different types of content.

content: 20% of users share more than 30% of all their content using SACLs.

Thus, we observe that SACLs are widely used by our users for sharing content, which encourages us to further explore the composition of SACLs and complexity of SACL specification in the following sections.

2. How many SACLs do users need to create? Having observed that SACLs are widely used, we now investigate how many different SACLs users create from amongst their friends. Figure 1(c) shows the cumulative distribution of the number of unique SACLs specified by each user. A large fraction (75%) of the users use less than 10 SACLs, and 20% of the users use only a single SACL. However, there are 5 heavy SACL users, who have used more than 100 unique SACLs. We find that these are heavy users of privacy settings and use different combination of a small number of lists and a set of handpicked friends to specify multiple SACLs for multiple pieces of content. Overall, most users only require a limited number of SACLs to share sensitive content; we leverage this finding later in Section 6 to reduce the user overhead in specifying SACLs.

3. Does SACL usage vary with content types? Facebook allows users to upload a variety of content types. Table 1 presents a breakdown of the total number of content items of different types, and the fraction of those items shared with SACLs. We are interested in understanding if users are biased towards a few types of content when using SACLs. The third and fourth columns of Table 1 show the number and fraction of each type of content shared using SACLs. We observe that SACLs are used across all nine different types of content. In fact, 10-20% of almost all types of content are shared with SACLs. Questions and music are shared least often with SACLs; we suspect that these types of content tend to be more public and are usually not privacy sensitive. This widespread use of SACLs across all types of content further justifies looking deeper into SACL membership and complexity, with the goal of increasing the usability of SACLs.

4. How are SACLs created? Facebook users can construct their SACLs in different ways. As mentioned in Section 2.1, while creating a SACL the user may allow or deny access to individual friends, or lists, or use a combination of friends and lists. Table 2 shows the distribution of number

	SACL created using only lists	only friends	both lists and friends
Number of Users	593	555	213
Percent of users using SACLs	75.9%	71.1%	27.3%
Percent of SACLs	33.5%	44.3%	22.2%
Percent of content shared with SACLs	61.4%	27%	11.6%

Table 2: Distribution of the number of users using different mechanisms to create SACLs while sharing contents.

of users using different mechanisms to create SACLs. We observe that more than 70% of users are creating at least one SACL by individually selecting their friends and more than 44% of SACLs fall in this category; this is surprising, as selecting friends individually is a somewhat tedious task. Interestingly, only 27% of SACL content is shared with such SACLs; users share the majority of their content with SACLs created using lists.

5. How many users are in SACLs? Next, we examine the size of SACLs (i.e., how many of a user's friends are in different SACLs). Figure 2 presents a CDF of fraction of the SACL owner's friends that the SACLs contain. We observe that the distribution exhibits three distinct regions, described below:

1. **Include only few friends:** The first region is highlighted in gray on the left side of the graph; this region contains SACLs with between 0% and 5% of the user's friends. This region contains 25% of all SACLs. In the remainder of the paper, we refer to these as include few SACLs.

2. **Exclude only few friends:** The next region is highlighted in gray on the right side of the graph; this region contains SACLs with between 95% and 100% of the user's friends. This region contains 26% of all SACLs. In the remainder of the paper, we refer to these as exclude few SACLs.

3. **Include subset of friends:** The final region is in the middle of the graph; this region contains between 5%

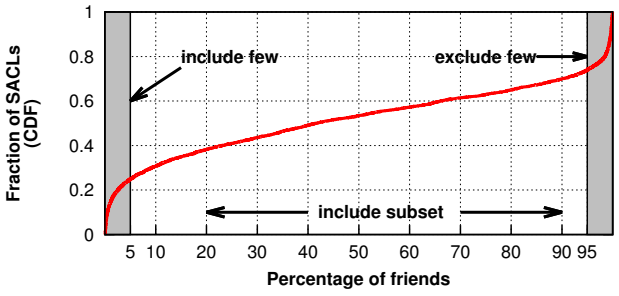

Figure 2: Percentage of friends included in SACLs. SACLs in the left gray region are the include few SACLs, SACLs in the right gray region are exclude few SACLs, and the SACLs in middle white region are the include subset SACLs.

and 95% of the user's friends. This region contains the plurality (49%) of the SACLs. In the remainder of the paper, we refer to these as include subset SACLs.

As we suggest below, the distribution of SACL sizes is very likely influenced by the interface for SACL specification. We use our categorization in the rest of the chapter when we try to characterize the SACL members across different features. However, for exclude few SACLs we also want to see whether we can characterize the excluded friends; for these, we also examine the *excluded members of exclude few SACLs*. The plurality of include subset SACLs shows that our users are not simply including or excluding a handful of their friends, but are often including large subsets of their friends. This result further motivates the need to understand how these subsets are selected.

Overall, our observations suggest SACLs are being widely used today by a majority of our users to control access to a non-trivial fraction of their content. SACLs are used at different rates by different users, but they do appear to be used to share many different types of content. Finally, SACLs show wildly different sizes, with many SACLs containing few or almost all of a user's friends. With this understanding, we turn to examine the membership of SACLs in the following section.

5. SACL MEMBERSHIP

We now take a closer look at the membership of SACLs. In other words, are the members of a given SACL distinguishable from the SACL creator's other friends? (e.g., do the members share a profile attribute?) This question is interesting to examine, as any automatic detection of SACL membership would only work if the SACL members were distinguishable. Moreover, existing work [16, 39, 48] hypothesizes that profile attributes, network structure, and user activity can help us to automatically detect clusters corresponding to SACLs; we aim to see if this is true using our dataset of real-world fine-grained privacy settings.

5.1 Methodology

Our analysis in this section explores the possibility of characterizing SACL members as a group across three features: (i) profile attributes, (ii) social network structure, and (iii) activity. In other words, we would like to see whether the SACL members form a distinct cluster among the friends of the user. To do so, we form clusters based on these three features and then examine how closely the SACLs of the user match our cluster (e.g., we form a cluster of all user's friends who attended the same high school and then look to see if this cluster matches any SACL). To compare our automatically detected clusters and the user's SACLs, we address three separate questions:

1. Do the automatically detected clusters match SACLs? Once we have the clusters of friends for a given feature, for each SACL, we try to find the best matching cluster. To compute the "goodness" of a match, we use the F-score metric [37] which provides a measure of detection accuracy. It is computed as the harmonic mean of precision and recall; F-score varies from 0 to 1, with 1 representing a perfect match.

Unfortunately, a low F-score does not necessarily imply that SACLs are not correlated with automatically detected groups. For example, the members of a SACL could be split between two automatically detected groups; in this case, the F-score for both groups would be quite low, but the F-score of the union of the groups would be quite high. Looking deeper into this issue takes us to our next question.

2. How distributed are the SACLs across clusters? In order to check how widely the SACL members are spread across the clusters we use the metric *entropy* [10]. For a given SACL and a cluster c from a set of automatically detected clusters C, we can compute $p(c)$, the probability of a SACL member belonging to c. Then we measure the *entropy* of the SACL as

$$-\sum_{c \in C} p(c) \log_2 p(c)$$

A higher value of entropy signifies more diversity within the SACL members (i.e., they are spread across more clusters).

To be able to compare across the SACLs which belong to different users (with different numbers of clusters and friends per cluster), we normalize the entropy using maximum possible entropy per SACL. A SACL will have maximum entropy when its members are uniformly distributed across all clusters [10]; in this case the entropy will be $\log_2 |C|$, where $|C|$ is the number of clusters in C. We therefore calculate *normalized entropy* as

$$-\frac{\sum_{c \in C} p(c) \log_2 p(c)}{\log_2 |C|}$$

Normalized entropy for a SACL ranges from 0 to 1. A normalized entropy close to 1 indicates that a SACL is uniformly spread across the maximum number of clusters and a normalized entropy close to 0 indicates that all or most of the SACL members are part of one cluster.

3. How are SACLs different from random groups? One outstanding issue remains: We are examining the entropy of SACLs, but we would really like to measure what's the likelihood of selecting the members of a SACL by pure chance. For example, suppose all of a user's friends attended the same high school; in this case, the "high school" cluster (all friends in a single cluster) would perfectly match any large SACL.

To measure the uniqueness of SACLs relative to random groups, we use the Adjusted Rand Index (ARI) [43] to determine the similarity between SACL and automatically detected clusters. ARI is a similarity metric normalized against chance and varies from -1 to 1. An ARI of 0 indicates no better similarity than a random group, a negative ARI implies worse similarity than a random group, and an ARI of 1 indicates exact similarity. For each SACL, we calculate at the ARI provided by the most similar cluster. If most of the SACLs have ARI close to 0, then the automatically detected clusters are no better in detecting the SACLs than simply using random groups.

5.2 SACLs and user attributes

We first explore whether SACL membership is correlated with a common profile attribute. To do so, we leverage the profile attributes provided by Facebook users, focusing on four attributes: workplace, education, current city, and family. We choose these attributes as they have been shown be most strongly correlated with groupings of users in social networks [40]. Using these attributes, we group the friends

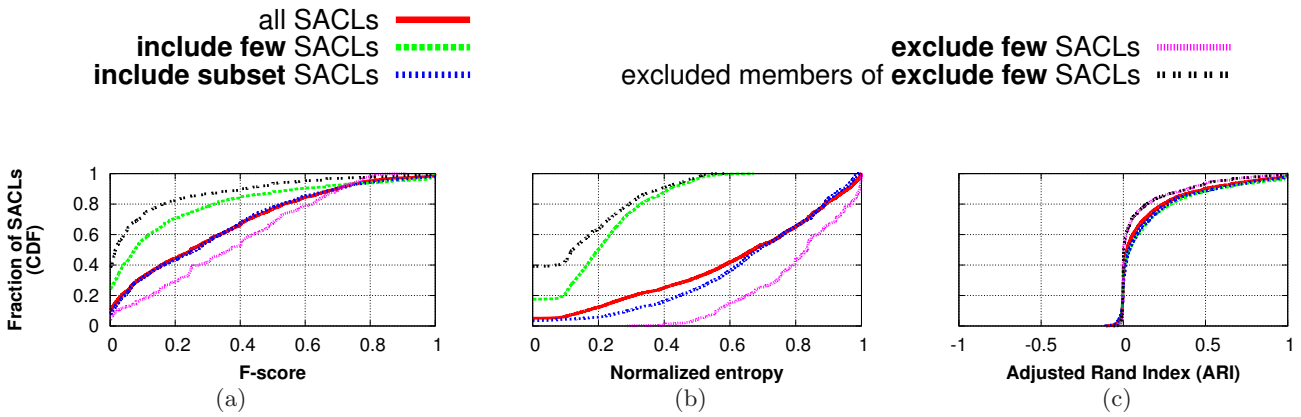

Figure 3: Correspondence between the attribute-based clusters and SACLs, with the cumulative distributions of (a) F-score, (b) Normalized entropy, and (c) ARI. Figure 3(a) shows only 15% of the automatically generated attribute-based communities have a F-score of more than 0.6, indicating low number of SACLs showing high match. Figure 3(b) shows that larger SACLs are spread across multiple such clusters and have higher normalized entropy. The reverse is true for include few SACLs and excluded members of exclude few SACLs. However, Figure 3(c) confirms that more than 40% of these SACLs show better similarity with attribute-based clusters than random groups.

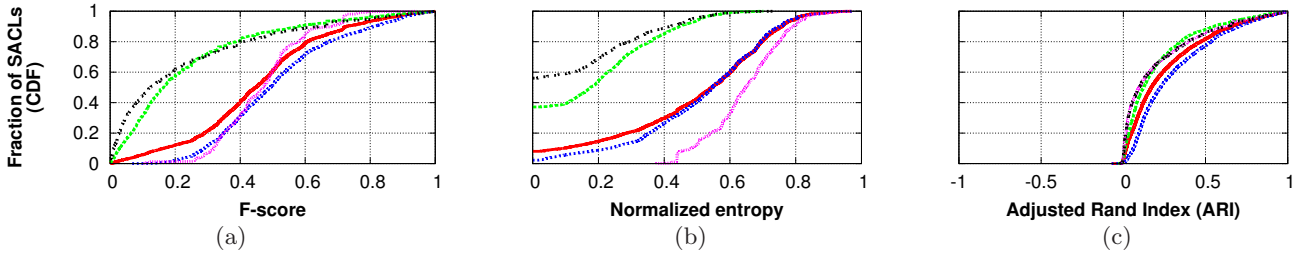

Figure 4: Correspondence between the network communities and SACLs. Figure 4(a) shows 21% of network communities have a F-score of more than 0.6, indicating a relatively poor match between network communities and SACLs. Figure 4(b) and Figure 4(c) confirms that though the larger SACLs have higher entropy (i.e., they are distributed across multiple communities), more than 90% of these SACLs show better similarity with network-based clusters compare to random groups.

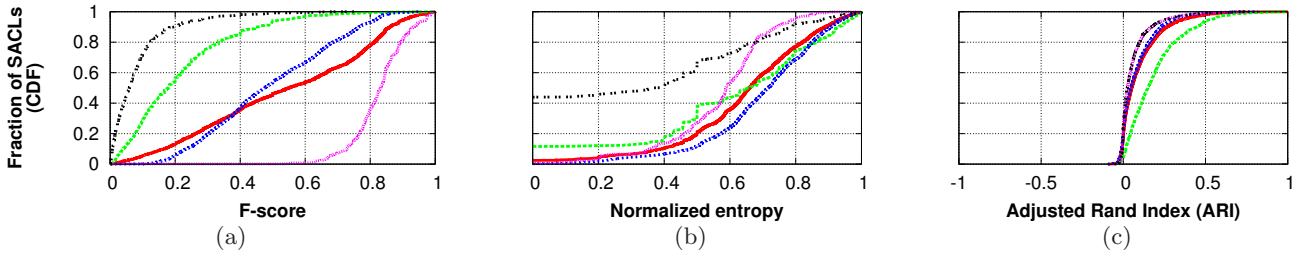

Figure 5: Correspondence between the activity-based clusters and SACLs. Figure 5(a) shows 47% of the automatic attribute clusters shows a F-score of more than 0.6, indicating comparatively strong match between activity communities and SACLs. However, Figure 5(b) shows that the larger SACLs have higher entropy, and Figure 5(c) shows that only 4% have ARI more than 0.3. As a result, random groupings of friends the same size as SACLs would likely show a degree of similar matching.

of a user into clusters who share a common attribute. We report results for all attribute groups in aggregate for brevity; the results are similar when considering each attribute type alone.

We begin by using the F-score metric to check how many of the SACLs exactly match the attribute-based clusters. We present the cumulative distribution of F-scores across all SACLs in Figure 3(a). The figure shows that only 15% of all SACLs have a F-score of more than 0.6, indicating a good match for a small subset of SACLs. The result is even worse for very small SACLs (include few or the excluded members of exclude few), with only 10% of such SACLs having a F-score more than 0.6.

We explore the reason for the low F-scores by analyzing the normalized entropy of these SACLs in Figure 3(b). The figure shows that the small SACLs have a low entropy (with 20% of include few SACLs with entropy 0) indicating they are mostly part of single attribute-based clusters (this is unsurprising, given that these SACLs are small). On the other hand, the larger SACLs show a high entropy with 35% of exclude few SACLs having an entropy of more than 0.8. These results suggest that our attribute clusters are overestimating the smaller SACLs (indicated by low entropy and low F-score) and underestimating the larger SACLs (indicated by high entropy and low F-score).

Finally, we examine whether SACLs match attribute clusters better than random groups using ARI. As mentioned in Section 5.1, an ARI of 0 indicates similarity no better than random groups. Figure 3(c) presents the cumulative distribution of ARI across all SACLs; we observe that 68% of all SACLs have ARI larger than 0, indicating they have more similarity with attribute-based clusters than a purely random set of friends.

Overall, our results suggest that only a small number of attribute clusters serve as a close match for SACLs. However, the SACLs do show some correlation with attribute groups when compared to random subsets of the user's friends. Next, we look into the correlation between SACLs and the social network to see if network-based clusters more closely approximate the SACLs.

5.3 SACLs and network structure

In order to explore whether the SACLs correspond to the network structure, we first identity clusters of the user's friends that are tightly connected in the social network (these clusters are often called *network communities*). Specifically, we extract all of the friendship connections between the user's friends, and then use a community detection algorithm that has been shown to work well in grouping a user's friends into a small set of clusters [35]. This algorithm is a combination of a global community detection algorithm [6] and a local community detection algorithm [40] to detect overlapping communities.[5]

We begin by examining how many of the SACLs exactly match one of the social network-based clusters. Figure 4(a) presents the cumulative distribution of F-scores across all SACLs. The figure shows that 21% of all SACLs have a F-score of more than 0.6, indicating a good match for 21% of SACLs. This is significantly higher than the attribute-based clusters in the prior section, but still does not show a strong correlation.

We next analyze the normalized entropy of these SACLs in Figure 4(b). Similar to the attribute-based clusters, the network communities tend to overestimate smaller SACLs and underestimate larger SACLs, but at a much lower rate. We verify these findings using ARI in Figure 4(c). We can observe that 98% of all SACLs have ARI more than 0, indicating almost all of the SACLs have more similarity with community based clusters than a purely random set of friends.

Overall, the network communities show better match with SACLs compared to the attribute clusters. However, still only a small fraction of SACLs have strong correlation with network communities, making it unlikely that network communications could be used to infer SACLs for sharing content. Next we will look into the correlation between activity-based clusters and SACLs.

5.4 SACLs and user activity

For our final user feature, we examine whether the membership of SACLs is correlated with the strength of the link between the user and their friends. As a proxy for link strength, we use *activity*; this is a common way to estimate how closely connected two users are [3]. For each user, we collected data about four different types of interaction between the user and their friends: (i) posting on the user's wall, (ii) liking the user's posts, (iii) commenting on the user's posts, and (iv) being tagged in the user's status and photos. We observe that 94% of the users who used SACLs have at least one such interaction with their friends.

Using this data, we cluster each user's friends by activity (i.e., frequency of interaction) and see whether the activity-based clusters matched the SACLs. We use the same algorithm as prior work [3] to find the activity clusters. The algorithm is essentially a k-means algorithm modified to automatically find the optimal number of clusters. As a result, all friends with a similar number of interactions will be put in one cluster. After running the algorithm, we find that the median number of clusters across all users is four.[6]

As before, we begin by examining the cumulative distribution of the F-score in Figure 5(a); we observe that 47% all SACLs have a F-score of greater than 0.6. This is even better than network-based communities. The larger SACLs (e.g., exclude few SACLs) show an even stronger match with high F-scores, but the match is considerably worse for smaller SACLs (e.g., include few SACLs), with only 3% of such SACLs having a F-score more than 0.6. Closely examining the activity-based clusters, we hypothesize that our method of creating activity communities often results in creating a large cluster containing all friends with low levels of interaction with the user. As a result, this single, large community alone overlaps with the large SACLs considerably, making their F-score quite high.

To confirm this hypothesis, we also calculate the cumulative distribution of normalized entropy (Figure 5(b)) and ARI (Figure 5(c)). We find a poor match between SACLs

[5] We note that there are a large variety of community detection techniques in the literature. To make sure our choice of algorithm did not bias the results of our analysis, we performed the same analysis with two additional community detection algorithms [9, 42] similar to ones used in earlier work on unsupervised detection of privacy settings [16, 39]. Our results were similar with these algorithms, and so we omit the results for brevity.

[6] Interestingly, this observation matches Dunbar's sociological study [12, 28] where the number of Dunbar's circles, the number of activity-based clusters in people's offline network is also four.

Figure 6: Facebook's interface for specifying SACLs.

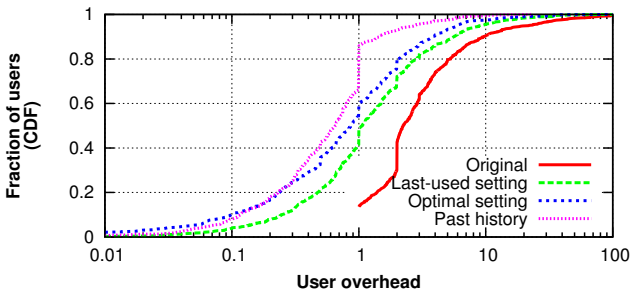

Figure 7: Cumulative distribution of overhead for specifying SACLs. Shown are the distributions of measured SACLs (Original), measured SACLs taking into account Facebook's last-used setting (Last-used setting), the optimal overhead for measured SACLs (Optimal), and the overhead of our proposed mechanism of presenting the user with the last 5 SACLs (Past history). Our proposed mechanism shows a substantial reduction in user overhead.

and activity-based clusters using both pieces of analysis; the ARI values for the SACLs are very close to 0 for almost all activity-based clusters (e.g., only 8% of the SACLs have ARI more than 0.3). This finding confirms that any random groups of the size of larger SACLs will show the same level of similarity with the activity-based clusters, thus making the clusters a poor mechanism for approximating SACL membership.

5.5 Summary

In this section, we examined the membership of SACLs by trying to correlate SACL members with attribute, network structure, and activity-based clusters. Our results show that very few of these clusters show a significant correlation with SACLs, suggesting that automatically detected SACLs-based on these features are unlikely to be very accurate. This finding is in opposition to the results from prior work [16, 39, 48], which suggest that it is possible to use automatically detected clusters to create SACLs. We believe this difference is due to the fact that these prior works were not able to evaluate their proposals against ground-truth SACLs. In fact, others have also found [33] that users are able to group their friends in meaningful groups, but find it difficult to choose the right group to share content with. Consequently, we explore alternative approaches to increase the usability of SACLs in the next section.

6. SACL SPECIFICATION

We have observed that SACLs appear to be quite difficult to infer automatically. We now examine the "overhead" (i.e., the amount of work that users must perform) in order to specify SACLs today. Then, we explore how we can increase the usability of these SACLs by reducing the user overhead and making SACLs easier to use. If successful, these approaches would make privacy easier for users to manage, thereby increasing the usability of OSNs in general.

6.1 SACL specification overhead

The act of specifying a SACL—choosing which friends to share content with—induces cognitive overhead on the user. While there may be multiple dimensions of this overhead (e.g., deciding whether to include a specific user, using the interface, etc), many of these are quite challenging to measure. As a first step, we define the *SACL specification overhead* to be the number of *terms* used to specify a SACL. Our reasons for doing so is that Facebook allows users to specify SACLs using an allow/deny interface, where users can select friends or lists to allow or deny access (a screenshot of Facebook's interface is shown in Figure 6). Thus, the amount of work the user has to do is proportional to the number of friends/lists that the user selects to allow or deny. Of course, we recognize there are dimensions of overhead that this measure fails to capture; we leave the task of characterizing those dimensions of user overhead to future work.

As an example, consider the screenshot shown in Figure 6. In this example, the user is choosing to allow friends Bob, Carol, the list Football Friends, and the Facebook list Family. The user is also choosing to deny the list Drinking Buddies. As a result, the SACL specification overhead for this SACL is five (A total of five terms appear in the allow and deny settings.) It is important to note that the size of the SACL is different from its specification overhead: Consider the case of a user only denying access to a single friend. In this case, the specification overhead is low, but the SACL has many users in it.

We define the *average user overhead* as the average of all SACL specification overheads for content shared by a given user. Formally, if a user specifies access to her content $\{c_1, c_2, ...c_n\}$ with privacy settings $\{p_1, p_2, ...p_n\}$, the average user overhead for this user is

$$\frac{\sum_i |p_i|}{n}$$

where $|p_i|$ denotes the SACL specification overhead of setting p_i. Note that the p_i settings are not necessarily distinct, as multiple pieces of content may be shared with the same SACL. The best-possible average user overhead is 1, meaning the user only used the SACLs with a single term when sharing her content.

We present cumulative distribution of the average user overhead in Figure 7 with the line "Original". We ob-

serve that users do have significant overhead when specifying SACLs: 86% of users have an overhead more than 1, and there are more than 150 users with overhead more than 5. This suggests there is significant potential to reduce the SACL specification overhead for users, making SACLs more usable.

6.2 Last-used setting

Facebook's default privacy setting for content is set to select the *last-used* privacy setting [20]. So, if a user selects a SACL for a newly uploaded piece of content, all future pieces of content will be shared with the same SACL until the user chooses a different privacy setting. We therefore modify our definition of average user overhead to capture this behavior; if a user selects the same SACL as the previous piece of content, we define this SACL specification overhead to be 0. As a result, a user's average user overhead may be less than one.

We present the cumulative distribution of the average user overhead, taking into account the last-used setting, in Figure 7 with the line "Last-used setting". We immediately observe a significant reduction in the measured overhead, which we believe more accurately captures the work a user must do. The figure shows that this simple technique of using last setting as the default can significantly reduce the user overhead: this technique lowers the overhead by more than 50% for almost half (48%) of the users. Thus, Facebook's choice to enable default last-used settings is useful in reducing user overhead. For the remainder of our analysis, we use average user overhead, taking into account the last-used setting, as our baseline.

6.3 Optimal overhead

It is important to note that there may be multiple ways of specifying a given SACL: For example, a user could specify the SACL by only allowing the friends in the SACL. Or, the user could use an existing list, and exclude the users not allowed to access the content. Or, the user could allow all friends, and then deny only the friends who should not be able to access the content. We now examine how close the user's chosen specifications are to the *optimal specification*, in terms of the SACL specification with the minimum overhead.

To do so, for each SACL we observe, we determine the optimal specification overhead.[7] We then present the cumulative distribution of average user overhead in the optimal case in Figure 7 with the line "Optimal setting". We observe that there is still room for improvement from using the last-used setting alone; many users could express their SACLs in a manner than involves fewer terms.

6.4 Using past history

In this section, we explore a generalization of the last-used setting, with the goal of further reducing the average user overhead. The results in Section 4 suggested that there are certain SACLs that users select to share content with a significant fraction of the time. Figure 8 plots the cumulative distribution of the percentage of content shared with the top k most frequently used SACLs for each user. For example we can see that if we allow each user to use their top 5 SACLs, this would cover over 80% of the content for the vast majority (90%) of users.

[7]Note that computing the optimal overhead of a setting is a modified version of the NP-hard set cover problem [32] where the setting is the universe and lists and individual friends are subsets of the universe. We use a brute force solution to the problem, which is feasible due to small number of subsets in this case.

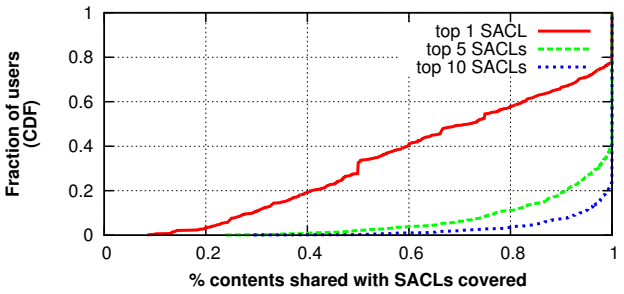

Figure 8: Cumulative distribution of percentage of content covered by the top k SACLs. Even if we set k=5, most of the content for the majority of users are covered.

This observation means that we may be able to significantly reduce the average user overhead by allowing users to choose from their k *most used* SACLs, rather than just the last-used SACL. To do so, we calculate the average user overhead, assuming one would have made it possible for the user to directly use the top 5 most frequently used settings while sharing content (Should the user re-use these settings, we calculate the overhead as 0). A cumulative distribution of the resulting overhead is shown in Figure 7 with the line "Past history". We immediately observe a dramatic reduction in user overhead (In fact, the overhead is lowered for 86% of users).

In summary, this approach of leveraging past history has the potential to significantly reduce the user overhead in specifying SACLs. An OSN operator can create these SACLs based on user's past history, and provide them as options to select from, when the user uploads a new piece of content. Should the user select one of the previously-used SACLs, it will reduce their overhead and make privacy specification more usable.

7. CONCLUDING DISCUSSION

Online social networks are increasingly popular and users are sharing ever more content on these services. In this paper, we focused on the most privacy-sensitive of these content: the content with hand-crafted privacy settings selected by the users. We found that these SACLs are surprisingly common (over 17.6% of all content is shared with SACLs), but that the membership of these SACLs shows relatively little correlation with the profile attributes, the social network structure, or the activity level of the members. As a result, there appears to be little hope of automatically detecting more than a few of these SACLs. We also found that the act of specifying SACLs is often complicated for users, but a simple technique like remembering a few of the most frequently used SACLs is likely to significantly reduce this burden in practice.

However, much work remains to be done. In the remainder of this section, we discuss a few of the limitations of our study, as well as future directions for exploration.

Understanding motivation for SACLs Our study explores the use of SACLs, but does not reveal *why* users create SACLs or how they choose the content to share with SACLs. Possible reasons include dissatisfaction with the default privacy settings, the sharing of highly privacy sensitive content, or using SACLs as a mechanism to choose the audience for a particular content.

Moving target We quantified the way users create in-use SACLs today, but Facebook is known for changing their privacy interface over time [18]; these changes are likely to impact the usage of SACLs for individual users. We aim to repeat our analysis as Facebook makes these changes, hoping to capture resulting changes in user behavior.

SACL accuracy It remains an unexplored question as to which of the friends users would *ideally* want to share their content with (i.e., who does the user want to be in a SACL, regardless of who is actually in the SACL). Prior work has shown that users often misunderstand other Facebook privacy settings [34], and we suspect that this would likely hold true for SACLs as well.

SACL overhead In our calculation of overhead, we took into consideration the number of terms specified by users explicitly while specifying SACLs, where a term can be a friend list or an individual friend. However, this quantification does not directly account for the mental effort required for a user when creating SACLs (e.g., certain SACLs may be easier or harder to create, even if they have the same number of terms). We leave a full exploration of this effect (possibly via detailed debriefing interviews of a small sample of Facebook users) to future work.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. We also thank our FLM users for enabling us to conduct this study. This research was supported in part by NSF grants CNS-1054233 and CNS-1319019, ARO grant W911NF-12-1-0556, and an Amazon Web Services in Education Grant.

8. REFERENCES

[1] S. Amershi, J. Fogarty, and D. Weld. Regroup: interactive machine learning for on-demand group creation in social networks. In *Proceedings of the 30th SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*, Austin, TX, May 2012.

[2] Amazon Mechanical Turk terms of service. `https://requester.mturk.com/mturk/help?helpPage=policies`.

[3] V. Arnaboldi, M. Conti, A. Passarella, and F. Pezzoni. Analysis of Ego Network Structure in Online Social Networks. In *Proceedings of the 4th ASE/IEEE International Conference on Social Computing (SocialCom'12)*, Amsterdam, The Netherlands, September 2012.

[4] M. S. Bernstein, E. Bakshy, M. Burke, and B. Karrer. Quantifying the invisible audience in social networks. In *Proceedings of the 31st SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*, Paris, France, April 2013.

[5] A. Besmer and H. R. Lipford. Moving Beyond Untagging: Photo Privacy in a Tagged World. In *Proceedings of the 28th SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*, Atlanta, GA, April 2010.

[6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of community hierarchies in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), 2008.

[7] D. Boyd and E. Hargittai. Facebook privacy settings: Who cares? *First Monday*, 15(8), 2010.

[8] S. Chen. Can Facebook get you fired? Playing it safe in the social media world. `http://www.cnn.com/2010/LIVING/11/10/facebook.fired.social.media.etiquette/index.html`, 2010.

[9] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70:1–6, 2004.

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[11] R. Dey, Z. Jelveh, and K. W. Ross. Facebook users have become much more private: A large-scale study. In *Proceedings of the 10th Annual IEEE International Conference on Pervasive Computing and Communications (perCom'12)*, Lugano, Switzerland, March 2012.

[12] R. I. M. Dunbar. The social brain hypothesis. *Evolutionary Anthropology*, 6(5), 1998.

[13] A. Etzioni. Despite Facebook, privacy is far from dead. `http://www.cnn.com/2012/05/25/opinion/etzioni-facebook-privacy/`.

[14] Facebook smart lists. `https://blog.facebook.com/blog.php?post=10150278932602131`.

[15] Number of active users at Facebook over the years. `http://finance.yahoo.com/news/number-active-users-facebook-over-years--214600186--finance.html`.

[16] L. Fang and K. LeFevre. Privacy Wizards for Social Networking Sites. In *Proceedings of the 19th International World Wide Web Conference (WWW'10)*, Releigh, NC, April 2010.

[17] Facebook API. `http://developers.facebook.com/docs/reference/api/`, 2011.

[18] Detailed History of Facebook Changes 2004-12. `https://www.jonloomer.com/2012/05/06/history-of-facebook-changes/`.

[19] The Evolution of Privacy on Facebook – Changes in default profile settings over time. `http://mattmckeon.com/facebook-privacy/`, 2010.

[20] When I post something, how do I choose who can see it? `https://www.facebook.com/help/120939471321735`.

[21] New Tools to Control Your Experience. `https://blog.facebook.com/blog.php?post=196629387130`.

[22] Making Photo Tagging Easier. `https://blog.facebook.com/blog.php?topic_id=203150980352`.

[23] Facebook's New Privacy Changes: The Good, The Bad, and The Ugly. `https://www.eff.org/deeplinks/2009/12/facebooks-new-privacy-changes-good-bad-and-ugly`.

[24] Tag Friends in Your Status and Posts. `https://www.facebook.com/notes/facebook/`

tag-friends-in-your-status-and-posts/109765592130.

[25] Friendlist Manager. http://friendlist-manager.mpi-sws.org/.

[26] More Privacy Options. https://blog.facebook.com/blog.php?post=11519877130.

[27] Facebook's New Groups, Dashboards, and Downloads Explained. http://www.fastcompany.com/1693443/facebooks-new-groups-dashboards-and-downloads-explained-video.

[28] R. A. Hill and R. I. M. Dunbar. Social network size in humans. *Human Nature*, 14(1), 2003.

[29] C. M. Hoadley, H. Xu, J. J. Lee, and M. B. Rosson. Privacy as information access and illusory control: The case of the Facebook News Feed privacy outcry . *Electronic Commerce Research and Applications*, 9(1), 2010.

[30] M. Johnson, S. Egelman, and S. M. Bellovin. Facebook and Privacy: It's Complicated. In *Proceedings of the 8th Symposium on Usable Privacy and Security (SOUPS'12)*, Washington, DC, July 2012.

[31] S. Kairam, M. Brzozowski, D. Huffaker, and E. Chi. Talking in circles: selective sharing in google+. In *Proceedings of the 30th SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*, Austin, TX, May 2012.

[32] R. M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, 1972.

[33] P. G. Kelley, R. Brewer, Y. Mayer, L. F. Cranor, and N. Sadeh. An Investigation into Facebook Friend Grouping. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-computer Interaction (INTERACT'11)*, Lisbon, Portugal, September 2011.

[34] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing Facebook privacy settings: User expectations vs. reality. In *Proceedings of the 11th ACM/USENIX Internet Measurement Conference (IMC'11)*, Berlin, Germany, November 2011.

[35] Y. Liu, M. Mondal, B. Viswanath, M. Mondal, K. P. Gummadi, and A. Mislove. Simplifying Friendlist Management. In *Proceedings of the 21st International World Wide Web Conference (WWW'12)*, Lyon, France, April 2012.

[36] M. Madejski, M. Johnson, and S. M. Bellovin. The Failure of Online Social Network Privacy Settings. Technical Report CUCS-010-11, Department of Computer Science, Columbia University, 2011.

[37] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[38] W. Mason and S. Suri. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1), 2011.

[39] A. Mazzia, K. LeFevre, and E. Adar. The PViz comprehension tool for social network privacy settings. In *Proceedings of the 8th Symposium on Usable Privacy and Security (SOUPS'12)*, Washington, DC, July 2012.

[40] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: Inferring user profiles in Online Social Networks. In *Proceedings of the 3rd ACM International Conference of Web Search and Data Mining (WSDM'10)*, New York, NY, February 2010.

[41] D. Noyes. The Top 20 Valuable Facebook Statistics. http://zephoria.com/social-media/top-15-valuable-facebook-statistics/.

[42] P. Pons and M. Latapy. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10(2), 2004.

[43] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336), 1971.

[44] People Manage Their Privacy On Facebook Naturally. http://www.sciencedaily.com/releases/2009/04/090420084957.htm.

[45] More Facebook Friends Means More Stress, Says Report. http://www.sciencedaily.com/releases/2012/11/121126131218.htm.

[46] F. Stutzman, R. Gross, and A. Acquisti. Silent Listeners: The Evolution of Privacy and Disclosure on Facebook. *Journal of Privacy and Confidentiality*, 4(2), 2012.

[47] J. D. Sutter. Some quitting Facebook as privacy concerns escalate. http://www.cnn.com/2010/TECH/05/13/facebook.delete.privacy/index.html.

[48] Q. Xiao, H. H. Aung, and K.-L. Tan. Towards ad-hoc circles in social networking sites. In *Proceedings of the 2Nd ACM SIGMOD Workshop on Databases and Social Networks(DBSocial'12)*, Scottsdale, AZ, May 2012.