Online Privacy of Personal Information - Perceptions v Reality

Diane Gan(✉) and Dennis Ivory

School of Computing and Mathematical Sciences, University of Greenwich, London, UK
{D.Gan,D.A.Ivory}@gre.ac.uk

Abstract. This empirical study investigates how (n = 252) users of online social networking sites perceived their online privacy and compares this to what can be collected by someone who has no connection to them in cyber-space. A survey was undertaken to determine each participant's perceived privacy awareness of their online personal information at different levels of distance from them, such as by a friend, friend of a friend or a complete stranger. Experiments were performed for each participant to retrieve as much personal information as possible using OSINT (open source intelligence) tools. For the majority of participants their personal information was collected in under two minutes by someone who had no connection with them in cyber-space. The results that were predicted by the participants was compared to what was actually found and are shown to support our hypothesis that the majority over-exaggerated how secure their personal information was.

Keywords: Online social networks · Online privacy · Data harvesting · Facebook · Twitter · Osint

1 Introduction

The volume of personal information published on online social networking sites (OSNs) increases each year (from 2.46 bn in 2019 to 4.2 bn in 2021 [1]) which provides an ever-growing source of information for criminals to search and abuse. Due to the huge numbers of profiles there will inevitably be many accounts with poor privacy settings from which information can be easily harvested [2]. Important personal details, publicly available online for anyone to view, can help criminals to launch successful targeted social engineering attacks [3]. The threat becomes greater the more information an individual has published about themselves on OSNs, often when they believe their information to be private.

Online privacy is defined as "he level of privacy protection an individual has while connected to the Internet. It covers the amount of online security available for personal and financial data, communications, and preferences." [4]. Personal data can include a person's full name, the names of family and friends, or more

W. Shi et al. (Eds.): SPNCE 2021, LNICST 423, pp. 346–368, 2022.
https://doi.org/10.1007/978-3-030-96791-8_26

personal data such as their date of birth or home address. Due to the fraudulent activity that is possible using this information it is vitally important that it is kept secure and that it is not viewable to anyone with a browser or at least only viewable by people with whom you choose to share this information with. Sometimes personal and sensitive information is revealed due to secondary leakage (covered in more detail in Sect. 2.2). This occurs when a person's information is leaked by a friend, a friend of a friend or a family member and often by someone who has been "riended" on the same platform who re-tweets or links a post to their own OSN [5].

Why is it important to keep your personal data private? There have been a number of celebrity users who have had their Twitter and facebook accounts hacked over the years. In 2008 during the US presidential elections Sarah Palin's private email was hacked. The email password was reset when hackers answered the three security questions, which were Mrs. Palin's date of birth, her home postal code and the location where she met her spouse. The answers to all these questions were found on her Facebook page, posted by Palin herself [6]. The founder of Facebook, Mark Zuckerburg, had his LinkedIn password stolen in 2016 and this was used to access his Pinterest and Twitter accounts as he had reused his password on these sites [7]. This demonstrates how the security settings of one user can affect the privacy settings of others when users link their OSNs together using "likes" or by sharing pictures or tweets. Users will continue to reuse their data such as name, picture, friends, etc. on multiple sites and share pictures and videos across those sites. It also demonstrates how complex it can be for a user to maintain their privacy and how aware they must be regarding their own privacy [8].

The focus of this work was to determine how individuals from a student population, perceived their own online privacy and compare this to the freely available information about them that can be found by anyone with a computer and an Internet connection. A survey was used to determine the personal information that individuals had shared knowingly and what they perceived as being private. Experiments were carried out to determine how accurate their perceptions were, by harvesting as much personal information as possible on each participant using a manual search.

An additional question posed is - can automated tools find more information than the manual search for a smaller subset of the participants? To answer this question, we carried out searches using automated tools on two smaller subsets, called the "minimal sharers" and "over sharers" (the bottom and top sharers from the survey respectively), see Table 5.

The process of harvesting each individual's information was timed and the results are compared to those obtained during the manual searches. The results of these were then compared to the individual participants' responses to the survey and also with the overall survey responses. We also compared these results to the whole survey when looking at the demographics for these subsets of participants. It was interesting to note that different information was located by the manual tools than by the automated tools.

The remainder of this paper is organised as follows. Section 2 discusses related literature. The methodology is described in Sect. 3. The results of the survey are presented in Sect. 4 and Sect. 5 analyses the experimental results. There is a discussion of the results in Sect. 6 and in Sect. 7.

2 Related Literature

2.1 Online Social Network (OSN) Investigations

People reveal information about themselves and their family through careless posts on sites such as Twitter and Instagram because they are unaware that it can be seen by anyone performing a simple search. Examples include posting pictures of new debit/credit cards on Twitter, both front and back, with the card number, security code and expiration date clearly visible [9], posting their UK driving license on Twitter [10] and birth certificates on Instagram displaying all the person's details [11]. These individuals have revealed significant personal information about themselves including their full name, place of birth, photo of the holder, signature and current address, none of which should be shared publicly for obvious reasons.

Ali et al. (2018) [12] identify Online Social Networks (OSNs) as a social graph. They also emphasise how OSNs leak a user's personal information when they publish information, upload photographs and videos which contain metadata and GPS coordinates. All this data has attracted the attention of attackers who harvest personal data during targeted attacks. They undertook a survey of undergraduate students to determine if the users were concerned about their online privacy. They found that many of the users had no idea about the privacy settings available to them. They also determined that nearly half of their participants used their real names and their pictures [12]. This differs from our research, as they have only used a survey, while our work goes on to verify the credibility of the users' perceptions of their OSN security.

Rathore and Tripathy (2020) [37] conducted a survey to determine the OSN users' awareness of their online security, which had 374 participants. The majority of these were male and aged 18–40. The survey determined that these users were worried about their privacy but did not actively modify the security available on the OSNs they used. They proposed a framework to increase security on OSNs. They found similar results to our work with their survey, except our participants were more evenly split between males and females and we had a larger age range [37].

A five year study of privacy perceptions undertaken by Tsay-Vogel, Shanahan and Signorielli (2018) [14] identified two types of users, those who were light users and those who were heavy users of social networks. At the beginning of the study the light users were more risk aware of the security issues, but as the study progressed, the heavy users became as aware as the previous group [14]. Our work also identified two similar groups of users, but our study differs in that we looked at a snap shot of the users.

Another study investigated the perceptions of over 200 non-student UK Facebook users, to measure their attitude to risk and their online behaviours though the use of an online survey. They also searched online for "16 hazards" related to the users' risk perceptions directly linked to their security and privacy settings, to determine how this modified the users' behaviour. They constructed risk profiles for each participant and concluded that the perceived risk claimed by them influenced how precautionary their online behaviour was [2]. In our study we have used a similar sized sample (n = 252) from a cross school student population and tested their security perceptions.

Wisniewski et al. (2017) [15] undertook an empirical analysis of the privacy awareness of Facebook users using an online survey based on 314 responses. This sample included a high percentage of undergraduate students, with a high proportion being female. They determined that these users had very varied perceptions of their privacy and online security. The main focus here was to determine "privacy behaviour and feature awareness" [15] by linking them using a Structural Equation Model (SEM). They determined that participants with the highest privacy settings tended to be very computer literate, while "privacy minimalists" covered the whole spectrum of users, showing that there is no simple linear relationship between these features. This paper has some similarities to our work in that we also identified different groups of users and in particular we also found a group of "privacy minimalists" that we identified as the "minimal sharers".

The majority of social network sites use an opt-out policy, which puts the onus on the user to check their own privacy settings, with many being unaware of the security implications of this [5]. Data leakage from OSNs often occurs due to the user's lack of knowledge and understanding of the privacy issues.

2.2 Secondary Leakage

Secondary leakage occurs when a user shares photos or links from friends without realising that they are actually making their friends more vulnerable to privacy breaches. Gan and Jenkins (2015) [5] identified that secondary leakage can be a serious problem which reveals a lot of information about a user to the point that their pattern of life can be determined.

Cascavilla et al. (2018) [16] demonstrated information leakage using an OSSINT (Open Source Social Network Intelligence) prototype tool that they developed, using 20 user profiles which had high privacy settings. This is an example of the use of an automated search tool which retrieves information that the user has designated as private. The tool uses the "victim's" friends list to retrieve information on all the friends, including friends of friends which are at a two hop distance. This information was then used to build a "friendship graph". They also used this to gain access to further private information such as place of work, education and their locations. A confusion matrix was used to measure the success of the OSSINT tool [16]. This work differs from our work in that we used a manual search method to retrieve personal and private information from users.

Privacy leakage was examined by looking at tweets which indicated "happy life events" [38]. They concluded that it is very easy to give away location information about yourself and the person you have tagged. They also emphasise the danger of tweeting details if the user is away from home, which can be used by burglars.

As parents become more aware of the privacy issues surrounding OSNs they also are concerned for their children's online privacy. The UK government warns about "the over-sharing of personal information" in their advice on "Child Safety Online: A practical guide for parents and carers whose children are using social media" [17].

2.3 Facebook Privacy Concerns

The Facebook and the political consultancy Cambridge Analytica incident (2017) [18] underlined how insecure our personal information can be online. Facebook estimated that approximately 305,000 people were affected because they installed the app "This is your digital life", which enabled Cambridge Analytica to harvest their personal data. What users did not realise was that Cambridge Analytica were also collecting additional information about their Facebook friends [18]. It is reported that "Facebook's 'platform policy' allowed the collection of friends' data to improve the user experience in the app but barred it being sold on or used for advertising." [19]. It was subsequently revealed that the data collected, included "emails, invoices, contracts and bank transfers" from over 50 million profiles of registered US voters, had in fact been sold on to third parties. Facebook then announced that any apps unused for three months will have their access to personal data restricted and future apps will also be restricted ("just their name, email address and profile photo") unless the user specifically grants permission by signing a contract [20].

Within Facebook, apps will no longer be able to "access certain information like religious or political views, relationship status and details, friend lists, education, and work history" [21]. These changes had a detrimental effect on some existing apps as demonstrated by the dating app Tinder, which reported that the changes that Facebook made to their policies had a knock-on effect for users of their app, as many had linked it to their Facebook profile [20].

2.4 Multimedia Privacy Concerns

The uploading and sharing of photos and videos are an integral part of OSNs. As the volume of multimedia uploaded daily to OSNs has increased, so have the risks to users, with Facebook having the most, with over 4.2 billion active users world wide, as of February 2021 [1]. Also there are approximately 995 images uploaded to Instagram per second [35], so it is not surprising that hackers utilise this huge resource for malicious purposes, such as hiding malware, spamming and identity theft.

Malware can be embedded within multimedia files and widely distributed through OSNs. GPS data, also called geotagging, can reveal the owner's location

as this information is embedded in the meta-data of uploaded images, videos, posts and tweets and can be used to exploit personal information in order to determine a participant's pattern of life for the purposes of stalking [5]. This has become less effective after Twitter changed the default privacy setting for geotagging on user accounts to "off". It now requires users to actively turn this on, but many still do this. There are still some sites that have geolocation turned on by default including Flickr, which embeds the longitude and latitude of the location where an image was taken into the EXIF metadata of the image [22]. This information is easily exploited, enabling anyone to pinpoint the location where a photograph was taken or a post was uploaded.

2.5 Summary

The literature identifies that there are still security issues around OSNs, which demonstrates the importance of highlighting these issues. The main concerns are the user's lack of awareness of their privacy settings and that a great many rely on the default settings. Another concern is that many users have accounts on multiple platforms, which are often linked together for ease of use, including the same personal details (name, picture, password), which makes it easier to collect data on them in cyber space. This facilitates secondary leakage between platforms when "friends" post pictures, "likes" or messages which circumvent privacy settings, regardless of the user's intended security settings. Locations are revealed through casual multimedia uploads which have GPS coordinates embedded in them, which the user may or may not be aware of and which hackers can utilise for malicious purposes.

3 Methodology

3.1 General Methodology

A survey was used to gather data from volunteer participants (for which they gave their explicit permission and were offered feedback on their online privacy). This data was then used to determine how they perceived their online privacy and to compare this to what could actually be harvested online.

The survey was sent out to all the students studying at the University of Greenwich Maritime campus and (n = 252) individuals responded. These included a mix of participants on Computer Science, Business, Humanities, History and Law degrees (undergraduate to postgraduate). The participants were asked to provide their email address, full name, their age range, what degree they were on and what year of study they were in, to determine the demographics of the participant group. They were also asked to identify which OSNs they regularly used. The survey asked them to identify what personal information they had knowingly published online, what they thought was freely available for anyone to find and view, and what personal information they thought was private. They were also asked what information they thought could be seen by someone

they were friends with on an OSN. To determine their awareness of secondary leakage, we also asked them to identify what information about themselves they thought could be found through a friend of a friend, i.e. someone that they personally had no direct connection with. There were nine options possible for each of these questions (full name, date of birth (DoB), exact home address, a picture of themselves, their email address, their phone number, names of a close family member, actual places visited in the last few days and their current (physical) location). The last two items enable anyone to determine places that they regularly visit, such as their workplace or gym and potentially where they live, if they use their OSNs from these locations [5]. There was also a free text box for anything else not covered in the list.

There are many tools available from the Internet used by companies to mine information from social networks sites for marketing purposes to identify a target audience for advertising. These applications are designed to work with Twitter as their main source of information often using the geotags present but will also search for information on a specific account from publicly visible information [5]. The functionality ranges from merely displaying someone's tweets to being able to view posted pictures from a particular locale or time period.

3.2 Manual Search Methodology

Only the participants' first and last names were used as the starting point for the searches using a search engine. In the majority of cases this returned result that included the various social networking sites that the participants were using. From these social networking accounts their personal information could be gathered, such as personal pictures and the names of family members by searching the friends lists in the accounts. Other information found was used elsewhere, such as the BT Phonebook [36](Other countries have their own equivalent resource) and FindMyPast [32]. Sites like this then enabled more personal data to be gathered, including information such as birth certificate records, current address and phone number, the manual method is about finding a piece of information from one source and then using this to narrow down a search for other information from another source.

3.3 Automated Search Methodology

The top ten "over sharers" and the bottom ten "minimal sharers" were investigated further using fully automated 'people search engine tools. These are readily available OSINT tools that automatically search through a variety of different sources (Social networking sites, Census/Birth/Marriage/Military/Death records and other online accounts) normally used for genuine purposes such as making contact with someone you didn't get contact details for or finding lost family or family from other countries amongst other uses.

The "minimal sharers" were of particular interest due to the lack of information found about these participants using the manual method (discussed in Sect. 3.2). We needed to determine if the automated tools would identify more

of their personal information. Some of these searches found information which was hidden behind a 'pay wall' requiring payment to access the information. Examples of this are websites that hold the census data, birth, death and marriage records (See Table 1). The automated tools that were used to search UK sources were pipl, 192, peopletraceuk, social-searcher and findmypast [28–32]. These tools are designed to take a small piece of information (normally a persons name) and collate as much information about this person. It normally does this in such broad stroke however that you will need to sift through a large number of false entries as there is generally more that one person in the world with the same name.

4 Results of the Survey

The survey asked the participants to give details of themselves and their usage of OSNs. 252 participants completed the survey and although all were students they were representative of a quite diverse population with age ranges from 18 to 60 (See Table 1) and on a wide variety of different degrees, although the majority are in the age range 18–25.

4.1 Demographics of Survey Participants

The majority of participants (91.6%) were undergraduate students (See Table 1). The group 20 to 30 was the largest with 164 participants (65%). These have been broken into smaller ranges in Table 1 for discussion. The largest number of respondents fell within the age ranges 18 to 23 (193, 76.6%), but there were also five people in the age range 51 to 60. The reason these age brackets were chosen was that it was anticipated that there would be a much larger number of respondents between the age of 18 and 25 as the participants in this study are university students, and due to this, smaller age brackets at the lower end of the age range were used to increase fidelity. Females were the majority of respondents being 58.4% of the total and 41.2% being male with one person not specifying their gender (See Table 1).

Participants were asked to identify their year of study. The largest number (36.9%) were in their first year of study and the smallest group (8.7%) were post-graduate students ranging from masters students to PhD students. This also included a small number of post-doc researchers. There were 6 people who were working (full-time and part-time) while studying and these are included in the Master's Degree, PhD student and Postdoc category. Only 42% of survey participants were on computer orientated degrees (BSc Computer Science, BSc Computer Security and Forensics, BSc Computer Systems and Networking and BSc Business Computing). This shows that the results are not biased by "computer savvy" participants studying computer science related degrees.

Table 1. Survey participants' demographics

	Category	Number	Percentage
Gender	Male	104	41.2%
	Female	147	58.4%
	Other	1	0.4%
Current status	1st year	93	36.9%
	2nd year	57	22.6%
	3rd year	81	32.1%
	Masters, PhD, Postdoc	21	8.4%
Age range	18–19	66	26%
	20–21	70	28%
	22–23	57	23%
	24–25	20	8%
	26–30	17	7%
	31–40	14	6%
	41–50	3	1%
	51–60	5	2%

4.2 Participants' Perceptions

The participants were asked to identify which social networks they regularly used. The majority (86.5%) used Facebook, 69.4% used Instagram, 48% used Twitter, 18.7% used Tumblr and 20.2% used other social networks which were not listed in the survey. The majority also used multiple OSNs. However, eleven participants did not use any OSNs at all, although personally identifiable information was still found on six of them during the searches.

Published Online. The participants were asked to identify all information that they had voluntarily published online (See Table 2). For full name, DoB, email address and current location our results approximately matched what they believed could be found by anyone. The exception was their exact home address and the phone number. The majority (88%) had published their full name, their date of birth (68%), their email address (64%) and a picture of themselves (88%) on at least one OSN.

Information Found by a Friend. When asked what information they thought could be found in this section, the expectation was that more information could be found by someone that they were "friends" with compared to a complete stranger. The number expecting that their DoB could be found rose from 61%, to 84% their picture was up from 88% to 94%, phone number was up from 32% to 62% and family members was up from 54% to 80%. This demonstrates that

Table 2. Information voluntarily published online

Category	Number of participants	Percentage
Full name	222	88.10%
Picture of yourself	222	88.10%
Date of birth	171	67.90%
Email address	160	63.50%
Visited locations	102	40.50%
Family members	78	31.00%
Current location	70	27.80%
Phone number	47	18.70%
Exact home address	25	9.90%

they were aware that there was a greater chance of someone obtaining more of their personal information if they were friended on a OSN (See Fig. 1).

Information Found by a Stranger. Participants were asked what information they thought could be found online by someone that they had no connection with and the results are shown in Fig. 1. Almost everyone supposed that their full name (90%) and a picture of themselves (88%) could be found by a stranger. However, the majority (70%) did not think that a stranger could locate their home address, find their phone number (68%) or their current location (70%).

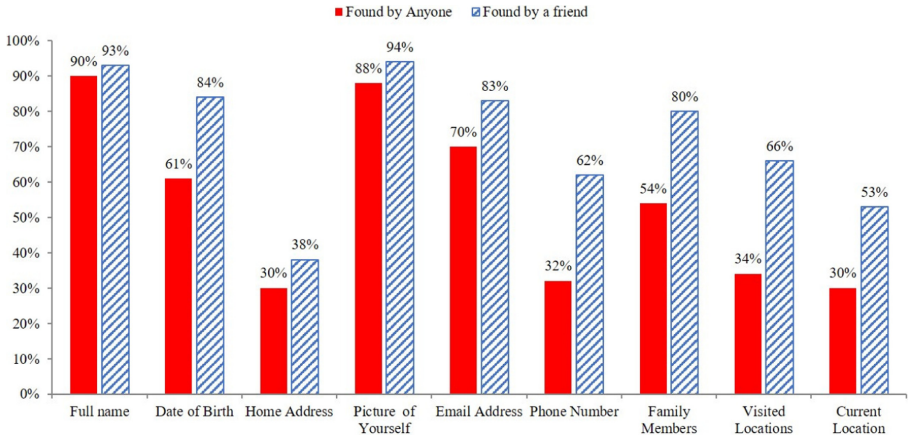

Fig. 1. Comparison of perceptions of personal information that could and could not be found by anyone and by a friend.

4.3 Secondary Leakage

The participants were asked if they thought that a friend of a friend would be able to find their personal information, to gauge their awareness of secondary leakage. The majority of participants did think that a friend of a friend could find most of their information, with the exception of where they lived, their phone number and their current location (See Fig. 2). These responses indicate that the majority had some awareness of secondary leakage. This type of secondary leakage can actually negate a person's privacy settings by leaking information which they may have set to be private.

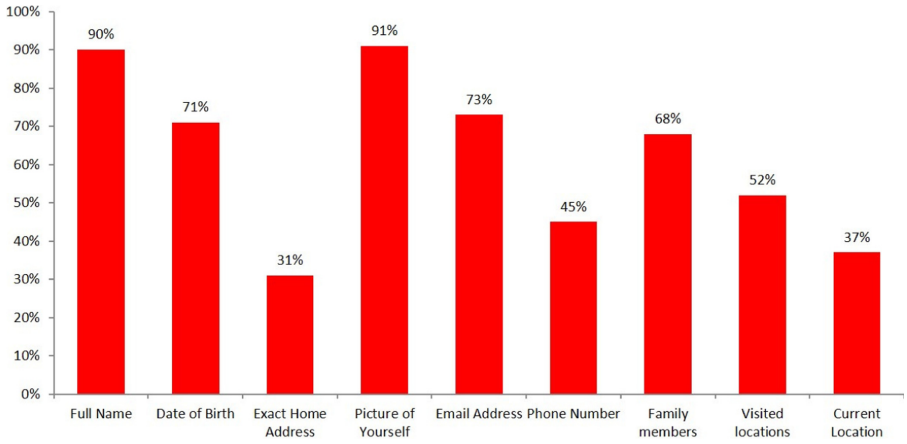

Fig. 2. Secondary leakage caused by a friend of a friend.

5 Overview of Experimental Results

A manual search method was used for all participants, and a number of automated tools (see Table 4) were used for searches on the top 10 and the bottom 10 (designated the "over sharers" and "minimal sharers" respectively, relative to the amount of data collected), see Sect. 5.4. There was a small amount of overlap of the information found by both methods, as can be seen in Fig. 3.

5.1 Results of Manual Harvesting

Using simple OSINT tools, at least one piece of information was found on 79% of the participants (n = 252), see Fig. 3. As the majority did have Facebook accounts (86.5%) it was relatively easy to find their names (82.1%) and pictures (72.2%), although not everyone used their actual photograph (some used cartoons or funny pictures). Some of these people also had their birthdays listed (20%).

Some had their birthdays listed on Twitter and in Facebook on posts which were viewable by anyone, although for some only partial (incomplete) DoBs were found. For those who had weaker privacy settings the names of close family and friends were also easily found.

Recent locations were identified through posted tweets using the geotag information and for a few (11%), their current location was also identified through recent posts on Facebook. LinkedIn was useful for obtaining email addresses as these are generally not hidden by the privacy settings due to the nature of the web site. The FindMyPast web site [32] gave access to a lot of information as this was used to search public records such as the census, birth, marriages, deaths, parish records and land and survey records. These web sites revealed home addresses using only the person's name, the area where they live and their approximate age. Census records also provided the names of other family members living at the same property. The phone numbers (land lines) were harder to find using the BT Phonebook web site, as it only revealed 12 phone numbers in total, probably due to the majority of respondents being students and only having mobile phones. Phones numbers are not a requirement when registering on most OSNs [36].

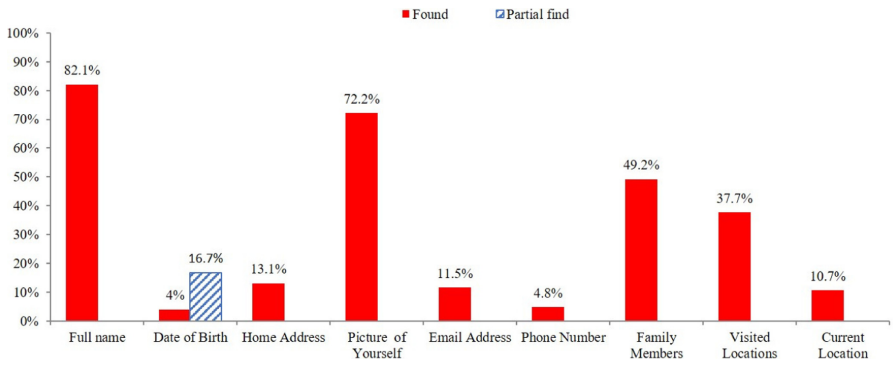

Fig. 3. Information that can be harvested by a stranger.

5.2 Timing the Manual Searches

The searches for the participants' data were undertaken in a timed experiment by someone who had no connection or link with them in cyber space. Table 3 shows the details, grouped in 1-minute time bands. It can be seen that for the majority (66%), their information was found in less than two minutes. In under three minutes 88.5% of the participants had their data exposed.

The average time taken for each search was 109 s. The fastest exploit took 14 s with the participant's full name, picture and the places that they had visited in the last few days all being identified in that time. The longest time was 460 s

(7 min 40 s) and for this participant their full name, partial DoB (month and year only), exact home address, picture and close family names were all found. This particular investigation took more time due to the participant's information being distributed across multiple web sites, which took longer to locate, see Table 3.

Table 3. Investigation in 1 min time bands

Time taken	Number of participants exploited
Less than 1 min	48
Between 1 to 2 mins	118
Between 2 to 3 mins	57
Between 3 to 4 mins	19
Between 4 to 5 mins	3
Between 5 to 6 mins	5
+7 min	2

5.3 Timed Use of Automated Tools

Using the ten least successful results (the "minimal sharers") and the top ten best results (the "over sharers"), five automated tools listed in tab:List of automated search tools used were used to find further personal information. This was then compared to the manual searches results. Table 4 also shows the personal data which these tools found and the related source.

Using the manual search method for these two groups took and average of 160.45 s, compared to the average time of 94.35 s for the search using the automated tools listed in Table 3. This was a decrease of 41.2% in the time taken for the searches. However, it was found that there was also a 37.9% decrease in the information that was found by the automated tools compared to the manual approach used, i.e. the manual searches were more successful.

5.4 Minimal Sharers vs Over Sharers

The "minimal sharers" were labelled P1 to P10 and the "over sharers" were P11 to P20. For the ten "minimal sharers", where no information at all was found using the manual method, the automated tools were more successful for 50% of them. Participants P1, P3, P4, P6 and P8 had no information found using either method. However, for participants P2, P5, P7, P9 and P10 two pieces and one partially complete piece of information was found (full name, their home address and a partial date of birth (month and year only)). So for these five people from the "minimal sharers" group the automated search was more successful.

Table 4. List of automated search tools used

Automated tools	Data found	Sources
Pipl [28]	Full name, email, current location, Recent location, photo	Online social networking sites
192 [29]	Full name, date of birth, telephone numbers, Full address, family members	Census/Birth/Marriage/ Military/Death Records
Peopletraceuk [30]	Full name, date of birth, telephone numbers, Full address, family members	Census/Birth/Marriage/ Military/Death Records
Social-searcher [31]	Full name, photo, email, family members, Current location, recent locations	Online social networking sites
Search.findmypast [32]	Full name, date of birth, telephone numbers, Full address, family members	Census/Birth/Marriage/ Military/Death records

For the ten "over sharers" some of the information from the manual searches was also found by the automated tools. However, overall the automated tools did not find as much data as the manual searches (average of 6.25 pieces of information found) compared to the automated tools average of only 2.05. Table 5 summarises the number of individual pieces of personal information found for each participant (P11 to P20) and also gives the overall average for each method. It should be noted that the 0.5.indicates a partial date of birth found. There was a 60.6% decrease in the information found by the automated tools.

The manual method found almost all of their information, except for a complete DoB, for participants P14 and P15, while the automated tools only found 3.5 and 2.5 pieces of information respectively. The least successful cases were P11 and P13 where the automated search only found 1.5 pieces of information for each. In the case of P17 and P18 the automated tools found no information at all, compared to 5.5 and 6.5 items of personal information respectively found using the manual method (See Table 5).

At this point it is worth mentioning the limitations of both the manual and automated searches that were used in this investigation. The automated tools selected for this investigation have limitations inherent within the design of the tools themselves. However, we still chose to use these automated tools as part of our methodology, as this study was carried out from the perspective of information gathering that could be conducted by someone with no prior knowledge and little to no experience in this area. In every case the manual method found more information. This demonstrates that automated tools are useful for finding information about people but are not nearly as effective as performing the searches manually, even when taking into consideration how much faster the tools are at performing the searches.

Table 5. Over sharers group data collected - manual search vs automated search

Participant #	Manual search results	Automated search results
P11	7	1.5
P12	4.5	3
P13	4.5	1.5
P14	8.5	3.5
P15	8.5	2.5
P16	6.5	3.5
P17	5.5	0
P18	6	0
P19	5.5	2.5
P20	6	2.5
Average	6.25	2.05

5.5 Participant's Perceptions Compared to Reality

Each participant in the survey had a view of their own personal privacy on the social network platforms that they used. So how did the results of the experiments, both manual and automated, compare to the participants' perceived online security?

Of the 20 participants examined in more detail only two thought that no information could be found about themselves online. One was in the "minimal sharers" group (P6) and the other, surprisingly, was in the "over sharers" group (P20). For P20 their perceptions were completely incorrect, as their full name, their picture, their email address, their phone number, a close family member and places that they had visited recently were all easily found. Only the perceptions of P6 were accurate, as no information could be found using either a manual or an automated search. Similarly, only two participants (P9 and P15), one from each group, thought that all their information could be found. P9 was incorrect, although as no information was found using the manual search, two pieces and a partial were found by the automated tools. Only P15 correctly identified that all their information could be found online and this was confirmed during the manual search. However, using the automated tools a full name, exact home address and a partial date of birth were identified for P9, P15 and P20. For P6 and P9 (minimal sharers) it can be assumed that they care about their data and keep it relatively safe. For the "over sharers" only P15 correctly identified the information that could be found. P20 was clearly naive regarding what of their personal information could be found online, with their Full Name, Photo, Email Address, Phone Number, Close Family and Recent Locations being found by the manual search. It is clear that they overestimated the privacy of their personal information and it is also important to consider how they may also be naive about the value of this personal information for impersonating or targeting an individual.

5.6 Evaluation

It was interesting to compare the results and the demographics of the "minimal sharers" and "over sharers" with the results as a whole. These two groups are a small sub-set, so no statistical analysis was undertaken, but it was interesting to compare the demographics of these two groups. The following questions were considered:- Who were these participants? Did the age range or the gender influence which group they were in? How did these participants compare to the total respondents as a whole?

For the "minimal sharers" (P1 to P10) the ratio of females to males was 50–50. This was interesting as the total number of female participants was slightly higher (58.4%). For this group 90% were aged 18 to 21 years, and only one person was in the range 51 to 60 years. These two age ranges 18–19 and 20–21 made up 54% of the total survey respondents. All were students and included one in employment. Only two of the students were studying a computer related degree. This meant that 80% of these participants were on non-computing related degrees. This is in direct contrast to the total survey where 57% were on non-computing degrees.

The "over sharers" (P11 to P20) had a majority of males (60%), which was also in direct contrast to the survey as a whole, where only 41% were male. This meant that males were over-represented in this group. The "over sharers" ages ranged from 18 to 25, which was not entirely unexpected as these three age ranges included the largest number of respondents. However, the majority (80%) fell into the range 21 to 25 and all were undergraduate students, with six on computer related degrees and only one female student included. This compares to 43% who were on computer related degrees for the total survey.

For participants P11 to P14, additional information was found as well as the nine items that had been targeted in this work. These included someone's job, their age, partner's and Mother's names, places of education and other family members' names (See Table 6). A number of these items could be useful to a hacker, for example, to reset passwords for online accounts where the answer to a secret question is required. In the case of P13 more information was found manually, although it took 7 min 40 s to retrieve this, compared to the automated search which only took 1 minute 54 s, although the automated search did not find exactly the same information (See Table 5).

A paired t-test was conducted on what was actually found by a stranger compared to the participants' perceptions for all participants. This was done to identify if there was a significant difference between our two subsets of participants and identify if this difference was due to random chance. The result is significant ($t = -10.621$, $p < .0001$, $df = 251$, 95% CI: -2.16, -1.49). The correlation was found to be 0.1228186, which also shows that there was very little connection between the two sample sets. The calculated effect size is therefore 0.6668677, which is reasonable. These results demonstrate that our hypothesis that the participants over-estimated the amount of information that could be found by a total stranger is correct. The p-value is significant, being less than alpha = 0.05, meaning there is a 95% chance that the result is not due to random

Table 6. Additional information found for the "over sharers" group

		Additional information found
P11	Manual	Birth certificate number
	Tools	Current job, places of education, current age with two-year margin, there was more data available but was locked behind a paywall
P12	Manual	
	Tools	The full names of their parents and siblings, there was more data available but was locked behind a paywall
P13	Manual	Their mothers and current partners full name, The month and year of their birth, Birth certificate number
	Tools	Mum's name, age range, there was more data available but was locked behind a paywall
P14	Manual	
	Tools	Full names of parent's siblings and other family members, Places they have been and are employed, Places of education, there was more data available but was locked behind a paywall

chance. A power analysis was calculated to be 1.000, which demonstrates that our sample size is sufficiently large. Using the same effect size, alpha, and power only a total sample size of 118 would need to be collected to achieve the same results and our sample was 252. There was no statistical analysis undertaken on the "minimal sharers" or the "over sharers" as these two samples are too small to get any meaningful results, but their demographics were examined.

6 Discussion

6.1 All Participants

The participants identified what information they had voluntarily shared online and their responses where compared to the information gathered. For the survey as a whole the majority of participants were aware that some of their personal information could be found online by a complete stranger. However, very few thought that a stranger could locate their home address (predicted 30% vs found 13%) or their current location (predicted 30% vs found 11%). Awareness amongst all participants regarding secondary leakage was high, as seen by their responses (See Fig. 2). In almost every case the information gathered was less than the participants' expectations. Only in the "visited locations" category was more information retrieved (See Table 7).

The date of birth was a very difficult piece of information to find, which the participants grossly over estimated, with 61% predicating that this could be found by anyone. In practice only 4% had their full DoB revealed and a further 16.7% had a partial DoB (month and year only) found. This was interesting as 68% claimed that they had voluntarily published this information online. An

email is one piece of information that most people would willingly give if someone asked for it and 70% of the survey respondents thought that their email could be easily found by anyone. However, during the investigation only 12% of the emails were found. This is because web sites tend to obfuscate email addresses to help prevent email harvesting for phishing attacks. More than half of the respondents (54%) presumed that a complete stranger could identify the names of their close family members. Reality was that 49% were actually identified including a photograph of the family member.

Table 7. Overview of perceptions vs Information found

	Perceptions by a friend	Perceptions by a stranger	Actually found
Full name	93%	90%	82%
Picture of themselves	94%	88%	72%
Email	83%	70%	11%
DoB	84%	61%	20%
Family members	80%	54%	49%
Visited location	66%	34%	38%
Phone number	64%	32%	48%
Current location	53%	30%	11%
Home address	38%	30%	13%

Additional information collected included Blackberry messenger, Skype and Snapchat accounts. Some participants also had personal web sites and blogs, which were easily located. These can reveal a lot of information about their hobbies, interests, aspirations and even their work place. The places they had visited in the last few days were interesting, as 34% of respondents thought that this could be found. In fact 38% were actually found, see Table 7. Figure 4 shows a graphic of the comparison of results for all participants.

It was note worthy that the initial email sent out inviting people to take the survey was actually sent from the email account of a senior member of the teaching staff at the university. The email message explicitly asked people to reply directly to the second author and his email was given in the text. However, six participants responded directly to that email without noticing that the email address did not come directly from the person requesting assistance. This is exactly how phishing attacks succeed, as the attackers are hoping that the recipient will not notice that the email address does not match. This was just over 2% of the total participants, but it illustrates that if enough emails are sent out in a phishing attack, then there will always some people who will respond.

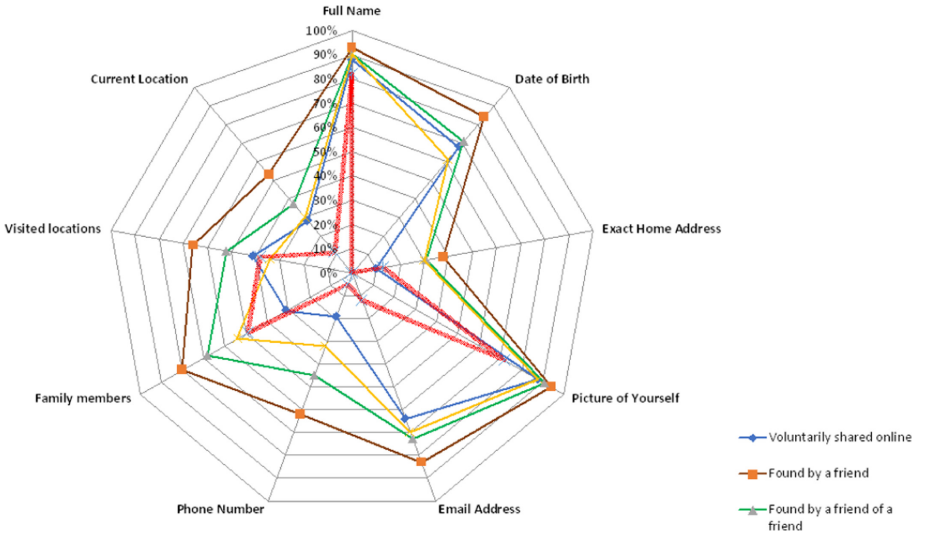

Fig. 4. Comparison of results for all participants.

6.2 Minimal Sharers vs Over Sharers

The combined average time taken to access the personal information using the manual method for the "minimal sharers" and "over sharers" was 160 s. With the automated tools this only took 94.35 s, but found 38% less information than the manual approach. Compare this to all participants, where each search took on average 109 s per person. The manual searches found more information on these participants than the automated tools. However a number of additional pieces of information were found by the automated tools which were not initially found during the manual search (See Table 6).

While statistical analysis of these two groups was not possible due to the small numbers, we did examine the demographics of each group and found significant differences and also when compared to the total population. This is highlighted in Table 8. In the "over sharers" group 60% were on computer related degrees compared to 43% of all participants, and only 20% within the "minimal sharers" group. So students studying computer related degrees are over-represented in the "over sharers" group.

The age range was split into two equal sizes (4 years each) for comparison here. The majority of the "minimal sharers" were in the younger 18 to 21 age band, with 90% in that age range, compared to the total population of 54%. The "over sharers" age range was 80% in the 22 to 25 band compared to 31% in the overall survey (see Table 8). So these two groups have different demographics to the population as a whole and between the two groups.

Gender related statistics showed that 36% of the "minimal sharers" were female and on computer related degrees, compared to only 17% in the "over sharers" group, so females were under-represented in this group. For the "mini-

Table 8. Comparison of total survey, minimal sharers and over sharers groups

		Total survey	Minimal sharers	Over sharers
Gender	Male	41%	50%	60%
	Female	59%	50%	40%
Age ranges	18–21	54%	90%	20%
	22–25	31%	10%	80%
Degree type	Non-Computing	57%	80%	40%
	Computing related	43%	20%	60%

mal sharers" only two participants (one female and one male) were on computer related degrees (one was on a BSc Computer Security and Forensics degree and the other was on BSc Business Computing degree). In total there were a higher proportion of females (50%) in the "minimal sharers" group regardless of the degree taken, compared to only 40% in the "over sharers" group. This is still lower than the total survey, in which the majority, 59% of participants were female. This implies that males are less concerned about their privacy or possibly that they are more confident in their ability to control their personal data (see Table 8).

7 Conclusion

In this work we investigated personal data leakage caused by the interaction of OSN users with friends, friends of friends and a complete stranger. We found that for the n = 252 participants, the majority of their personal information could be accessed in under two minutes by someone who had no connection with them in cyberspace. This work demonstrated that there are still people who publish too much personal information on OSNs (over sharers), either intentionally or accidentally with no regard for who can see this or appreciation of how this information may be abused. It also highlights the very real risks and threats that people can face due to their lack of awareness of online privacy. These results show that the majority of participants did not perceive their online privacy accurately, including students on computer related degrees. The paired t-test showed that the results of the survey compared to what was actually found by us to be statistically significant. Our hypothesis that the majority inaccurately predicted how secure their personal information is, was confirmed.

The manual searches were more effective at finding information than the automated tools. However, the use of automated tools for data harvesting makes it easier for someone with no computer skills to find information about a particular target. While both search methods did find some similar information, they also found different pieces of personal information for the majority, see Tables 4 and 5. It cannot be underestimated how little time it took to manually harvest the majority of the their personal data effectively - less than 2 min for two thirds

of the participants. However, for these experiments it was the amount of information that was found that was more significant rather than the actual time taken, although it was not anticipated that the time taken for the majority of searches would be so low. For future work it would be interesting to see the results if this work was repeated on a more diverse survey population, i.e. non-students.

This study underlines the importance of continuing to educate users on how their existing behaviours can present a serious threat to the security of their personal information and that of their virtual "friends" on OSNs and emphasises the need for greater security awareness.

References

1. Statista Social Media Statistics. https://www.statista.com/topics/1164/social-networks/. Accessed 8 Mar 2021
2. Van Schaik, P., Jansen, J., Onibokun, J., Camp, J., Kusev, P.: Security and privacy in online social networking: risk perceptions and precautionary behaviour. Comput. Hum. Behav. **78**, 283–297 (2018)
3. Irshad, S., Soomro, T.R.: Identity theft and social media. Int. J. Comput. Sci. Netw. Secur. **18**, 43–55 (2018)
4. What is the Definition of Online Privacy?. https://www.winston.com/en/legal-glossary/online-privacy.html. Accessed 8 Mar 2021
5. Gan, D., Jenkins, L.: Social networking privacy-who's stalking you? Future Internet **7**, 67–93 (2015)
6. Palin's Email Account Hacked. https://www.huffingtonpost.co.uk/entry/palins-email-account-hack_n_127184?guccounter=1. Accessed 8 Mar 2021
7. Hacker group targets Mark Zuckerberg's online accounts - again. https://www.zdnet.com/article/hacker-group-targets-mark-zuckerberg-online-accounts-again. Accessed 8 Mar 2021
8. How to [ALMOST] keep your tinder private from your Facebook. http://www.knowyourmobile.com/apps/tinder/22254/how-keep-your-tinder-private-your-facebook. Accessed 8 Mar 2021
9. Debit card (@NeedADebitCard). https://twitter.com/needadebitcard. Accessed 8 Mar 2021
10. Driving Licence posted on Twitter. https://www.google.co.uk/search?q=twitter+picture+driving+licence+uk. Accessed 8 Mar 2021
11. Alex Jones on Instagram: Teddy Thomson is official. https://www.instagram.com/p/BRFtCCHDL_O/?hl=undefined. Accessed 8 Mar 2021
12. Ali, S., Islam, N., Rauf, A., Din, I., Guizani, M., Rodrigues, J.: Privacy and security issues in online social networks. Future Internet **10**, 114 (2018)
13. Islam, M.B., Watson, J., Iannella, R., Geva, S.: A greater understanding of social networks privacy requirements: the user perspective. J. Inf. Sec. Appl. **33**, 30–44 (2017)
14. Tsay-Vogel, M., Shanahan, J., Signorielli, N.: Social media cultivating perceptions of privacy: a 5-year analysis of privacy attitudes and self-disclosure behaviors among Facebook users. New Media Soc. **20**, 141–161 (2016)
15. Wisniewski, P.J., Knijnenburg, B.P., Lipford, H.R.: Making privacy personal: profiling social network users to inform privacy education and nudging. Int. J. Hum.-Comput. Stud. **98**, 95–108 (2017)

16. Cascavilla, G., Beato, F., Burattin, A., Conti, M., Mancini, L.V.: OSSINT - open source social network intelligence. Online Soc. Networks Media **6**, 58–68 (2018)
17. Child Safety Online: a practical guide for parents and Carers whose children are using social media. https://www.gov.uk/government/publications/child-safety-online-a-practical-guide-for-parents-and-carers. Accessed 8 Mar 2021
18. Cambridge analytica: the story so far. https://www.bbc.co.uk/news/technology-43465968. Accessed 8 Mar 2021
19. Revealed: 50 million Facebook profiles harvested for Cambridge analytica in major data breach. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election. Accessed 8 Mar 2021
20. The Cambridge analytica and facebook data scandal: how to tell if your data was shared. https://www.techradar.com/news/us-uk-investigating-facebooks-role-in-cambridge-analytica-data-breach. Accessed 8 Mar 2021
21. Facebook's new data-sharing policies are crashing tinder. https://www.wired.com/story/facebook-policies-tinder-crashing. Accessed 8 Mar 2021
22. Albrecht, K., Mcintyre, L.: Psst...your location is showing!: metadata in digital photos and posts could be revealing more than you realize. IEEE Consum. Electron. Mag. **4**, 94–96 (2015)
23. Pontes, H.M., Taylor, M., Stavropoulos, V.: Beyond Facebook addiction: the role of cognitive-related factors and psychiatric distress in social networking site addiction. Cyberpsychology, Behav. Soc. Netw. **21**, 240–247 (2018)
24. Pegg, K.J., O'donnell, A.W. Lala, G., Barber, B.L.: The role of online social identity in the relationship between alcohol-related content on social networking sites and adolescent alcohol use. Cyberpsychology, Behav. Soc. Netw. **21**, 50–55 (2018)
25. Hendriks, H., Den, Van, Putte, B., Gebhardt, W.A.: Alcoholposts on social networking sites: the Alcoholpost-typology. Cyberpsychology, Behav. Soc. Netw. **21**, 463–467 (2018)
26. Jensen, M., Hussong, A.M., Baik, J.: Text messaging and social network site use to facilitate alcohol involvement: a comparison of U.S. and Korean college students. Cyberpsychology. Behav. Soc. Netw. **31**, 311–317 (2018)
27. Ying, Q.F., Chiu, D.M., Venkatramanan, S., Zhang, X.: User modeling and usage profiling based on temporal posting behavior in OSNs. Online Soc. Netw. Media **8**, 32–41 (2018)
28. Pipl - identity information search and API. https://pipl.com/. Accessed 8 Mar. 2021
29. Search for people, businesses and places in the UK. http://www.192.com/. Accessed 8 Mar 2021
30. Finding people the right way. www.peopletraceuk.com/. Accessed 8 Mar 2021
31. Real-time social media monitoring. https://www.social-searcher.com/. Accessed 8 Mar 2021
32. Find my past. https://search.findmypast.co.uk/search-world-records. Accessed 8 Mar 2021
33. The top 10 worst social media cyber-attacks. https://www.infosecurity-magazine.com/blogs/top-10-worst-social-media-cyber/. Accessed 8 Mar 2021
34. Tweet location FAQs. https://help.twitter.com/en/safety-and-security/tweet-location-settings. Accessed 8 Mar 2021
35. Instagram by the numbers. https://www.omnicoreagency.com/instagram-statistics/. Accessed 8 Mar 2021
36. BT - find a person. https://www.thephonebook.bt.com/person/. Accessed 8 Mar 2021

37. Rathore, N.C., Tripathy, S.: AppMonitor: restricting information leakage to third-party applications. Soc. Netw. Anal. Min. **10**(1), 1–20 (2020). https://doi.org/10.1007/s13278-020-00662-7
38. Kekulluoglu, D, Magdy, W., Vaniea, K.: Analysing privacy leakage of life events on twitte. In: 12th ACM Conference on Web Science, 287–294 (2020)