Snooping on Mobile Phones: Prevalence and Trends

Diogo Marques,[1] Ildar Muslukhov,[2] Tiago Guerreiro,[1] Konstantin Beznosov[2] and Luís Carriço[1]

[1] LaSIGE, Faculdade de Ciências
Universidade de Lisboa, Lisbon, Portugal
[dmarques, tjvg, lmc]@di.fc.ul.pt

[2] Department of Electrical and Computer Engineering
University of British Columbia, Vancouver, Canada
[ildarm, beznosov]@ece.ubc.ca

ABSTRACT

Personal mobile devices keep private information which people other than the owner may try to access. Thus far, it has been unclear how common it is for people to snoop on one another's devices. Through an anonymity-preserving survey experiment, we quantify the pervasiveness of *snooping attacks*, defined as "looking through someone else's phone without their permission." We estimated the 1-year prevalence to be 31% in an online participant pool. Weighted to the U.S. population, the data indicates that 1 in 5 adults snooped on at least one other person's phone, just in the year before the survey was conducted. We found snooping attacks to be especially prevalent among young people, and among those who are themselves smartphone users. In a follow-up study, we found that, among smartphone users, depth of adoption, like age, also predicts the probability of engaging in snooping attacks. In particular, the more people use their devices for personal purposes, the more likely they are to snoop on others, possibly because they become aware of the sensitive information that is kept, and how to access it. These findings suggest that, all else remaining equal, the prevalence of snooping attacks may grow, as more people adopt smartphones, and motivate further effort into improving defenses.

1. INTRODUCTION

Mobile phones are not just phones anymore, they are interfaces to much of users' social lives, and keep records which, in all likelihood, include intimate, sensitive, or confidential information. As long as those records are interesting to anyone, there is a risk that they might try to obtain them.

The speed and extent to which mobile devices are being adopted has created new opportunities for remote, sophisticated adversaries. Phenomena like mobile malware, surveillance by state-sponsored actors, and personal data tracking for commercial purposes, have entered into public discourse, and became, reasonably so, a point of concern [38]. However, in their daily lives, users face a more immediate threat: people with whom they have close social ties can infringe on their privacy just by picking up their devices and browsing through their data. Those social *insiders* [29] can act opportunistically, without having any special skills or abilities. Such may happen when devices are left unattended, or handed over with the expectation of limited use. Often, social insiders can achieve their objectives just by undertaking what we will refer to as a *snooping attack*, that is, by looking at information that was not intended for them, without a primary intent to extract data or make changes. If we conceive of privacy as the ability to have control over the ways others know us [33], being snooped on by people whose opinion we care about is a violation of privacy in its most fundamental sense.

There are technological defenses against snooping attacks, most notably authentication mechanisms. However, it has become clear that people very often do not use them [12, 17, 23]. While there is debate over why people make such a choice, and over if and how they could be encouraged to choose differently, users remain in a situation where there are more opportunities for snooping than there could otherwise be. More opportunities, however, do not necessarily translate into more actual offenses. This uncertainty about whether people's phones are commonly, or only rarely, being snooped on, casts doubt over the importance and/or urgency of securing their devices against third parties that are, at first sight, trusted.

In this paper, we bring new evidence into this conversation, by measuring actual successes in conducting snooping attacks, from the attacker's perspective. From a security standpoint, it is of special importance to know how successful snooping attacks are, because high degrees of success indicate that existing defenses, both behavioral, like keeping the device on oneself at all times, and technological, like device locking, are inadequate. We thus aimed to measure the proportion of people, in population with a large degree of mobile device adoption, that successfully snooped on someone else's device, and to explore the pervasiveness of the phenomenon, or lack thereof, across population groups. We selected the U.S. adult population as a target, because it is easily accessible and well characterized in terms of mobile device adoption.

The main challenge with obtaining such data is methodological. If we were to field a survey asking people whether they had snooped on someone else's device, we could not reasonably expect honest responses, because such behavior is commonly deemed to be censurable. Thus, we employed the list experiment (e.g., [27]), a technique in which participants are asked to look at a list of items, and indicate how many (not which) they identify with. In list experiments, one group of participants receives a list of control items, and another group a list of the same control items plus an item of interest. An aggregate estimate of positive response to the item of interest can be calculated by the difference between groups, without knowing the true answer for each respondent. A more detailed description of the technique, and the rationale for its selection over other techniques, is provided in Section 3.

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

List experiments are understood to provide less biased estimates of response to sensitive questions, in comparison with direct self-reporting, but require careful design. The way in which list questions are worded, and the way in which surveys are administered, can have significant impact on measurement error [27]. We conducted two empirical studies to address these issues.

In a first study, conducted with Google Consumer Surveys (GCS), we selected the control and sensitive items to include in the list. For the control items, we measured, with direct questioning, the prevalence of previously reported behaviors that relate to privacy and security. Based on 1,140 responses, we selected a mix of items that prevents ceiling and floor effects. For the sensitive item, special consideration was given to how it was framed, because the specific wording would be the operational definition of the construct that we wanted to measure – in this case, successful snooping attacks. We tested 4 alternative ways of wording the concept such that it was easy to understand and mapped to the security issue at stake. Based on 1,086 responses, we concluded that the most adequate wording, among the alternatives, was "looked through someone else's phone without their permission". This study is reported in Section 4.

A second methodological challenge arose from a decision, made at the outset, to field the survey in Amazon Mechanical Turk (MTurk). MTurk is commonly used to target large participant pools [34], but doubts have been raised about its appropriateness for survey research [14], since participants, and especially those with low reputation, may engage in satisficing [32, 35]. To validate that list experiments on MTurk produce reliable measurements, we ran a list experiment with one control group and two treatment groups (n = 434), who received extra items with known prevalence of ~100% (having opened eyes in the morning) and ~0% (having travelled in interplanetary space). We were then able to compare the known prevalence to the one estimated by the list experiment, across 3 groups of MTurk participants, with distinct reputation levels. We concluded that list experiments appropriately estimated expected proportions, without the need to control for participant reputation. This finding, which is reported in Section 5, is a secondary generalizable contribution of this work.

Taking these findings into consideration, and making conservative design choices, we deployed a list experiment to MTurk to measure the prevalence of snooping attacks (n = 1,381). In Section 6, we describe the final design, the data collection process, and report on the proportion of people who, in 1 year, successfully engaged in snooping attacks on others' mobile phones, offering both a point estimate of prevalence, and predictors of such behavior. We provide estimates for the MTurk sample, which is often taken as being representative of the Internet population, and further project it into the U.S. adult population, by post-stratification weighting. The main findings are as follows:

- An estimated 31% of participants had "looked through someone else's phone without permission," in the 12-month period before the survey was conducted.
- Adjusting the younger and more male MTurk sample to the U.S. adult population, the 1-year prevalence was estimated at 20%.
- Engaging in snooping attacks does not seem to be strongly related to gender, level of education, or geographical region.
- Younger participants were notably more likely to have engaged in snooping attacks, to the extent that the behaviour was estimated to be prevalent (52%) among those between 18 and 24 years of age.
- Those who own smartphones are much more likely to snoop on others.

Although this study could not establish mechanisms by which the observed trends emerged, the fact that the youngest participants and those who used smartphones were more likely to snoop on others suggested a common cause. It has been noted that smartphone users often engage in a pattern of adoption in which the phone mediates important aspects of their private social life [9, 39]. In a follow-up study (n = 653), with a similar design to the previous, we examined whether, among smartphone users, depth of adoption predicted the prevalence of snooping attacks. We confirmed that the more people use their smartphones in ways that generate privacy-sensitive data, the more likely they are to snoop on others, even when controlling of age. A compelling explanation for these findings is that, as people learn by their own usage what kinds of sensitive information is kept on smartphones, they gain a better sense of what they could have access to if they were to snoop. This final study is reported in Section 7.

Overall, these results indicate that snooping on other people's devices must be relatively easy, to be so common. Furthermore, the population trends that we found suggest possible growth of the phenomena. If it is the youngest, and those who adopt smartphones to a larger extent, that are more likely to snoop on others, then growth may come from aging of the cohort, or from more people adopting smartphones in ways that make them aware of the private data that is kept. The situation calls for additional efforts in providing adequate defenses against socially-close adversaries, and for a re-examination of assumptions of trust in mobile security threat models.

2. RELATED WORK

It has been widely documented that smartphones are used very differently than either regular phones or computers, and, as a result, store a great deal of sensitive information, including access codes, personal communication, call and text logs, contacts, pictures, videos, and location records (e.g., [2, 13, 28]). Users have been found to be concerned about the risks to their privacy that have therefore emerged [9, 36]. Events have not proved them wrong.

In the last few years, there has been much discussion about phenomena like mobile malware, government surveillance, and personal data gathering for commercial purposes (e.g., [13, 37, 38]). Threats such as these, in which adversaries are technologically sophisticated, and act remotely, have traditionally been seen as the potentially most damaging. However, end-users are very rarely affected in a practical sense, and, when they are, the impact on their lives has been somewhat limited, mainly taking the form of unsolicited advertising [13].

Recently, as spearfishing and insider threats have gained more attention in the computer security community, so have socially-close adversaries been recognized as a threat to personal mobile computing [29]. Younger users, the so-called digital natives, are indeed more concerned about insiders: they are more aware of threats with a social context (like those arising from loss, theft, snooping or shoulder-surfing) than of threats with a technical connotation (like those arising from malware or network attacks) [23].

In a recent Pew survey [36], 12% of US mobile phone owners reported having had another person access the contents of their phone

in a way that made them feel their privacy was invaded. This statistic can be seen as an indirect measure of snooping attack success, but one that is likely inaccurate. For instance, many people may have had their smartphones snooped on but not know about it. Conversely, the fact that someone felt that their privacy was invaded does not mean that there was an explicit intention by the person accessing the device.

Corroborating that finding, in a recent survey with an MTurk sample, 14% of participants reported being targets of snooping ("Someone used my mobile phone without my permission with intention to look at some of my data"), and 9% reported being attackers ("I used someone's mobile phone without owner's permission to look into his/her data") [29]. This is, as far as we know, the first measurement of successful snooping attacks from the attacker's perspective. This measurement, however, is not generalizable, for two reasons. First, because it was meant to be a sample summary, not a population estimate, as part of a study with a broader objective. Second, because the questions were asked directly, and thus the number of people willing to identify with behaviour that can be seen as offensive is expected to be biased by the social desirability effect, as suggested by a 6 percentage point mismatch between reported targets and attackers.

We aimed to measure how often people actually succeed in conducting snooping attacks, taking into consideration that they might not be willing to admit it. Furthermore, we were interested in an estimate bounded in time, namely one year, to allow periodical comparisons. By measuring 1-year prevalence periodically, it is possible to discern any changes, which could, for instance, indicate adoption of new defenses. In contrast, if participants are periodically asked if they *ever* snooped, changes might not be observable until there is a sufficiently large proportion of new entrants into the population.

Comparing our results to previous statistics, the problem does seem to have been underestimated. We found that 20% of U.S. adults engaged and succeeded in snooping attacks in a year, while only 12%, over their lifetime, report having had the contents of their devices accessed [36]; and, for a comparable MTurk population, using the list experiment procedure, we estimated 1-year prevalence of snooping attacks (31%) to be approximately 3 times as high as the previous lifetime prevalence estimate obtained with direct questioning (9%) [29]. Unless there was a very large upward shift in prevalence that would explain these differences, it seems that indeed many people never come to learn that they were snooped on, and that when asked directly, people who have snooped on others often do not admit to it.

3. ASKING SENSITIVE QUESTIONS

Studies of attitudes, opinions and behaviors run into measurement error whenever self-reports can not be trusted. One classic example is that men consistently report having had a far greater number of sexual intercourse partners than women, which, if true, would defy logic [42].

One source of measurement error is social desirability bias [41]. When questions are sensitive, respondents tend to give answers that they understand to be the right ones, and not necessarily the truth. Questions that pertain to protecting one's privacy are known to be subject to that bias. It has been shown that the mere addition of privacy wording in surveys makes respondents much more likely to give socially desirable responses [6].

Indirect survey techniques to reduce social desirability bias have emerged in the last few decades. Their main principle is assurance of response confidentiality by design, not policy. Respondents have strict guarantees that their individual answer will not be revealed, and are therefore more likely to answer truthfully. The cost to researchers is that they will not know the response of each individual, only aggregate estimates.

Two main types of such survey instruments have received attention. One is the *randomized response technique* (RRT) [5]. In its simplest form, respondents are shown a sensitive question and asked to privately flip a coin. If it lands on one side, participants must answer "yes", regardless of truthfulness, and if it lands on the other side, they must answer truthfully, "yes" or "no". Each individual respondent is thus assured that answering "yes" does not reveal their true response, as long as no one else knows on which side the coin landed. But knowing that the probability of a coin landing heads or tails is equal, the total proportion of positive responses can be calculated by assuming that half the positive responses are a consequence of the coin toss, and the remaining are truthful.

The other technique is the *list experiment* (sometimes called unmatched count technique, or item count technique, or unmatched block design), which we have employed. List experiments are a kind of survey experiment [30], which involve dividing a sample into two groups, the control and the treatment. As an example, in a recent study [40], where researchers addressed the puzzle of why a particular ballot initiative failed to pass when opinion polls indicated otherwise, the control group was asked the following question:

Here is a list of four things that some people have done and some people have not. [. . .] Do not tell me which you have and have not done. Just tell me how many:
- Discussed politics with family or friends;
- Cast a ballot for Governor Phil Bryant;
- Paid dues to a union;
- Given money to a Tea Party candidate or organization.
How many of these things have you done in the past two years?

The treatment group saw the question with the following extra item:

- Voted 'YES' on the 'Personhood' Initiative on the November 2011 Mississippi General Election ballot

With this technique, participants do not have to reveal their truthful answer to the extra item, which is the one actually being measured. Yet, the proportion can be estimated by comparing the mean number of items selected by respondents in control and treatment groups. All the rest being equal, a difference in means can be attributed to the presence of the extra item. The difference in means is thus the estimate of proportion of positive responses to the sensitive item.

It has been shown that both the list experiment and the RRT reduce response bias. In the mentioned validation study [40], which tested both approaches, it was found that an RRT survey predicted almost exactly the actual vote. A list experiment survey considerably reduced the bias, but still underestimated the actual vote share.

For online surveys, however, application of the RRT is problematic. Since the procedure is complex, respondents have to expend considerable time to understand it, and often they have trouble believing their true answers are not revealed [11]. As we intended to deploy the survey on MTurk, where participant attention is already scarce (e.g., [35]), and extra time is costly, we opted for a list experiment. Even if list experiments provided estimates that were overly conservative, on the issue of snooping, it was best to err on the side

of caution. If even a conservative estimate was relevant, than surely a higher estimate would have at least the same consequence.

The list experiment procedure seldom appears in HCI research (with one exception that we know of [1]). With this paper we also wanted to call attention to the growing tool belt of survey research methods for sensitive topics, which can help untangle the often found discrepancy between self-reports and actual behaviour in privacy-related studies.

4. STUDY 1: ITEM SELECTION

List experiments aim to reduce the measurement error that would occur if sensitive questions were asked directly. For them to be effective, careful consideration has to be given to the composition of the list. The perception of confidentiality can be jeopardized when lists are not credible, or when truthful answers would reveal that respondents had answered positively to the sensitive item. With this first empirical study, we aimed to compose a list of items that would minimize the chances of obtaining unreliable measurements from a full-scale survey experiment.

The danger of unreliable measurement can be mitigated by following common advice on designing list experiments (e.g., [4, 11, 15, 27]), which includes:

1. **Avoid ceiling effects** A ceiling effect happens when all the control items are so common that many participants would, if answering truthfully, identify with all items, thus revealing their positive answer to the sensitive one.
2. **Avoid floor effects** A floor effect occurs when the control items are so uncommon that, for many participants, the only item they could credibly report as identifying with would be the sensitive one.
3. **Avoid lists that are too short** Short lists increase the likelihood of a ceiling or floor effect.
4. **Avoid lists that are too long** Long lists increase variance and demand more attention from participants.
5. **Avoid contrast effects** If the sensitive item is too salient, respondents might worry that any non-zero answer to the list is indicative of identification with it. The list should therefore include control items that are on the same topic as the sensitive item, which itself should be worded in neutral language.

Taking this advice into account, we decided to run surveys on individual behaviors to obtain prevalence estimates, so we could select a combination of control items, and a wording for the item pertaining to snooping attacks, that would make confidentiality plausible.

4.1 Procedure

To build the list of items, we ran direct question surveys on several candidate items using Google Consumer Surveys (GCS).

For each candidate control item, we aimed at a target sample of 100 participants. For candidate sensitive items, we targeted a sample of 250 participants, as we expected lower sensitivity, due to social desirability bias. The actual number of participants is often different than the target, because of the particular way in which GCS samples [26].

For the control items, to avoid contrast effects with the sensitive item, we selected candidates among previously documented behaviors or situations related to mobile privacy [13] and online privacy [38], shown in Table 1, rows 1 to 8.

For the sensitive item, that pertains to snooping attacks, we tested four ways of wording the behavior, shown in Table 1, rows 9 to 12. The formulations avoid the word "snooping", which we deemed to have a too-negative connotation, and instead test a maliciousness dimension, with "used" vs. "looked through" wording, and an egregiousness dimension, with "without knowledge" vs. "without permission" wording.

4.2 Results

4.2.1 Control item selection

Our surveys did not find privacy-relevant behaviors or situations that can be said to be of high prevalence, but items of low prevalence were abundant. In part, such could be explained by the existence of social desirability bias for some of the controls.

Nevertheless, taking the measured prevalences for candidate items as indicative of true differences in the population, results indicated it would be trivial to avoid ceiling effects (advice 1) even with a short list, by selecting among the items with very low prevalence.

Avoiding floor effects (advice 2) was more challenging, as we did not find highly prevalent items. We decided to include 4 control items in the final list, at the cost of possible lower precision in estimates (advice 4). With 4 control items rather than 2 or 3, there were, we reasoned, enough guarantees of confidentiality. Even if respondents answered "1" it would be plausible enough that they were referring to one of the controls that is not abundantly privacy-sensitive, such as receiving spam.

We finally selected the items from surveys 1, 2, 4 and 5, which are the ones with the highest and lowest prevalence, that still pertain to mobile security, and thus generate less contrast (advice 5) with the sensitive item.

4.2.2 Sensitive item selection

For the item conveying the "snooping attack" construct, the surveys we conducted did not show any appreciable differences as a result of different wording. A Chi-squared test did not provide evidence that the wording had an overall effect on the rate of positive answers ($\chi^2(3) = 5.36$, $p = 0.1471$, Cramer's V = 0.07), nor that wording conveying either egregiousness or maliciousness had significant effects in isolation ($\chi^2(1) = 2.610$, $p = 0.1062$, Cramer's V = 0.05, and $\chi^2(1) = 1.192$, $p = 0.2749$, Cramer's V = 0.04, respectively). In a logistic regression model of positive or negative answer as a function of egregiousness or maliciousness wording, we also did not find either factor to be a significant predictor at the 0.05 significance level, and the model accounted for very little of the deviance (null deviance 751 on 1085 d.f. vs. residual deviance 746 on 1083 d.f.).

We could have expanded the sample to get more precise estimates and possibly establish minute differences between wording choices, but given the observed effect sizes, and the likelihood that social desirability bias was already introducing measurement error, any differences, even if statically significant, were unlikely to be of practical importance. We thus concluded that, for the purpose of our main survey, we should use the wording that, on its face, represented an egregious violation of an access policy with malicious intent: having *looked through* someone else's cell phone without their *permission*.

4.3 Discussion

Based on the results of direct question surveys, we composed a list of items that included a mix of controls which were low to medium

Table 1: Results of single question surveys conducted in Google Consumer Surveys: 1 to 8 for candidate control items for the list experiment question (1-5 behaviors related to mobile security, 6-8 behaviors related to online privacy), 9 to 12 for different ways to word the item conveying the concept of snooping attacks. Items selected for the list in bold.

Survey	Question	Yes	No	Participants
1	**In the past 12 months, have you purposefully made phone calls or sent text messages to 1-900 numbers from your cell?**	6%	94%	183
2	**In the past 12 months, have you lost or had your cell phone stolen from you?**	11%	89%	191
3	In the past 12 months, have you sent a text message to the wrong person by mistake?	17%	83%	155
4	**In the past 12 months, have you shared pictures from your cell phone on social media (for instance, Facebook or Twitter)?**	27%	73%	108
5	**In the past 12 months, have you received at least one text message/IM with unsolicited advertising (spam) on your cell?**	42%	58%	173
6	In the past 12 months, have you been asked to create a new password for an online service?	37%	63%	110
7	In the past 12 months, have you at least once cleared your cookies or browsing history?	54%	46%	113
8	In the past 12 months, have you at least once deleted / edited something you posted online?	26%	74%	107
9	In the past 12 months, have you used someone else's cell phone without their knowledge?	9%	91%	250
10	In the past 12 months, have you used someone else's cell phone without their permission?	11%	89%	335
11	In the past 12 months, have you looked through someone else's cell phone without their knowledge?	10%	90%	250
12	**In the past 12 months, have you looked through someone else's cell phone without their permission?**	15%	85%	251

prevalence, and an item of interest that referred to a "snooping attack" with mild language.

One limitation of this study is that an entire baseline list wasn't tested, but only individual items. Although unlikely, interactions may exist between items that increase likelihood of ceiling or floor effects. Another limitation is that this study is not informative as to the possibility of contrast effects between the selected controls and the candidate items of interest, which could hurt credibility.

Either of these two limitations could potentially result in estimates of prevalence that were lower than the true population proportion. This was deemed acceptable, as it represented a conservative design choice for the main survey.

5. STUDY 2: CONDUCTING LIST EXPERIMENTS ON MTURK

Once we had selected items for the list question, we considered the problem of satisficing in MTurk list experiments.

In surveys, respondents sometimes select answers that are acceptable and easy to give, regardless of their truthfulness [31, p. 244]. Previous research suggests that satisficing is indeed a concern for survey research with MTurk samples [14, 22, 32].

There was reason to suspect that this concern extended to list experiments. List questions are cognitively more demanding than short, direct ones [11], taking more time and effort to answer thoughtfully. Yet, MTurk workers have incentives to maximize compensation per time unit [34]. For studies in which groups of observational units are compared, as is the case of list experiments, there are concerns that MTurk samples, especially those with non-naive participants, may provide measurements with greater error, leading to underestimation of effect sizes [8] and, at worst, to not finding effects when they are present (type II error).

One popular way to counteract satisficing is using attention check question (ACQs) [32, 35]. ACQs are questions whose right answer is known in advance, such as logic puzzles, trick questions, and direct instructions to answer a certain way. Although their use is well accepted and built on evidence (e.g., [35]), MTurk workers are now very much aware of this practice, and may have therefore adjusted. It has been suggested that some workers may scan for ACQs, answer them attentively, and rush through the remaining questions [18].

Another way to mitigate satisficing is restricting participation to high-reputation workers. When posting a task to MTurk, it is possible to restrict participation on a set of criteria. Two such criteria are commonly used as proxies for reputation: the total number of tasks that participants have completed in the past, and the proportion of their submitted work that was accepted by requesters. Previous research indicates that filtering participation to workers with at least 95% acceptance rate is sufficient to obtain good quality data [35]. But, based on our own experience conducting studies on MTurk, and expert opinion we had solicited, we came to believe that a 95% acceptance rate was now relatively easier to attain than at the time in which that research was conducted. There's indication that requesters have grown weary of refusing work, as it might affect their own reputations, which are disseminated in platforms like Turkopticon [21].

Since satisficing, and the measurement error associated with it, would affect the reliability of the estimates we were to obtain in our main study, we aimed to understand if list experiments in MTurk could be made trustworthy by restricting participation based on reputation and using ACQs. We devised a between-subjects experiment where surveys were administered to MTurk workers with distinct degrees of reputation (3 levels). Participants in each reputation group would be randomly assigned to receive a question with only the control items, or with the control items plus an item with ~0% expected prevalence, or with the control items plus an item with ~100% expected prevalence. Thus, we could compare the expected prevalence to the one estimated by the difference-in-means between groups.

5.1 Procedure

We configured an online questionnaire to randomly assign participants to receive a list question with one of the following lists:

Control The 4 control items derived from Study 1 (Table 1, items in bold).

Treatment-0 Control items, plus: "In the past 12 months, I've been to space, aboard an interplanetary vessel that I built myself" (~0% true prevalence).

Table 2: Number of participants, and mean items selected, by level of reputation and question version.

	Control		Treatment-0		Treatment-1	
	n_c	Mean	n_{t0}	Mean	n_{t1}	Mean
Low	51	1.71	54	1.61	44	2.59
Medium	46	1.13	47	1.51	42	2.43
High	57	1.46	33	1.45	60	2.50
Overall	154	1.44	134	1.54	146	2.51

Table 3: Prevalence estimated by the difference-in-means between groups, by level of reputation and question version.

	Treatment-0 - Control		Treatment-1 - Control		Treatment-1 - Treatment-0	
	Estimate	SE	Estimate	SE	Estimate	SE
Low	-9 %	0.190	88 %	0.186	98 %	0.189
Medium	38 %	0.195	130 %	0.201	92 %	0.206
High	-0.2 %	0.182	104 %	0.177	105 %	0.201
Overall	10 %	0.110	107 %	0.110	97 %	0.115

Treatment-1 Control items, plus: "In the past 12 months, I've opened my eyes in the morning at least once (for instance, after waking up)" (~100% true prevalence).

The attention check items were created by us, and, as far as we know, not previously used in MTurk surveys. In this way, we intended to minimize the effect of respondents detecting them without expending much mental effort, or using automated tools.

The rest of the questionnaire had the same structure and questions as the one to be used in the main survey. We posted it as a task on MTurk 3 times, assuring no repeated participation by the custom qualifications method [25]. Each time we posted it, we enforced system-level qualifications that limited participation to workers in the US, and created the following three reputation groups:

High Approval rate of 98% or higher, and at least 10,000 completed tasks.

Medium Approval rate of 95% or higher; at least 5,000, and no more than 10,000 completed tasks.

Low No minimum approval rate, and at most 5,000 completed tasks.

We targeted 150 participants by reputation group, with randomization expected to assign approximately 50 to each version of the list questions.

5.2 Results

5.2.1 Effect of reputation

Table 2 shows the average number of items that participants selected, discriminated by levels of reputation and version of questionnaire.

We found no evidence that the mean number of selected items was different depending on reputation, when the list question was either the Treatment-0 or Treatment-1 versions (columns 5 and 7; one-way ANOVA for Treatment-0: $F(2) = 0.305$, $p = 0.737$; for Treatment-1: $F(2) = 0.292$, $p = 0.747$). Only those that received the Control version, which had no attention check items, were found to have answered differently according to reputation level (column 3, $F(2) = 5.053$, $p = 0.00751$). Particularly, those in the (*Medium reputation x Control version*) condition selected, on average, 1.13 items, which was the lowest among those that received either the Control version or the Treatment-0 version.

5.2.2 Comparison to ground truth

Table 3 shows the estimates, by the difference-in-means, of positive answers to "been to space" (Treatment-0) and "opened eyes in the morning" (Treatment-1) items.

The difference between the means of the Treatment-0 group and the Control group was expected to be 0 if participants were answering attentively, since they had the same number of items they could identify with. If, on the other hand, participants were choosing at random, those that received the Treatment-0 version would have selected, on average, more items, because there is one more option – a truly random response pattern in both groups would yield a difference-in-means of 0.5. The difference we actually found, not taking into account level of reputation, was 0.1, which is non-negligible, as it would mean that 10% of our sample had travelled in space. We also observed an inconsistent pattern across reputation groups, with the abnormally low mean in the (*Medium reputation x Control version*) condition inducing a difference-in-means of 0.38, thus closer to 0.5 than the expected 0.

For differences between Treatment-1 and the two possible baselines, Control and Treatment-0, the same principle applies: attentive participation should yield a difference-in-means of 1.0, and random response 0.5. Either the Control or Treatment-0 can be baselines because one item in the Treatment-0 version has true prevalence of 0%. What we found was that when the baseline was Control, the overall difference-in-means, regardless of reputation, was 1.07, and when the baseline was Treatment-0, it was 0.97. The comparison between the groups that received attention checks, Treatment-0 and Treatment-1, was the closest to yield the expected proportion of 1.0. Furthermore, that comparison did not overestimate the true proportion, as did the comparison between Treatment-1 and Control.

Thus, the attention checks we had crated seemed to elicit enough attention from participants as to prevent degrees of satisficing that would jeopardize the validity of difference-in-means estimates. The feedback form that we included in the task provided some anecdotal indication that they generated goodwill among workers. As an example, participant 208 (low reputation group, Treatment-0 questionnaire version) commented: *"That was a funny attention check. I wish I could have answered as having done that."*

5.3 Discussion

Although we could not exclude that there were workers who engaged in satisficing, we did not uncover evidence of a pattern of misreporting that could be attributed to reputation, as measured by work history. The estimates by difference-in-means generally approached the expected 0% and 100% proportions. However, the Control group, which did not receive attention check items in their questions, was seemingly less consistent.

The differences-in-means between Treatment-1 and Treatment-0, both of which contained attention checks, were very close to the expected 100%, suggesting that the attention checks indeed mitigated the effect of satisficing.

We thus decided not to use reputation criteria to exclude partici-

pants in the main survey, as well as to add both the attention checks items. Inclusion of attention checks in both conditions of the main survey was the conservative design choice, as we had observed that their absence had, in this experiment, led to overestimation.

6. STUDY 3: MEASURING SNOOPING ATTACKS

6.1 Design

Having selected the list of items, and validated that a deployment to MTurk could provide good quality data, we proceeded to design and deploy the main survey.

We opted to create a very short questionnaire, with only the list question, and six other questions on personal characteristics, none of them open-ended. The questions are shown in Appendix B. The decision to not include more questions was made for two reasons. First, we had started with very concise research question, and broadening the scope before that question was answered could be a waste of time. Second, with more questions, or questions that were more probing, there was a risk that participants might feel that anonymity was reduced. For instance, they could reasonably suspect that their identity could be triangulated with responses to other surveys.

For that reason, we chose questions on personal characteristics carefully, for instance not including questions about level of income or race, which are very common in surveys, but that participants may feel to be very personal. We also asked for state of residency, but not city; and asked for level of education in broad categories.

Another design choice was the ordering of questions. We chose to show the list question at the beginning of the survey, to maximize attention and decrease incomplete responses. Since the question is cognitively heavy, it would be more frustrating to answer it after having cruised through simple demographics questions. We also inquired about personal characteristics in what we reasoned to be an increasing level of identifiability, to keep the sense of anonymity strong, as long as possible.

The list question included the control items and the item of interest selected in Study 1, and the two attention checks used as treatment manipulations in Study 2. The main purpose of including the attention checks was not to "catch" inattentive participants but to engage participants when thinking of the answer.

6.2 Fielding

We put the questionnaire online on a private web server, and configured it to randomly assign participants to either the treatment or the control group, each receiving the corresponding version of the list question. The survey proper was preceded by an informed consent form. We posted the survey several times as a task in MTurk, so that it would re-appear on the front page. Repeated participation was prevented by the custom qualification method [25]. MTurk qualifications were also used to restrict participation to residents in the United States. No other restrictions regarding past performance were enforced, as we found them to be superfluous in Study 2. Participants were paid $0.20, regardless of them giving valid responses. The survey took 1 to 2 minutes to complete attentively.

6.3 Data cleanup

We received a total of 1,481 responses to the survey. Of those, 84 (6%) were incomplete, and were removed from the dataset. Additionally, 16 responses (1%) were eliminated for being obviously invalid: 8 for responding “none” to the list question, and 8 for responding “all”. The following analysis is based on the remaining 1,381 responses.

Table 4: Summary of participant demographics, overall and by group, in the survey containing the list experiment question.

	Control (n_c = 688)	Treatment (n_t = 693)	Total (n = 1381)
By gender			
Female	43.2 %	42.3 %	42.7 %
Male	56.4 %	57.6 %	57 %
Other	0.4 %	0.1 %	0.3 %
By age group			
18-24	26 %	26 %	26 %
25-34	46.2 %	47.3 %	46.8 %
35-44	15.4 %	14.6 %	15 %
45-54	6.8 %	8.5 %	7.7 %
55-64	5.4 %	3 %	4.2 %
65 +	0.1 %	0.6 %	0.4 %
By level of education			
Less than high school	0.6 %	0.9 %	0.7 %
High school	28.3 %	27.4 %	27.9 %
Other college degree	18.8 %	19.9 %	19.3 %
Bachelor’s degree	41.4 %	39 %	40.2 %
Masters or PhD	9.6 %	11.4 %	10.5 %
Other	1.3 %	1.4 %	1.4 %
By region			
Midwest	23 %	21.1 %	22 %
Northeast	19.5 %	21.2 %	20.3 %
South	35.2 %	33.8 %	34.5 %
West	22.4 %	24 %	23.2 %
By ownership status			
Doesn’t own smartphone	12.4 %	10.1 %	11.2 %
Owns smartphone	87.6 %	89.9 %	88.8 %

Following Pew’s approach [39], we computed smartphone ownership status combining responses from two questions on ownership, SMART1 and SMART2. Whenever the response to the question “Is your cell phone, if you have one, a smartphone?” was “Not sure”, or “No, it is not a smartphone”, we referred to the next question, “Which of the following best describes the type of cell phone you have”, and assumed participants to be smartphone users if they selected either “iPhone”, “Android”, “Windows Phone” or “Blackberry”. There were 12 (1%) such cases.

Responses to the question about state of residency were binned into the 4 statistic regions defined by the US Census Bureau: Northeast, Midwest, South and West. For some of the analysis, ages were binned into commonly used age groups.

6.4 Dataset

6.4.1 Demographics

Table 4 summarizes the personal characteristics of the sample, segregated by control and treatment groups. A logistic regression of characteristics as predictors, and membership to either control or treatment group as outcome, did not reveal any significant differences between groups. Applying stepwise elimination of variables, starting with a model with AIC = 1926.1 and no significant predictors, the final model marginally improved AIC to 1916.45, with the elimination of all variables. In the final model, the remaining term was not a significant predictor ($Z = 0.135$, $p = 0.893$).

Therefore, as expected from randomized assignment, there was no evidence to suggest existence of a priori differences between the control and treatment groups, which would hurt the validity of the

Table 5: Number and proportion of respondents who selected each option in the list experiment item (adjusted for 4 control items).

	Control	Treatment
0	88 (12.8%)	76 (11%)
1	258 (37.5%)	204 (29.4%)
2	249 (36.2%)	239 (34.5%)
3	84 (12.2%)	122 (17.6%)
4	9 (1.3%)	43 (6.2%)
5	-	9 (1.3%)

prevalence estimates obtained through this list experiment. The demographics were similar across experimental groups, and any possible confounds could reasonably be expected to be equally distributed among them.

6.4.2 Attentive participation

We investigated if there were any indications that answers were inattentive. For that we looked at the relationship between how much time it took to answer the list question, and the actual response. If participants were rushing through the question, it would be expected that they had selected one of the first options, and hence that there would be a negative correlation between the time to complete the task and the number of behaviors that participants reported as having engaged in.

The correlations for either group were close to 0 (treatment: r = -0.0015 with 95% CI -0.0760 to 0.0730; control: r = 0.0185 with 95% CI: -0.0563 to 0.0931), and, for both, the hypothesis of the true correlation being 0 could not be excluded (treatment: t (691) = -0.402, p = 0.968; control: t (686) = 0.484, p = 0.6284). We therefore found no evidence that participants chose one of the first options that were available.

The possibility remains that participants chose an answer at random. Given the random assignment to groups, the noise created by responses at random should be equally distributed among groups, thus affecting the error, but not the difference-in-means.

6.4.3 Response to list experiment question

Table 5 shows the raw distribution of responses to the list experiment question for both groups. The vast majority of participants selected an answer between 1 and 3 (85.9% in the control group, 81.5% in the treatment group). Thus, the presence of appreciable ceiling or floor effects was unlikely.

We then investigated the possibility that the sensitive item changed how participants in the treatment group identified with the control items. For instance, participants could be more willing to identify with having called a 1-900 number because it appeared to be less censurable when compared to snooping. Blair and Imai [4] describe a statistical procedure to check for such an effect. Constructing the prescribed tabulation of estimated proportions of types of responses, we found no negative estimate. We therefore concluded that there wasn't evidence of a design effect.

Taking all this evidence together, we concluded that the design of the study and its deployment yielded a sound dataset.

6.5 Prevalence estimate

We defined (1-year) prevalence as the proportion of people in the population who internally identified as having had looked through someone else's cell phone without their permission. Prevalence was estimated by the difference-in-means between groups in a list experiment.

Table 6 summarizes the estimated 1-year prevalence for the sample and further breaks it down by segments of personal characteristics. For the overall sample (line 1), the 12-month estimate of prevalence was 31%. Our sample was not, however, a fair reflection of the U.S. population. Participants, on average, were younger, attained a higher level of education, and predominately identified as being male, which is expected in MTurk convenience samples [7]. We adjusted the data to the U.S. population estimates from the 2010 Census, and obtained an estimate of 20% for the U.S. adult population (see Table 7).

The data was adjusted with cell-based post-stratification weighting. We created weights for strata which, from the sample subset summaries, we found to have appreciably different prevalence estimates between levels. Using every possible demographics criteria to stratify would create cells with two few observations. Even the combination of gender, age group and region yielded marginal frequencies of 0. Moreover, using demographics criteria for which there weren't diverging differences between strata would have little impact on the overall prevalence estimate. We therefore decided to use weights based on the cross-tabulation of only age group and gender. At that granularity, the number of observations for some (AGE * GENDER) subsets was still too low to obtain reasonable weights. Recoding the 3 older age groups into one (45+), we were able to obtain more adequate weights, shown in Appendix C. As with any adjustment of this type, we obtained a more representative estimate, at the cost of increasing standard error. The national population statistics and diagnostics are shown in Table 7, and were computed with the R "survey" package, which implements Lumley's [24] weighted analysis instruments.

6.6 Trends

Although the overall 1-year estimates are informative by themselves, having a large sample allows us to look at differences between cohorts that can help explain the phenomenon. Table 6 suggests that in all demographic criteria, except for level of education, the estimates of prevalence are considerably different between subsets, but more detailed analysis is required to discern if demographic criteria can *predict* lower or higher prevalence.

It is, however, impractical and uninformative to try to understand the underlying demographics of snooping behavior based on all possible criteria. We therefore sought to find the demographic variables that better explained the list experiment outcomes, and only then to model the prevalence according to those variables.

6.6.1 Variable selection

To find relationships between demographic criteria and prevalence, we first constructed linear regression models of the number of items participants selected as a function of each available variable (gender, age, level of educations, region, and ownership), controlling for assignment to control or treatment group. Table 8 summarizes those models with the R-squared and F statistic, and shows comparisons to a smaller model in which the group assignment is the only predictor. Coefficients of each model are reproduced in Appendix D.

Regarding gender, for respondents who identified as being female, the prevalence estimate in the sample was 38%, whereas for the ones who identifies as male, it was 26% – a difference of more than 10 percentage points (Table 6, lines 2 and 3). However, the model with the both gender and experimental group as predictors, indicated that the gender variable explained very little of the vari-

Table 6: Estimated 1-year prevalence in the sample, as estimated by the difference in means between experimental groups. The table shows estimates for overall sample and for subsets based on personal characteristics. No estimations were made for subsets in which there were less than 20 observations in either experimental group, except for the age 65+ subset, which was binned with the 54-65 subset into the 55+ level. *P*-values from a t-test with the null hypothesis that there was no difference between experimental groups, with alpha set at 0.05. Bonferroni-adjusted significant differences in bold.

	Control group mean (SE)	Treatment group mean (SE)	Prevalence (SE)	*P*-value
Overall	2.517 (0.035)	2.825 (0.042)	30.8 % (0.055)	**<0.00001**
By gender				
Male	2.500 (0.046)	2.759 (0.057)	25.9 % (0.073)	**0.00043**
Female	2.542 (0.053)	2.918 (0.063)	37.6 % (0.083)	**0.00001**
By age group				
18-24	2.631 (0.067)	3.156 (0.086)	52.4 % (0.109)	**<0.00001**
25-34	2.522 (0.051)	2.820 (0.062)	29.8 % (0.080)	**0.00023**
35-44	2.509 (0.089)	2.644 (0.096)	13.4 % (0.131)	0.30730
45-54	2.362 (0.116)	2.407 (0.124)	4.5 % (0.169)	0.79038
55+	2.158 (0.158)	2.240 (0.202)	8.2 % (0.257)	0.75036
By level of education				
High school	2.482 (0.061)	2.789 (0.087)	30.7 % (0.106)	**0.00396**
Other college degree	2.667 (0.085)	2.949 (0.096)	28.3 % (0.129)	0.02889
Bachelor's degree	2.526 (0.054)	2.826 (0.067)	30.0 % (0.086)	**0.00053**
Masters or PhD	2.318 (0.110)	2.633 (0.105)	31.5 % (0.153)	0.04102
By region				
Midwest	2.494 (0.071)	2.699 (0.092)	20.5 % (0.117)	0.07989
Northeast	2.515 (0.078)	2.776 (0.093)	26.1 % (0.122)	0.03290
South	2.566 (0.060)	2.915 (0.072)	34.8 % (0.094)	**0.00024**
West	2.468 (0.073)	2.855 (0.086)	38.8 % (0.113)	**0.00067**
By ownership status				
Doesn't own smartphone	1.800 (0.093)	1.914 (0.093)	11.4 % (0.131)	0.38513
Owns smartphone	2.619 (0.036)	2.928 (0.044)	30.9 % (0.057)	**<0.00001**

Table 7: Proportion of U.S. adults who snooped on mobile phones in a 12 month period, as estimated by the difference in means between groups in a list experiment. Sample adjusted by cell-based post-stratification weighting to the 2010 Census by age and gender. *P*-value from a design-based t-test of the difference in means.

	Control group	Treatment group	Prevalence	*P*-value
Adjusted mean	2.41	2.61	20%	0.01515
SE	0.055	0.061	0.081	

Table 8: Linear regression models of number of items selected in the list experiment question. The first row indicates the proportion of variance explained by being in the treatment or control group. In the remaining rows, a variable is added to that model. F statistic from an ANOVA of the smaller and larger models.

Predictor variables	R^2	ΔR^2	*F*	D.f.	*P*-value
GROUP	0.022				
GROUP + GENDER	0.025	0.003	1.87	2	0.1542
GROUP + AGE	0.053	0.031	44.78	1	**<0.0001**
GROUP + EDUCATION	0.031	0.009	2.47	5	**0.0306**
GROUP + REGION	0.025	0.003	1.32	3	0.2671
GROUP + OWNER	0.100	0.077	118.38	1	**<0.0001**

ance in either group. This model did not significantly improve on the smaller model, with just the experimental group as predictor, explaining only an additional 0.003 of the variance (Table 8, line 2). Gender, therefore, did not seem to have strong relationship with snooping behavior, or at least not strong enough to justify including it in a model with other predictors.

Age (modelled as continuous variable, not by age group), on the contrary, significantly contributed to selecting more items. Looking at the details of the model, each additional 10 years predicted selecting, on average, less 0.18 items ($p < 0.0001$), in addition to the effect of group membership. Age, was therefore, considered a good candidate variable for a larger model.

The results of the model of level of education were mixed. Level of education can be thought of as an ordered variable, raising the question of whether more education could predict selecting a greater or lower number of items. Looking into the estimates of that regression, we found no clear evidence. Taking post-graduate education as a baseline, the model indicated that those with a college or Bachelor's degree selected a higher number of items (+ 0.33 with $p = 0.0016$, and + 0.20 with $p = 0.0347$, respectively), but there wasn't evidence of an effect for other levels of education. We expected to find that greater predicted difference in number of selected items would be associated with the greater differences in level of education, but that was not the case. Without an interpretation for that pattern, we concluded that this variable was not a good candidate for a larger model, despite the fact that adding it modestly improved the smaller model.

Region, like gender, did not seem to have a relationship with prevalence, on the basis that the model including it as a predictor did not significantly improve on the smaller model. We found it, therefore, to not be a good candidate.

Finally, regarding ownership status, the model suggested that those who owned smartphones selected more items from the list, even when controlling for membership in either control or treatment group. Adding ownership status to a model of only group membership explained 7.7% more of the variance, the greatest difference we found. Looking at the estimates of the model, we found the additional effect of owning a smartphone to be selecting 0.91 more items ($p < 0.0001$). Thus, ownership was clearly judged as candidate variable for a larger model.

6.6.2 Model

Having identified gender and smartphone ownership status as variables of interest, we finally aimed to understand how they predicted the probability of engaging in snooping attacks. For variable selection, we had used number of items selected, controlled by group membership, as an indicator of higher probability. For the final model, we wanted to look at actual predicted probability, while using both variables as predictors, and accounting for possible non-linear relationships.

Recently, it has been noted that although list experiments cannot reveal what each participant responded to the sensitive item, it is still possible to estimate conditional and joint proportions [10, 15], and thus model the joint probability distribution [4, 20]. Using the R "list" package [3] to that end, we created a model of the proportion of respondents identifying with the sensitive item, as a function of age and ownership status.

Appendix E.1 shows the coefficinet of that model, and Figure 1 depicts it graphically. It shows two clear trends:

- There is a sharp, concave decline in likelihood of snooping as people get older. Each additional year of age disproportionately decreases the likelihood of snooping on others.
- Those that own smartphones are more likely to engage in snooping. The difference is attenuated, and eventually disappears, as people get older.

The model also suggests that the youngest participants who are smartphone owners, are more likely to have snooped on others than to have abstained from it. Thus, for some groups, conducting snooping attacks, as we have defined them, may be the norm, not the exception.

6.7 Discussion

Summarizing, through a list experiment, we estimated the 1-year prevalence of successful snooping attacks to be 30.8% in an online sample. With post-stratification weighting, we generalised that finding to a national population, estimating that 20% of US adults had engaged in snooping in a 1-year period. Looking at specific subsets of the sample, some apparent trends emerged, but, due to the nature of list experiment data, comparisons between raw subsets can be misleading. Expanding our analysis, we did not find gender, level of education, or geographical sub-region to be strongly related to snooping behavior. We did however find that being young, and owning a smartphone, was independently linked to the likelihood of engaging in snooping. In the sample, those that did not own smartphones were, indeed, much less likely to have engaged in snooping attacks (11% 1-year prevalence), while those that were younger were more likely (52% 1-year prevalence in the 18-24 age group).

It should be noted, however, that being young and owning a smartphone is very much related: in the US, 85% of those between 18 and 29 own a smartphone, whereas for those that are 65 or older, the proportion is 27% [39]. In our sample, there is also a notable relationship between the two variables ($r_{\text{point-biserial}} = 0.28$). This fact suggests that there other variables, which we did not examine, relating to both age and ownership.

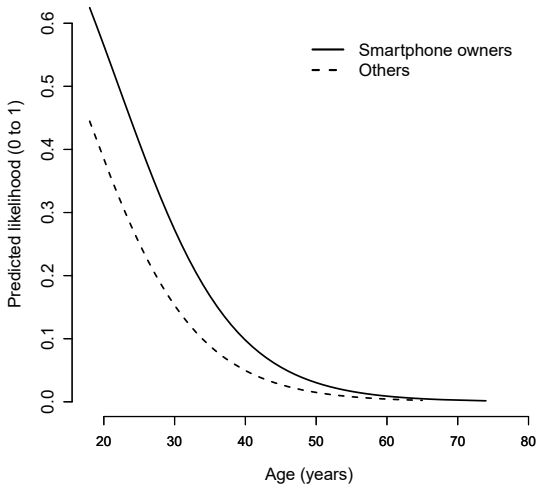


Figure 1: Predicted 1-year likelihood of having engaged in snooping attacks, by age and smartphone ownership status. Predictions from a list experiment regression model, shown in Appendix E.1.

7. STUDY 4: SNOOPING ATTACKS AND DEPTH OF ADOPTION

Being young and owning a smartphone, variables which the model suggests to be indicative of higher likelihood of engaging in snooping attacks, are also the typical characteristics of "digital natives." This population is known to be much more aware and concerned about threats within social context, such as snooping [23]. Where does that concern stem from? We hypothesize that those who use smartphones intensively as gateway to their social lives, thus producing privacy-sensitive information, become, by their own experiences, more aware of what they would have to gain, or loose, with a snooping attack. Thus, they would be more concerned about others snooping on their devices, and they would also be more likely to snoop on others.

In a final list experiment, we examined the likelihood of engaging in snooping attacks among smartphone users. Specifically, we explored how that likelihood is influenced by age, and by the degree to which people use their devices for personal purposes, in ways that may leave a trace of potentially privacy-sensitive data.

7.1 Procedure

We created a new online survey, similar to the one used in Study 3. The questions about gender, level of education, geographical region, and smartphone ownership were removed. The list experiment question, the question about age, and the question about the kind of smartphone the participant had were kept (the latter without the option "I do not have a cell phone").

An additional question group, shown in a second page, was added. This question group was a Likert scale of depth of adoption for

privacy-sensitive purposes, with 10 questions. For each, participants rated their perceived degree of frequency of use, from "Never" (1) to "All the time" (7). As an example, one item was "I use my smartphone to to look up information about health conditions". Items were based on behaviors of smartphone users that were reported in a Pew survey [39]. The scale is reproduced in Appendix F.

The survey was fielded in MTurk, following the same procedure as Study 3. The advertisement (HIT) asked specifically for smartphone users, both in the title ("Survey of smartphone users") and the description ("[...] Do not accept this HIT if you do not regularly use a smartphone"). Data cleanup was done also as described in Study 3, resulting in the exclusion of 7 responses (1%). All participants were paid $0.25.

There were 653 valid responses, 314 of which in the control group, and 339 in the treatment group. The majority of participants (56%) reported having an Android smartphone, followed by an iPhone (41%), Windows Phone (3%) and Blackberry (<1%). No participants selected the option "I do not have a smartphone", that was kept to exclude responses in case of inattentive reading of the advertisement.

7.2 Results

7.2.1 Depth of adoption and age

Responses to the depth of adoption scale, whose possible values are between 10 and 70, ranged from 16 to 70, and where somewhat skewed toward the higher end. The middle point of the scale is 40, and the mean response was 44.66 (SD = 10.6). Details about the distribution of responses, for the scale and individual questions, can be found in Appendix F.2.

Responses to the depth of adoption scale were, as expected, negatively correlated with age (r = -0.18, $t(651)$ = -4.78, $p < 0.00001$). This correlation, however, was not strong (according to Cohen's effect size criteria, it falls between small, 0.1, and medium, 0.3). Because depth of adoption, as it was measured, was relatively independent of age, it could more easily be interpreted as a predictor of likelihood of engaging in snooping attacks.

7.2.2 Depth of adoption as predictor

Using the same procedure as in Study 3, we created a model of likelihood of having engaged in a snooping attack, based on age and depth of adoption. The model predictions are depicted in Figure 2, and coefficients shown in Appendix E.2. In the left panel, the predictions are shown as a function of age, with a trend line representing a reduced model, with only age as predictor. In the right panel, the predictions are shown as a function of depth of adoption, with the corresponding reduced model line.

If there were noticeable differences in the pattern of dispersion in relation to the lines, such could be interpreted as one variable being a stronger predictor than the other (the stronger predictor should show less dispersion, or none at all). What is observable, however, is that neither the age or depth of adoption variables explain the other away.

The model with both variables has Log-likelihood of -868.786, which is higher than either the reduced models for age (-880.458) and depth of adoption (-873.834), indicating that it's a better fit. Predictions of both reduced models are strongly correlated to the ones of the larger model (age: r = 0.75, $t(651)$ = 28.6, $p < 0.00001$; depth of adoption: r = 0.71, $t(651)$ = 25.7, $p < 0.00001$). They are also correlated amongst themselves, as would be expected from the correlation of the variables, but no strongly (r = 0.14, $t(651)$ = 3.7, p = 0.00022). Again, these correlations indicate that neither variable explains the other away, and both contribute independently to the larger model.

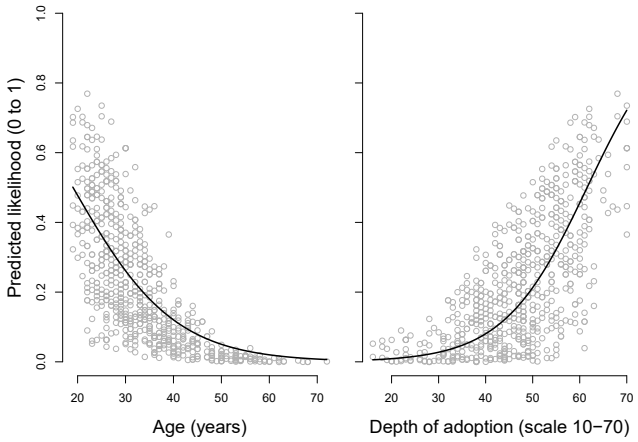


Figure 2: Predicted 1-year likelihood of having engaged in snooping attacks, by age (left panel) and depth of privacy-sensitive adoption (right panel). Dots represent per-participant predicted likelihood based on a list experiment regression model, with both age and depth of adoption as predictors. Trend lines represent the respective single predictor regression model. Regression coefficients are shown in Appendix E.2.

7.3 Discussion

We find evidence supporting the theory that people that use their smartphones in ways that may lead to privacy-sensitive information being kept, are more likely to snoop on others. Higher depth of adoption, as measured by a short scale we developed, predicts higher likelihood of identifying with the list experiment item indicating having "looked through someone else's cell phone without their permission" in the last 12 months, even when controlling for age.

However, depth of adoption does not explain away the effect of age that we had found in Study 3. Our scale, which was not thoroughly validated, may not have captured the factor it attempted to measure correctly. In fact, the scale does not accurately measure the frequency of certain behaviors, but how people *feel* about the frequency, which may be a weaker proxy for the construct of depth of privacy-sensitive adoption. Alternatively, there may also be, and we believe there are, other factors, related to age, which weren't measured but also play a role in predicting higher likelihood, like tech-savvy, or degree of volatility of social relationships.

8. CONCLUSIONS

8.1 Summary of findings

In this paper, we shown that the prevalence of snooping attacks on mobile devices is considerably higher than previously estimated. We found new evidence supporting that the problem is related to depth of adoption of mobile technology, and thus, that it is the youngest, those who use smartphone, and particularly those that use smartphones in ways that it stores privacy-sensitive data, that are more likely to snoop on others. In some segments of the population, people were more likely to "have gone through someone else's phone without permission", than not, in a period of one year.

To obtain these findings, we conducted a series of empirical studies. In the first two studies, we designed items for a list experiment, and validated the use of that methodological approach in MTurk. Our finding that list experiments in MTurk produce reliable data, as long as there are appropriate attention checks, is a secondary contribution of this work.

In the latter two studies, we conducted list experiments that inform on the prevalence of snooping attacks. Employing conservative design choices, that may have had the effect of underestimating prevalence, we were still able to estimate 1-year prevalence rates for the MTurk population, and, by weighting, for the U.S. adult population, that are much higher than previous lifetime prevalence indicators. Furthermore, we uncovered predictors of the likelihood of engaging in snooping attacks, and discerned independent population trends related to age and adoption of smartphones. We hypothesize that one mechanism for the observed trends is that users learn by their own experiences the kinds of valuable information kept on smartphones, which makes them more capable of engaging in snooping attacks.

8.2 Implications

This state-of-affairs can and should be addressed. There is room to improve privacy-preserving technologies that still impose too much effort on users, like mobile authentication. In recent year, biometric authentication on mobile devices, especially fingerprint authentication, has become more available and usable. There have also been extensive research efforts in making secret-based authentication more usable. Trends such as these indicate that defenses may be catching up.

However, two considerations should be given to the authentication approach of defense. First, as usable as authentication is made to be, it is not unreasonable to think that, for many people, it will never be attractive. Potential users of secret-based authentication may continue to think that it's a hassle. Potential users of biometric authentication may have privacy concerns. Defenses against snooping attacks for those people are few, if any.

A second consideration it that innovations in authentication should include snooping attacks in their threat models, because snooping attacks are likely to be attempted. Some adaptive authentication methods that have been proposed can reduce authentication requirements when devices are in "trusted places", like at home or at work (for instance, Android's Smart Lock [16]). It should now be clear that, in face of the pervasiveness of snooping attacks, that increase in usability will likely come at the cost of increased security risk.

Another possible road to improve the current situation is education and awareness-building. In that respect, however, it should be noted that in the realm of security, there has been little success in getting expert's messages across to users [19]. Specifically in the case of snooping, the reality if that many people are already aware of the risk, and want to secure against it, but fail to find practical ways to do it [28].

We hope this work plays a role in helping builders of interactive systems, educators, and policy-makers, to consider, when reasoning about mobile security, how prevalent it is for users' privacy to be violated by people they know.

8.3 Snooping as an attack

We have abstained throughout this paper from making judgements on whether snooping on others is justified. The use of the word *attack*, common in security lingo, should not be taken as having legal or moral connotations. It is an *attack* in the sense that actions were taken by an agent to circumvent an access policy; as much as one would call a *brute-force attack* to a situation where a mobile device owner who, upon forgetting their own PIN, ran a script to try out all possible combinations. We are aware that some people think it is acceptable for parents to go through their children's devices, or for romantic partners to go through one another's devices, and we do not dispute those opinions.

We note, however, that people who hold the opinion that their unauthorized access is acceptable, should also not be greatly impacted by social desirability bias. Thus, they should be expected to trend towards answering truthfully to a direct question on the topic. In the first study here reported (Section 4), and in previous studies [29], between 9% and 15% of respondents admit to having had snooped when asked directly. However, we found, for a comparable sample, that 2 to 3 times more people (~31%) self-identify with the behaviour when asked indirectly. The gap can be explained by participants themselves finding their actions censurable. We must conclude that a large portion of the population engages in a behavior that they know to be, from their own personal perspective, an attack, in the common sense of the word.

8.4 Future work

Security risks are often seen as being a function of the probability that they materialize and the severity of their consequences. This series of studies is informative as to the first factor, probability. We have, in this paper, focused on an overall measure of probability, and its relationship to demographic and usage factors. It would now be important to find other factors, especially ones related to the relationship between the attacker and the attacked (like social distance and motivation), and factors related to the context that creates the opportunity for the attack (like physical environment and circumstance). Both would be important for evaluating the effectiveness of new or existing defenses.

The other factor of which risk is a function is the severity of the consequences. We did not explore severity in this paper, but note that theory (e.g., [33]) predicts that the loss of control over what people that matter to us know about us, is likely to have considerable impact. We also note that one practical challenge in assessing severity is that people may not associate negative outcomes in their lives with someone having had snooped through their device, because, as our data suggests, they may never find out that it happened. Still, it is possible to gage how people *think* they would feel, or how they felt in the instances they know about, and find distinctions related, again, to context or social relationship between parties.

Both a fine-grained understanding of probability and of severity requires additional research, which we leave for future work. The quantitative approach we have employed here is not appropriate for a wide exploration of possible explanations, and possible outcomes, of snooping attacks. Finding factors requires breadth, and calls for a more qualitative approach. We believe that the fact that snooping attacks are much more common than previously thought justifies such an effort.

9. ACKNOWLEDGMENTS

This work was partially supported by FCT through funding of a PhD studentship, ref. SFRH/BD/98527/2013, and of the LaSIGE Research Unit, ref. UID/CEC/00408/2013. Special thanks to Serge Egelman, Kristy Milland, to several anonymous members of TurkerNation.com, and to the MTurk workers who participated in our surveys.

10. REFERENCES

[1] J. Antin and A. Shaw. Social desirability bias and self-reports of motivation. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, page 2925, New York, New York, USA, May 2012. ACM Press.

[2] N. Ben-Asher, N. Kirschnick, H. Sieger, J. Meyer, A. Ben-Oved, and S. Möller. On the need for different security methods on mobile phones. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services - MobileHCI '11*, page 465, New York, New York, USA, Aug. 2011. ACM Press.

[3] G. Blair and K. Imai. list: Statistical methods for the item count technique and list experiment. Available at The Comprehensive R Archive Network (CRAN) `http://CRAN.R-project.org/package=list`.

[4] G. Blair and K. Imai. Statistical Analysis of List Experiments. *Political Analysis*, 20(1):47–77, Jan. 2012.

[5] G. Blair, K. Imai, and Y.-Y. Zhou. Design and analysis of the randomized response technique. *Journal of the American Statistical Association*, 110(511):1304–1319, 2015.

[6] A. Braunstein, L. Granka, and J. Staddon. Indirect content privacy surveys. In *Proceedings of the Seventh Symposium on Usable Privacy and Security - SOUPS '11*, page 1. ACM Press, 2011.

[7] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1):3–5, Feb. 2011.

[8] J. Chandler, G. Paolacci, E. Peer, P. Mueller, and K. A. Ratliff. Using nonnaive participants can reduce effect sizes. *Psychological Science*, 26(7):1131–1139, 2015.

[9] E. Chin, A. P. Felt, V. Sekar, and D. Wagner. Measuring user confidence in smartphone security and privacy. *Proceedings of the Eighth Symposium on Usable Privacy and Security - SOUPS '12*, July 2012.

[10] D. Corstange. Sensitive Questions, Truthful Answers? Modeling the List Experiment with LISTIT. *Political Analysis*, 17(1):45–63, Feb. 2008.

[11] E. Coutts and B. Jann. Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). *Sociological Methods & Research*, 40(1):169–193, Feb. 2008.

[12] S. Egelman, S. Jain, R. S. Portnoff, K. Liao, S. Consolvo, and D. Wagner. Are You Ready to Lock? Understanding User Motivations for Smartphone Locking Behaviors . In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS '14*, pages 750–761, 2014.

[13] A. P. Felt, S. Egelman, and D. Wagner. I've got 99 problems, but vibration ain't one: A survey of smartphone users' concerns. In *Proceedings of the 2nd ACM Workshop on Security and Privacy in Smartphones and Mobile Devices*, pages 33–44, New York, New York, USA, Oct. 2012. ACM Press.

[14] U. Gadiraju, R. Kawase, S. Dietze, and G. Demartini. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pages 1631–1640, New York, New York, USA, 2015. ACM Press.

[15] A. N. Glynn. What Can We Learn with Statistical Truth Serum?: Design and Analysis of the List Experiment. *Public Opinion Quarterly*, 77(S1):159–172, Feb. 2013.

[16] Google. Google Smart Lock. Online, Retrieved Jan 19, 2016. `https://get.google.com/smartlock/`.

[17] M. Harbach, E. V. Zezschwitz, A. Fichtner, A. D. Luca, and M. Smith. It's a Hard Lock Life: A Field Study of Smartphone (Un) Locking Behavior and Risk Perception. In *Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 213–230, Menlo Park, CA, July 2014. USENIX Association.

[18] D. J. Hauser and N. Schwarz. Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48(1):400–407, 2016.

[19] C. Herley. More is not the answer. *IEEE Security & Privacy*, 12(1):14–19, Jan.-Feb. 2014.

[20] K. Imai. Multivariate Regression Analysis for the Item Count Technique. *Journal of the American Statistical Association*, 106(494):407–416, June 2011.

[21] L. C. Irani and M. S. Silberman. Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 611–620, New York, NY, USA, 2013. ACM.

[22] A. Kapelner and D. Chandler. Preventing Satisficing in Online Surveys : A " Kapcha " to Ensure Higher Quality Data. In *Proceedings of the 2010 CrowdConf*, 2010.

[23] S. Kurkovsky and E. Syta. Digital natives and mobile phones: A survey of practices and attitudes about privacy and security. In *International Symposium on Technology and Society, Proceedings*, pages 441–449. IEEE, June 2010.

[24] T. Lumley. Analysis of complex survey samples. *Journal of Statistical Software*, 9(8):1–19, 2004.

[25] S. Maldonado. Using mTurk Qualifications to prevent workers from participating in an experiment multiple times. Online, Retrieved Jan 19, 2016 `http://sgmaldonado.com/main/content/using-mturk-qualifications-prevent-workers-participating-experiment-multiple-times`.

[26] P. McDonald, M. Mohebbi, and B. Slatkin. Comparing Google Consumer Surveys to existing probability and non-probability based internet surveys. Google Whitepaper, Retrieved Jan 19, 2016. `https://www.google.com/insights/consumersurveys/static/consumer_surveys_whitepaper_v2.pdf`.

[27] S. McNeeley. Sensitive Issues in Surveys: Reducing Refusals While Increasing Reliability and Quality of Responses to Sensitive Survey Items. *Handbook of Survey Methodology for the Social Sciences*, pages 377–396, 2012.

[28] I. Muslukhov, Y. Boshmaf, C. Kuo, J. Lester, and K. Beznosov. Understanding users' requirements for data protection in smartphones. In *Proceedings - 2012 IEEE 28th International Conference on Data Engineering Workshops, ICDEW 2012*, pages 228–235. IEEE, Apr. 2012.

[29] I. Muslukhov, Y. Boshmaf, C. Kuo, J. Lester, and K. Beznosov. Know your enemy: the risk of unauthorized access in smartphones by insiders. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services - MobileHCI '13*, page 271, New York, New York, USA, Aug. 2013. ACM Press.

[30] D. C. Mutz. *Population-Based Survey Experiments*. Princeton University Press, 2011.

[31] H. Müller, A. Sedley, and E. Ferrall-Nunge. Survey Research in HCI. In J. S. Olson and W. A. Kellogg, editors, *Ways of Knowing in HCI*, pages 229–266. Springer New York, 2014.

[32] D. M. Oppenheimer, T. Meyvis, and N. Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4):867–872, 2009.

[33] L. Palen and P. Dourish. Unpacking "privacy" for a networked world. In *Proceedings of the conference on Human factors in computing systems - CHI '03*, number 5, page 129, New York, New York, USA, 2003. ACM Press.

[34] G. Paolacci and J. Chandler. Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, 23(3):184–188, 2014.

[35] E. Peer, J. Vosgerau, and A. Acquisti. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, 46(4):1023–1031, Dec. 2014.

[36] Pew Research Center. Privacy and Data Management on Mobile Devices. Report. Retrieved Jan 19, 2016 `http://www.pewinternet.org/2012/09/05/privacy-and-data-management-on-mobile-devices/`, 2012.

[37] Pew Research Center. Anonymity, Privacy, and Security Online. Report. Retrieved Jan 19, 2016 `http://pewinternet.org/Reports/2013/Anonymity-online.aspx`, 2013.

[38] Pew Research Center. The Future of Privacy. Report. Retrieved Jan 19, 2016 `http://www.pewinternet.org/2014/12/18/future-of-privacy/`, 2014.

[39] Pew Research Center. The Smartphone Difference. Report. Retrieved Jan 19, 2016 `http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/`, 2015.

[40] B. Rosenfeld, K. Imai, and J. N. Shapiro. An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions. *American Journal of Political Science*, 2015.

[41] R. Tourangeau and T. Yan. Sensitive questions in surveys. *Psychological Bulletin*, 133(5):859–883, Sept. 2007.

[42] M. W. Wiederman. The truth must be in here somewhere: Examining the gender discrepancy in self-reported lifetime number of sex partners. *The Journal of Sex Research*, 34(4):375–386, 1997.

APPENDIX

A. REPRODUCTION MATERIALS

Minimized datasets and R analysis code can be found at `https://github.com/diogomarques/snooping-paper`.

B. SURVEY QUESTIONS

List of questions in online survey reported in Section 6 (Study 3).

The first is a list experiment question, here shown in the version distributed to participants in the treatment group. Participants in the control group received the same question without sensitive item, in bold. The second and sixth items are attention checks.

LIST EXPERIMENT Below is a list of experiences you might have had in the past 12 months. To preserve your anonymity, we ask you only to indicate HOW MANY, not WHICH ONES, apply to you.

- In the past 12 months, I've shared pictures from my cell phone on social media.
- In the past 12 months, I've opened my eyes in the morning at least once (for instance, after waking up).
- In the past 12 months, I've purposefully made phone calls or sent text messages to 1-900 numbers.
- In the past 12 months, I've received at least one text message with unsolicited advertising (spam) on my cell phone.
- **In the past 12 months, I've looked through someone else's cell phone without their permission.**
- In the past 12 months, I've been to space, aboard and interplanetary vessel that I built myself.
- In the past 12 months, I've lost or had my cell phone stolen from me.

Please count how many you have had and indicate below.

◯ 0 (None) ◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6 ◯ 7 (All)

AGE How old are you (years)?

␣␣␣

GENDER What is your gender?

◯ Male ◯ Female ◯ Other

EDUCATION What is your highest level completed education?

◯ Less than High School
◯ High School
◯ Community College or Professional School (College degree)
◯ University (Bachelor's)
◯ Graduate School (Master or PhD)
◯ Other: _____

STATE In which state do you reside?

◯ Alabama ◯ Alaska ◯ Arizona ◯ Arkansas [...]

SMART1 Some cell phones are called "smartphones" because of certain features they have. Is your cell phone, if you have one, a smartphone?

◯ Yes, it is a smartphone.
◯ No, it is not a smartphone.
◯ Not sure if it is a smartphone or not.
◯ I do not have a cell phone.

SMART2 Which of the following best describes the type of cell phone you have, if you have one?

◯ iPhone
◯ Android
◯ Windows Phone
◯ Blackberry
◯ Something else
◯ I do not have a cell phone

C. WEIGHTS

Weights used in post-stratification adjustment, based on the difference between Study 3's (Section 6) sample and the U.S. adult population, as measured by the 2010 Census.

Weights reveal that the sample was younger and had a greater proportion of males than the general population.

Gender	Age group	Proportion of US population	Proportion of respondents	Weight
Female	18-24	6.4%	10.4%	0.6162
Female	25-34	8.7%	19.0%	0.4596
Female	35-44	8.8%	6.5%	1.3459
Female	45+	27.6%	7.0%	3.9534
Male	18-24	6.7%	15.6%	0.4276
Male	25-34	8.8%	27.7%	0.3171
Male	35-44	8.7%	8.5%	1.0254
Male	45+	24.3%	5.3%	4.5923

D. VARIABLE SELECTION MODELS

Coefficients of linear regression models of number of items selected in the list experiment question, in Study 3 (Section 6). Models used for identifying candidate predictors of likelihood of having had engaged in snooping attacks.

The first model has a single predictor: assignment to either treatment or control group.

The remaining models add each of the other variables (gender, age, level of education, region, and smartphone ownership), controlling for assignment to control or treatment group.

Differences between models reported in Table 8.

Variables	Estimate	SE	*t*	*p*
Intercept	2.51744	0.03885	64.806	<0.00001
GROUP	0.30795	0.05484	5.616	<0.00001
RSE(1379) = 1.109; R^2 = 0.02236				

Variables	Estimate	SE	*t*	*p*
Intercept	2.57587	0.04999	51.531	<0.00001
GROUP	0.30797	0.05483	5.617	<0.00001
GENDER: Male	-0.10050	0.05545	-1.812	0.0702
GENDER: Other	-0.40287	0.51104	-0.788	0.4306
RSE(1377) = 1.018; R^2 = 0.02501				

Variables	Estimate	SE	*t*	*p*
Intercept	3.08289	0.09275	33.24	<0.00001
GROUP	0.30305	0.05399	5.613	<0.00001
AGE	-0.01784	0.00267	-6.692	<0.00001
RSE(1378) = 1.003; R^2 = 0.05313				

Variables	Estimate	SE	*t*	*p*
Intercept	2.32081	0.08951	25.929	<0.00001
GROUP	0.30991	0.05474	5.662	<0.00001
EDU.: Bachelor's	0.20050	0.09483	2.114	0.0347
EDU.: Some coll.	0.33175	0.10484	3.164	0.0016
EDU.: H. School	0.16002	0.09906	1.615	0.1065
EDU.: Less H.S.	-0.00675	0.33226	-0.02	0.9838
EDU.: Other	0.46345	0.24794	1.869	0.0618
RSE(1374) = 1.016; R^2 = 0.03108				

Variables	Estimate	SE	*t*	*p*
Intercept	2.44412	0.06408	38.14	<0.00001
GROUP	0.30814	0.05485	5.618	<0.00001
REGION: NE	0.08432	0.545	0.586	
REGION: S	0.1418	0.07478	1.896	0.0582
REGION: W	0.06479	0.0816	0.794	0.4274
RSE(1376) = 1.019; R^2 = 0.02516				

Variables	Estimate	SE	*t*	*p*
Intercept	1.72178	0.08209	20.975	<0.00001
GROUP	0.2875	0.05268	5.458	<0.00001
OWNER: Yes	0.90782	0.08344	10.88	<0.00001
RSE(1378) = 0.9781; R^2 = 0.0997				

E. LIST EXPERIMENT REGRESSIONS

E.1 By age and ownership status

Coefficients from a list experiment regression model where the sensitive item is whether someone "looked through someone else's cell phone without their permission" in the last 12 months. Data from Study 3 (Section 6).

Regression using Maximum Likelihood (ML) estimation with the Expectation-Maximization algorithm [4]. Control group parameters not constrained to be equal.

Variables	Sensitive item		Control items $h_0(y; x, \psi_0)$		Control items $h_1(y; x, \psi_1)$	
	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	2.014	1.714	-1.167	0.194	-3.529	4.567
Age	-0.124	0.057	-0.002	0.004	-0.024	0.018
Owner	0.732	0.953	0.832	0.122	3.824	4.542

E.2 By age, depth of adoption and both

Coefficients from list experiment regression models where the sensitive item is whether someone "looked through someone else's cell phone without their permission" in the last 12 months. Data from Study 4 (Section 7).

Regression using Maximum Likelihood (ML) estimation with the Expectation-Maximization algorithm [4].

Variables	Sensitive item		Control items	
	Estimate	SE	Estimate	SE
Intercept	1.80821	1.48669	-0.17064	0.17876
Age	-0.09492	0.05080	-0.00728	0.00474

Variables	Sensitive item		Control items	
	Estimate	SE	Estimate	SE
Intercept	-6.95467	4.36000	-1.00872	0.21807
Depth adop.	0.11296	0.07714	0.01315	0.00446

Variables	Sensitive item		Control items	
	Estimate	SE	Estimate	SE
Intercept	-1.48857	4.23927	-0.88936	0.37492
Age	-0.11248	0.06047	-0.00457	0.00536
Depth adop.	0.07617	0.06999	0.01360	0.00505

F. PRIVACY-SENSITIVE ADOPTION

F.1 Scale

Scale used in Study 4. Each item indicates the perceived frequency of a type of smartphone use that can leave potentially sensitive information on the device. It attempts to measure, in a range from 7 to 70, the depth of privacy-sensitive adoption of smartphones.

PROMPT Here are some statements about smartphone usage for personal purposes.

Please answer on a scale from 1 to 7, where a 1 means that the statement indicates something you *feel like* you never do, and a 7 means that the statement indicates something you *feel like* you do all the time.

You can also use the values in-between to indicate where you fall on the scale.

[RANDOMIZE]

Item-1 I use my smartphone to check my personal email account.

Item-2 I use my smartphone to take pictures of myself or of people close to me.

Item-3 I use my smartphone to go on social networks (like Facebook, Twitter, Snapchat) with my personal account.

Item-4 I use my smartphone to exchange instant messages with people that are close to me.

Item-5 I use my smartphone to to look up information about health conditions.

Item-6 I use my smartphone to do online banking on my personal accounts.

Item-7 I use my smartphone to look up jobs or submit job applications.

Item-8 I use my smartphone to look up government services or information.

Item-9 I use my smartphone to look up directions to places, or to get turn-by-turn navigation.

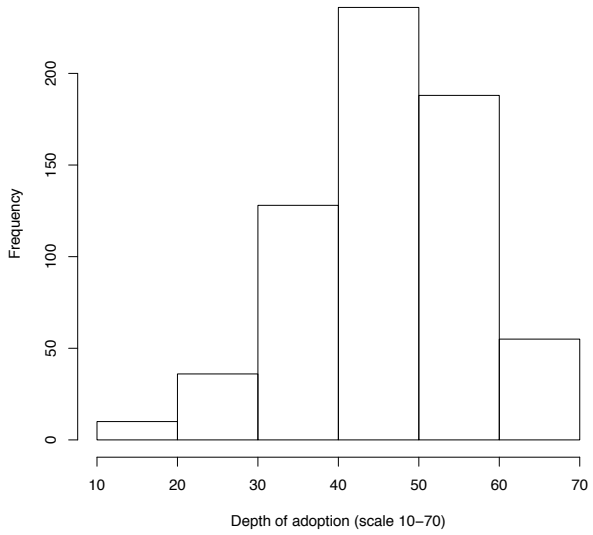
Item-10 I use my smartphone to organize personal affairs (for instance, access personal notes, calendar or shopping list).

F.2 Responses

Distribution of responses to scale and individual items in Study 4 (Section 7).

F.2.1 Scale

Sum of ratings to individual items.



F.2.2 Items

Frequency of response to scale items, each rated 1 (Never) to 7 (All the time).