Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook

Pre-proceedings version. Privacy Enhancing Technologies Workshop (PET), 2006

Alessandro Acquisti[1] and Ralph Gross[2]

[1] H. John Heinz III School of Public Policy and Management
[2] Data Privacy Laboratory, School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213

Abstract. Online social networks such as Friendster, MySpace, or the Facebook have experienced exponential growth in membership in recent years. These networks offer attractive means for interaction and communication, but also raise privacy and security concerns. In this study we survey a representative sample of the members of the Facebook (a social network for colleges and high schools) at a US academic institution, and compare the survey data to information retrieved from the network itself. We look for underlying demographic or behavioral differences between the communities of the network's members and non-members; we analyze the impact of privacy concerns on members' behavior; we compare members' stated attitudes with actual behavior; and we document the changes in behavior subsequent to privacy-related information exposure. We find that an individual's privacy concerns are only a weak predictor of his membership to the network. Also privacy concerned individuals join the network and reveal great amounts of personal information. Some manage their privacy concerns by trusting their ability to control the information they provide and the external access to it. However, we also find evidence of members' misconceptions about the online community's actual size and composition, and about the visibility of members' profiles.

1 Introduction

"Students living in the scholarship halls [of Kansas University] were written up in early February for pictures on facebook.com that indicated a party violating the scholarship halls alcohol policy" [1]. "'Stan Smith' (not his real name) is a sophomore at Norwich University. He is majoring in criminal justice even though he admits to shoplifting on his MySpace page" [2]. "Corporations are investing in text-recognition software from vendors such as SAP and IBM to monitor blogs by employees and job candidates" [3]. Although online social networks are offering novel opportunities for interaction among their users, they seem to attract non-users' attention particularly because of the privacy concerns they raise. Such concerns may be well placed; however, online social networks are no longer niche phenomena: millions of people around the world, young and old, knowingly and willingly use Friendster, MySpace, Match.com, LinkedIn, and hundred other sites to communicate, find friends, dates, and jobs - and in doing so, they wittingly reveal highly personal information to friends as well as strangers.

Nobody is literally forced to join an online social network, and most networks we know about *encourage*, but do not *force* users to reveal - for instance - their dates of birth, their cell phone numbers, or where they currently live. And yet, one cannot help but marvel at the nature, amount, and detail of the personal information some users provide, and ponder how informed this information sharing is. Changing cultural trends, familiarity and confidence in digital technologies, lack of exposure or memory of egregious misuses of personal data by others may all play a role in this unprecedented phenomenon of information revelation. Yet, online social networks' security and access controls are weak by design - to leverage their value as network goods and enhance their growth by making registration, access, and sharing of information uncomplicated. At the same time, the costs of mining and storing data continue to decline. Combined, the two features imply that information provided even on ostensibly private social networks is, effectively, public data, that could exist for as long as anybody has an incentive to maintain it. Many entities - from marketers to employers to national and foreign security agencies - may have those incentives.

In this paper we combine survey analysis and data mining to study one such network, catered to college and high school communities: the Facebook (FB). We survey a representative sample of FB members at a US campus. We study their privacy concerns, their usage of FB, their attitudes towards it as well as their awareness of the nature of its community and the visibility of their own profiles. In particular, we look for underlying demographic or behavioral differences between the communities of the network's members and

non-members; we analyze the impact of privacy concerns on members' behavior; we compare members' stated attitudes with actual behavior; and we document the change in behavior subsequent information exposure: who uses the Facebook? Why? Are there significant differences between users and non-users? Why do people reveal more or less personal information? How well do they know the workings of the network?

Our study is based on a survey instrument, but is complemented by analysis of data mined from the network before and after the survey was administered. We show that there are significant demographic differences between FB member and non-members; that although FB members express, in general, significant concern about their privacy, they are not particularly concerned for their privacy *on* FB; that a minority yet significant share of the FB population at the Campus we surveyed is unaware of the actual exposure and visibility of the information they publish on FB; and we document that priming about FB's information practices can alter some of its members' behavior.

The rest of the paper is organized as follows. In Section 2 we discuss the evolution of online social networks and FB in particular. In Section 3 we highlight the methods of our analysis. In Section 4 we present our results. In Section 5 we compare survey results to network data.

2 Online Social Networks

At the most basic level, an online social network is an Internet community where individuals interact, often through profiles that (re)present their public persona (and their networks of connections) to others. Although the concept of computer-based communities dates back to the early days of computer networks, only after the advent of the commercial Internet did such communities meet public success. Following the SixDegrees.com experience in 1997, hundreds of social networks spurred online (see [4] for an extended discussion), sometimes growing very rapidly, thereby attracting the attention of both media and academia. In particular, [5], [6], and [7] have taken ethnographic and sociological approaches to the study of online self-representation; [8] have focused on the value of online social networks as recommender systems; [4] have discussed information sharing and privacy on online social networks, using FB as a case study; [9] have demonstrated how information revealed in social networks can be exploited for "social" phishing; [10] has studied identity-sharing behavior in online social networks.

2.1 The Facebook

The Facebook is a social network catered to college and high school communities. Among online social networks, FB stands out for three reasons: its success among the college crowd; the amount and the quality of personal information users make available on it; and the fact that, unlike other networks for young users, that information is personally identified. Accordingly, FB is of interest to researchers in two respects: 1) as a mass social phenomenon in itself ; 2) as an unique window of observation on the privacy attitudes and the patterns of information revelation among young individuals.

FB has spread to thousands of college campuses (and now also high schools) across the United States, attracting more than 9 million (and counting) users. FB's market penetration is impressive: it can draw more than 80% of the undergraduate population in many colleges. The amount, quality, and value of the information provided is impressive too: not only are FB profiles most often personally and uniquely identified, but by default they show contact information (including personal addresses and cell phone numbers) and additional data rarely available on other networks.

FB requires a college's email account for a participant to be admitted to the online social network of that college. As discussed in [4], this increases the expectations of validity of the personal information therein provided, as well as the perception of the online space as a closed, trusted, and trustworthy community (college-oriented social networking sites are, ostensibly, based "on a shared real space" [11]). However, there are reasons to believe that FB networks more closely resemble *imagined* [12] communities (see also [4]): in most online social networks, security, access controls, and privacy are weak *by design*; the easier it is for people to join and to find points of contact with other users (by providing vast amounts of personal information, and by perusing equally vast amounts of data provided by others), the higher the utility of the network to the users themselves, and the higher its commercial value for the network's owners and managers. FB, unlike other online networks, offers its members very granular and powerful control on the privacy (in terms of searchability and visibility) of their personal information. Yet its privacy default settings are very permeable: at the time of writing, by default participants' profiles are *searchable* by anybody else on the FB network, and actually *readable* by any member at the same college and geographical location.

In addition, external access to a college FB community (e.g., by non-students/faculty/staff/alumni, or by non-college-affiliated individuals) is so easy [4], that the network is effectively an open community, and its data effectively public.

3 Methods

Our study aims at casting a light on the patterns and motivations of information revelation of college students on FB. It is based on a survey instrument administered to a sample of students at a North American college Institution, complemented by analysis of data mined from the FB network community of that Institution.

3.1 Recruiting Methods

Participants to the survey were recruited in three ways: through a list of subjects interested in participating in experimental studies maintained at the Institution where the study took place (and containing around 4,000 subscribed subjects); through an electronic billboard dedicated to experiments and studies, with an unknown (to us) number of campus community subscribers; and through fliers posted around campus. The above two lists are populated in majority by undergraduate students. The emails and the fliers sought participants to a survey on "online networks," and offered a compensation of $6, plus the possibility to win a $100 prize in a lottery among all participants.

Around 7,000 profiles were mined from the FB network of the same Institution. In order to automate access to the Facebook we used Perl scripts [13], specifically the Perl LWP library [14], which is designed for downloading and parsing HTML pages. The data was mined before and after the survey was administered.

3.2 Survey Design

The survey questionnaire contained around forty questions: an initial set of screening questions; a consent section; a set of calibration questions (to ascertain the respondents' privacy attitudes without priming them on the subject of our study: privacy questions were interspersed with questions on topics such as economic policy, the threat of terrorism, same-sex marriage, and so on); and, next, FB-related questions. Specifically, we asked respondents to answer questions about their usage, their knowledge, and their attitudes towards FB. Finally, the survey contained a set of demographics questions.

Only respondents currently affiliated with the Institution were allowed to take the survey (students, staff, and faculty). Respondents received somewhat different questions depending on whether they were current FB members, previous members, or never members. The survey is available on request from the authors.

3.3 Statistical Analysis

We analyzed survey results using STATA 8.0 on Windows and other *ad hoc* scripts. The study was performed on dichotomous, categorical (especially 7-point Likert scales), and continuous variables. We performed a number of different tests - including Pearson product-moment correlations to study relations between continuous variables, χ^2 and t tests to study categorical variables and means, Wilcoxon signed-rank test and Wilcoxon/Mann-Whitey test for non-normal distributions, as well as logit, probit, and linear multivariate regressions.

4 Results

A total of 506 respondents accessed the survey. One-hundred-eleven (21.9%) were not currently affiliated with the college Institution where we conducted our study, or did not have a email address within that Institution's domain. They were not allowed to take the rest of the survey. A separate set of 32 (8.2%) participants had taken part in a previous pilot survey and were also not allowed to take the survey. Of the remaining respondents, 318 subjects actually completed the initial calibration questions. Out of this set, 278 (87.4%) had heard about FB, 40 had not. In this group, 225 (70.8%) had a profile on FB, 85 (26.7%) never had one, and 8 (2.5%) had an account but deactivated it. Within those three groups, respectively 209, 81, and 7 participants completed the whole survey. We focus our analysis on that set - from which we further removed 3 observations from the non-members group, since we had reasons to believe that the responses had been created by the same individual. This left us with a total of 294 respondents.

4.1 Participants

In absolute terms, we had exactly the same number of male participants taking the survey as female participants, 147. We classified participants depending on whether they were current members of the FB campus network (we will refer to them as "members"), never members, or no longer members (we will often refer to the last two groups collectively as "non-members").

A slight majority of FB members in our sample (52.63%) are male. Our sample slightly over-represents females when compared to the target FB population, whose data we mined from the network (male represent 63.04% of the Institution's FB network, but it is important to note that the gender distribution at the Institution is itself similarly skewed). However, we know from the information mined from the network that 79.6% of all the Institution's undergraduate males are on the FB (91.92% of our sample of male undergrads are FB members) and 75.5% of all the Institution's undergraduate females are on the FB (94.94% of our sample of female undergrads are FB members). In other words (and expectably), our total sample of respondents slightly over-represents FB members.

The gender distribution of our sample is reversed among respondents who were never or are no longer members of FB: 56.46% are female. This gender difference between current members and current non-members is not statistically significant (Pearson $\chi^2(1) = 2.0025$, $Pr = 0.157$). However, when we test usage by contrasting actual FB users and non-members plus members who claim to " I never login/use" their profile, the gender difference becomes more radical (54.19% of *users* are male, but only 40.66% of *non users* are) and significant (Pearson $\chi^2(1) = 4.5995$ $Pr = 0.032$). See Figure 1 for the gender distribution in the three FB member groups.

User	What is your gender Male	Female	Total
Current FB Member	110	99	209
	52.63	47.37	100.00
	74.83	67.35	71.09
Former FB Member	3	4	7
	42.86	57.14	100.00
	2.04	2.72	2.38
Never FB Member	34	44	78
	43.59	56.41	100.00
	23.13	29.93	26.53
Total	147	147	294
	50.00	50.00	100.00
	100.00	100.00	100.00

Fig. 1. Gender distribution of the survey participants for the three FB member groups.

There is no significant difference among the distributions of undergraduate versus graduate students in our sample and in the overall FB population.

Overall, sixty-four percent of our respondents (64.29%) are undergraduate students; 25.17% are graduate students; 1.36% are faculty; and 9.18% are staff. We did not consider alumni in our study. This distribution slightly oversamples undergraduate students when compared to the actual Institution's population (total student population in 2005/06: 10,017. Undergraduate students: 54.8%). This was expected, considering the available recruiting tools and the comparatively higher propensity of undergraduate students to take paid surveys and experiments. However, when checking for current FB membership in our sample, we find that undergraduates dominate the picture (84.21%), followed by graduate students (14.35%) and staff (1.44%). These numbers are comparable to the distribution of the target population discused in [4] when correcting for alumni (91.21% were undergraduate students on the Facebook network).

Again, the distribution of non-members is reversed: graduate students dominate (51.76%), followed by staff (28.24%). The distributions of user types (undergraduates, graduates, staff, or faculty) by FB membership status are significantly diverse (Pearson $\chi^2(3) = 135.3337$ $Pr = 0.000$). See Figure 2 for a breakdown of the academic status of survey participants across the three FB groups.

User	Are you... Undergrad	Graduate	Faculty	Staff	Total
Current FB Member	176	30	0	3	209
	84.21	14.35	0.00	1.44	100.00
	93.12	40.54	0.00	11.11	71.09
Former FB Member	2	4	0	1	7
	28.57	57.14	0.00	14.29	100.00
	1.06	5.41	0.00	3.70	2.38
Never FB Member	11	40	4	23	78
	14.10	51.28	5.13	29.49	100.00
	5.82	54.05	100.00	85.19	26.53
Total	189	74	4	27	294
	64.29	25.17	1.36	9.18	100.00
	100.00	100.00	100.00	100.00	100.00

Fig. 2. Distribution of survey participant status for FB members, non-members and people who never had a FB account.

Unsurprisingly, age is a strong predictor of membership (see Figure 3). Non-members tend to be older (a mean of 30 years versus a mean of 21) but their age is also more broadly distributed (*sd* 8.840476 vs. *sd* 2.08514). The difference in the mean age by membership is strongly significant (t = -14.6175, $Pr<t$ = 0.0000).

usercomb	group_age 17-24	25-34	35-44	45+	Total
Current member	204	4	1	0	209
	97.61	1.91	0.48	0.00	100.00
	88.31	9.30	9.09	0.00	71.09
Current non member	27	39	10	9	85
	31.76	45.88	11.76	10.59	100.00
	11.69	90.70	90.91	100.00	28.91
Total	231	43	11	9	294
	78.57	14.63	3.74	3.06	100.00
	100.00	100.00	100.00	100.00	100.00

Fig. 3. Distribution of age for FB members and non-members.

4.2 Privacy Attitudes

Age and student status are correlated with FB membership - but what else is? Well, of course, having heard of the network is a precondition for membership. Thirty-four participants had never heard of the FB - nearly half of the staff that took our survey, a little less than 23% of the graduate students, and a negligible portion of the undergraduate students (1.59%).

However, together with age and student status (with the two obviously being highly correlated), another relevant distinction between members and non-members may arise from privacy attitudes and privacy concerns.

Before we asked questions about FB, our survey ascertained the privacy attitudes of participants with a battery of questions modelled after the Alan Westin's studies [15], with a number of modifications. In particular, in order not to prime the subjects, questions about privacy attitudes were interspersed with questions about attitudes towards economic policy and the state of the economy, social issues such as same-sex marriage, or security questions related to the fear of terrorism. In addition, while all instruments asked the respondent to rank agreement, concern, worries, or importance on a 7-point Likert scale, the questions ranged from general ones (e.g., "How *important* do you consider the following issues in the public debate?"), to more and more specific (e.g., "How do you *personally* value the *importance* of the following issues for your own life on a day-to-day basis?"), and personal ones (e.g., "Specifically, how *worried* would you be if" [a certain scenario took place]).

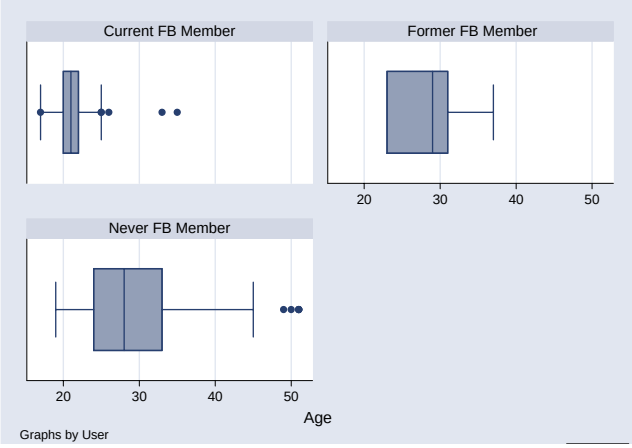

Fig. 4. Box-plots of age distribution for different membership status

"Privacy policy" was on average considered a highly *important* issue in the *public debate* by our respondents (mean on the 7-point Likert scale: 5.411, where 1 is "Not important at all" and 7 is "very important"; *sd*: 1.393795). In fact, it was regarded a more important issue in the public debate than the threat of terrorism ($t = 2.4534$, $Pr{>}t = 0.0074$; the statistical significance of the perceived superiority was confirmed by a Wilcoxon signed-rank test: $z = 2.184$ $Pr{>}|z| = 0.0290$) and same sex marriage ($t = 10.5089$, $Pr{>}t = 0.0000$; Wilcoxon signed-rank test: $z = 9.103$ $Pr{>}|z| = 0.0000$); but less important than education policy (mean: 5.93; *sd*: 1.16) or economic policy (mean: 5.79; *sd*: 1.21). The slightly larger mean valuation of the importance of privacy policy over environmental policy was not significant. (These results are comparable to those found in previous studies, such as [16].)

The same ranking of values (and comparably statistically significant differences) was found when asking for "How do you personally value the *importance* of the following issues for your *own life* on a *day-to-day basis*?" The mean value for the importance of privacy policy was 5.09. For all categories, subjects assigned slightly (but statistically significantly) more importance to the issue in the *public debate* than in *their own life* on a day-to-day basis (in the privacy policy case, a Wilcoxon signed-rank test returns $z = 3.62$ $Pr > |z| = 0.0003$ when checking the higher valuation of the issue in the public debate).

Similar results were also found when asking for the respondents' *concern* with a number of issues directly relevant to them: the state of the economy where they live, threats to their personal privacy, the threat of terrorism, the risks of climate change and global warming. Respondents were more concerned (with statistically significant differences) about threats to their personal privacy than about terrorism or global warming, but less concerned than about the state of the economy.

Finally, we asked how *worried* respondents would be if a number of specific events took place in their lives. The highest level of concern was registered for "A stranger knew where you live and the location and schedule of the classes you take" (mean of 5.78, with 45.58% of respondents choosing the 7th point in the Likert scale, "very worried," and more than 81% selecting Likert points above 4). This was followed by "Five years from now, complete strangers would be able to find out easily your sexual orientation, the name of your current partner, and your current political views" (mean of 5.55, with 36.39% - the relative majority - choosing the 7th point in the Likert scale, and more than 78% with points above 4), followed, in order, by the 'global warming' scenario ("The United States rejected all new initiatives to control climate change and reduce global warming"), the security scenario ("It was very easy for foreign nationals to cross the borders undetected"), the 'contacts' scenario ("A friend of a friend that you do not even know knew your name, your email, your home phone number, and your instant messaging nickname"), and the 'same-sex' scenario ("Two people of the same sex were allowed to marry in your State").

Privacy Attitudes and Membership Status Privacy concerns are not equally distributed across FB members and non-members populations: a two-sided t test that the mean Likert value for the "importance" of privacy policy is *higher* for non-members (5.67 in the non-members group, 5.30 in the members group)

is significant (t = -2.0431, $Pr<t$ = 0.0210). Similar statistically significant differences arise when checking for the level of *concern* for privacy threats and for *worries* associated with the privacy scenarios described above. The test becomes slightly less significant when checking for member/non-member differences in the assigned importance of privacy policy on a day-to-day basis.

Importantly, in general no comparable statistically significant differences between the groups can be found in other categories. For example, worries about the global warming scenario gain a mean Likert valuation of 5.36 in the members sample and 5.4 in the non-members sample. (A statistically significant difference can be found however for the general threat of terrorism and for the personal worry over marriage between two people of same sex: higher values in the non-members group may be explained by their higher mean age.)

We also used two-sample Wilcoxon rank-sum (Mann-Whitney) tests to study the distributions of the sensitivities to the various scenarios. We found additional evidence that the sensitivity towards privacy is stronger among non-members than members. In the "A stranger knew where you live and the location and schedule of the classes you take" scenario, concerns are higher in the non-member population - the Mann-Whitney test that the two distributions are the same returns z = -3.086 $Pr>|z|$= 0.0020. Similar results are found for the "Five years from now, complete strangers would be able to find out easily your sexual orientation, the name of your current partner, and your current political views" scenario (z = -2.502 $Pr>|z|$= 0.0124), and the "A friend of a friend that you do not even know knew your name, your email, your home phone number, and your instant messaging nickname" scenario. Importantly, no such differences were found to be significant for the same sex marriage scenario, the illegal aliens scenario, and the US rejecting initiatives to control climate change scenario.

Overall, the distributions of reported intensity of privacy concerns tend to be more skewed towards higher values, and less normally-distributed for non-members. For the most invasive scenarios, however, *both* members and non-members' distributions are not normal, with the distribution for non-members more skewed towards the higher values on the right (see Figure 5). These results do not change after accounting for people who do not know about FB - the t tests simply become more significant.

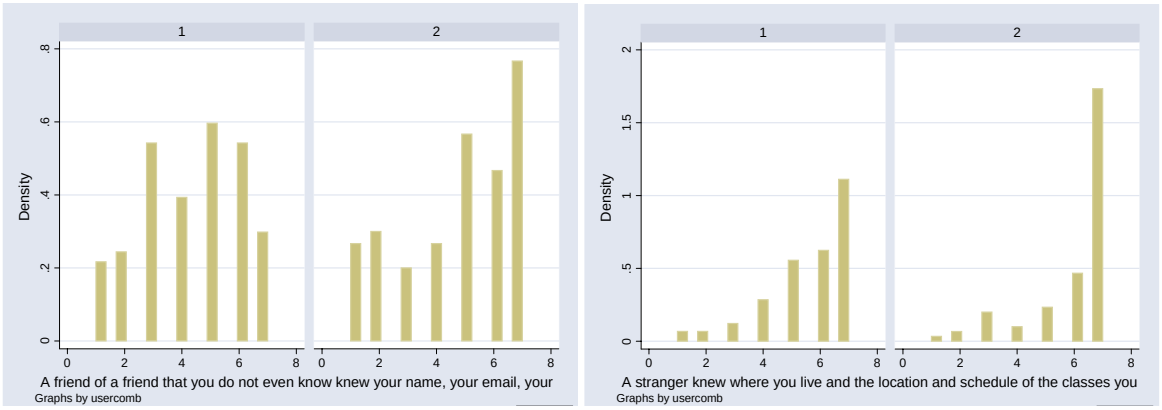

Fig. 5. Distribution of privacy attitudes for FB members (columns marked "1") and non-members (columns marked "2"; this set includes both people that never had a profile and those who had a profile but deactivated it) for an exemplary scenario.

Disentangling Age, Student Status, and Privacy Concerns An obvious hypothesis about FB membership is that individual privacy concerns will be inversely correlated with the probability of joining FB. However, while non FB members seem to have higher average privacy concerns than members (over the scenarios we tested), we cannot directly conclude that the higher one's general privacy concerns, the less likely he will be a FB member. Figure 6, for instance, shows the distribution of levels of concern for privacy threats for both FB members and non-members. A measure of correlation provided by Pearson $\chi^2(6)$ is not significant ($\chi^2(6)$=8.0467, Pr = 0.235). (Pearson χ^2 is significant when studying the "stranger knows where you live scenario:" $\chi^2(6)$ = 16.5665, Pr = 0.011; likelihood-ratio $\chi^2(6)$ = 17.4785, Pr = 0.008.)

usercomb	Threats to your personal privacy 1	2	3	4	5	6	7
Current member	6	15	20	28	46	52	42
	2.87	7.18	9.57	13.40	22.01	24.88	20.10
	100.00	71.43	80.00	68.29	71.88	76.47	60.87
Current non member	0	6	5	13	18	16	27
	0.00	7.06	5.88	15.29	21.18	18.82	31.76
	0.00	28.57	20.00	31.71	28.13	23.53	39.13
Total	6	21	25	41	64	68	69
	2.04	7.14	8.50	13.95	21.77	23.13	23.47
	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Fig. 6. Distribution of levels of concern for threats to personal privacy for FB members and non-members.

In addition, privacy concerns may also be correlated with gender,[3] and status (undergraduate, graduate, faculty, staff).[4] This makes it difficult to understand the actual impact of privacy attitudes and concerns and various other personal characteristics on FB membership.

For instance, when we focus on the undergraduate respondents in our sample, we find that even the undergraduates who expressed the highest level of concern for threats to their personal privacy are still in vast majority joining the Facebook: 89.74% of them. We also find that the mean level of concern is not statistically different between undergraduates who are members and those who are not. (Among undergraduate students, 2 were former members who were no longer members at the time of the survey; their expressed level of concern for threats was 5 and 7; one user was still a member but claimed to never login - his concern level is 6). On the other hand, among respondents who are *not* undergraduates, the mean concern level of non-members (controlling for those who have heard about the FB) is 5.41; the mean for members is 4.81. A two-side Student t test shows that the difference is mildly significant: $Ha : diff < t =$ -1.5346 and $Pr < t = 0.0646$). In fact, the ratio of members to non-members decreases with the intensity of concern.

In order to disentangle these complex relations between age, respondent type, and privacy concerns - that we hypothesize are all factors affecting FB membership - we employed multivariate regression analysis.

In a first approach, we used k-means multivariate clustering techniques [17] to cluster respondents according to their privacy attitudes: we used all the 7-Likert scale responses relevant to privacy (from importance assigned to privacy policy, to worries about specific scenarios) and created a new categorical variable called *_Iprivacy_*. We employed that variable in logistical regressions (logit and probit) over the dependent variable *user_logit*, a dichotomous variable representing membership to the FB network (*user_logit*=1; or lack thereof, *user_logit*=0). We also used age (*age*), a dummy variable representing gender (male if *gender*=1), and a dummy variable representing student status (undergraduate if *undergrad*=1) as independent variables. We restricted the analysis to respondents who had heard about FB. The results of the regression are reported in Figure 7. The model has a good fit, explaining more than half of the variance between members and non-members of FB. As expected, age and undergraduate status are significant while gender is not. The signs of the regression are as expected: being an undergraduate increases the probability of being a member, and age decreases it. Interestingly, at least one of the categorical clusters for privacy attitudes (represented by the variables *_Iprivacy_∼2,3,4,5,6,7*, measured against the base cluster *_Iprivacy_∼1* - the one with the highest level of concerns) is significant, with a large positive impact on the probability of being a member.

In a second approach, we took the means of all the 7-Likert scale responses relevant to privacy and constructed a new categorical variable (*privacy_at∼n*), that we used in the second regression reported in Figure 8.

The results are comparable to those from the previous regression. Both regressions show that even when controlling for age, status, and gender, one's privacy concerns have *some* impact on the decision to join the network, and the student status has some impact independent from age.[5] However, and importantly, this impact really only exists for the non undergraduate population: when restricting the analysis to the undergraduate population, neither the privacy cluster nor the privacy mean variables are significant. They are, however, significant ($Pr > z : 0.024$) when focusing on the non undergraduate population. In other words: privacy concerns may drive older and senior college members away from FB. Even high privacy

[3] For instance, female respondents in general report statistically significantly higher average concerns for privacy over the various scenarios and instruments we discussed above.

[4] We did not find a significant correlation between age and a number of indicators of privacy concerns in our sample; however, our sample cannot be considered representative of the population of age over 25.

[5] As noted above, in our sample age alone is not significantly correlated with privacy concerns.

Logit estimates — Number of obs = 260; LR chi2(9) = 130.27; Prob > chi2 = 0.0000; Log likelihood = -63.57221; Pseudo R2 = 0.5061

user_logit	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.5632942	.1385778	-4.06	0.000	-.8349017	-.2916866
undergrad	.9950484	.5923704	1.68	0.093	-.1659763	2.156073
Iprivacy~2	4.906447	2.119911	2.31	0.021	.7514989	9.061395
Iprivacy~3	.2758868	.5935938	0.46	0.642	-.8875356	1.439309
Iprivacy~4	1.043322	.8747547	1.19	0.233	-.6711654	2.75781
Iprivacy~5	1.598968	1.113793	1.44	0.151	-.5840251	3.781961
Iprivacy~6	.6435271	.6934733	0.93	0.353	-.7156556	2.00271
Iprivacy~7	1.451756	1.173762	1.24	0.216	-.8487741	3.752287
gender	.1879386	.4760592	0.39	0.693	-.7451203	1.120997
_cons	12.81519	3.278655	3.91	0.000	6.389147	19.24124

Fig. 7. Results of logit regression on FB membership using demographical characteristics and k-means clustered privacy attitudes (unstandardized effect coefficients).

Logit estimates — Number of obs = **260**; LR chi2(**4**) = **121.80**; Prob > chi2 = **0.0000**; Log likelihood = **-67.804697**; Pseudo R2 = **0.4732**

user_logit	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	**-.4953475**	**.1236441**	**-4.01**	**0.000**	**-.7376856**	**-.2530095**
undergrad	**1.025618**	**.5752887**	**1.78**	**0.075**	**-.1019268**	**2.153164**
privacy_at~n	**-.5152155**	**.2212539**	**-2.33**	**0.020**	**-.9488651**	**-.0815658**
gender	**.1798292**	**.4576217**	**0.39**	**0.694**	**-.7170929**	**1.076751**
_cons	**14.69864**	**3.322535**	**4.42**	**0.000**	**8.186591**	**21.21069**

Fig. 8. Results of logit regression on FB membership using demographical characteristics and mean privacy attitudes (unstandardized effect coefficients).

concerns, however, are *not* driving undergraduate students away from it. Non-members have higher generic privacy concerns than FB members. These results suggest that FB membership among undergraduates is *not* just a matter of their not being concerned, in general, about their privacy - other reasons must be explored.

4.3 Reported Facebook Usage

In order to understand what motivates even privacy concerned individual to share personal information on the Facebook, we need to study what the network itself is used for. Asking participants this question directly is likely to generate responses biased by self-selection and fear of stigma. Sure enough, by far, FB members deny FB is useful to them for dating or self-promotion. Instead, members claim that the FB is very useful to them for learning about and finding classmates (4.93 mean on a 7-point Likert scale) and for making it more convenient for people to get in touch with them (4.92), but deny any usefulness for other activities. Other possible applications of FB - such as dating, finding people who share one's interests, getting more people to become one's friends, showing information about oneself/advertising oneself - are ranked very low. In fact, for those applications, the relative majority of participants chooses the minimal Likert point to describe their usefulness (coded as "not at all" useful). Still, while their mean Likert value remains low, male participants find FB slightly more useful for dating than female.

And yet, when asking participants to rate *how often*, on average, their *peers* use FB for the same activities, the results change dramatically: learning about classmates and the convenience factor of staying in contact are still ranked very highly, but now "Showing information about themselves/advertising themselves," "Making them more popular," or "Finding dates" suddenly become very popular. See how the distributions almost invert in Figure 9.

Information Provided What information do FB members provide, and of what quality? Many members are quite selective in the type of information they provide - for instance, most publish their birthdays but

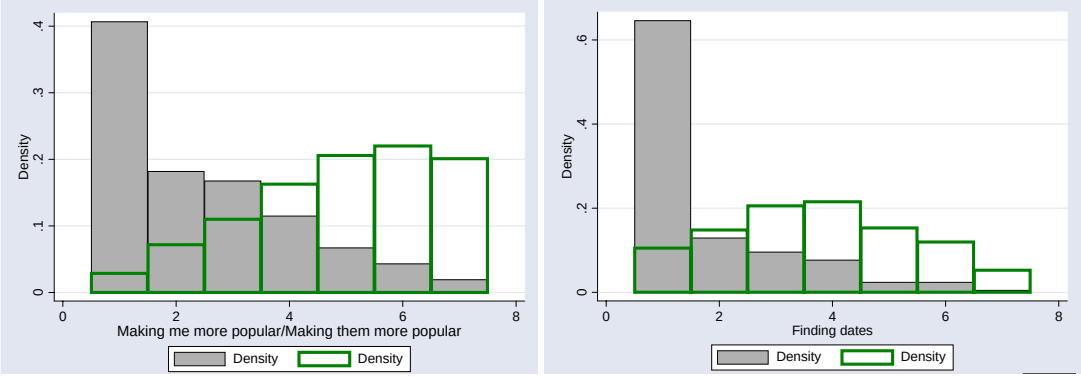

Fig. 9. Do as I preach, not as I do - How useful is FB for you (grey boxes) vs. how often do you believe other members use FB for (transparent boxes)?

hide their cell phone numbers. However, interestingly, our survey participants' answers imply that if a certain type of information is provided at all, it is likely to be of good quality: complete and accurate (see Figure 10).[6]

When controlling for participants who have abandoned the Facebook, we find out that as users they were less likely than continuing members to provide information such as their birthday (85.71% do not provide this information, while 86.12% of current members claim they *do* provide it - Pearson $\chi^2(2) = 33.9440$ $Pr = 0.000$), AIM (Pearson $\chi^2(2) = 14.2265$ $Pr = 0.001$), cellphone number, home phone number, personal address, political orientation, sexual orientation, and partner's name (the differences between non-members and members across the last six categories however are not statistical significant).

What personal information do you provide on the FaceBook and how accurate is that information?			
	I don't provide this information	I provide this information and it is complete and accurate	I provide this information but it is intentionally not complete or not accurate
Birthday	12% (29)	**84% (195)**	3% (8)
Cell phone number	**59% (138)**	39% (90)	2% (4)
Home phone number	**89% (207)**	10% (24)	0% (1)
Personal address	**73% (169)**	24% (55)	3% (8)
Schedule of classes	**54% (126)**	42% (97)	4% (9)
AIM	24% (56)	**75% (173)**	1% (3)
Political views	42% (97)	**53% (122)**	6% (13)
Sexual orientation	38% (88)	**59% (138)**	3% (6)
Partner's name	**71% (164)**	28% (65)	1% (3)

Fig. 10. Information provided by FB members.

Female members are not more or less likely than male members to provide accurate and complete information about their birthday, schedule of classes, partner's name, AIM, or political views. However, they are much *less* likely to provide their sexual orientation (Pearson $\chi^2(2) = 11.3201$ $Pr = 0.003$), personal address (Pearson $\chi^2(2) = 10.5484$ $Pr = 0.005$), and cell phone number (Pearson $\chi^2(2) = 10.9174$ $Pr = 0.004$). This confirms the results reported in [4], where less than 29% of females were found providing cell phone information, compared to 50% of male.

[6] Also such survey answers that elicit personal admissions about the quality of the data provided on FB may be, in turn, biased. However, since survey participants were not asked to disclose the actual information whose quality they were asked to evaluate, we have no reason to believe that their incentives to offer inaccurate answers were significant.

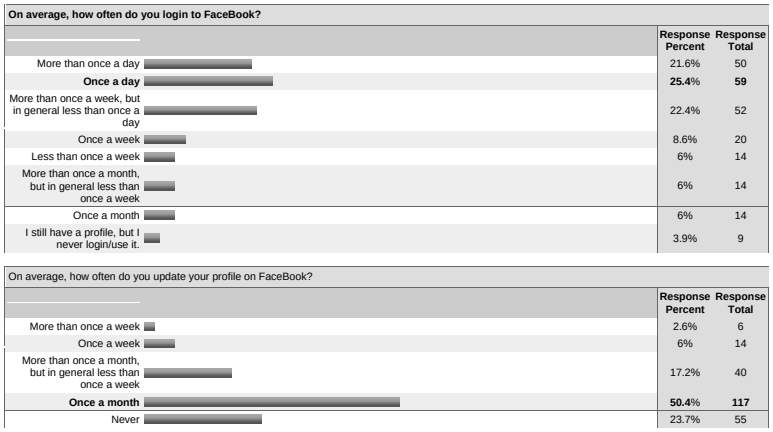

Fig. 11. Frequency of login and profile update.

Self-selection Bias? Often, survey participants are less privacy conscious than non participants. For obvious reasons, this self-selection bias is particularly problematic for survey studies that focus on privacy. Are our respondents a biased sample of the Institution's FB population - biased in the sense that they provide more information than the average FB members?

We did not find strong evidence of that. Since we mined the network before the survey was administered, we were able to compare information revelation by survey participants and non survey participants. It is true that, on average, our survey takers provide slightly more information than the average FB member. However, the differences in general do not pass a Fisher's exact test for significance, except for personal address and classes (where non participants provide statistically significant less information) and political views (in which the difference is barely significant).

Attitudes vs. Behavior We detected little or no relation between participants' reported privacy attitudes and their likelihood of providing certain information, even when controlling, separately, for male and female members. For instance, when comparing the propensity to provide birthday and the Likert values reported in the answers to the privacy threat question described at the beginning of Section 4.2, no statistically significant difference emerged: Pearson $\chi^2(12) = 5.2712$ $Pr = 0.948$. Comparable results were found when testing sexual orientation (Pearson $\chi^2(12) = 10.7678$ $Pr = 0.549$), partner's name (Pearson $\chi^2(12) = 15.1178$ $Pr = 0.235$), cell phone number (Pearson $\chi^2(12) = 19.0821$ $Pr = 0.087$), or personal address.

We obtained the same results when using the cluster variable that summarizes each respondent's privacy attitudes (see Section 4.2), both when using standard Pearson's χ^2 as well as when using Student's t test (the latter was used when comparing the mean privacy concern across respondents who provided or did not provide accurate information about various data types).

Combined with the results discussed in 4.2, the above evidence may suggest that privacy attitudes have some effect on determining who joins the network, but after one has joined, there is very little marginal difference in information revelation across groups - which may be the result of perceived peer pressure or herding behavior.

If anything, we found new confirmations of a privacy attitude/behavior dichotomy [18]. Almost 16% of respondents who expressed the highest concern (7 on the Likert scale) for the scenario in which a stranger knew their schedule of classes and where they lived, provide nevertheless both pieces of information (in fact, almost 22% provide at least their address, and almost 40% provide their schedule of classes).

Similarly, around 16% of respondents who expressed the highest concern for the scenario in which someone 5 years from now could know their current sexual orientation, partner's name, and political orientation, provide nevertheless all three types of information - although we can observe a *descending* share of members that provide that information as their reported concerns *increase*. Still, more than 48% of those with the highest concern for that scenario reveal at least their current sexual orientation; 21% provide at least their partner's name (although we did not control for the share of respondents who are currently in relationships); and almost 47% provide at least their political orientation.

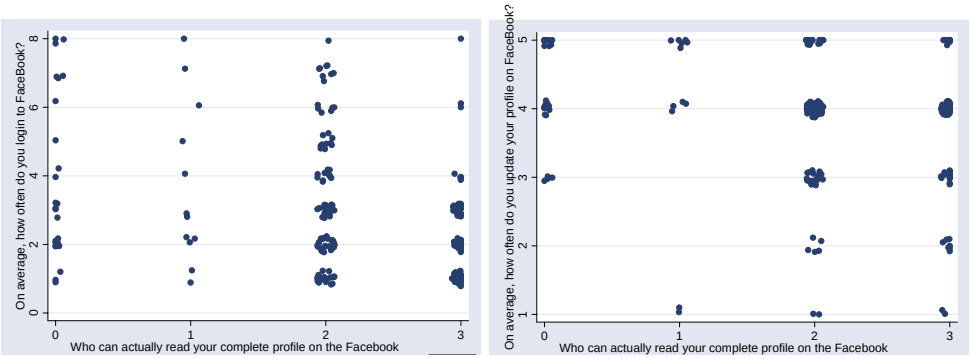

Fig. 12. Self-awareness of ability to control who can see one's profile, by frequency of login (left) and frequency of update (right). On the x-axis, the value 0 means "Do not know" if there is any way to control; 1 means "No control"; 2 means "Some control" and 3 means "Complete control." On the y-axis, higher values mean *less* frequent login or update.

4.4 Awareness of Facebook Rules and Profile Visibility

How knowledgeable is the average FB member about the network's features and their implications in terms of profile visibility?

By default, everyone on the Facebook appears in searches of everyone else, and every profile at a certain Institution can be read by every member of FB at that Institution. However, the FB provides an extensive privacy policy and offers very granular control to users to choose what information to reveal to whom. As mentioned above, relative to a FB member, other users can either be friends, friends of friends, non-friend users at the same institution, non-friend users at a different institution, and non-friend users at the same geographical location as the user but at a different university (for example, Harvard vs. MIT). Users can select their profile visibility (who can read their profiles) as well as their profile searchability (who can find a snapshot of their profiles through the search features) by type of users. More granular control is given on contact information, such as phone numbers.

And yet, among current members, 30% claim not to know whether FB grants any way to manage who can search for and find their profile, or think that they are given no such control. Eighteen percent do not know whether FB grants any way to manage who can actually read their profile, or think that they are given no such control. These numbers are not significantly altered by removing the 13 members who claim never to login to their account. In fact, even frequency of login does not explain the lack of information for some members. On the other hand, members who claim to login more than once a day are also more likely to believe that they have "complete" control on whom can search their profile.

Awareness of one's ability to control who can see one's profile is not affected by the frequency of login, but *is* affected by the frequency of update (a Pearson $\chi^2(12) = 28.9182$ $Pr = 0.004$ shows that the distribution is significant): see Figure 13. Note the difference between the two graphs and, specifically, the distribution by frequency of update for respondents who answered "Do not know" or "No control" (graph on the right).

Twenty-two percent of our sample do not know what the FB privacy settings are or do not remember if they have ever changed them. Around 25% do not know what the location settings are.

To summarize, the majority of FB members claim to know about ways to control visibility and searchability of their profiles, but a significant minority of members are unaware of those tools and options.

Self-reported visibility More specifically, we asked FB members to discuss how visible and searchable their own profiles were. We focused on those participants who had claimed never to have changed their privacy settings (that by default make their profile searchable by everybody on FB and visible to anybody at the same Institution), or who did not know what those settings were.

Almost every such respondent realizes that anybody at their Institution can search their profile. However, 24% incorrectly do not believe that anybody on FB can in fact search their profile. Misunderstandings about visibility can also go in the opposite direction: for instance, 16% of current members believe, incorrectly, that *anybody* on FB can read their profile.

In fact, when asked to guess how many people could search for their profile on FB (respondents could answer by selecting the following possible answers from a drop-box: a few hundred, a few thousands, tens of thousands, hundreds of thousands, millions), the *relative* majority of members who did not alter their default settings answered, correctly, "Millions." However, *more than half* actually underestimated the number to tens of thousands or less.

In short, the majority of FB members seem to be aware of the true visibility of their profile - but a *significant minority* is vastly underestimating the reach and openness of their own profile. Does this matter at all? In other words, would these respondents be bothered if they realized that their profile is more visible than what they believe?

The answer is complex. First, when asked whether the current visibility and searchability of the profile is adequate for the user, or whether he or she would like to restrict it or expand it, the vast majority of members (77% in the case of searchability; 68% in the case of visibility) claim to be satisfied with what they have - most of them do not want more or less visibility or searchability for their profiles (although 13% want less searchability and 20% want less visibility) than what they (correctly or incorrectly) believe to have. Secondly, as we discuss further below in Section 4.5, FB members remain wary of whom can access their profiles, but claim to manage their privacy fears by controlling the information they reveal.

4.5 Attitudes Towards the Facebook

So far we have glanced at indirect evidence of a number of different reasons for the dichotomy between FB members' stated privacy concerns (high) and actual information hiding strategies (mixed, but often low also for members with high stated concerns). Those reasons include peer pressure and unawareness of the true visibility of their profiles.

Another possible reason is the level of trust FB members assign to the network itself. On average, FB members trust the system quite a bit (and in general trust its members *more* than members of comparable services, like Friendster or MySpace - see Figure 13).

This happens notwithstanding the fact that almost 77% of respondents claimed not to have read FB's privacy policy (the real number is probably higher); and that many of them mistakenly believe that FB does not collect information about them from other sources regardless of their use of the site (67%), that FB does not combine information about them collected from other sources (70%), or that FB does not share personal information with third parties (56%). (We note that having read, or claiming to have read, the privacy policy, does *not* make respondents more knowledgeable about FB's activities.)

How much do you trust	Do not trust at all						Trust completely	N/A	Response Average
The FaceBook (the Company)	5% (10)	8% (17)	17% (38)	24% (53)	**25% (56)**	14% (32)	4% (8)	3% (7)	**4.20**
Your own friends on FaceBook	1% (3)	1% (3)	3% (6)	7% (15)	26% (57)	**38% (84)**	22% (49)	2% (4)	**5.62**
CMU Facebook users	3% (7)	6% (13)	14% (32)	26% (58)	**32% (70)**	15% (33)	3% (7)	0% (1)	**4.35**
Friends of your friends on FaceBook	6% (13)	8% (17)	15% (33)	**27% (60)**	24% (54)	14% (31)	4% (9)	2% (4)	**4.17**
On average, FaceBook users not connected to you	13% (28)	17% (38)	23% (51)	**27% (60)**	12% (27)	5% (10)	1% (3)	2% (4)	**3.29**
If you use or know about MySpace, MySpace users not connected to you	15% (33)	9% (19)	6% (14)	12% (26)	4% (9)	1% (2)	1% (3)	**52% (115)**	**2.78**
If you use or know about Friendster, Friendster users not connected to you	16% (35)	8% (17)	8% (17)	12% (27)	5% (12)	2% (4)	0% (1)	**49% (108)**	**2.82**

Fig. 13. How FB members assign trust.

While respondent are mildly concerned about who can access their personal information and how it can be used, they are not, in general, concerned about the information itself, mostly because they control that information and, with less emphasis, because believe to have some control on its access. Respondents are fully aware that a social network is based on information sharing: the strongest motivator they have in providing more information are reported, in fact, as "having fun" and "revealing enough information so that necessary/useful to me and other people to benefit from FaceBook."

However, psychological motivations can also explain why information revelation seems disconnected from the privacy concerns. When asked to express whether they considered the current public concern for privacy on social network sites such as the FaceBook or MySpace to be appropriate (using a 7-point Likert scale,

from "Not appropriate at all" to "Very much appropriate"), the response average was rather high (4.55). In fact, the majority of respondents agree (from mildly to very much) with the idea that the information *other* FB members reveal may create privacy risks to *those* members (that is, the other members; average response on a 7-point Likert scale: 4.92) - even though they tend to be less concerned about their *own* privacy on FB (average response on a 7-point Likert scale: 3.60; Student's t test shows that this is significantly less than the concern for other members: t = -10.1863, P <t = 0.0000; also a Wilcoxon matched pair test provides a similar result: z = -8.738, Pr <|z| = 0.0000).

In fact, 33% of our respondents believe that it is either impossible or quite difficult for individuals not affiliated with an university to access FB network of that university. "Facebook is for the students" says a student interviewed in [19]. But considering the number of attacks described in [4] or any recent media report on the usage of FB by police, employers, and parents, it seems in fact that for a significant fraction of users the FB is only an *imagined* community.

5 Survey and Network Data

In order to justify conclusions informed by a survey, the validity of the answers provided by the subjects has to be addressed. For this study we were in the *unique* position to be able to *directly* compare the answers provided by the participants with visible FB profiles to the information they actually provide in the profile (downloaded and archived immediately before the survey was administered). This section compares the survey responses with profile data and examines survey impacts in the form of changes to FB profiles of survey participants.

5.1 Comparison between Reported Answers and Actual Data

In order to gauge the accuracy of the survey responses, we compared the answers given to a question about revealing certain types of information (specifically, birthday, cell phone, home phone, current address, schedule of classes, AIM screenname, political views, sexual orientation and the name of their partner) with the data from the actual (visible) profiles. We found that 77.84% of the answers were exactly accurate: if participants said that they revealed a certain type of information, that information was in fact present; if they wrote it was not present, in fact it was not. A little more than 8% revealed more than they said they do (i.e. they claim the information is not present when in fact it is). A little more than 11% revealed less then they claimed they do. In fact, 1.86% claimed that they provide false information on their profile (information is there that they claim is intentionally false or incomplete), and 0.71% have missing false information (they claimed the information they provide is false or incomplete, when in fact there was no information).

We could not locate the FB profiles for 13 self-reported members that participated in the survey. For the participants with CMU email address, 2 of them did mention in the survey that they had restricted visibility, searchability, or access to certain contact information, and 3 wrote that not all CMU users could see their profile.

5.2 Survey Impacts

For this analysis, we eliminated the survey responses for users whose profile we could not locate on the network, ending up with 196 profiles out of the 209 self-proclaimed FB members participants. We downloaded information from the network immediately before and after administering the survey, both for users who responded to it and those who did not, and then compared the profiles.

First, we found a statistically significant difference in the byte size of the resulting files. The mean byte size decreased in both the experiment and the control group, but the experiment group changed significantly more than the control group (paired t test Pr <t = 0.0060). See Figure 14 for histograms of the file size changes for both groups. However, no significant changes were found when evaluating *individual* data fields: 5 survey participants reduced the information they provided compared to 4 profiles in the control group that similarly removed specific information.

After further investigation, we found that what happened was the following: the 9 profiles with the highest byte change (all >10kb) were in fact the ones that completely changed the visibility of their profile. They represent slightly more than 5% of our sample of current FB members (whose profile before the survey was visible). Out of this group 6 were female and 3 male. In the control group only 2 profiles changed visibility. This difference is statistically significant (χ^2 Pr<0.05).

While the difference is significant and somewhat surprising, the magnitude in terms of number of members that changed their behavior is relatively small. One should note that this change happened even without

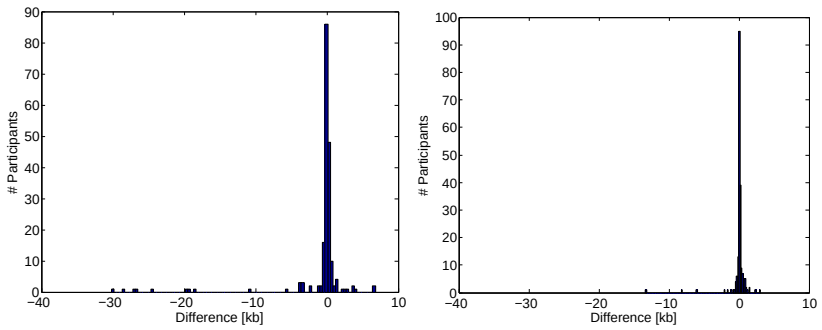

Fig. 14. Changes in profile sizes for survey participants and a control group. The sizes for the survey participants changed significantly more.

us providing the survey participants with a real threat scenario. In addition, although privacy concerned individuals are on FB, only a fraction of them may have such high concerns to be induced to abandon the network just by questions about its privacy implications. In fact, we found that this group of "switchers" have higher means in terms of average privacy attitudes, and their distributions of privacy attitudes are skewed towards the right (that is, towards higher concerns) - than non "switchers," although such differences are not statistically significant.

6 Discussion and Future Work

Online social networks offer exciting new opportunities for interaction and communication, but also raise new privacy concerns. Among them, the Facebook stands out for its vast membership, its unique and personally identifiable data, and the window it offers on the information revelation behavior of millions of young adults.

In our study we have combined survey instruments with data mined from a FB community at a North American college Institution. We looked for demographic or behavioral differences between the communities of the network's members and non-members, and searching for motivations driving the behavior of its members. Our analysis is going to be complemented by other experiments, but we can summarize here a number of initial results.

Age and student status obviously are the most significant factors in determining FB membership. However, we observe that privacy attitudes also play a role, but only for the non undergraduate population. In fact, most of highly privacy concerned undergraduates still join the network. While a relative majority of FB members in our sample are aware of the visibility of their profiles, a significant minority is not. The 'aware' group seems to rely on their own ability to control the information they disseminate as the preferred means of managing and addressing their own privacy concerns. However, we documented significant dichotomies between specific privacy concerns and actual information revelation behavior. In addition, misunderstanding or ignorance of the Facebook (the Company)'s treatment of personal data are also very common.

It is interesting to note that a pilot study we ran in September 2005 provided similar results, but also small, yet significant differences in terms of members' awareness of their profile visibility and their ability to control it: respondents a few months ago appeared less aware of privacy risks and of means of managing their own profiles. This evidence may suggest that the widespread public attention on privacy risks of online social networks is affecting, albeit marginally, some of their users.

7 Acknowledgements

This research was supported by CMU Berkman Faculty Development Fund and CMU CyLab. We would like to thank Lorrie Cranor, Charis Kaskiris, Julia Gideon, Jens Grossklags, and Bradley Malin for helpful insights and suggestions in the development of the survey protocol.

References

1. Parker, R.: Alcohol policy violated. Kansan.com **February 28** (2006)
2. Youngwood, S.: Networking by the 'book'. The Times Argus **February 26** (2006)
3. Kharif, O.: Big brother is reading your blog. BusinessWeek online **February 28** (2006)
4. Gross, R., Acquisti, A.: Privacy and information revelation in online social networks. In: Proceedings of the ACM CCS Workshop on Privacy in the Electronic Society (WPES '05). (2005)
5. d. boyd: Reflections on friendster, trust and intimacy. In: Intimate (Ubiquitous) Computing Workshop - Ubicomp 2003, October 12-15, Seattle, Washington, USA. (2003)
6. d. boyd: Friendster and publicly articulated social networking. In: Conference on Human Factors and Computing Systems (CHI 2004), April 24-29, Vienna, Austria. (2004)
7. Donath, J., d. boyd: Public displays of connection. BT Technology Journal **22** (2004) 71–82
8. Liu, H., Maes, P.: Interestmap: Harvesting social network profiles for recommendations. In: Beyond Personalization - IUI 2005, January 9, San Diego, California, USA. (2005)
9. Jagatic, T., Johnson, N., Jakobsson, M., Menczer, F.: Social phishing. Communications of the ACM **Forthcoming** (2006)
10. Stutzman, F.: An evaluation of identity-sharing behavior in social network communities. In: Proceedings of the 2006 iDMAa and IMS Code Conference, Oxford, Ohio. (2006)
11. Sege, I.: Where everybody knows your name. Boston.com **April 27** (2005)
12. Anderson, B.: Imagined Communities: Reflections on the Origin and Spread of Nationalism. Revised edn. Verso, London and New York (1991)
13. Wall, L., Christiansen, T., Orwant, J.: Programming Perl. 3rd edn. O'Reilly (2000)
14. Burke, S.: Perl & LWP. O'Reilly (2002)
15. Westin, A.F.: Harris-equifax consumer privacy survey (1991). Technical report, Equifax, Inc., Atlanta, GA (1991)
16. Acquisti, A., Grossklags, J.: Privacy and rationality in decision making. IEEE Security & Privacy **January-February** (2005) 24–30
17. Berry, M., Linoff, G.: Data Mining Techniques for Marketing, Sales and Customer Support. Wiley, New York (1997)
18. Acquisti, A.: Privacy in electronic commerce and the economics of immediate gratification. In: Proceedings of the ACM Conference on Electronic Commerce (EC '04). (2004) 21–29
19. Kornblum, J., Marklein, M.B.: What you say online could haunt you. USA Today **March 8** (2006)