THE UNIVERSITY
of EDINBURGH

Analysing Privacy in Online Social Media

Dilara Keküllüoğlu

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2022

Abstract

People share a wide variety of information on social media, including personal and sensitive information, without understanding the size of their audience which may cause privacy complications. The networked nature of the platforms further exacerbates these complications where the information can be shared without the information owner's control. People also struggle to achieve their intended audience using the privacy settings provided by the platforms. In this thesis, I analyse potential privacy violations caused by social media users and their networks, as well as the usage and understanding of privacy settings. I focus on Twitter which has rather simplistic privacy settings with binary states.

The first part of my studies includes investigating personal information disclosures by networks using congratulatory messages. I analyse these messages and detect various types of life events including relationships, illness, familial matters, and birthdays. I show that public replies are enough to infer the content of the original message, even if the event subject hides or deletes the message. I further focus on birthdays which is one of the most popular life events and the potential date of birth disclosure has security implications besides the privacy ones. I show that over 1K users have their date of birth exposed daily, where 10% of these users have protected their tweets. I also show that users react positively to these congratulatory messages even though these posts potentially disclose personal and sensitive information.

In the second part of my thesis, I focus on privacy settings on Twitter. I quantify the usage patterns of privacy settings and investigate the reasons for changing these settings between public and protected by conducting a mixed-method study. I show that there is a set of users who frequently utilize the privacy settings provided by the platform. I also show that users turn protected to share personal content and regulate boundaries while they turn public to interact with others in ways prevented by being protected.

In the last stage of the thesis, I investigate the user understanding of information and tweet visibility of different account types by conducting a user survey. I show that the users are aware of the visibility of their profile information and individual tweets. However, the visibility of followed topics, lists, and interactions with protected accounts is confusing. Less than a third of the survey participants were aware that a reply by a public account to a protected account's tweet would be publicly visible. Surprisingly, having a protected account did not result in a better understanding of the information or tweet visibility.

Actual functionalities and the user understanding of them should align so that users can take the right actions for desired levels of privacy protection in online social networks. I show that even with simplistic privacy settings, users have difficulty understanding the reach of their posts. Implications of interactions between users need to be clearly relayed. I give design suggestions to increase this awareness and for users to have better tools to manage their boundaries. I conclude the thesis by giving general implications around the studies conducted and possible future directions.

Lay Summary

People share a wide variety of information on social media, including personal and sensitive information, without understanding who can see these posts. This may cause privacy problems. A user's connections over social networks can also share information about the user. Social media platforms provide privacy settings to users to protect their privacy. However, these settings are not easy to configure as intended. In this thesis, I analyse potential privacy violations caused by social media users and their connections, as well as the usage and understanding of privacy settings. I focus on Twitter which has rather simplistic privacy settings with two states.

Through a series of studies, I show that a user's connections can leak personal information about the user, including the date of birth. Some people use the privacy settings of the platform frequently to protect their privacy. However, users do not understand the visibility of interactions by different accounts on Twitter. Even having a protected account did not result in a better understanding of the information or tweet visibility. I conclude my thesis by giving implications of my studies and design suggestions to overcome these challenges.

Acknowledgements

I am grateful to everyone who supported and helped me during this journey. Firstly, I would like to thank my supervisors, Dr Walid Magdy and Dr Kami Vaniea. Without their support and understanding, I would not be able to finish this journey. I feel lucky to be given the opportunity to work with them and be their student. In terms of advising, mine was a special case where both my supervisors took equal workload and our meetings were mostly all together. I feel grateful and fortunate that I got to experience the guidance of two supervisors fully. They supported me throughout the process and stood by me when things got difficult and unpredictable. They guided me in a time where everything had to be remote and everyone had to adjust to a new way of mentoring, advising, and research. They gave me their time whenever I needed help while also making sure I had the space to grow as a researcher. They were kind, compassionate, understanding, and really amazing. Thank you.

I want to thank Dr Nadin Kökciyan who was a committee member of my annual reviews. Nadin is probably the person who was there the most to witness my academic growth. She was there when I was doing my master's degree and she has been with me throughout most of my academic journey. She has been an amazing supporter, a reliable mentor, and also a great role model from Boğaziçi to Edinburgh. I want to thank Dr Yasemin Acar for accepting me as an intern and creating a welcoming environment for an unusual research group with people from all over the world working remotely. Yasemin has been empowering, supportive, and kind throughout the internship and beyond. I also want to thank my collaborator Dr Maria Wolters for her contributions and also helping me connect with friends in the Forum. I want to thank my thesis committee, Björn Ross and Norah Abokhodair, for agreeing to be on the committee, reading my thesis, and providing great feedback. Thank you.

I want to thank my labmates and friends from SMASH and TULIPS, as well as fellow friends on the PhD journey, Abeer, Adam, Alex, Ameer, Duncan, Florian, Ibrahim, Julie-Anne, Kahraman, Kholoud, Lucia, Mohammad, Nicole, Rabia, Sara, Silviu, Tarini, Youssef, and many others. I also want to thank the friends I made in Edinburgh for life who did not hesitate to help me when I was at my lowest, away from my family, and they became like a second family to me in Edinburgh. Last but not the least, my friends (dostlarım) who have been with me for so long, some from my undergrad years and some even from childhood, thank you for being there for me, listening to my rants, and supporting me from far away. Thank you.

I want to thank my family who were there for me in many ways until now, my parents who supported my education fully without restraint, as well as my extended family who were nothing but supportive. They probably believe in me more than I do myself. I want to thank my mom who is one of the strongest people I know. She is smart, capable, compassionate, caring, and someone I can always rely on when things get difficult. I thank my dad who is accepting, understanding, adventurous, loving, and someone I can always seek advice and discuss anything. They are both very hard-working and they did not hesitate to do everything they could to support me and my sister. I thank my younger sister, who has been my rock, where I can share my thoughts and worries freely without fear of judgement and seek comfort as we go through life and hardships. I don't think anyone can understand me better than her. Thank you.

This journey has been extraordinary and difficult in many ways. I lost my dear grandfather a month after I came to Edinburgh to start my PhD and my mom had a heart attack close after. I received support from many people to have the energy to come back to Edinburgh and keep going. After 1.5 years in Edinburgh, I had to move back to Istanbul to stay with my family during the pandemic and continued my studies from there. Throughout these hardships, I felt supported and cared for with the help of many people including my family, supervisors, friends I made in Edinburgh, long-time friends from Turkey, and many I might have missed mentioning. I feel fortunate that I had the pleasure to meet all these amazing people. Thank you all from the bottom of my heart.

Alhamdulillah.

Declaration

I declare that the thesis has been composed by myself and that the work has not be submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated below. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others. All included contributions here are confirmed by the co-authors.

The work presented in Chapter 3 was previously published in 12th ACM Conference on Web Science (WebSci'20) as "Analysing Privacy Leakage of Life Events on Twitter" by **Dilara Keküllüoğlu**, Walid Magdy (supervisor), and Kami Vaniea (supervisor) [83]. This study was conceived by all of the authors. I carried out the data collection, data analysis, and wrote the first draft of the paper. WM and KV assisted in the study design and provided feedback throughout the project. All authors contributed to writing the paper.

The work presented in Chapter 4 was accepted for publication in 16th International Conference on Web and Social Media (ICWSM'22) as "From an Authentication Question to a Public Social Event: Characterizing Birthday Sharing on Twitter" by **Dilara Keküllüoğlu**, Walid Magdy (supervisor), and Kami Vaniea (supervisor) [85]. An early version of this paper was presented as a poster at 6th International Conference on Computational Social Science (IC^2S^2'20) [84]. This study was conceived by all of the authors. I carried out the data collection, participant recruitment, data analysis, and wrote the first draft of the paper. I designed the user survey with KV. KV was the second coder for the qualitative analysis. WM and KV assisted in the study design and provided feedback throughout the project. All authors contributed to writing the paper.

The work presented in Chapter 5 was previously published in CHI Conference on Human Factors in Computing Systems (CHI'22) as "Understanding Privacy Switching Behaviour on Twitter" by **Dilara Keküllüoğlu**, Kami Vaniea (supervisor), and Walid Magdy (supervisor) [86]. This study was conceived by all of the authors. I carried out the data collection, participant recruitment, data analysis, and wrote the first draft of the paper. I designed the user surveys with KV. KV was the second coder for the qualitative analysis. KV and WM assisted in the study design and provided feedback throughout the project. All authors contributed to writing the paper.

The work presented in Chapter 6 is conducted by **Dilara Keküllüoğlu**, Kami Vaniea (supervisor), Maria K. Wolters, and Walid Magdy (supervisor). This study

was conceived by all of the authors. I carried out the participant recruitment, data collection, data analysis, and wrote the first draft of the paper aside from the statistical results part in Section 6.4.2. I designed the user survey with KV. I worked with MW to design the statistical questions and put the data into an appropriate format. MW conducted the statistical analysis and subsequently wrote the statistical results. KV and WM assisted in the study design and provided feedback throughout the project. All authors contributed to writing the paper.

Dilara Keküllüoğlu

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Online Social Networks (OSN) are widely used to share information with others where the flow of the information depends not only on the understanding and the actions of the information owner but also all those who interact with their posts. OSNs are platforms where people can create accounts, form connections with other accounts, and observe the activities of these connections & others on the platform [23]. People use OSNs to share their experiences, interact with each other, as well as to gain social capital [45]. Some people may use OSNs in professional contexts [107] and to build reputation [150]. Others may share their daily lives and life events on social media; such as birthdays, marriages, buying a new car, or acceptance to their dream university. They also use social media as a source of support during difficult life events such as pregnancy loss, serious and chronic illnesses [12, 173]. While some OSNs allow users to form closed networks where communications are only visible to select users, these communications can also be shared publicly for everyone to see. Information shared on social media by users and their networks can lead to unintended disclosures. Some disclosures may result in discomfort such as colleagues learning that the user is moving houses, leaving their job, or getting a divorce. Some posts can result in serious damage, e.g. location sharing may signal the burglars that the house is vacant. These disclosures can also have privacy implications that are not obvious to users. For example, posts about surgeries and illnesses may result in insurance premium increases as insurance companies monitor social media posts to make decisions [58, 102]. Identity thieves can use publicly shared lower-value information to trick companies into giving up higher-value information, as a Wired reporter found out when someone used public

information about him to remotely format his Mac and take over his social media account [60].

Privacy is a dynamic process where people selectively disclose and conceal personal information [10]. Humans are social creatures and are well accustomed to this continuous shifting of how they present themselves and how they manage their boundaries around personal privacy, self presentation, and access to self. People regulate this information flow using their implicit privacy rules and shared information becomes co-owned with the people who received the information [123]. These co-owners gain the ability to further share the information with others as they wish. The introduction of social media platforms though can lead to challenges in boundary management where many of the natural high-granularity actions taken in the physical world must be translated into the more rigid privacy setting options used by platforms [45]. Hence, configuring these privacy settings to match an intended audience is not trivial for the users. These settings can be hard to understand and time-consuming to configure [100, 105, 62]. Privacy settings change depending on the platform where some provide only simple configurations, e.g. account-based settings, others provide more granular settings, e.g. different settings per post. Hence, users need to learn the configurations for each platform they use. These settings can also be updated over time with the introduction of new features which adds to the complexity of configuring them as intended.

The networked nature of the social media further complicates the information flow management and causes information leaks that are not easily controlled by the information owner. Even if a user configures the privacy settings as they wanted, their networks can leak information about them [68]. Users' age, gender, religion, diet, and personality can be inferred only using the posts mentioning their account name [77]. Analysing a user's network of connections can disclose attributes about them such as age, gender, location, political orientation, and sexual orientation [7, 176, 70, 8, 106]. Trusting people in the user's network to properly protect their privacy might not be sufficient [129], privacy expectations of a user and their network might be different [75], and priming the network might even backfire, leading them to share more [11].

Privacy protection on OSNs is a collective work that is based on both the *actions* and the *understanding* of all involved users, i.e. both the information owner and members of their personal networks in platforms. Engagements on social media posts have the potential to increase visibility of the post and disseminate the information to other users that are not originally in the information owner's network. Hence, it is neces-

sary to understand what steps users are currently taking to protect privacy online and how well the implications of these measure are understood by the users who share the information and the users who engage with it.

Currently, there are 17 online social networks with more than 300 million active users [136]. These platforms can have varying rules for account creation and network formation. For example, Facebook [47] and LinkedIn [98] only allow users to open accounts with their real names, while users can create pseudonymous profiles on Twitter [153] and Reddit [133]. Some platforms may have an intended focus on the relations between users and contents of the posts too. LinkedIn is a platform where people connect and post with the aim of having a professional work presence and potentially find jobs. Instagram [65] only allows picture or video posts with accompanying texts.

Most of the prior studies mentioned are conducted on platforms with granular privacy settings, e.g. Facebook, where the audience of each post can be configured separately, customized groups can be created for different kinds of posts, and so on. However, some platforms have limited low fidelity privacy controls, such as Twitter, where the privacy settings apply to the whole account, once changed the settings will affect future tweets as well as the past ones. Simpler controls can have the potential to make their privacy easier for users to understand. However, the functional reality is that the way controls combine across time and between accounts may make them deceptively confusing. There is a research gap on the user understanding and utilization of privacy settings in platforms with simpler controls. This thesis aims to fill this gap by investigating the disclosures with potentially private information and user reactions to such disclosures. It also measures the user understanding and usage of privacy settings on a platform with such low fidelity privacy controls. I select Twitter as the focus of my thesis which has a relatively simplistic binary privacy settings where the privacy settings provided are account-based which may lead to privacy leaks when a user changes their account visibility on Twitter since the setting changes the visibility of all their tweets, past and future. In addition, the visibility of posts is only dependent on the tweeter's account type (i.e. public or protected). Hence, all posts from a public account are publicly visible, even if the tweet mentions a protected account. The result is that a public account can easily breach the privacy of a protected account even if the breach is unintentional.

1.2 Research Questions and Contributions

In this thesis, I study unintended privacy leaks, either shared by users or their networks on Twitter. I aim to understand the types of information shared on the platform, as well as the reactions to such leaks by the users. I investigate the utilization of privacy settings and the user understanding of information visibility with different types of accounts Twitter provides.

To do so, I firstly focus on personal information disclosed by networks on Twitter. People self-disclose information using social media to express themselves, manage their identity, and seek social validation and approval [16]. These posts can elicit responses which in turn can disclose personal information about the data owner that is not easily manageable. Additionally, this personal data sometimes can be sensitive and used maliciously. Hence, after analyzing the personal information disclosure by networks, I continue to analyze the reactions to such potential data disclosures and the comfort levels around public celebrations. One of the strategies people can employ to minimize the data leak and protect their privacy is using the platform-provided settings. However, account-based binary privacy settings provided by Twitter might restrict users to manage their information flow as they want. Hence, some users find circumventions and employ strategies that were not intended by the platform. My third research question concerns one of these behaviours, i.e. the phenomenon of Twitter users changing their privacy settings back and forth, and the reasons behind these changes. These people use the privacy settings frequently. However, even when a user actively utilizes privacy settings, they might not be fully-aware about the implications of the settings. Hence, for my last research question, I chose to research the user understanding of visibility regarding the provided privacy settings.

I give my research questions with summarized motivations as follows;
A user's network can disclose information about the user by mentioning them or replying to them on Twitter. The disclosed information can be personal or sensitive depending on the situation. My first research question concerns these types of personal information disclosed online by the personal networks.

RQ1 What kind of personal information are disclosed online by networks on Twitter?

My main focus for the RQ1 is the information disclosed by users' networks on the platform. These disclosures cannot be controlled by the mentioned users, even when they choose to have their own tweets protected. Understanding user reactions to these kinds of leaks can inform the design of the platforms.

RQ2 How do users react (e.g. like, retweet, or reply) to tweets that share their personal information when they are mentioned in them and how comfortable are they with such information disclosure?

Twitter provides binary privacy settings for users to manage the visibility of their tweets. These settings can be used by users and their networks to decrease the privacy leaks investigated by the previous RQs. Some users choose to use these settings actively and might change their settings depending on different situations.

RQ3 What reasons do users have to change their privacy settings on Twitter and how do the reasons differ between changing the account setting to protected versus public?

Users find it cumbersome to configure privacy settings to their intended audience in platforms with detailed post-based settings [140]. Twitter has simpler privacy settings compared to these platforms which may lead to better user understanding. It is also expected that users who actively utilize these settings would have higher awareness about the visibility of their information and tweets configured by these settings.

RQ4 How well do Twitter users understand the visibility of user information and tweets in relation to different privacy settings?

I focus on these four main research questions in this thesis along with more detailed research sub-questions given in the next chapters. For RQ1 and RQ2, I focus on life events that prompt celebrations to analyze the network-based personal information disclosure. Sharing positive events with others increases positive affect face-to-face [92], as well as through interpersonal media [33] including social media. Hence, people tend to share them over social media [139, 94] and these posts can receive many responses, effectively leaking personal information. I further focus on birthday wishes to analyze the tension between the celebrations vs possible privacy and security consequences.

Following on the research questions, this thesis makes contributions below:

- **Finding life events shared on Twitter by only using the replies given to tweets.**

  I show that replies to tweets sharing life events, which is personal information, are enough to infer the types of the events. Hence, even the life events protected accounts share with only their followers can be leaked by their public networks. I compare the different types of life events in terms of quantity and the account types of the user who share them.

- **Collecting birthday messages and analysing the date of birth exposure from such tweets.**
  Birthdays are one of the most celebrated life events where even social media platforms encourage users to send congratulatory messages to their networks. The date of birth is also a personal information that is commonly used in secondary authentication by organizations. I analyze the birthday celebrations shared publicly on Twitter, as well as investigating the frequency of sharing ages in these tweets, which effectively discloses the date of birth of the birthday person.

- **Measuring the reactions and comfort levels of users regarding the personal information disclosure over Twitter.**
  I gather reactions of the mentioned users on the platform (i.e. likes, retweets, replies) to tweets disclosing life events including birthday congratulations. I compare the reactions given to different life events. I conduct a user survey to measure the comfort levels regarding public birthday messages over Twitter, including the messages that possibly disclose the date of birth of the birthday person. I find that users are comfortable with public congratulatory messages by others even when these posts disclose personal information. The majority of the tweets received positive interaction from the mentioned user.

- **Characterizing the frequent utilization of privacy settings on Twitter.**
  I quantify the privacy setting changes over a set of Twitter users showing that some people change their settings frequently between public and protected. I compare tweets of users when they were public vs. protected. I conduct two user surveys to further understand the reasons for the changes and quantify them.

- **Measuring the user understanding of privacy settings.**
  I conduct a user survey about information and tweet visibility with regards to different privacy settings to measure their awareness about the Twitter functionality. I show that the visibility of interactions between different account types are not clearly understood. It is shown that the account type of the user does not necessarily mean a better understanding of the visibility of information and interactions.

1.3 Thesis Outline

The rest of this thesis is organized as follows;

Chapter 2: In this chapter, I give a background about Twitter functionality and privacy settings. I also present a literature review around privacy on social media; user awareness of data dissemination on platforms, understanding of visibility of information and privacy settings, the networked nature of privacy protection, and inferral of data from social media posts by users and their networks.

Chapter 3: People use Twitter in different contexts such as finding people with similar interests, having a professional presence, as well as sharing their daily lives and happenings. In this chapter, I investigate the types of life events users are sharing on Twitter by analysing the congratulatory replies to such tweets. I collect tweets that mention a user and contain the phrase "happy for you". I find the themes of the tweets by using LDA topic modeling and group the tweets to resulting life event topics. I analyse the features of these tweets and compare the reactions to such congratulatory tweets towards public and protected accounts. This chapter contributes to the answering of RQ1 and RQ2.

Chapter 4: In this chapter, I focus on birthday celebrations over Twitter which is one of the most popular life events celebrated as shown in Chapter 3. I collect birthday celebration tweets sent to Twitter users to analyze the date of birth exposure. I also collect the reactions (i.e. replies, likes, and retweets) to these birthday celebrations on the platform. I design and conduct a user survey to measure the comfort levels of people around the public birthday celebrations on Twitter. This chapter also aims to answer RQ1 and RQ2.

Chapter 5: Previous chapters focused on the types of information disclosure by a user's network and the comfort levels around those situations. Platforms provide settings to prevent these kinds of privacy violations. In this chapter, I focus on users who actively utilize these privacy settings. I curate a dataset consisting users that changed their tweet visibility settings on Twitter at least once. I check the tweet visibility changes of these users and collect their tweets. I cluster users according to their privacy settings usage and compare the tweets they sent while protected vs. public. I also design and conduct two user surveys to understand the reasons behind the changes. I provide implications for the findings including a design implications section. This chapter's focus is answering the RQ3.

Chapter 6: People share information about themselves and others in social media platforms, including sensitive and private information that could be used in malicious ways. Platforms provide privacy configurations to users so they can protect their information and manage their boundaries. I show that some users utilize these settings actively in Chapter 5. However, users might not understand these settings and their implications. Prior work has shown that users are confused with privacy settings in platforms with granular settings. In this chapter, I focus on the user understanding of Twitter privacy settings which are rather simplistic with a binary state. I design a user survey around information and tweet visibility with different privacy settings. I recruit participants with different privacy settings usage and conduct the user survey. I measure the user awareness of privacy settings and information visibility. I collaborated with a statistical expert to investigate the factors affecting the user understanding by conducting a statistical analysis [1]. This chapter answers RQ4.

Chapter 7: In this chapter, I give summaries of findings discussed in the previous four chapters. I found that interactions can leak personal information that is not easily controlled by the information owners and the visibility of interactions are not well understood. Even users who frequently utilize the provided privacy settings and actively manage their boundaries are confused by who could see the interactions between different types of accounts. I discuss the overall implications of the studies and conclude my thesis.

1.4 Publications

- Dilara Keküllüoğlu, Walid Magdy, and Kami Vaniea. *"Analysing Privacy Leakage of Life Events on Twitter."* 12th ACM Conference on Web Science (WebSci'20). 2020. [83] detailed in **Chapter 3**

- Dilara Keküllüoğlu, Walid Magdy, and Kami Vaniea. *"From an Authentication Question to a Public Social Event: Characterizing Birthday Sharing on Twitter."* 16th International Conference on Web and Social Media (ICWSM'22). 2022. [85] detailed in **Chapter 4**

[1] I prepared the survey data for statistical calculations and supported the expert during the analysis. Section 6.4.2, which explains the preparation of the data, the statistical analysis, and the results, is written by the statistical expert who is also a co-author in the resulting paper.

- Dilara Keküllüoğlu, Kami Vaniea, and Walid Magdy. *"Understanding Privacy Switching Behaviour on Twitter"* CHI conference on human factors in computing systems (CHI'22). 2022. [86] detailed in **Chapter 5**

- Dilara Keküllüoğlu, Kami Vaniea, Maria K. Wolters, and Walid Magdy. *"Twitter has a Binary Privacy Setting, are Users Aware of How It Works?"* under submission in ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW). **Chapter 6**

1.4.1 Extended Abstracts

- Dilara Keküllüoğlu, Walid Magdy, and Kami Vaniea. *"Happy 18th Birthday! Measuring Birth Date Disclosure on Twitter."* Poster Presention at 6th International Conference on Computational Social Science (IC^2S^2'20). 2020. [84]

1.5 Outreach

- *"Korumalı Twitter hesabına sahip olmak kullanıcılara tam bir koruma sağlayamıyor"* (tr. Protected Twitter accounts do not offer full protection to users) covered by Yılmaz Yeniler, published in Boğaziçi University News Portal about the paper titled "Analysing Privacy Leakage of Life Events on Twitter." August 2020 [83].

- *"Birthday wishes inadvertently give away private information online"* covered by Matthew Sparkes, published in NewScientist about the paper titled "From an Authentication Question to a Public Social Event: Characterizing Birthday Sharing on Twitter." January 2022 [85].

Chapter 2

Background and Related Work

This chapter provides literature review about information flow management and privacy in online social media. It starts with a background on different privacy definitions over time with a focus on social media. Then surveys studies around user awareness of the potential reach of social media posts. Addition of the users' networks further complicates the information diffusion and controlling it to the information owner's desires becomes harder. Hence, literature review around networked privacy is given, followed by the protection strategies users employ to control the reach of their information. Lastly, works around privacy settings understanding and utilization are discussed. The chapter is concluded by giving a background on the current state of Twitter features and functionalities to contextualize the studies conducted in this thesis better.

2.1 Privacy Definitions

Defining privacy is not a trivial matter since it has implications for various domains ranging from everyday life to legal issues. Boundaries of privacy are not agreed upon [109]. The perception of privacy also changes throughout the time and with the introduction of new technologies. One of the early definitions of privacy as a legal concept is "The right to be let alone" by Brandeis and Warren [24]. Posner [125] interpreted this as selective concealment rather than total seclusion. Information that could be used against a person is desired to be hidden.

Altman [10] describes privacy as a dynamic process where people select what part of themselves they want to show, when they want to show it, and who they want to show it. More privacy does not always translate to a better case. Every person has a desired level of privacy and an actual one. People strive for the ideal level of privacy

where the desired and actual levels are equal. Privacy is also not independent from cultural influences. According to Altman, privacy can be considered in individual and collective levels.

Taking Altman's privacy regulation theory as a basis, Petronio developed Communication Privacy Management (CPM) Theory [123] regarding the management of private information, both individually and collectively. CPM proposes that people use privacy rules when regulating their privacy as well as privacy of others. People believe they own their information and they should control the flow of it. Once shared, the information becomes co-owned with the selected audience. If the privacy rules of the individual and co-owners do not match, then turbulence happens where the information flow is not controlled as the owner's preference.

Along with the introduction of the internet and social networks, privacy theories are expanded to contextualize these online spaces. Aside from the applications to interpersonal relations offline, Petronio's CPM theory is also applied to computer-mediated communication such as blogging and online social media. Child et al. [31] developed a privacy management measure for bloggers using CPM as a foundation. Jin [73] investigates the information disclosure and withholding motivations in the e-health context with a CPM lens. Child and Westermann [32] also utilizes CPM theory to analyze how young adults manage parental friending requests on Facebook.

This thesis looks at the information flow that might not be intended by the owner and co-owners on social media. It also investigates the utilization of privacy settings provided by the platforms to apply personal privacy rules. People want to take their intrinsic privacy rules and apply them to the online social media posts. However, it is difficult for users to manage the information flow as they desire. One of the reasons is understanding the reach and managing the audience of their posts. In the next section, we focus on the perceived audience of social media posts and the user awareness of the potential reach of these posts.

2.2 Awareness of Reach on Social Media

Online Social Network (OSN) users want to both share information and control its reach. However, given that users underestimate audience size [17], do not fully understand the visibility to third parties [87], and have difficulty understanding that information shared online can result in other types of information being inferred [3], an argument could be made that truly controlling information flow is quite challenging

for an OSN user. Bernstein et al. [17] looked at the true audience reach of 220,000 Facebook users as well as surveyed users about their perceived audience. They found that the imagined size of the audience was only 27% of the true size.

Acquisti and Gross [3] surveyed 294 participants to investigate their awareness about the platform, information sharing, and privacy attitudes of members and non-members. This study was conducted in the early days of Facebook where the platform was only open to college and high school communities. While majority of the Facebook users in the survey correctly guessed their profile visibility, there was a significant minority who underestimated the reach of it. They also found that most of their participants incorrectly thought Facebook did not collect and combine information about them from other sources, or share them with third parties.

King et al. [87] conducted a survey with 516 Facebook users to investigate their understanding around the access of third-party applications to their profile information. They found that 42% of the participants was wrong or unsure about the meaning of "pulling" information by applications. Most of the participants also did not realize that the applications they downloaded gain access to the information of their networks too. This suggests that they are also not aware that their information could be reached when their networks download the application.

OSN users' privacy awareness has increased from the early days of Facebook [152], particularly in regards to their understanding of the visibility of their data to the public and the visibility to their connections. However, even with these improvements, people still struggle to understand how broad the reach of their posts are. Understanding of the visibility of content to companies, as opposed to people, is also somewhat challenging for users. This situation makes some sense considering that the majority of privacy settings offered by companies, like Facebook, control the visibility of posts to other people, with only a small subset of these settings are about controlling the usage by the company itself.

2.3 Networked Privacy

Even when a user is perfectly aware of the reach of their posts, their connections can leak unwanted information. Networked nature of the online social media requires connections of the user to understand the implications of their interactions for privacy protection. These connections over social media cause protecting privacy to be a collective work [21, 112]. An example would be a photo which is uploaded by a user

onto an OSN and then tagged with the people in it. Each member of this collective (photo taker, photo subjects, event space owner) has some privacy stake in the photo and therefore its management is a collective issue. However, most OSNs give the right to manage the privacy settings of a content only to the uploader. While this enables users to control their self-disclosures to some degree, they have no say over what others share about them. Connections can share/re-share posts about users, disclosing information that is not easily controlled by the user [13, 77].

Marwick and boyd [112] conducted 166 semi-structured interviews with teenagers to understand how they navigate the increasingly networked online social spaces. They found that teenagers control the flow of their information by employing both online strategies, e.g. privacy settings, and offline ones, e.g. socially negotiating what can be done with their information. Privacy controls provided by platforms assume individual ownership on the information where there could be multiple stake-holders. Teenagers try to manage this shortcoming by trusting their networks to uphold social norms. Marwick and boyd argue that the focus of privacy understanding should shift from individualistic to collectivist perspective.

Trusting privacy protection to the users' network might not be sufficient [129] and disagreements over what is private can also lead to unwanted disclosures [75]. Priming the network might even backfire, leading them to share more [11]. Amon et al. [11] conducted an online experiment with 379 participants to understand photo sharing decisions of users with respect to different scenarios. The participants were shown 98 pictures with different contexts and accompanying short text explaining it. All of the pictures had people in them and participants were divided into three groups and were asked if they would share the photo (1) without any relation to the people depicted (baseline), (2) if they were in the photo, and (3) without any relation to the people depicted but prompted to think of the privacy of them. Surprisingly, they found that sharing likelihood increased compared to the baseline condition when participants were primed about privacy. Follow-up study they conducted to investigate this result also confirmed the finding.

Some users take drastic measures and deactivate/delete their accounts to protect themselves from getting tagged by others [15]. However, this does not prevent creation of shadow profiles where information about the user is shared by their networks. Analysing a user's connection network can disclose attributes about them such as age, gender, location, political orientation, and sexual orientation [7, 176, 70, 8]. Jurgens et al. [77] showed that an analysis of tweets mentioning a user is enough to deter-

mine their gender, age, religion, diet and personality traits. Magdy et al. [106] inferred users' gender and age with 92% accuracy from their network interaction and comments, which allowed them to spot "fake" accounts that might be used for catfishing on adult social networks. Similarly, Garcia et al. [51] found that using networks of people on Twitter, allows detecting the physical location of a user with median error of 68.7km, and identify the city the user lives in with 32% accuracy.

Unlike offline interactions, there is a certain permanence to online posts and interactions. Hence, users might want to edit or delete some of their posts [174]. However, some parts of the interactions can stay in the platform and leak information. For example, Twitter keeps the replies to a protected/deleted tweet visible in the platform. Mondal et al. [114] quantitatively investigated the longevity of information shared over Twitter. They curated a set of tweets sent over the six years on the platform and analyzed their visibility status. They found that 28% of the older tweets were withdrawn from the public by either tweet deletion, account deletion, or account protection. The withdrawal rate usually decreased for more recent tweets. They also found that the contents of the withdrawn tweets can be inferred by observing the residual activity around them. Hence, deletion or changing privacy settings are not enough to protect privacy because of the networked publics.

Jia and Xu [72] commented on the lack of framework in place for designers to use when designing collective privacy management systems. They developed a survey instrument and deployed it in two surveys. The questions were on three topics: collaborative ownership management – if they discuss what to post about their interactions or who can view the post; collaborative access management – if they hide/lock/delete old posts because they reveal too much information; and collaborative extension management – if they tend to keep their interactions in the online group between themselves. They find that people do manage privacy collectively, but the effect is much stronger if the group engages in in-group information sharing resulting in individual members building trust in the group's ability to manage privacy issues. In other words, a locus of control shift to the group resulted in effective collaborative privacy management. There are some early proof-of-concept research to solve these collaborative privacy situations automatically [63, 82, 88], but they have yet to be adopted by any OSN providers.

2.4 Privacy Protection Strategies

To prevent potential privacy violations that can be caused by the reach of the posts or the networked nature of social media, people utilize various methods including self-censorship, limiting connections, creating multiple accounts for different purposes, deactivating, and so on. Lampinen et al. [91] interviewed 27 participants about online sharing behaviors and boundary regulation tactics. They divided their identified set of strategies into *preventative* and *corrective*. Preventive strategies included not sharing particular contents, or sharing content only with a targeted audience. Corrective strategies included deleting an already shared post or presenting it as a joke.

Das and Kramer [37] analyzed self-censorship on Facebook using the number of posts a user started to write but ultimately did not post. They found that 71% of the users in their dataset censored themselves at least once in the 17 days data collection period. They suggest that the reason for this self-censorship is related to the lack of clarity on who can see the posts. Sleeper et al. [140] followed a qualitative approach to study self-censorship on Facebook. They conducted a diary study followed by interviews with 18 participants to understand the reasons behind the self-censoring and the types of posts that were not shared. They found that entertainment related posts, personal opinions, and updates were some of the frequently censored topics by their participants. They also found that participants would be willing to share nearly half of the posts if there was a way to perfectly configure the intended audience.

Online content and interactions has a persistence over time that is not present in offline interactions. People can choose to delete their posts to protect their privacy. Wang et al. [169] conducted a study with 569 participants to investigate regrets on Facebook. They found that one of the main reasons for regret is unintended audiences for the posts. Deletion and untagging unwanted photos were some of the actions taken to handle these regrets. Unlike self-censorship, which happens before a post is shared, deletion happens after the post is uploaded. Even if these posts are deleted, they can leave interactions on the platforms that could be used to infer the content [114].

Johnson et al. [74] conducted a study with 260 Facebook users to understand their privacy concerns and protection strategies. They found that participants used custom lists with different privacy settings, deleted their posts, and created multiple accounts to protect their privacy. Their participants also managed their networks by either denying a connection request or breaking the already established connection by deleting them.

Other studies also found that people are using multiple accounts to protect their

privacy [163, 146, 172]. Stutzman and Hartzog [146] interviewed 20 participants who have multiple accounts to understand the motivations for the practice. Privacy was one of the four main motivations for creating multiple accounts. Their participants cited wanting to separate their personal and professional image. Some of them felt safer by maintaining a protected environment and sharing personal information there. Xiao et al. [172] found that Instagram users create alternative accounts, i.e."finsta", to create a private space for themselves to be freer and more intimate without the pressure of maintaining their social image.

People sometimes take drastic measures such as deactivating [22] or stopping usage completely [15, 55, 90] to protect their privacy. boyd and Marwick [22] found that users deactivate their accounts to limit their interactions especially when they cannot immediately monitor them. Baumer et al. [15] surveyed 400 participants who decided to leave Facebook and their motivations for the decision. 127 of their participants reported they deactivated or deleted their accounts. More than quarter of the participants cited privacy concerns as a motivation to stop using Facebook.

2.5 Usage of Privacy Settings

In addition to the strategies mentioned in the previous section, people utilize privacy configurations given by platforms to protect their privacy online too [112]. Social network platforms provide privacy settings for users to adjust a range of configurations including setting the audience of their posts, regulating interactions, changing the visibility of the information they share, and so on. Ellison et al. [45] investigated the relation between privacy and social capital which included the usage of privacy settings. They conducted two studies; a user survey with 299 university students and an interview study with 18 adult Facebook users. They found that 78% of the survey participants used provided privacy settings to manage the audience of their posts. Interviews also confirmed that privacy settings were used by participants. There were also some participants, especially older adults, who were not familiar with these settings and not sure how to use them.

Effective utilization of these privacy settings may be achieved by increasing general internet skills [27] and privacy literacy [14]. However, users find these privacy settings cumbersome [155], either preferring to stick to defaults or setting them only once at the beginning. Fiesler et al. [48] collect nearly 11K Facebook posts of 1,815 users and found that 37% of the posts were left in the default setting provided by the platform.

Similarly, Liu et al. [100] found that default settings were not changed for 36% of the content they collected to investigate privacy settings. Strater et al. [145] interviewed 18 Facebook users about the personal information they decide to share on the platform and the reasons behind the disclosures. They also analyzed the privacy settings usage and found that users often configured the privacy settings of their posts when creating their accounts and did not change them after that. Some users choose not to change their settings even when there were posts shown to be violating their privacy [104, 115]. Madejski et al. [105] measured the discrepancies between privacy settings and the intentions of the users on Facebook. They created a Facebook application to conduct a study with 65 participants. They showed participants posts where the privacy settings did not match their intended audiences and asked them whether they want to change the settings. They found that 42% of the participants chose to not take action for any of the shown posts. Nearly all of the participants had one such post where they wanted to keep as it is.

Configuring privacy settings to accurately reflect the user's intended audience is not trivial [100, 105, 62]. Privacy violations in social media may be caused by unclear permissions or people not knowing how social media permissions work correctly. Liu et al. [100] analyzed the privacy settings set by Facebook users and compared them to the desired settings to understand the discrepancies. They recruited 200 participants and collected over 116K content items they shared on the platform. They found that majority of the time users' expectations did not reflect the actual settings of their posts and the size of the audience was mostly larger than expected. This discrepancy persisted even with the exclusion of content with default settings, showing users who actively utilize the settings are also finding them difficult to configure correctly. Madejski et al. [105] also found that every participant reported at least one sharing violation in the previously mentioned study they conducted to investigate privacy settings and intentions. Hoyle et al. [62] studied LinkedIn users' understanding of "viewed by" feature including what controls they had available around it. They found that most people do not understand how the permission works. These violations can result in unintended consequences, users may lose their job [169], or insurance companies can increase their premiums [108].

Context collapse where users' different social circles, such as friends, family, and colleagues, are all on the same social network [119, 111, 161] makes it hard for users to control the flow of information to different social circles. The temporal persistence of social media posts further complicates the situation [25, 64]. Intended audiences

Figure 2.1: Public profile information that can be seen by anyone even when the account is protected.

can also change overtime, e.g. a user regretting their decision to post [141]. Hence, it is immensely difficult to configure privacy settings correctly and satisfy the privacy rules of individuals with all these factors in mind.

2.6 Twitter Functionality

This thesis focuses on Twitter as a platform, and in this section we introduce the current features of Twitter, as well as the Twitter API (Application Programming Interface) endpoints and objects used in the studies. Social network platforms are dynamic systems; they add new features and update existing ones frequently. Hence, the snapshot of the current features and settings is needed to understand the studies in this thesis. Following information gives the state of Twitter when the last study was conducted (Summer 2021). There are few changes already such as the ability to remove followers which is only added in Autumn 2021.

Twitter is a social platform where people can share content with each other using *tweets* which are short 280 character text statements that can contain links to other content, such as news sites, videos, and pictures. Every user has a unique username and a public profile that contains information about the user. The user can select which information they want to share. Profiles can have header photo, profile photo, a short

Figure 2.2: A tweet example from Twitter.

bio, location, website, and birth date information. Aside from the birth date, all of these data points are always public. The users can choose who can see the birth date or parts of it. Figure 2.1 shows the profile of an account. This account has profile and header photos, as well as bio, location, and website information.

Users can follow others and get followers but this relation is not necessarily bidirectional. Twitter supports explicitly mentioning other accounts in tweets using "@" symbol before the name of the account (i.e. "@username"). Users can interact with tweets by replying, liking, retweeting, or quote tweeting them. They can also block other users and mute words to curate their timelines. Users can follow Topics, create Lists of other users, and create Spaces to have voice-based conversation with other users. Following are some description of the functionalities referred in this thesis. Figure 2.2 shows a tweet example with the main components annotated.

Timeline: This is the main page of the service where the user can traverse around the tweets and retweets of the users they follow. Recently, Twitter introduced a new timeline where sometimes liked tweets from the followed accounts are also shown. Timeline can also include promoted tweets and suggested tweets from topics the user follows or might want to follow.

Retweet: A user can forward tweets to their followers as it is. The tweet will retain the original sender. This tweet can be seen in the profile page of the user too.

Quote tweet: This is similar to retweets with only difference is the added commentary to the original tweet.

Like: This is a way to indicate that the user likes the tweet.

Reply: A user can reply to another user's tweet and users can create conversations by replying to each other.

Mention: A user can mention another by using their username "@username" and the mentioned user will get a notification. It is important to note that the replies are also mentions. If a mention is at the start of a tweet then that tweet usually will not be shown in the followers' timelines. However, tweets with mentions any other place in the tweet will be shown in the timelines.

Topics: When a user follows a Topic, the tweets related to the topic will be shown to the user. User can also traverse the topic tweets by going to the topic pages. It is a way to gather set of tweets without having to follow everyone who posts about the topic.

Lists: Lists are similar to topics in a way that a user can create lists without following individual users. Lists have their own timelines with the added members' tweets. Anyone can add public accounts to a list and only way for a user to remove themselves from a list is blocking the list owner.

Mute: A user can mute other users or specific keywords to prevent seeing them in their timelines. If a user mutes a person they follow, the user will only get the direct interaction notifications from the muted account (e.g. replies to their tweets). Muted person will still be followed.

Block: A user can block another from following them, seeing their tweets, and adding them to lists. Blocked accounts can mention the user but no notification will be sent.

Twitter has a binary privacy configuration where accounts are either public (default) or protected. Figure 2.3 shows the settings page to change tweet visibility for the current design of Twitter. A *protected* account has a small padlock near the display name of the account (e.g. as seen in Figure 2.1). All tweets associated with a *public* account can be seen by anyone on the Internet, while all tweets associated with a *protected* account can only be seen by the followers of that account. Changing an account's setting changes the visibility of all tweets associated with this account, including past tweets. Any Twitter user can follow a *public* account, while following a *protected* one first requires the account owner's approval. A *protected* account's tweets

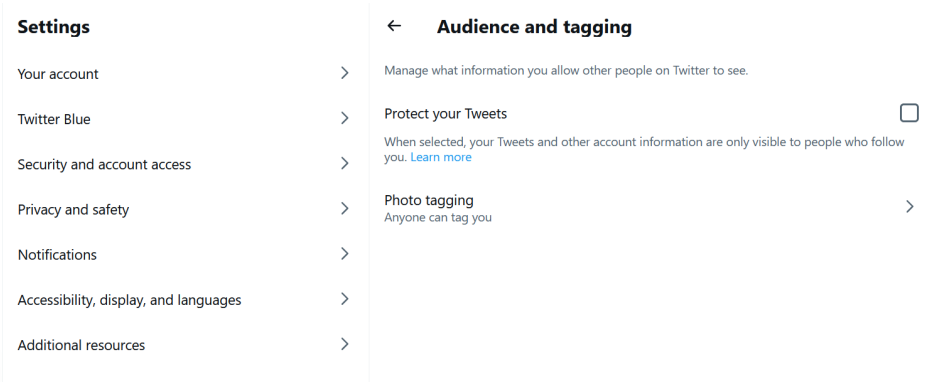

Figure 2.3: Settings to change tweet visibility.

can be liked by the followers but they cannot be retweeted or quote tweeted. Interactions of *protected* accounts can only be seen by the followers. Hence, if these *protected* accounts reply to a public tweet of a non-follower, that will not be visible to that non-follower. This also applies to mentioning users by using their handles (@username), quote tweeting them, as well as liking and retweeting them. Hence, *protected* accounts must change to *public* if they want their interactions with non-followers to be seen, but doing so also makes their full tweet history public. On the other hand, if a *public* account mentions a *protected* one, anyone on the Internet can see the content since the owner of the tweet is *public*.

2.6.1 Twitter API

Twitter provides an API to developers, researchers, and companies. Twitter API V1 [154] was used to collect Twitter data for the studies in this thesis. This API provides endpoints to collect public tweets in real-time with filtering options according to keywords, language, location, and so on. The collected tweets have various attributes including status text, the user who sent it, location, whether if the tweet is a reply to another tweet or not, mentioned users, shared pictures, number of likes/retweets, and so on. Public profile information about users can also be collected regardless of the users' account type. Two main API objects collected in this thesis are tweets and users.

2.6.1.1 Tweet Object

Following are some description of the fields of the tweet object commonly used in the thesis.

id: Unique identifier of the tweet. It is given both in integer and string forms.

text: Main text of the tweet. It is in UTF-8. Sometimes can be truncated. In that case, the *full_text* in the *extended_tweet* field must be retrieved.

user: The user object of the tweet owner.

in_reply_to_status_id: The id of the replied tweet if the collected tweet is a reply.

in_reply_to_user_id: The id of the replied tweet's user if the collected tweet is a reply.

is_quote_status: Boolean value to indicate whether the collected tweet is a quote tweet.

entities: The users mentioned in the tweet, as well as hashtags, urls, and media shared in the tweet.

lang: Machine-detected language of the tweet.

2.6.1.2 User Object

Following are some description of the fields of the user object commonly used in the thesis.

id: Unique identifier of the user. It is given both in integer and string forms.

screen_name: The username of the user. These can be changed while user id stays the same.

location: The location information shared on profile.

url: The website information shared on profile.

description: The bio information shared on profile.

protected: Boolean value to indicate whether the account is protected or not.

verified: Boolean value to indicate whether the account is verified by Twitter or not.

Chapter 3

Unintended Privacy Leaks by Networks

3.1 Overview

In this thesis, we study unintended privacy disclosures by users and their networks. We particularly focus on information leak by interactions (RQ1) and the user reactions to such leaks on Twitter. We then continue onto investigating the privacy settings usage (RQ3) and information visibility understanding (RQ4). In this chapter, we focus on our first two research questions: (RQ1) "What kind of personal information are disclosed online by networks on Twitter?" and (RQ2) "How do users react (e.g. like, retweet or reply) to tweets that share their personal information when they are mentioned in them and how comfortable are they with such information disclosure?". To do so, we focus on finding happy life events shared on Twitter by collecting congratulatory tweets with user mentions and users' reactions to these tweets.

Users can *mention* each other in tweets (e.g. "@username"), which makes it easy to link tweet content to specific accounts. While simple, Twitter's privacy model can lead to public accounts having conversations with protected accounts where half the conversation is world readable. For example, a protected account might say "its my birthday!" and the public account might respond "happy birthday @alice!" disclosing the protected account's birthday publicly. Using Twitter, we can identify users' life events at scale, potentially even for protected accounts.

In particular, we are interested in tweets where the poster says that they are happy for an explicitly mentioned user regarding a life event. We ask the following research questions:

RQ1.1 What kind of life events can be detected in tweets that express happiness for a mentioned user?

RQ2.1 How do users react (e.g. like, retweet or reply) to such tweets when they are mentioned in them?

RQ2.2 How do protected (private) accounts react (e.g. like, retweet) relative to public accounts in these cases?

To answer these questions, we collected 1.4 million tweets/retweets between July and October 2019 containing the exact phrase "happy for you". We removed all posts involving *verified* accounts, as these are held by famous people or organizations and tend to have a large number of Twitter users discussing their life events, which would likely skew the data. We also removed retweets, and tweets that mention no users, resulting in 635k tweets. We then used Latent Dirichlet Allocation (LDA) [18] to detect topics in the dataset, resulting in 12 identified life events topics; including positive events like having a new baby, marriage, and graduation, as well as sensitive topics such as cancer, surgery, and mental health. However, as expected, not all tweets in our corpus corresponded to life events. Out of the original 635k tweets, only 59k belonged to a life event related topic. Looking only at the life event tweets, 51k mention only a single user, providing a clear indication of who is experiencing the life event. Out of these 51k tweets, 4k mention a single protected user, potentially breaching that user's privacy by making their life event public. The majority of protected and public account users reacted to the life event tweet mentioning them.

Our results suggest that it is possible to automatically identify Twitter users' key life events, even if they have a protected account. The outcome has implications for privacy, particularly around the impact of the sharing decisions of connections.

3.2 Related Work

3.2.1 Harms

Disclosing events like vacations and illnesses can have unwanted results. Vacations may signal that the tweeter's home is vacant and burglars can use this information (e.g. `PleaseRobMe.com`). Sharing events that include high-risk activities like sky-diving may result in increased premiums by insurance companies [137] or sharing the events attended may jeopardize the long-term disability benefits of a person [102].

Mao et al. [108] looked for tweets with vacation, illness and drinking topics by using keyword-based data filtering. They classify these tweets as sensitive or non-sensitive using Naive Bayes with bag-of-word model.

3.2.2 Detecting and inferring life events

Detecting life events using social media posts is not a new research area. Researchers have looked at a range of feature sets, types of life events, and approaches in an effort to accurately automatically identify these events from messy social media data.

Simple keyword queries were tried by De Choudhury et al. [38] to find women who were new mothers. They used keywords curated from birth announcements of local newspapers to find accounts of potential new mothers, then used lexicon-driven gender inference to identify women (as opposed to new fathers), with a 83% accuracy rate. Finally, they used crowdsourcing to label the accounts as new mothers or not, to have a high precision dataset.

Other works use keywords to gather a life event themed corpus, crowdsourcing to annotate it, and then use the annotated data to build a model that can automatically associate tweets with a life event. Dickinson et al. [42], focused on five life events psychologists have identified to be the most prominent in peoples' lives: "Starting School", "Falling in Love", "Getting Married", "Having Children", and "Death of a Parent". They were able to use the content features of tweets such as n-grams, mentions, and number of retweets, user, semantic, and interaction features to build an effective classifier using labeled data from crowdsourcing. Similarly, Akbari et al. [5] focused on personal wellness events (diet, exercise, and health). They used keywords to collect tweets from Twitter, manually labeled them, and built a classifier.

Instead of starting with a specific set of life events, some research starts with a very broad corpus and identifies the life events that exist within it. Li et al. [94] collected tweets using the broad keywords "congratulations" and "condolences" and used LDA [18] to find topics in the data set from which they focused on the life event topics. In their approach, they start with tweets identified through keyword matches, then look for any parent tweets, combining the parents and children into one document of only verbs and nouns. They used bootstrapping to find phrases other than "congratulations" and "condolences" used in the tweets such as "have fun" and "my deepest condolences" to expand their dataset. They repeated this process for four times and found 30 different phrases alongside with 42 event types.

Our work also uses broad keywords to gather an interesting corpus and then identify the life events found in it. However, rather than focus on general posts, we attempt to identify life events posted by people other than those experiencing the event. This focus allows us to look not only at public users, but also get a sense of the life events experienced by protected accounts.

3.2.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation is an unsupervised probabilistic topic modelling algorithm. Given a set of documents, the LDA model aims to find underlying "topics" and the distribution of probability for each topic in each document. LDA works with the assumption that each document is a mixture of topics and each topic is a mixture of words. LDA considers documents as "bag of words" where the order of the words are not important in the model. The number of topics is must be given to the model for it to work.

Aside from fields such as news [28], literature mining [80], and bioinformatics [99], LDA also used to extract topics from Twitter too. Prior work applies LDA to tweets in various research domains such as misinformation [165], mental health [113], and politics [79]. As mentioned in the last section, LDA is also used to extract life events from Twitter [94], which is the approach we are taking in this chapter. We used gensim's ldamallet [134] implementation to apply LDA to our data.

3.3 Detecting Life Events From Tweets

For this study, we collected tweets that contain the phrase "happy for you". Those tweets were then divided according to the life event discussed in them. We analyzed these tweets to understand whether they are mentioning protected or public users. We describe our data collection, analysis and topic modeling and assignment below.

3.3.1 Tweets Collection

Using the Twitter streaming API [154], we collected tweets that contain the words "happy for you" resulting in all tweets that have these three words but not necessarily in consecutive order. So, we filtered the tweets to only those that have the exact phrase.

We streamed tweets for four months between July and October 2019. A set of 1.4 million tweets/retweets were collected that contain the phrase "happy for you". Mul-

tiple filtering steps were then applied to the collected tweets. Initially, we filtered out all retweets, since we are only interested in the original tweet. In addition, we filtered out tweets that have no explicit mentioned accounts, since we are only interested in the tweets directed to specific users indicating they were about them. We then checked the type of the mentioned accounts within our tweets and removed any tweet mentioning verified accounts which indicates that the event is probably about a famous person, and thus privacy is less of a concern. After all these filtering steps, we had 634,590 tweets mentioning a total of 777K Twitter accounts. We refer to this dataset as "*HFY*" tweets dataset.

We divided the tweets according to the number of accounts mentioned in them and the conversation type of the tweets. We call the tweets that mention only one account *single* tweets, and the ones who mention more than one account is called *multiple* tweets. A tweet that mention another user has one of the following conversation types; a *reply* to an existing tweet, *directed to a user*, or *other*. *Reply* tweets are direct replies to another existing parent tweet; *directed to a user* tweets are not replies but they mention another account at the start of the tweet (e.g. "@username heard about the new addition to the family, I'm happy for you"). *Other* comprises all the tweets that mention at least one user but at the middle or end of the tweet. We use these terms throughout this chapter. The majority of the tweets in *HFY* (607,703 tweets, 96%) are replies to another parent tweet.

3.3.2 Finding Life Events

Since all our tweets have the phrase "happy for you" for a given mentioned user, our next task was to infer the event that the mentioned users are being congratulated for. Manually checking each tweet is not feasible, since our *HFY* dataset contains 635k tweets. Hence, we used Latent Dirichlet Allocation (LDA) topic modeling [18] to divide tweets into topics. By setting an input n as a suggested number of topics, LDA modeling assumes that each document in the corpus is a mixture of topics and each topic is a mixture of words. It uses a bag-of-words approach where the order of the words are ignored. First, we cleaned the tweets, lemmatized them and used the python implementation of gensim's ldamallet [134] to find the topic models.

Tweet Pre-processing: Given a tweet, we firstly converted it to lower case. We removed URLs, mentioned accounts ("@username"), as well as the # character in hashtags. Then we tokenized the tweets with the NLTK tweet-tokenizer [103]. After this

step, we removed the words "happy", "for" and "you", since they were common in all tweets in our collection. We also removed emoji. We used bigram-phraser provided by the gensim Python package to combine words that co-occur consecutively more than 100 times in the dataset. For example "safe travel", "speedy recovery", and "health pregnancy" are some of the bigrams we have in the dataset. Lastly, we lemmatized the tokens and removed the ones that are not nouns or verbs using spaCy [61] following the approach of Li et al. [94]. After all these steps, we stored these lemmatized tweets for use by the LDA model.

Creating the LDA Model: We firstly created a dictionary from the words where each word is represented by an ID. Then we created our corpus with each tweet represented by list of IDs created using the dictionary. We used ldamallet [134] to create our LDA model. A fixed random seed was used to be able to reproduce the model, and we experimented modeling our dataset with different number of topics n. We examined the following number of topics n=$\{10, 20, 25, 30, 40, 70, 100, 120\}$. After looking at the distribution of the topic keywords and themes with each topics number n, we decided to continue with n=100 topics, which based on manual inspection, seems to be the optimal number of topics to produce many clean groups related to the life events discussed in the tweets.

Topics Selection: For each of the 100 topics, we extracted the 30 most representative words and 10 most representative tweets, where "most representative" means the highest probability of belonging to that topic. A researcher looked through all the keywords and representative tweets and labeled topics as involving life events or not, resulting in 22 topics that involved life events. During the process, 8 topics were identified that contained a mix of life events because they had formed around an activity, such as prayer, that touches on many different life events. We therefore chose to exclude these topics. Three of the topics involved different activities associated with having a baby such as a baby shower, so we combined these three into one "having a baby" topics. The result was 12 life event topics, which are shown along with examples in Table 3.1.

Assigning Topics to the Tweets: LDA assumes that each document is a mixture of topics. Some of these topics are more probable in the document and some of them less. We assumed that each tweet only belongs to the most probable topic assigned by the LDA model. This way, we were able to divide the tweets to topics.

After all these steps, only 58,801 tweets from *HFY* were assigned to the 12 life event topics. 86.3% of these tweets in were *single*, i.e. mentioning only one account.

Topic	Keywords	Example tweet
Having a baby	family, congratulation, news, blessing, member, addition, baby, girl, mommy, daddy, pregnancy, delivery, healthy_pregnancy, motherhood, boy, shower, gender_reveal	@username I'm so happy for you! Can't wait for the baby shower!
Travel	enjoy, time, rest, trip, weekend, summer, travel, visit, vacation, relax, holiday, london, japan, flight, safe_travel, korea, europe, germany, chicago, italy	@username so happy for you enjoy your trip [..]
Relationship start	person, relationship, boyfriend, girlfriend, bf, partner, keeper, gf, long_distance	A boyfriend like this is a keeper I'm happy for you sis @username
Cancer (in family)	mom, family, sister, dad, fight, brother, cancer, die, beat, lose, stage, grandma, uncle, cousin, warrior, nephew, aunt, fighter, niece, battle, survivor, monster, treatment, grandpa	@username So happy for you! May God keep you cancer free.
Birthday	day, birthday, today, celebrate, gift, bday, present, belated_birthday	@username CONGRATS AND HAPPY BIRTHDAY IM SO HAPPY FOR YOU
Surgery	hope, continue, pray, stay, recovery, health, take_care, heal, recover, speedy_recovery, rest, improve, surgery	@username [..] hope you have ease in your recovery.
Graduation	congratulation, success, work, achievement, accomplishment, celebrate, future, earn, cheer, graduation	[..] Happy graduation @username
Mental health	deal, pain, struggle, problem, doctor, issue, anxiety, mental_health, fear, therapy, overcome, depression, surgery, brain, stress, relief, med	@username I really hope you overcome your anxiety [..]
Familial matters	parent, kid, son, child, daughter, mother, family, mom, father, dad, bear, wife, raise, age, miracle, birth, husband, awareness, sibling, carry, grandmother, adopt	@username [..] that you have a grandchild too. [..]
Marriage	congratulation, wedding, marry, wife, husband, marriage, invite, day, engage, dress, bride, honor, ring, engagement, hubby, anniversary, party, honeymoon, honour, propose, divorce, fiancé	@username [..] The wedding will be fantastic though! [..]
Moving	move, place, home, house, leave, visit, fall, room, space, settle, city, town, area, apartment, land, pack	@username Everyone needs to leave home at one point. I feel you sis. [..]
LGBTQ-related	speak, people, woman, part, trust, realize, process, lie, power, figure, faith, community, truth, accept, idea, pride, gay, doubt, gender, embrace, tran	Congratulations to my favorite lesbians![..]

Table 3.1: Life event topics from *HFY-LE* and keywords selected from the 30 most probable words for each topic. Example tweets shown with usernames blinded and some content removed for privacy.

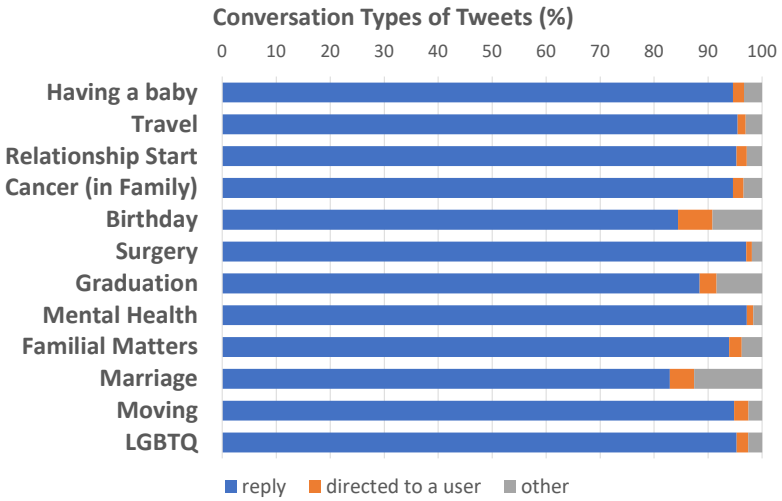

Figure 3.1: Topics divided by the conversation types; *reply*, *directed to a user*, and *other*.

In cases where only one account is mentioned, we can safely assume that the event is about that user, but for tweets with multiple mentioned users it is impossible to accurately infer who is the event subject. Therefore, we focus on *single* tweets in our analysis. We call this subset *HFY-LE* for Life Event. 8% of *HFY-LE* mention protected accounts.

3.3.3 Resulting Tweet Groups

Of all the tweets in *HFY-LE*, 47,342 (93%) were in *reply*, 2% were *directed to a user* and remaining were *other*. This shows that nearly all tweets were sent in response to an existing tweet. However, some topics received more tweets as *reply* than others. 97% of tweets with "mental health" and "surgery" topics were *reply* whereas this rate was 83% for "marriage". Tweets with sensitive topics related to health, sexual orientation, and so on were more likely to be replies to existing tweets. On the other hand, commonly celebrated things like marriage, birthday, and graduation are more frequently tweeted as a stand-alone tweet rather than in reply. The rates of the conversation types for each topic is shown in Figure 3.1.

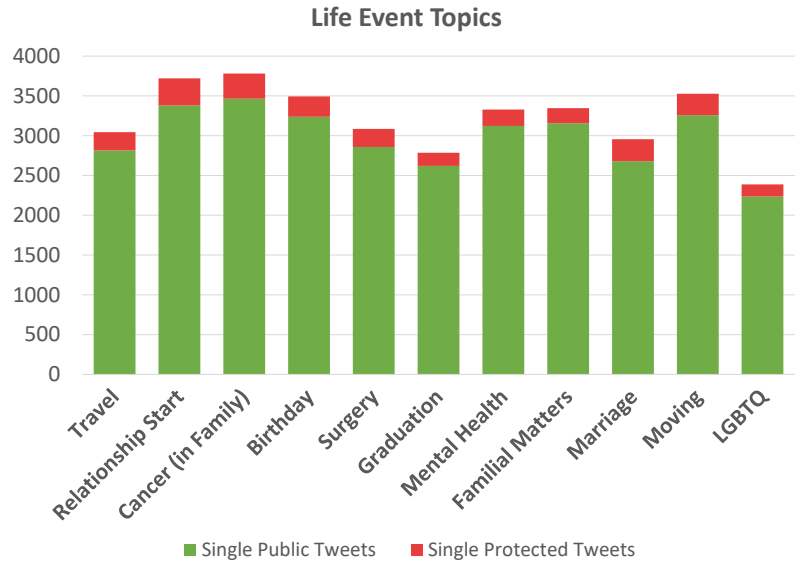

Figure 3.2: Number of tweets broken by the type of the mentioned user (“having a baby” not shown).

The largest topic is “having a baby” with 15,289 tweets since it is a combination of three groups from the original topics. The smallest topic is “LGBTQ-related” with only 2,387 tweets. In Figure 3.2, we provide the number of tweets from each topic broken by the account type of the mentioned users, we do not display “having a baby” since it is four times the second largest topic. “Having a baby” has 13,912 tweets mentioning public accounts and 1,377 mentioning protected ones. The topic with the most protected tweets is “marriage” with 9%, whereas “familial matters” tweets are the least common with 6%.

3.4 Reactions from Mentioned Users

After determining tweet topics, we collected the reactions from mentioned users such as likes, retweets, and replies to understand how users react to having their life-events disclosed by their friends.

3.4.1 Collection of Reactions

Four months after the last tweet was collected, we extracted the mentioned accounts and gathered the reactions to the tweets in the *HFY-LE*. Firstly, we tried to retrieve the reactions (like, retweet, or reply) to each tweet in *HFY-LE*. Some of these tweets were not available for various reasons, for example the tweet was deleted or the user protected their account.

3.4.1.1 Collection of Likes/Retweets

For the tweets we could reach, we checked whether the mentioned user liked or retweeted the tweet; however, the provided Twitter API retrieves this data very slowly. Hence, we used Twitter user interface (UI), which shows how many times a tweet is liked/retweeted. We also retrieved the list of who liked/retweeted via the UI. If the mentioned user's screen name is in the list, then we conclude that the user liked/retweeted the tweet. One drawback of this approach is that we can only get 25 people from the list. Hence, if the tweet is popular and has more than 25 likes/retweets, then we cannot decide whether the tweet was liked/retweeted by the mentioned user. However, this is a rare situation that happened in only 229 tweets from the 51k tweets in *HFY-LE*.

Protected accounts' tweets and likes are hidden, so we cannot see if they interacted with tweets. If a protected account likes/retweets a tweet, it will not show in the UI list but will still be counted towards the total likes/retweets. The difference between the like/retweet count and the number of users in the list indicates the number of protected accounts who liked/retweeted the tweet. Thus, we used this information to estimate if a mentioned protected account has likes/retweeted the tweet. While we cannot be sure that the protected account likes/retweeted a tweet is the mentioned one, we believe it is reasonable to assume so. The same drawback mentioned earlier also applies here, if a tweet is liked/retweeted more than 25 times we cannot be sure about the hidden like/retweet counts. We base our assumption around hidden likes/retweets belonging to the mentioned protected user on three main points; (1) tweets with mentions send notifications to the mentioned user alerting them of the tweet, (2) the tweets we analyze reactions have low interactions (i.e. not more than 25 likes/retweet), and (3) the share of protected users on Twitter is lower than the public users [101] lowering the probability of a like or retweet from a non-mentioned protected user.

3.4.1.2 Collection of Replies

Next we collected user replies to the tweet by collecting the timelines of the mentioned users that were not protected and scanning them for replies. Since Twitter API did not have a feature that gives the replies to a tweet at the time of the study, we had to get the timelines of each mentioned user to check whether there is a reply to the tweet mentioned in them in *HFY-LE*. We check every tweet of the mentioned user between the time of the original tweet and the response collection time. The Twitter API only allows us to get the last 3200 tweets from a user's timeline. Hence, for some of the users we were not able to decide whether they replied or not since they tweeted more than 3200 after the original tweet was sent.

We could not apply the same method for protected accounts, since their timeline is inaccessible. Thus, reactions of protected accounts by replying to tweets mentioned them is unfortunately not included in our analysis. Similarly, we could not retrieve tweets from users who were suspended or deleted their accounts or were unreachable for other reasons.

3.4.2 Analysing Users' Reaction

When collecting likes and retweets, we found that 5,910 (12%) of the tweets could no longer be viewed. However, since we had the ID and the mentions of the tweet, we could still check if there was a reply from the mentioned users. We couldn't reach the tweets from each topic with similar rates; between 10% ("moving") and 13% ("familial matters"). On the other hand, aside from the 4,005 protected accounts, we couldn't reach further 3,084 (7%) mentioned users to collect reactions. These rates are between 4% and 8% for the most of the topics, while the rate for "surgery" is 13%. This might mean that these users delete their profiles more than other mentioned users in other topics.

From the ones we could reach, 24,047 (62%) of the tweets were liked by public mentioned uses. Similarly, 2,221 (6%) of the tweets were retweeted by the mentioned user. On the other hand, 2,319 (68%) of the tweets that mentioned a protected account had hidden likes and 203 (6%) of them had hidden retweets. While we cannot be sure all of these hidden likes/retweets were from the mentioned users, it gives us some idea about the interactions. 11,941 (42%) of the users with public accounts replied to the tweets they were mentioned in. The average time to reply was 5 hours while the longest was just over three months. In total 27,545 of the mentioned users showed at

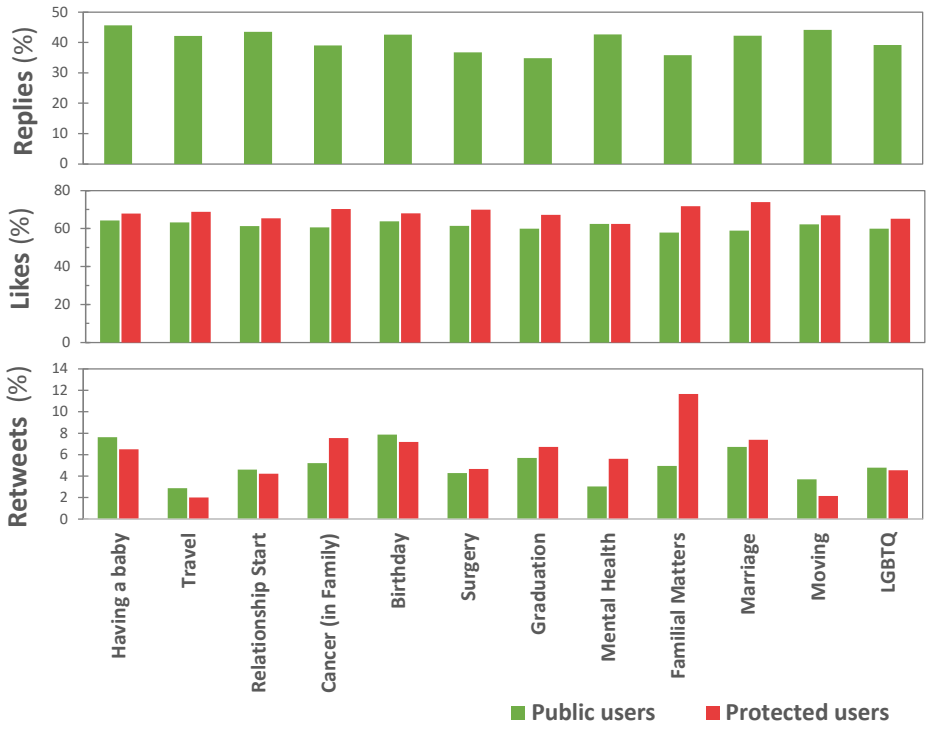

Figure 3.3: Reactions (reply, like, and retweet) by the mentioned users in the *HFY-LE* tweets.

least one type of reaction. 941 of them reacted using all three ways to the tweets (i.e. like, retweet, and reply). 7,256 did not give any reaction.

Figure 3.3 shows the reactions from the mentioned user for each topic. All of the topics had similar rates of likes from the mentioned user. “Familial Matters”, “marriage”, and “LGBTQ-related” topics were on the lower end while “having a baby”, “ birthday”, and “travel” were on the higher end. For retweets, “travel” and ”mental health” were on the lower end while “marriage” and “having a baby” on the higher end.

In every topic, the rate of likes from the protected users are more than the like rates of the public users. As mentioned, this may be a result of counting every tweet that mentions a protected account with hidden likes as a reaction. Still they show similar patterns with the public tweets.

3.5 Discussion

The purpose of this study is to investigate our main thesis’ first two research questions around personal information disclosure by networks and the reactions towards

these disclosures. In this chapter, we particularly focus on how users' privacy can be breached from social media posts by people happy about their life-event. To measure this, we used the general phrase "happy for you" to collect potential tweets that might communicate with other users about happy events in their life. We managed to collect a large number of tweets mentioning users, including protected accounts. Using LDA, we inferred the discussed life-event in around 10% of the collected tweets. We managed to identify 12 life-events that we included in our analysis. This result relates to our first research question, where life-events about new born baby, marriage, graduation, and mental health could be inferred from the tweets. Investigating our second and third research questions, it was interesting to find the most users react positively to tweets that disclose their life-events. More surprisingly, we noticed that tweets were liked more often by the protected accounts, who could be assumed to have more privacy concerns but may consider their *protected* status sufficient protection.

3.5.1 Implications of Findings

The stated purpose of online social media is to help people connect with one another through a shared medium. It is therefore unsurprising that users would use such platforms to share life events since sharing is a common way people build social groups. However, the highly public nature of Twitter also means that information shared is open to a world-wide audience, a fact that may be technically known by users, but hard for them to conceptualize since many people believe that they themselves are not sufficiently interesting for an attacker to bother with [170]. But this view does not necessarily match how groups use data at scale.

The life events we identified are similar to those Janssen and Rubin [69] found when asking adults from Netherlands about the seven most important events that will happen in an ordinary Dutch child's life. "Having children" is the most frequently mentioned life event by Dutch people which is similar to our results where the largest topic was "having a baby". Our other topic also line up well with their results, indicating that our identified "happy for you" life event topics do align with common important life events, suggesting that such data can be automatically extracted from Twitter.

Life events are big money for companies. Target famously hired analysts to predict which of their customers were pregnant so they could market to them before the birth announcement because new parents tend to purchase large quantities of items [43]. In

our topics we identified multiple tweets from marketers who were targeting people experiencing life events. For example, several wedding planning groups were replying to user tweets where they talk about their engagement or upcoming wedding. A Twitter user posted about getting their venues booked for an upcoming wedding, resulting in the reply: *"@username We're SO happy for you, Kayla! Can we help with planning? [url] Plan with this Free Sample Kit: [url]"*. The threat of life events being automatically identified off of Twitter and used to target users is very real and currently used threat vector.

Life events are not just a privacy or marketing problem. They are also useful for attackers who want to cause harm or steal financial assets. Targeted attacks on valuable people, sometimes called *whaling*, often start with the attacker spending time on company people pages and friending them on social media. The attacker can then use the online trove of personal information as part of social engineering to trick a user or system into providing more valuable access. For example, what some companies consider to be public non-sensitive data is used by other companies for authentication, which means that an attacker can start with seemingly low sensitive data and work their way to remotely resetting a Wired journalist's Mac [60].

Personal privacy management is also challenging on Twitter due to the ease of creating *shadow* profiles where information about a user is available via other peoples' tweets. For example, one friend may tweet "happy birthday @ProtectedUser" creating a public shadow record of the protected user's birthday, then another public account tweets "looking forward to our trip" again creating public data. Together these public tweets create a shadow profile of the protected user which they cannot easily control. The design of Twitter also facilitates the creating of shadow profiles through the use of course access control at profile level and the culture of replying and retweeting posts to followers. While only 8% of the life event tweets were mentioning a protected account, that still amounts to 4k users having their life events exposed. Our research also indicates that the types of life events exposed for protected and public accounts are very similar, suggesting that public posters are equally willing to post about a range of life events for both protected and public accounts.

While sharing and long term existence of information is a problem, we also noticed the opposite where tweets vanished and users changed privacy settings between the time when we collected them and when we went back to get reactions. 12% of tweets containing a life event vanished between the collections, with each topic having roughly the same percentage of tweets vanish. More interestingly, 6.6% of mentioned

accounts were deleted or suspended in that time frame, with 12.8% of accounts mentioned in surgery life events vanishing. This observation suggests that people are taking some actions that protect theirs and others' privacy.

3.5.2 Limitations

By using Twitter we were able to get a large sample of social media posts to work from, but our data set and analysis still have some important limitations to consider. Our analysis is limited to tweets containing the phrase "happy for you". While this is useful to collect positive life events, it likely has few examples of common negative life events. It also focuses our analysis on a life event that someone else might want to provide positive commentary on.

To group the tweets we applied LDA which uses word co-occurrence at the document level to discover topics. However, tweets are very short documents which may inhibit LDA from performing as well as it does with longer documents. LDA also requires us to state the number of topics in advance, which is obviously not a known number. We used the standard approach of selecting several possible topic numbers, running LDA with each setting, and manually checking the coherence of the resulting topics. We also assume that each tweet has only one topic and therefore assign each only to the most probable topic. While we did read through many tweets during this process, we did not attempt to manually label tweets. The labels of the life event topics were assigned by only one researcher.

We gathered the reactions from the mentioned users four months after the last tweet was collected. The time delay meant that we could accurately collect reactions, but it also meant that some tweets and accounts vanished. Also in some cases where the mentioned users were very active we were not able to retrieve their older tweets because the Twitter API limits per-user tweet retrieval to 3,200 tweets.

While collecting the likes and retweets from the mentioned users, we use the hidden likes and retweets as an indicator of interaction from the mentioned protected user. We assume that if a tweet mentions a protected account and has a hidden interaction, this hidden interaction is from the mentioned user. While this may give us an idea about possible interactions, we cannot be sure whether the hidden like/retweet is actually by the mentioned user.

3.6 Summary

In this chapter, we investigated our first two main research questions about disclosures of personal information from users' networks and the user reactions. To do so, we collected 635k tweets containing the phrase "happy for you" that mention at least one user. We used LDA topic modeling to group the tweets, resulting in 12 life event topics with 51k of *single* tweets belonging to one of these topics. "Having a baby" was the largest topic while "LGBT-related" was the smallest. 8% of the tweets mention protected users and the rate of protected user mentions in topics ranged between 6% and 9%.

The majority of tweets received reactions from mentioned users. The most common reaction was liking the tweet, followed by replying. Retweeting was the least common reaction. The rates for likes/retweets/replies were fairly consistent between topics. Protected accounts tended to like the tweets that mention them more often than public accounts with no major variation between topics.

A user can protect their own tweets but the tweets that mention them can only be controlled by the tweeter. In addition, tweets from public accounts replying to protected account tweets can be seen by anyone and even if the parent tweet is deleted, replies will stay visible. One of the most popular topics in our dataset was birthday celebrations. These celebrations can disclose the date of birth of the mentioned person which is a sensitive personal information. We focus on the public birthday celebrations and investigate the possible date of birth exposure on Twitter in the next chapter.

Chapter 4

Personal Information Sharing Over Twitter : Birthdays

4.1 Overview

In the previous chapter, we showed that personal information around life events can be inferred only by the analysis of replies. We also found that most of the information owners reacted to these tweets positively, e.g. like the tweet. In this chapter, we focus on birthdays since it is one of the most popular life events celebrated in the previous chapter and it has security implications in addition to the privacy ones since date of birth is commonly used as a part of authentication by organizations.

One of the ways humans build relationships is through sharing of personal information to build trust, therefore it is unsurprising that they use social networks to do so. OSNs thrive on getting users to share information about themselves and they encourage this by adding prompts like "Wish Pat a happy birthday today!" or "Tell your friends about your new shoe purchase". Even LinkedIn, which is an OSN focused on professional networking, has prompts about birthdays.

Historically, date of birth (DOB) was considered as private information; this is why it has been used widely in authentication. Even today, some organizations such as phone companies and banks still use DOB as one of several authentication questions when users phone in [81, 93, 138]. The treatment of DOB as private data can also be seen in the European GDPR regulation where DOB is legally considered to be personal data [36] and is also regularly reported in data breach reports to the public as important personal information that may or may not have been lost during the breach [67]. Still, the use of DOB in authentication these days is significantly lower compared to a couple

of decades ago before the spread of social media.

Early research by Rabkin [131] surveyed the password recovery mechanisms of 20 banks with the aim of showing how vulnerable they were. They found that DOB is used in the process of password recovery of some of these banks, and highlighted that this information can be inferred using public data found in OSNs, which opens the accounts to automatic attacks. This early study was the first to highlight that the trend of how DOB is seen is changing in the era of social media, shifting from a fact that is shared only with close friends and family to a fact that is publicly shared with complete strangers online.

In this chapter, we aim to answer our first two research questions which focuses on the kinds of personal information disclosed by networks (RQ1) and the user reactions and their comfort levels around these disclosures (RQ2) with the specialized focus on birthdays and date of birth. We characterize the disclosure of birthdays/DOBs on the Twitter platform and reactions of the users to such celebrations. We explore the tension between birthdays being open celebrations and birth dates as private information and investigate behavior by measuring the disclosure of birthday wishes on Twitter and users' reactions to them; and attitudes through a user survey asking Twitter users about their thoughts on the topic. Our main research question for this chapter is: *Do social media users see their birthdays/DOB as private information anymore?* More precisely we investigate the following sub-research questions:

RQ1.2 How do people share public birthday wishes on Twitter? What percentage of those indicate the mentioned person's exact date of birth?

RQ1.3 Do protected accounts, who are theoretically more privacy concerned, see less disclosure of birth days/dates than public accounts?

RQ2.3 How do Twitter users react to these public wishes? Is there a difference in the reactions of public and protected accounts?

RQ2.4 How aware/comfortable are Twitter users with having their birth day/date disclosed online?

To answer our research questions, we collected over 18 million tweets/retweets mentioning "happy birthday" over 45 days. Of those, 2.8 million tweets directly mention one non-verified user account. The number of tweets shows just how many birth-

days are being disclosed over a single OSN[1]. Interestingly, we found that over 66K of these tweets likely disclosed the age of almost 50K unique users (e.g. "*Happy 16th birthday @user*"), which makes easy to directly infer the exact date of birth (DOB) of the user. Some of the mentioned accounts were "*protected*", where users have explicitly indicated that their tweets should be kept private and only visible to an approved list of people; yet these users still had their birthday, and sometimes DOB, publicly disclosed by their followers. While public account holders' ages are tweeted more often than the age of the users with protected accounts, still over 5K protected account holders' DOB were likely exposed within the 45 days of our collection period.

Finally, our user survey measured Twitter users' opinion/awareness on birth day/date exposures through celebration on the platform. 48% of the participants were comfortable with others tweeting publicly about their birthdays including their ages.

Our findings indicate that indeed Twitter users are publicly expressing birthday wishes, sometimes also exposing the full DOB, even for protected accounts. The majority of the users are reacting positively (e.g. retweeting and liking) to having their birthday disclosed publicly, which may lead to the inference of the date of birth. These findings show that the view of social media platform designers is the closest to the reality; a large number of users do not think that birthday and DOB are sensitive information anymore. This finding should be taken into account by the organizations that still use this piece of information in their authentication process.

4.2 Related Work

4.2.1 Finding and inferring personal data

Several studies have looked at the types of private information shared on OSNs as well as how to use that data to infer information which has not been shared. Mao et al. [108] looked at the information shared deliberately on Twitter; they find that events such as vacation, illness, and drinking are shared. One type of information can also be used to infer more information. For example, burglars can use the above types of tweets to know that a user's house is vacant. Insurance companies could also increase their premiums if they share their high-risk hobbies on the social media [137]. Jain et al. [68] studied phone numbers posted publicly on Twitter and Facebook in India.

[1]For context, around 0.85% of English tweets on Twitter contain the term "birthday". Based on a two week collection of Twitter's "Sampled Stream" which is a 1% random sample of all tweets on Twitter.

They found that most of the phone numbers were intentionally posted by their owners. However, they were also able to use the phone numbers to find the name of the owner, voter ID, family details, age, home address, and father's name. By adding the numbers to WhatsApp they were able to get further information such as their US numbers, relationship status, and so on.

A user's connections on OSNs can also be used to learn quite a bit about the user, even if that user has "locked down" their account using settings. Analysis of OSN friend networks has shown that knowing information about a user's friends is sufficient to accurately infer attributes such as age, gender, location, political orientation, and sexual orientation [7, 176, 70, 76].

4.2.2 Birth dates in the authentication process

Best practices advise against using knowledge-based questions in the authentication process [56] which includes asking for the birth dates as security questions. Usage of DOB should especially be avoided since it is considered easily discoverable information [122], especially with the spread of social media [131, 66]. Even with these warnings, some organizations such as banks [81, 116, 142], wireless carriers [93], and email service providers [95, 6] still use DOB while authenticating users.

Against the best security practices [46], birth dates are also commonly used by people while constructing passwords [26, 20, 167]. People also use DOB in their PINs which make them easier to be predicted. According to Bonneau et.al. [20], lost or stolen wallets will lead thieves to correctly guess PINs up to 8.9% of the time and the primary reason for that is the identification cards with DOB found in the wallets.

4.2.3 Secondary authentication

Most online accounts allow users to reset their password if they have forgotten it, to do so they use something called a "secondary authentication", where an alternative authentication method is used. A common secondary authentication approach is to ask the user a set of pre-setup questions. The challenge of creating a good set of secondary authentication questions is that they must: 1) apply to most users, 2) have a large set of possible answers, 3) be easy to remember, and 4) have a good distribution across the possible answers (entropy) [131]. It should also be the case that no one else could answer all the questions in the set. Birth dates fit all these requirements, except perhaps the last one, as all humans technically have a DOB, and the set of all possible

birth dates is large. It is therefore easy to see why an organization might select DOB as a secondary authentication question option.

Rabkin [131] surveyed the password recovery mechanisms of 20 banks with the aim of showing how vulnerable they were. Of the 20 banks reviewed, 15 used secondary authentication approaches on password reset. Four of those required customers to enter DOB, three of them required a ZIP code and one of them required mother's maiden name. All three of these questions can be inferred using public data in OSNs as discussed above.

Bonneau et al. [19] studied security and memorability of Google account recovery questions. They found that if a question has common answers, statistical attacks can be considered a risk. For example, favorite food for English-speaking users can be guessed with a success rate of 18% on the first guess. Similarly, city of birth is easily guessed from the first time for Korean-speaking users, with 12% success. They also found that users have hard time remembering answers to more secure questions such as library card number or frequent flyer number. Finding a personal knowledge question that is both memorable and secure is an open research problem.

4.3 Data Collection and Analysis Methodology

We collected tweets containing the words "happy" and "birthday" and then analyzed them in regards to the amount of disclosure, type of account (public, protected), age disclosure, and engagement by the mentioned person. In the following, we describe our data collection and annotation methodology that enables our initial quantitative analysis.

4.3.1 Collecting tweets

We used the Twitter streaming API [154] to collect public tweets in real time. We filtered for English language tweets that contained both the words "happy" and "birthday", resulting in only tweets containing both those words, but not necessarily in consecutive order.

We collected tweets for 45 days between January and March 2019 resulting in nearly 18 million tweets and retweets. We filtered out the 11 million retweets as we are only interested in the initial birthday mention. In addition, we filtered out 2.3 million tweets that had no mentioned account along with 630K tweets mentioning multiple

users where it was unclear whose birthday was disclosed. For some accounts, Twitter will *verify* the identity of the account holder and add a blue tick beside their user name. These tend to be owned by public figures rather than average users. Hence, we also removed 1 million tweets that mentioned verified accounts as well as 3K tweets where the user mentioned themselves. After this cleaning process, we ended up with a set of around 2.8 million tweets that use the words "happy" and "birthday", as well as mention only one non-verified account. We refer to this dataset as "*BD*" tweets dataset.

Two days after the last tweet was collected, we batch processed all *BD* tweets by: 1) identifying any mentioned accounts, 2) checking if the mentioned accounts are public or protected. We excluded 44K tweets where the mentioned account could not be reached (e.g. deleted or suspended) at the time of processing. We then labeled each tweet in the collection with two labels in terms of:

1. account status of the mentioned accounts: either mentioning a public account (*mPublic*) or mentioning a protected one (*mProtected*).

2. tweet conversation type: either a *reply* or, *directed to a user*. A *reply* is in response to an existing parent tweet, such as when a user tweets about their own birthday and a follower replies. A tweet *directed to a user* is a new tweet without a parent, mentioning the user (e.g. "@username Happy 21st birthday"). These are likely to be wishes by friends of the mentioned user who already know their birthday.

Its worth mentioning that *protected* account tweets are visible to their approved followers only and cannot be retweeted or quoted by other users. However, if a *public* account replies to a *protected* account's tweet, the reply can be seen publicly. This is also the case for any tweet mentioning a protected account. Protected accounts can also be mentioned by non-followers.

4.3.2 Gathering reactions on BD tweets

We also measure the reaction of the mentioned user accounts to birthday tweets. For measuring the reactions, we collected the engagement of the mentioned account with these tweets either by replying to or liking the tweet.

Inspecting all these tweets individually to check interaction with them was impractical, due to Twitter API limitations. Thus, we randomly sampled a set of 10,000

	# tweets (unique mentions)	% reply to a user	% directed to a user	% with two digits
mProtected	202K (88K)	30.5	69.5	3.2
mPublic	2.6m (636K)	42.7	57.3	2.7
Total	2.8m (724K)	41.8	58.2	2.8

Table 4.1: Overview of *BD* dataset broken out by the type of account - mProtected and mPublic.

Tweets (5000)	% Liked	% Retweeted	% Replied	Avg time to reply (h)	% Any Reaction	% not accessible
mProtected	51.6	13.8	-	-	54.1	12.4
mPublic	56.1	19.9	43.6	3.5	66.6	8.5

Table 4.2: Responses by the mentioned accounts. mProtected estimates were computed using the count of hidden interactions.

	Accounts	No info	BD	BY	DOB
Protected	4159	3603 (86.6%)	487 (11.7%)	30 (0.7%)	39(0.9%)
Public	4363	3474 (79.6%)	763 (17.5%)	43 (1%)	84 (1.9%)

Table 4.3: Birthday information sharing patterns by the type of account - protected and public.

tweets (5,000 mPublic, 5,000 mProtected) from the BD dataset. To avoid bias, we took samples equally from each day. We refer to this sample of our dataset as *BD-react*.

Twenty days after the last tweet in our main data set was collected, we measured the amount of engagement tweets in *BD-react* had experienced. The average time to reply to a tweet in our set was 3.5 hours, so we are fairly confident that the majority of engagement will have happened within our 20+ day time period. For protected accounts, it is not possible to see if the account has interacted with a tweet via API. However, how many protected accounts have retweeted or liked a tweet is visible via Twitter's user interface (UI). Thus, for mProtected tweets, we scraped if they have been liked or retweeted by a protected account. Note that we can only understand whether *a* protected account interacted with the tweet but we cannot get the usernames of those users to check whether the interaction was by the mentioned protected account. To determine if any of the mentioned accounts had replied to the tweet, we collected the tweets of the mentioned account and searched for replies to our recorded tweet. Doing so was necessary because the Twitter API did not have a method to collect replies

to a particular tweet at the time of the study. It should be noted that this was only possible for mentioned accounts that were public at the time of the collection. After this process, each tweet in our *BD-react* was labeled as being liked, retweeted, and/or replied to (in the case of mPublic) by the mentioned user.

4.3.3 Gathering Birthdays on Profiles

While our main focus is on birthday disclosure by others, users themselves might be self-disclosing the information publicly on their profile pages. In this case, the user may be fine with birthday exposure and others might feel encouraged to tweet about their publicly visible birthday. To see whether users shared their birthday or date information in their profiles, we collected the public birthday information from each account. This information could be gathered for both public and protected accounts. We collected the self-disclosed birthday information for all accounts in our *BD-react* collection. We have applied this process a few months after our initial collection, which led to losing access to some of the accounts due to deletion, deactivation, or suspension; resulting in getting the information of only 4364 public and 4159 protected unique accounts.

4.3.4 Tweets disclosing user's age

Some tweets explicitly mention the age of the person. An example from our BD dataset (username anonymized): "Happy 40th Birthday to @username. Have a great day and night". If the age is combined with the date the tweet was posted, it becomes trivially possible to reconstruct the full birth date. To understand the scope of this disclosure, we further analyzed the tweets to extract those containing a two-digit number between 10 and 99. We looked for all instances of two digits on their own or in combination with an ordinal indicator (i.e. "st", "nd", "rd", "th"). We selected the 10-99 range, because numbers below 10 might mean something other than the age, and technically Twitter does not allow users younger than 13 years old. Similarly, few people live to over 99, so the number of errors in this numeric range is expected to be large compared to the number of true ages. The percentage of tweets that contain two digits with our criteria are shown in Table 4.1.

To verify if the tweets containing two-digit numbers are referring to the user's age, we manually labeled a random sample of of 4000 tweets (2000 mPublic, 2000 mProtected) from the tweets that had two-digit numbers. We took samples equally

	Accounts	Received two-digit tweet	Age likely known	% age likely known
Protected	88K	5.5K	5K	5.7
Public	636K	51K	44K	7

Table 4.4: Estimations of the number of accounts where the full birth date and year can be determined. Estimations are based on a combination of the number of tweets containing a two-digit number and the observed rates of age prediction from the Appen annotations.

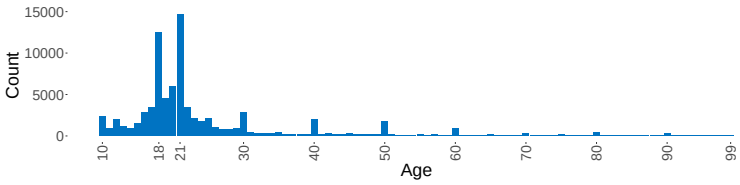

Figure 4.1: Distribution of the two-digit numbers in *BD*. Notable spikes at key ages: 18, 21, and multiples of 10.

from each day as we did with *BD-react*. We refer to this sample of tweets as *BD-age*.

For the annotation, we used the online crowdsourcing platform Appen[2] participants were asked: "Can we tell that this person: @username has their age disclosed in the tweet?" where @username was replaced with the mentioned person's account from the actual tweet. Each tweet was judged by three trusted workers and we used majority voting to label tweets. A test-set of 64 pre-labeled tweets, that we manually annotated, was provided for quality control of the annotation. If a worker got more than 20% of the pre-labeled tweets incorrect, their annotations were discarded. The final inter-annotator agreement rate was 93%.

4.4 The Share of Birthday Wishes on Twitter

4.4.1 BD dataset statistics

Table 4.1 shows the overview of our *BD* dataset, which contains 2.8m original tweets, broken out by the type of the mentioned account - mProtected, mPublic. These 2.8m tweets were directed to 724K unique accounts. The majority (56%) of these accounts

[2]Appen is formerly known as Figure-Eight and Crowdflower. `https://www.appen.com/`.

received only one birthday celebration tweet. 99% of them received 30 or less birthday wishes. Protected accounts tended to get less birthday wishes on average with 99% of them receiving 15 or less tweets. One public user in our collection received 3934 birthday wish tweets, while the most popular protected user received 394 birthday wishes.

We also check the percentage of tweets that were replies to other tweets or written to a user directly. We see that tweets were more likely to be directed to a user without replying to an existing tweet. Only 42.7% of the tweets were a reply to an existing tweet, whereas 57.3% of them were directed to a user without the user tweeting about their birthday. This difference is even higher for mProtected, where 69.5% were tweets directed to them and only 30.5% were replies (Table 4.1).

Finally, looking at the tweets containing two-digit numbers (including ordinal indicators) mProtected tweets had higher percentage (3.2%) than the mPublic tweets (2.7%). These numbers can be an indication of the mentioned person age, which can lead to easily inferring the person's exact DOB. We provide further analysis for the meaning of these numbers later.

4.4.2 Twitter users' reactions to birthday wishes

We carried out another analysis on *BD-react* in which the tweets were processed in more depth to analyze the reaction of the mentioned users to the BD tweets mentioning them (Table 4.2). At the time of processing, 10.5% of these were no longer accessible due to various reasons such as the deletion of the tweet, the author protecting their account, and so on. mPublic tweets received high interaction from the mentioned accounts. We observed that 56.1% of the mPublic tweets had likes from the mentioned account. For mProtected tweets, we only know that a protected account interacted with the tweet, not which account. However, hidden interactions can give us an idea. 51.6% of mProtected tweets had hidden likes while 13.8% of them had hidden retweets. 66.6% of the public mentioned accounts interacted with the tweets mentioning them while 54.1% of the mProtected tweets had hidden interactions. This result shows that people frequently interact with the tweets that wish them a happy birthday in a positive way such as liking and retweeting, regardless of those people's accounts being public or protected. In addition, the large number of interactions can indicate that the tweets are seen by other people who might not be necessarily following the birthday person.

4.4.3 Sharing Birthdays on Profile

By checking the birthday information on the profiles of the 8522 reachable accounts, we found 7077 (83%) shared no birthday information, 1250 (14.7%) shared the birthday (BD), 73 (0.9%) shared only the birth year (BY), and 123 (1.4%) shared their full DOB. Public accounts were more likely to share information on their birthday (890, 20.4%) than protected accounts (556, 13.4%). In total only 196 (2.3%) of the users disclosed their birth year. We report the birthday sharing behavior on profiles broken out by the account type in Table 4.3. Users who shared their birthday information reacted similarly to the birthday tweets with those who did not. This was also the case for protected users.

4.4.4 DOB leakage on Twitter

Regarding the *BD-age* tweets, 82.9% of the tweets with two-digit numbers refer to the mentioned person's age, according to Appen annotators. The percentage is slightly higher for the mentioned protected accounts (84.4%) than public ones (81.3%). We noticed that this percentage becomes higher (95.3%) if the two-digit number is followed by ordinal indicator (st, nd, rd, th). Using these rates, we extrapolated to the whole data set, taking into account the total number of tweets containing two-digit numbers, the results are shown in Table 4.4.

We look at the unique accounts mentioned in the *BD* dataset to understand the potential DOB disclosure for birthday people. There were 56K (8%) accounts in total that received at least one birthday tweet that contained a two-digit number, of those 33K received at least one tweet accompanied by an ordinal indicator. 51K of them were public accounts, while 5.5K of them were protected accounts. Based on the results of the annotation, we can estimate that the actual age of the person is exposed for over 49K accounts which when combined with the date of the tweet, likely exposed the full birth date and year. This is 6.8% percent of the accounts that were mentioned in the tweets we collected.

The mean of the two-digit numbers we found is 25 with median 21. The most celebrated ages were 18 and 21, followed by ages at multiples of ten (Figure 4.1). From our collection, we see that over 1K accounts receive birthday wishes that exposes their DOB every day, where 10% of those are protected accounts. While these users are mostly young adults, there are also users who are teenagers and elderly. Public accounts got more age exposing tweets than the protected accounts which suggests

that people treat accounts differently depending on their type. Interestingly, accounts that shared no birthday info got more birthday messages with two-digits.

Combining these results with the reaction of those users on the tweets, it becomes necessary to understand how Twitter users see this phenomena and if they perceive the DOB as private information.

4.5 Measuring Users' Opinions and Awareness

We conducted a survey to better understand how Twitter users think about the public sharing of birthday wishes on Twitter (RQ2.4), as well as their understanding of tweet visibility settings. We advertised the survey on Prolific Academic (PA) [128] as "Wishing a Happy Birthday on Social Media". The advertisement limited participants to Twitter users from the United States or United Kingdom to ensure similar culture and English label proficiency. We followed our University's ethics protocol in the design and running of the survey. Participants were compensated £0.5 (£8.34 per hour).

4.5.1 Survey Instrument

The survey started with informed consent followed by a screening question about if they had a Twitter account and if they used it more or less than once a month. Those without a Twitter account were screened out. We then asked if their primary Twitter account, was public, protected, or sometimes protected where they change the settings, followed by if they associated their Twitter account with their "real identity", and if they had their birthday publicly visible on *any* social media account.

To gauge understanding of Twitter setting impacts, we asked what would happen in two scenarios where public and protected accounts interact. We also asked if they can tell that a poster's account is public or protected when replying to a tweet, and if they look to see if the account is protected when engaging with tweets (reply, mention, retweet).

To understand their comfort with public birthday and date disclosure we asked them how comfortable they would be with friends and family publicly tweeting about their birthday with and without age information. We also asked how they might engage with such a tweet (like, retweet, reply, direct message (DM), ask to remove). We then asked them a similar question around the participant tweeting about a friend or

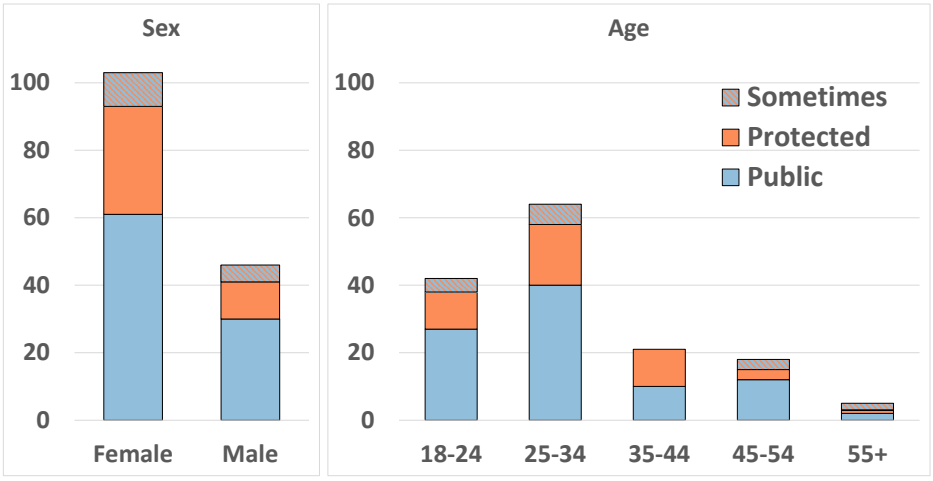

Figure 4.2: Sex and age distributions of survey participants by account type. "Sometimes" refers to accounts switching between public and protected.

family member's birthday with and without age. Finally, to gauge participants' understanding of the positive and negatives of public birthday wishing we asked them two free-text questions: *"Give at least one example of a good thing that could happen if someone knew your birthday and age."* and the same question with *"bad thing"*. The survey ended with an optional comment box. As PA provides common participant demographics to researchers, we did not directly ask for any demographics. The survey for this study is given in Appendix A.

4.5.2 Survey results

Participant demographics The survey had 151 participants, 118 from the United Kingdom (UK) and 33 from the United States (US). Their average age was 31.2 years ($\sigma = 10.6$) with a median of 28.5. Respondents were primarily female (n=103, 68.2%) vs. male (46, 30.5%) with 2 preferring to not respond.

For context, Twitter users from US have median age of 40 and half of them are female [171]. Only 44% of the UK Twitter users are female [144] with more than half of the users older than 35 [143]. Our participants are generally younger than the general Twitter population and have a higher percentage of female representation.

Account types Most participants had public accounts (92, 61%) with the rest having protected accounts (44, 29%) or switching between public and protected (15, 10%). Figure 4.2 illustrates sex and age distributions of the survey participants broken out by their account types. 84 (55.6%) participants associated their Twitter accounts with their real identity. Of those, 27 (32%) had a protected account or switched between

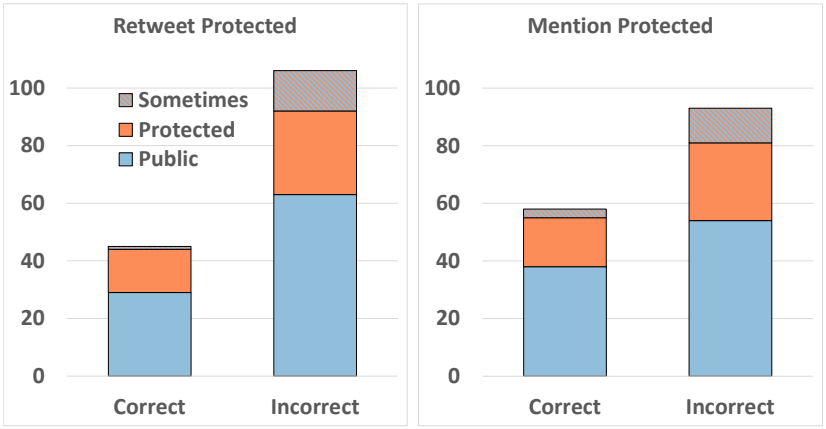

Figure 4.3: Participants' answers when asked what will happen if a public account retweets a tweet by a protected account or mentions them.

public and protected (9, 11%). In other words, roughly half of the people whose accounts were linked with a real identity were also protected. Publicly listing a birthday on at least one social media account was common, with 96 (63.6%) publicly listing a birthday, 46 (30.5%) not listing, and 9 (6%) not sure. 65% of the participants with public accounts shared their birthday on at least one social media account, whereas 57% of the protected accounts shared it. 75% of the users who switch between public and protected shared their birthday on social media.

13% of the adult US Twitter users have protected accounts [171], which is much lower than the share of protected accounts in our survey participants.

Visibility of protected accounts Participants were asked about two scenarios: imagine that *"Alice (Public) retweeted one of Bob's (Protected) tweets using the Twitter website?"* and *"Alice (Public) tweeted at Bob (Protected) using his handle (@bob) in the tweet?"* We then asked them if Twitter would: not allow Alice, warn Alice, allow Alice but restrict visibility to Bob's friends, allow Alice with no restrictions, or they didn't know. For retweets, the Twitter website does not allow public accounts to retweet protected accounts. Only 45 (29.8%) of participants gave this answer. The most common answer was that Alice could retweet but only Bob's followers could see it (70, 46.4%), which is incorrect.

For tweeting at protected accounts, the Twitter website allows public accounts to mention protected ones with no visibility restrictions. Participants were split on this question with 58 (38.4%) thinking that the tweet would be visible (correct), and 58 (38.4%) thinking that it would be visible only to Bob's followers (incorrect). Figure 4.3 and Figure 4.4 shows participants with different types of accounts and demographics

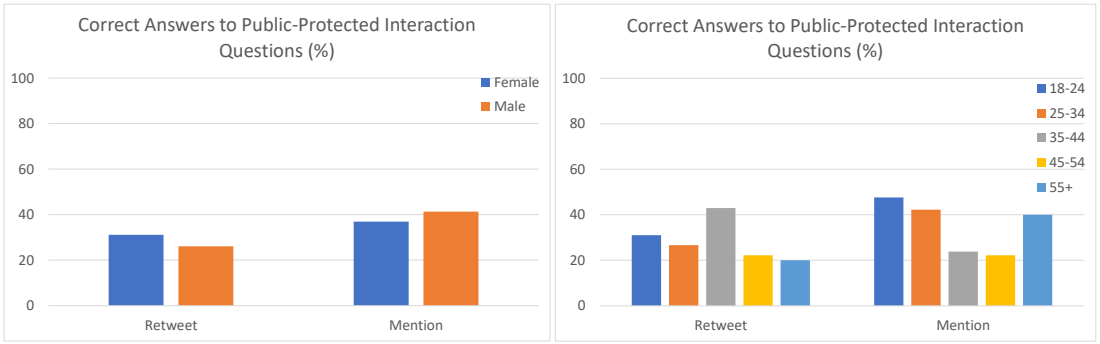

Figure 4.4: Percentage of participants answering correctly when asked what will happen if a public account retweets a tweet by a protected account or mentions them, broken by sex and age.

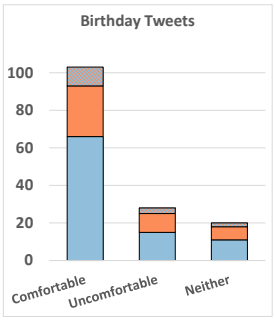

Figure 4.5: Comfort levels of participants if their friends or family members tweet birthday wishes without age and wishes with DOB disclosure

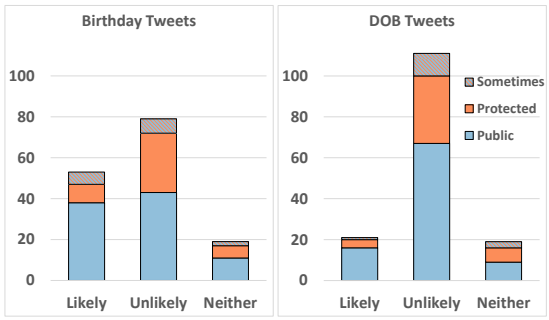

Figure 4.6: Likeliness of participants tweeting birthday wishes to their friends or family members without age and wishes with DOB disclosure

with their answers. The confusion over how Twitter protected accounts work is further highlighted by comments from participants about why they switch their accounts between protected and public. "*If I post things I don't mind everyone seeing I changed it to public but if [I] post photo that I only want my followers to see I make it private.*" The comment highlights a potential misconception that the protections are per-post instead of per-account.

When an account is protected, a padlock appears beside the username. However, when asked, only 33.8% of the participants agreed that they can easily see whether the Twitter account they are replying to is protected or not. Even less check the type of account they are replying to (24.5%).

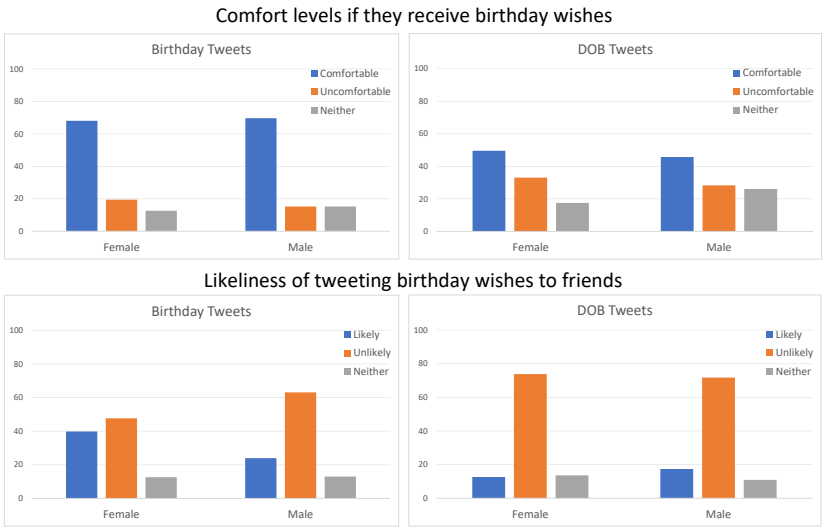

Figure 4.7: Comfort levels of participants if their friends or family members tweet birthday wishes without age and wishes with DOB disclosure (first row) , Likeliness of participants tweeting birthday wishes to their friends or family members without age and wishes with DOB disclosure (second row), percentages broken by sex.

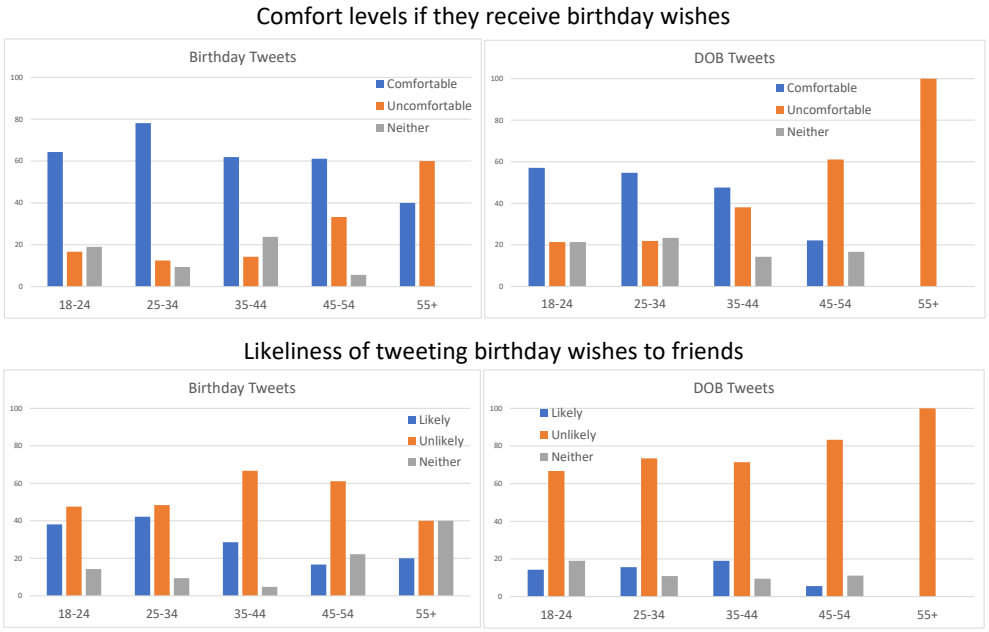

Figure 4.8: Comfort levels of participants if their friends or family members tweet birthday wishes without age and wishes with DOB disclosure (first row) , Likeliness of participants tweeting birthday wishes to their friends or family members without age and wishes with DOB disclosure (second row), percentages broken by age.

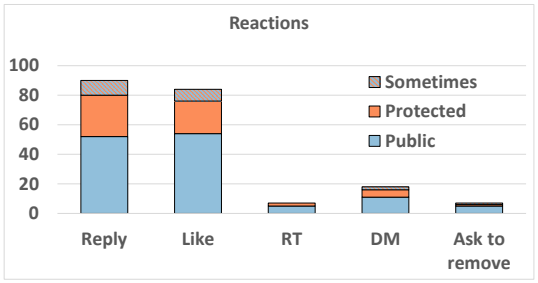

Figure 4.9: Reactions if friends or family members where to share the participant's birthday publicly.

Birthday tweet opinions Looking at participants' comfort with friends and family tweeting about their birthday, the majority of participants were comfortable with their birthday being tweeted (103, 68.2%). Most gave the same answer for birthday and birthday with age (90, 59.6%) , indicating that the addition of age had no impact on their comfort. A further 58 (38.4%) indicated that they would be less comfortable with birthday tweets containing an age. We show comfort levels of participants with their types of accounts in Figure 4.5, as well as comfort levels by sex and age in Figure 4.7 and Figure 4.8. While female and male participants showed similar levels of comfort, older participants feel less comfortable with birthday wishes that include ages compared to younger ones.

Looking at participants' likeliness of tweeting about a friend or family member's birthday, the majority of participants indicated they would be unlikely to do so (79, 52.3%), vs likely to do so (53, 35.1%). The majority of participants (80, 53%) gave the same answer for both birthday and the birthday with age, again indicating that the addition of age had no impact on tweeting likelihood. The rest (69, 45.7%) were less likely to post a birthday tweet containing age. We show likeliness levels of participants with their types of accounts in Figure 4.6, as well as likeliness levels by sex and age in Figure 4.7 and Figure 4.8. Similarly with comfort levels, female and male participants showed similar levels of likeliness levels while older participants expressed that they would be less likely to share birthday wishes that include ages compared to younger ones.

Regarding their reactions to birthday wishes online, 90 (60%) participant said they would reply with a thank you, or like the tweet (84, 56%). Replying via direct message was less common (18, 12%). And it was rare to ask the person to remove the tweet

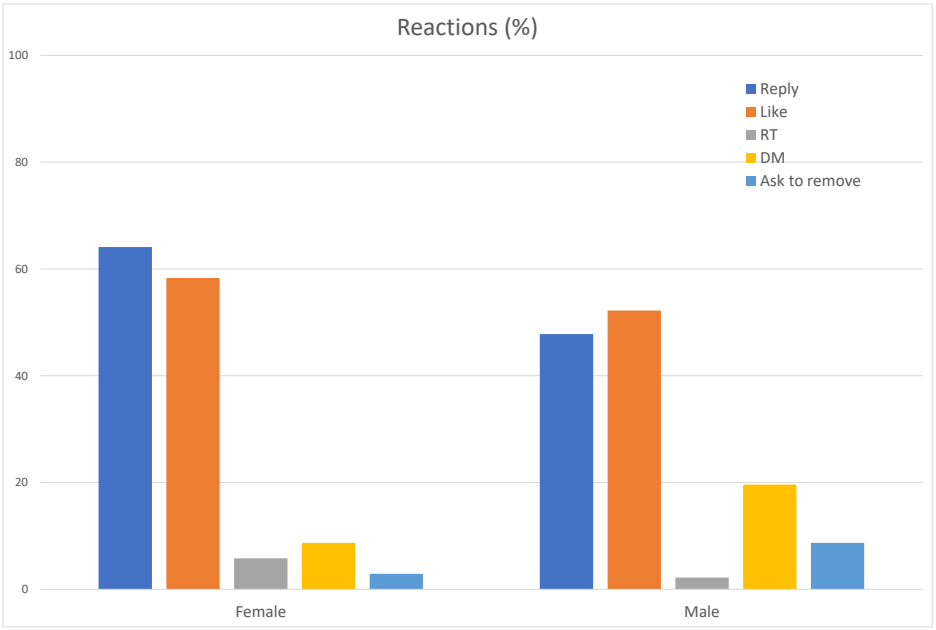

Figure 4.10: Reactions if friends or family members where to share the participant's birthday publicly, by sex.

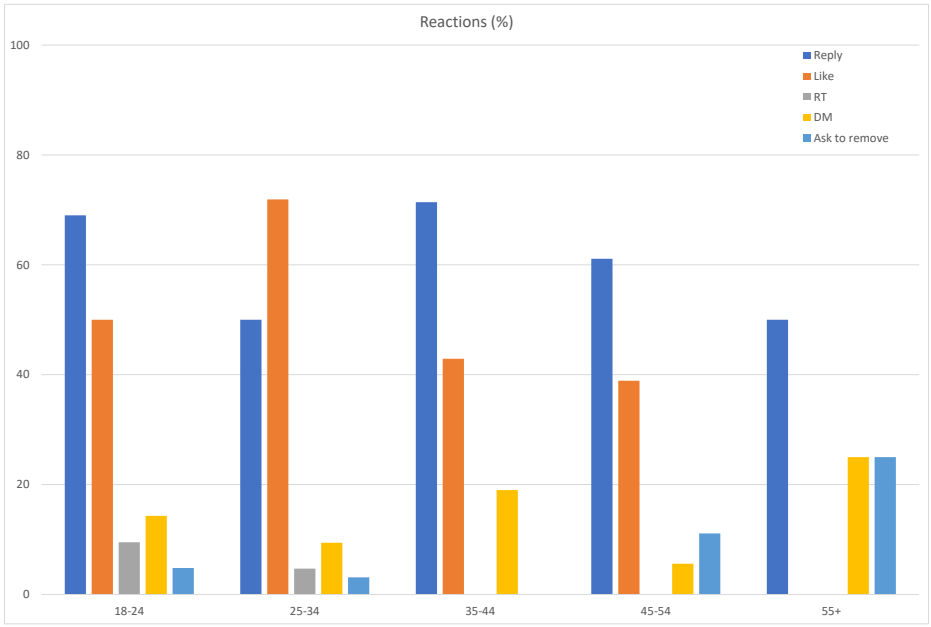

Figure 4.11: Reactions if friends or family members where to share the participant's birthday publicly, by age.

(7, 4.6%) or retweet it (7, 4.6%). Figure 4.9 shows the reactions of the participants with their types of accounts. Figure 4.10 and Figure 4.11 shows the reactions of the participants by their sex and age. Those with public and protected accounts indicated similar reactions to tweets. These findings are similar to Table 4.2 where likes and replies were the main reactions to birthday wishes.

Good & bad impacts of sharing We asked our participants to list some of the good and bad things that could happen if someone knew their birthday and age. Two researchers read through all the free-text answers and jointly grouped them into themes, discussing throughout to reach agreement. The most common good things mentioned were getting birthday wishes (45%) and gifts (31%). For the bad things, the most prominent worry was identity theft (37%), and the next most mentioned was being harassed, ridiculed or harmed (19%). 8.6% of the participants said nothing bad would happen.

The responses of participants on our survey showed that users are aware of the role birthday wishing can have in connecting them with others as well as some of the dangers of birthday and DOB disclosure can cause. Nevertheless, nearly half of them are still comfortable with their friends sharing their age along with birthday tweets online.

4.6 Discussion

4.6.1 Summary of findings

In this study, we focused on answering the first two research questions of the thesis: "What kind of personal information are disclosed online by networks on Twitter?" and "How do users react (e.g. like, retweet or reply) to tweets that share their personal information when they are mentioned in them and how comfortable are they with such information disclosure?". In this chapter, we aim to answer these questions in the context of birthday celebrations and the possible date of birth disclosure.

Our first research sub-question (RQ1.2) concerns how people share birthday wishes on Twitter and how many include information revealing the DOB of the mentioned user. We collected 2.8 million happy-birthday tweets mentioning 724K unique users over 45 days, which shows that a large number of happy-birthday tweets can be easily linked to a specific account. Further, we found that around 7% of those accounts received at least one tweet that disclosed their age, allowing for easy combination of

posting date and age to compute DOB. Considering only around 2% of the users in our dataset share their birth years on their profiles, DOB is being actively exposed for users who did not proactively share it.

Our second research sub-question (RQ1.3) concerns how birthday wishes differ towards protected and public accounts. 88% of all the birthday wishes we collected were towards public accounts. These accounts also received more birthday wishes per account on average than protected ones. We found that public accounts are also slightly more likely to receive birthday tweets that reveal their DOB than protected accounts, 8% vs. 6.2% respectively (Table 4.4). The result indicates that having a protected account does not prevent online disclosure of birthday or date information.

Finally, we measured users' reaction behavior (tweet interaction - RQ2.3) as well as their attitudes (survey - RQ2.4). We found that 66.6% of tweets mentioning public accounts were reacted to by the mentioned account and 54.1% of tweets mentioning a protected account were likely reacted to, though we have limited visibility of protected accounts. Both Twitter and survey data show that liking and replying are popular ways to react to birthday tweets. 56% of the mPublic tweets in *BD-react* received a like from the mentioned person. Similarly, 56% of the survey participants said they would like a birthday tweet they receive on Twitter. 60% of the survey participants would reply to the birthday tweet, while 44% of the mPublic tweets in *BD-react* received replies from the mentioned user. Our Twitter data collection (20%) and user survey (5%) differ on the cases of retweets. 5% of our survey participants selected that they would ask the person to remove the tweet if they received a birthday celebration over Twitter. While it is not possible to measure this reaction from the Twitter data directly, we recorded 428 (9%) cases where the tweet or the user was deleted while collecting the replies to the tweets in *BD-react*. However, there is no way to differentiate from the API response whether it was the tweet that was deleted or the user's account. There is also no way to make sure that the birthday person requested the tweet deletion in any case.

These findings suggest that users are aware of happy birthday tweets and react to them positively. Our survey verifying our tweet analysis findings show that Twitter users are comfortable with public celebrations of birthdays, both with and without explicit mention of their ages. This result is evident both in the scale of current birthday wishing on Twitter as well as the attitudes of survey respondents. However, most of the respondents were less likely to publicly celebrate birthdays of their friends and family with tweets containing their ages.

Our survey also showed that Twitter users might not be fully understanding who

can see their tweets when they mention protected accounts and sometimes not aware of account types of users they interact with. We further discuss these findings in the implications section below.

4.6.2 Limitations

Our research analysis was limited to tweets on Twitter which explicitly mentioned both the words “happy” and ”birthday”. While the term “happy birthday” is culturally quite common in English speaking countries, there are many other ways to express the sentiment. Additionally, we did not look for common misspellings or abbreviations such as “hbd”, “happy bday”. Therefore, the numbers presented in this work should be seen as a lower bound, or what an opportunistic data gatherer might be able to locate easily.

Another limitation is that we only look at the two-digit numbers that have a leading space and no trailing alphanumeric characters other than the ordinal indicators. Because of this, we are missing some age exposing tweets. Some tweets may also have the age spelled instead of writing with numbers like “twenty first”, or “sixteen”. We ran a test on *BD* data set (Table 4.1), to understand if people spelled out ages. We looked at a commonly celebrated birthday “twenty one” or “twenty first” and compared the occurrences with the numeric “21” and “21st”. We found that $< 0.01\%$ of our *BD* tweets spell out 21, likely due to the character limit pressures imposed by Twitter. Hence, we expect the effects of excluding spelled out numbers to be minimal.

The breakdown analysis with age and sex demographics for the user survey detailed in Section 4.5 was conducted posthoc.

4.6.3 Implications

Popularity of birthday wishing Twitter is clearly considered a suitable platform to wish someone else a “happy birthday”, which is further reinforced by verified account reactions as well as Twitter itself [29]. In our initial data collection we observed about a million tweets wishing a verified account a happy birthday. While we excluded these from analysis, the size of the dataset speaks to the public reinforcement that birthday wishing is a normal public Twitter activity. Twitter itself also encourages birthday celebrations by displaying a balloon animation on their birthday when others visit it. Other OSNs, such as Facebook, also actively encourage people to wish others a “happy birthday”, which can impact the number of birthday wishes [50].

Birthday wishing was also common among non-verified accounts, accounting for nearly three quarters of the wishes. As an OSN, one of Twitter's roles is in helping people maintain weak and strong ties through sharing information [162] which can have benefits on their mental health and sense of belonging [44]. These ties are also useful at helping people gain access to prospects such as jobs and opportunities [62], so maintaining them has value. Viswanath et al. [160] found that users who do not interact on social media frequently, mostly only exchange birthday messages. Hence, birthday wishes can support and encourage users to maintain their social ties.

Disclosure control Even if a user is inclined towards not sharing their DOB on Twitter, they have limited control over their network. Most OSNs only provide control to the poster, not the data subject. Hence, *networked privacy* means that the control over who can see what data is not solely in the hands of the person whose data it is. While some OSNs allow subjects to remove their tags from a post, they do not allow complete removal of the post. Consequently, if a user would prefer their birth date not be known, they would need to ask each poster to remove their post. Such a request may be socially challenging, especially if the majority of users feel that wishing someone a happy birthday is a good thing and nothing to be concerned about. Such a request might also risk being labelled as "paranoid" [53].

Controlling information disclosure is also dependent on Twitter users and their networks' accurate understanding of tweet visibility. Previous research on social media with granular privacy options show that users find it hard to comprehend and configure these settings [105], while also underestimating the audience size [17]. Twitter has a relatively simplistic access-control approach for a modern OSN. An account is either public or protected. If public, anyone on the internet can see the posts, if protected, then only a selected set of users can see them. Yet even this simplistic model confused our participants. In this study we found that 38.4% of participants thought that posts in reply to a protected account would also be protected. Findings suggest that users' mental models of Twitter protections might be inaccurate which may lead them to disclose information about protected accounts while honestly believing that the posts are not publicly visible.

Authentication question vs. celebration There are surprisingly few questions that work well for authenticating identity. A good authentication question should: apply to nearly everyone, have a large number of possible answers, have equal distribution of answers, be easy to remember, and should not change. Based on those requirements, it is obvious why facts like birthdays were regularly used as a part of

authentication. However, with birthdays celebrated publicly on OSNs and reaching more people than before, organizations justifiably shifted their use of DOB.

DOB is also used as one of the attack vectors in social engineering and re-identification methods [149, 89]. People often use their DOB when constructing passwords [26]. Bonneau et.al. [20] showed that up to 8.9% of the time, lost or stolen wallets will lead to thieves to correctly guess PINs. A primary reason is that DOB can be obtained using identification cards found in the wallet. However, as we can also see from our Twitter data and the following user survey, birthdays are not seen as secret by the general public. Similar with our findings, Markos et al. [110] also found that DOB was considered as a low-privacy segment data along with e-mails by their participants compared to mother's maiden name, home addresses, and phone numbers. This attitude by public leads to a tension between data privacy and the reality of cultural sharing. One of the main recommendations that we can learn from this study is that organizations, such as banks [81, 116], email service providers [95, 6], wireless carriers [93], that still consider DOB sensitive information should withdraw from using it as a part of their authentication system. Users should also not incorporate their birthdays into constructed passwords.

Twitter design implications Birthday sharing is widespread on Twitter and users are comfortable with it. Encouraging this behavior might help users to feel valued and appreciated, as well as maintain friendships [162]. In order to encourage birthday wishing, Twitter could notify followers of a user on their birthday if the user wants to receive such messages. Displaying a small indicator (e.g. balloon, cake) next to the username of a person having their birthday on their tweets may also act as a reminder for the birthday and encourage a message. Providing users personalized messages and collating the birthday tweets in one thread will help the birthday person to feel special and also allow easy access for replying to those messages.

On the other hand, some people might not want their birthdays to be celebrated publicly. However, Twitter accounts cannot manage the tweets that mention/quote/retweet them. Some users solve this by asking other users explicitly not to interact. For example, some users state that they do not want other users to quote them in their profiles or usernames. Information about protected accounts can also be revealed through public replies to their tweets. Twitter recently introduced a feature to select user groups who could reply to specific tweets. While this is a positive step, these tweets are still publicly visible. Another recently added feature is the option to hide replies, however, it is easy to access these tweets with an extra click. They are also reachable by search. In

addition, users might be socially uncomfortable to hide celebratory messages that leak information about them. Hence, users should be given control on the visibility of the tweets that mention them.

Another point is the confusion over the visibility of the posts when public and protected accounts interact. Our survey showed that users' mental models of tweet privacy might be quite different from the actual Twitter functionality. Our participants expected these tweets to be visible only to the followers of a protected account when a public follower interacts with the protected account. Hence, it is essential to disambiguate the interactions between them. This can be achieved by having an indication when interacting with protected accounts and let users know who can see the tweet if posted. The account types of the users mentioned in a tweet should be easy to check. As of now, when replying to a protected tweet, there is a padlock near the username indicating the status. However, when drafting a new tweet, the public or protected status is obscured.

4.7 Summary

In this chapter, we aimed to answer our first two main research questions concerning the personal information disclosures by users' networks and the reactions to these disclosures. To do so, we investigated the sharing of birthday wishes on Twitter, how they can reveal the date of birth of some users, and privacy concerns of the Twitter users regarding the DOB disclosure. Our objective was to provide an in-depth analysis of how social media users see their DOB, as a private personal information, or as a happy event to be celebrated publicly. Our aim was to assist designers in the security and social media field to get a clear answer of how to treat this information about users when designing their systems. We both conducted an analysis of 2.8 millions tweets sharing birthday wishes, and a survey of Twitter users to understand their opinions around public celebration of birthdays on the platform. We found that birthday celebrations are common over Twitter and over 1K tweets disclose the DOB of the mentioned account daily, where 10% of those are protected. While the majority of those accounts do not share their birthday publicly, they still seem to be comfortable with others sharing birthday tweets publicly regardless of their account type, even when it discloses their DOB. We show that birthdays and DOB are not considered as sensitive information by users; they are celebrated publicly. Our findings should move any organization that is still using DOB as a part of their authentication process to phase its use out.

In the user survey, we noticed that aside from the users with binary privacy settings, there were users who switched between the settings to a degree where they classified themselves as “Sometimes protected”. In the next chapter, we focus on these users who change their privacy settings to quantify the phenomenon and understand the reasons behind the practice.

Chapter 5

User Utilization of Privacy Settings

5.1 Overview

Previous two chapters focused on the types of personal information disclosed by user's networks by looking at replies and mentions of the user. We found that information around life events of users can be leaked by the their networks, as well as their date of birth. We also show that users reacted mostly positive to these disclosures through the data collected from Twitter and the user survey.

In this chapter, we aim to answer our third research question: "What reasons do users have to change their privacy settings on Twitter and how do the reasons differ between changing the account setting to protected versus public?". We look at Twitter's tweet visibility setting feature which is particularly rigid with only two options, public and protected, which apply to the user's entire tweet stream, including historical tweets. Yet even in this highly binary and low fidelity situation we see a range of how people use the system to create a rich boundary management to suit their needs. Twitter is a particularly interesting platform to study in regards to privacy management because of the limited control options and also because Twitter is quite blunt about the impact of switching from protected to public: "Unprotecting your Tweets will cause any previously protected Tweets to be made public" [30]. The implication is that Twitter accounts are meant to be public or protected and are generally assumed to stay in one of those states with changes between being rare. Existing work on privacy settings on Twitter mostly focus on the differences between users who are protected and public [34, 96]. These studies check the settings of user accounts only once to decide if it is public or protected. However this assumption may be wrong and it may be the case that users are changing their account visibility often enough to count as having a more

complicated visibility status.

Studies in other platforms, like Facebook, have found that even when privacy controls do not well match the effects users are trying to create, people are quite good at using the technology in unexpected ways to create the effects they want. A good example is work on how teenagers manage context in social networks by doing things like deactivating their accounts when they are not logged in [112]. While not immediately intuitive, deactivating an account prevents others from interacting with it, enabling the user to only allow others to interact with the account when they themselves are logged in. The point is that privacy permission management is not necessarily a single set-it-and-forget-it setting. People use these settings in boundary management, which is a continuously changing state between people.

While users may need to manage interpersonal relationships with granularity, the main control over privacy is still the visibility of the account content. On Twitter, that means the visibility of the account and its tweets. Twitter users have a relatively small set of options to protect their tweets and those users that change their account to protected also face some restrictions on the types of interactions they can have. For example, if a protected user mentions a non-follower, that person cannot see the mention and therefore cannot respond to it. Other than changing tweet visibility, users can also delete their tweets, block other users to prevent them from interacting, and deactivating their account.

It is also an open question what effect users are attempting to create when they change visibility settings. The most obvious assumption is that they are trying to hide their tweets from public consumption. But Twitter also has other issues, such as harassment [71], that may cause users to switch to protected [158]. This observation opens an interesting question about how users are changing their behavior when they are public vs when they are protected. A user simply trying to avoid harassment may actually not change behavior when they are protected, because it is themselves they are trying to protect, not the tweets. Users that are trying to protect their tweets may also have specific types of tweets that they are more inclined to post while protected as opposed to when public.

In this study, we focus on users who change their privacy settings in order to understand the behaviour and motivations behind the changes, as well as the differences in self-disclosure between when they are public versus protected. More precisely, we investigate the following research questions:

RQ3.1 How frequently do Twitter users change their tweet visibility settings?

RQ3.2 Do users employ different posting strategies when they are public vs. protected?

RQ3.3 Why do users change their visibility settings and how do the reasons differ between changing to protected versus public?

RQ3.4 What other strategies do switching users employ to control their audience and interactions?

To answer these questions, we curated a set of 107K users, whose accounts were protected at the start of the data collection, and analyzed their account visibility changes for three months. Of these, 45K changed their visibility to public at least once allowing us to collect their tweets to determine if their tweeting behaviour changes when they are protected compared to when they are public. We also conducted two user surveys to get insights into why users decide to change their tweet visibility back and forth.

Our findings show that large portion of protected users do change their privacy setting to public, where around 40% of the protected accounts we inspected changed to public at least once within three months. A quarter of those accounts changed more than ten times within this period. Analysis of tweet data shows that users' accounts have less tweets during times when they are protected. They also mention other users less compared to when they are public. Our survey results show that users mostly change their tweets to protected to regulate their boundaries but change to public to overcome the platform's technical constraints. Our findings provide a unique perspective into the usage of privacy settings in a platform where options are binary and in an account-level, where trade-off between privacy and functionality must be calculated with the historical posts in mind. We also provide design implications built upon our findings and prior literature to help users and platforms minimize potential privacy leaks.

5.2 Related Work

As mentioned in Chapter 2, users employ different strategies to manage their boundaries and flow of their information. We shared related work around privacy settings usage, self-censorship, multiple account usage, and so on in the Chapter 2. Here we focus more in detail on some of these protection strategies and their effectiveness.

Users can choose to delete some or all of their posts for various reasons such as regret [141], typos [9], to prevent temporal context collapse [64], and so on. However,

the residual activity around the deleted/hidden tweets can still be used to infer the content [114, 9]. Almuhimedi et al. [9] followed 292K users for a week to analyze deleted tweets. They find that around half of the users deleted at least some tweets during the week. The content of the tweets that were deleted were not substantially different than the ones that were left on the platform. Mondal et al. [114] analyzed the accessibility of tweets longitudinally by trying to recollect tweets from older datasets that date back to six years. Unlike Almuhimedi et al. [9], they also included the tweets of users who turned their accounts to protected to their analysis. They found that while most of the users did not withdraw any of their tweets, 8.3% of them deleted some tweets, 16% deleted their account, and 10% protected their tweets. They found that they can recover the topics of interests of the withdrawn tweets by analyzing residual interactions around them which shows that even deletion might not prevent possible privacy risks.

Users also try to sanction interactions without harming their relationships to regulate interactions. Rashidi et al. [132] interviewed 23 young adults to analyze the sanctions enforced by them on popular social media. They grouped the sanctions using three dimensions: on-site and off-site, individual and collaborative, visible and invisible sanctions. The found that people prefer using invisible sanctions like muting rather than blocking, unfriending, or deleting content.

Instead of sanctioning, users may choose to deactivate their accounts to protect their privacy, concentrate on their work [15], and to limit interactions [112]. Platforms provide deactivation as an alternative to deleting accounts and users are allowed to get their accounts back if they decide to do so. For example, Twitter gives a 30 day grace period during which users can reactivate their deactivated accounts. After that time the accounts will be permanently deleted and the username can be claimed by someone else. Ng [118] describes the users who temporarily deactivated their accounts as intermittent discontinuers and found that social media fatigue was one of the reasons of the deactivation. They also found that some users change their tweets to protected instead of deactivating.

5.3 Monitoring Switching Behaviour on Twitter

To understand the phenomenon of users switching their privacy settings on Twitter, we initially curated a large set of users who had protected accounts at the start of the data collection and monitored their behaviour over three months to record any possi-

	0	1	2	3 - 5	6 - 9	10+	Total
# Users	64,460	7,630	9,195	8,599	5,910	11,121	106,915

Table 5.1: Number of switches users made in three months. Even number of switches means being protected by the end of 3 months.

ble switch to their status between protected and public. We analyzed these switching patterns and compared their tweets using features such as sharing of hashtags, media, links, and so on. We followed our University's ethics protocol in the design and running of this work, including the social media data collection and following surveys. In this section, we describe the Twitter API as well as our data collection and analysis.

5.3.1 Collecting Switching Users

As mentioned, Twitter accounts are by default public and users need to actively change their account to protected if they want to limit their tweet visibility. Since our aim is to study users who switch between public and protected, we collected users with protected accounts which indicates that they have changed their tweet visibility at least once. To obtain our sample of protected users, we first sampled the Twitter stream without any filters and collected a list of mentioned protected users from these tweets for 60 hours starting on the 29th June 2020. During this period, we managed to collect tweets that contain mentions to 3.15M users. We inspected the account visibility of each of those users using the Twitter user API, and found that around 4% of them are protected accounts at the time of data collection.

Our initial pool of protected users contain around 107K accounts. We applied a frequent check to get the account visibility of these protected accounts every 30 minutes for three months from 17th July until 17th October 2020[1]. Out of the 107K users that were protected on the first check, 64.5K (˜60%) of them stayed protected during all three months. The remaining 42.5K (˜40%) have switched their account visibility to public at least once during this period, including 7.6K (7.1%) who changed to public and stayed that way. Table 5.1 shows the number of switches users made in the three months we collected their account visibility.

A user changing their tweet visibility settings frequently does not necessarily mean

[1]During this period, there were some instances where we could not reach some users because they were deactivated, suspended, or there were temporary problems with the Twitter API. There was only one instance where we could not reach any accounts because of a service disruption.

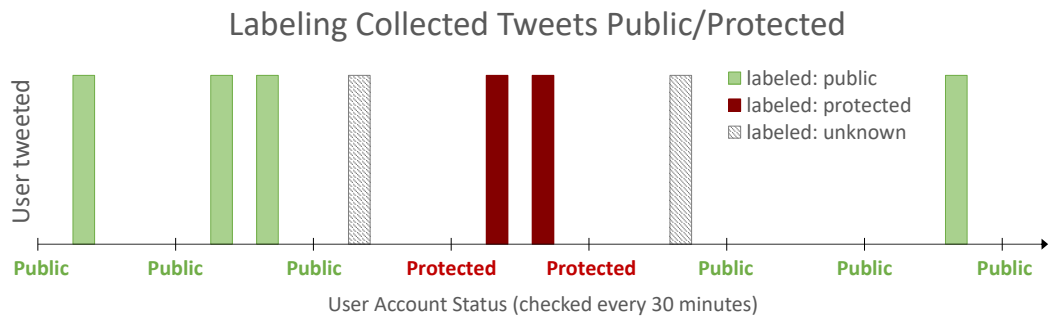

Figure 5.1: Example of the labeling process of the tweets of users in the dataset.

they stay protected/public equally. They can predominately stay protected or public most of the time and only switch for short time periods. Hence, we calculated how long our users stayed protected/public during the three months. For each user, we compared the number of 30 minutes blocks when they were protected to the number of blocks they were public. 14.6K (34.5%) of the users who changed their account visibility at least once preferred to stay protected more than 95% of the time. More than half of the switchers stayed protected 80%+ of the time.

5.3.2 Collection and Labeling of User Tweets

We collected the tweets of the switching users when they were in public settings. While it is possible to collect the last 3200 tweets of a user via Twitter API, we only collected the tweets posted during the account type check period (i.e. between 17th July and 17th October 2020). We managed to collect the tweets of 25.6K (60.4%) users out of the 42.5K switching accounts in our collection. Some of these users change their account visibility frequently and stay public for a short time. Hence, we could not get any tweets from the remaining 16.9K users.

After collecting the tweets, and since we have the visibility status of the account every 30 minutes, we labeled them according to the collected account visibility of the user when the tweet was posted. If a tweet was sent at time t, we compared the user's account visibility collected just before and after t. If the account visibility were the same, we labeled the tweet with that. In cases where the account visibilities were different or we could not reach one of them, we discarded those tweets and did not include them in our analysis. Figure 5.1 illustrates an example of the labeling process of the tweets in a users timeline. For simplification purposes, we call tweets we labeled as public t_o and the ones we labeled as protected t_x. Table 5.2 gives the notations and

Notation	Definition
t_o	Tweets labeled as public by our system, i.e. user was public in our account type checks preceding and succeeding the tweet.
t_x	Tweets labeled as protected by our system, i.e. user was protected in our account type checks preceding and succeeding the tweet.
u_o	Users who stayed public 90%+ of the time we collected the account types.
u_b	Users who stayed protected between 10% and 90% of the time we collected the account types.
u_x	Users who stayed protected 90%+ of the time we collected the account types.

Table 5.2: The notations and the definitions we use for users and labeled tweets of the Twitter data collection.

% time as protected	[0-10]%	(10-50]%	(50-90)%	90+%
# Users	2.6K	9.8K	10.3K	3K
# t_o	5.3M	14.3M	6M	375K
# t_x	214K	3.4M	7.1M	2.5M
# $t_o + t_x$	5.5M	17.7M	13.1M	2.9M

Table 5.3: User counts by the percentage of time the account was protected during the three months of data collection. Also the total number of t_o and t_x collected for each group of users. Only the users we could collect tweets from are reported.

definitions for the tweets we labeled . In the end, we were able to label nearly 39.2M tweets (including retweets and replies) collected from the 25.6K switching accounts in the three months, where 26M (66.2%) of them are t_o while the remaining 13.2M (33.7%) are t_x. Nearly 9M (23%) of these tweets were labeled as English by Twitter, and the remaining 77% covered 64 other languages including Japanese, Portuguese, Indonesian, and Korean in the order of size after English in our dataset.

Table 5.3 shows the user counts by percentages of the time the account was protected during the three months, as well as the number of total tweets we collected from these users. There is a positive relation between the time users stay protected and the number of tweets they send while they are protected. However, some users tend to tweet less when they were protected even if they stayed protected most of the time.

5.3.3 Comparing tweets characteristics of users

To understand the reasons why users switch their account visibility, we divided them into three groups by their privacy settings usage and analyzed them separately. These groups are u_x, the users who stay *protected* 90% of the time or more, u_o, the users who stay *public* 90% of the time or more, and u_b, the remaining users who have more *balanced* duration of staying public or protected compared to u_x and u_o. In this section, we compare 11 features of t_x and t_o tweets of u_x, u_b, and u_o. Table 5.2 shows the notations and the definitions we use for users and tweets of the Twitter data collection.

As shown in Table 5.3, out of the users we collected tweets from, 3K (11.7%) of them were u_x, 20.1K (78.3%) of them were u_b, and 2.6K (10%) of them were u_o. Users in our dataset have the majority of their tweets with the same visibility state as the dominant setting.

Here we compare the tweeting frequencies, language-independent tweet features, and English tweeting behaviour of the three user groups when they tweet under public or protected settings. Looking at tweet features can give us insights on why users might change their visibility settings back and forth. Hence, for each tweet we look at the following features: mentions (tweet has at least one mentioned username), verified mentions (mention of verified user), non-follower mentions, reply (the tweet is a reply to another existing tweet), retweet (RT), quote tweet (QT), URLs, hashtags, media (e.g. photo and video), and whether the tweet is in English using the language label given by the Twitter API. We did not take the geo-location as a feature since only a small portion of the users had it enabled. We collected the user profiles of the mentions to check whether if they were verified. We also collected the followers of our users to check whether if they mention non-followers in their tweets.[2] We extract these features from the tweets of u_x, u_b, and u_o. We then calculate the percentage of prevalence in t_x and t_o for each of the features and compare them for every user. We also divide the number of tweets a user sent while public/protected to the duration of staying public/protected so we can calculate the tweeting frequencies. We compare these features using paired sample t-test and apply Bonferroni correction post-hoc with p-values divided by 14.[3]For this analysis, 168, 675, and 227 users from u_x, u_b, and u_o respectively were discarded, since they had tweets of one visibility only.

[2]The data collection for both of these features are done at a later time than the initial data collection. Considering Twitter networks are dynamic, the state of the network might have changed between these two collection times.

[3]We made at most 14 comparisons for each user group but only report 11 of them in this study.

		$\overline{t_x}$	95% CI	$\overline{t_o}$	95% CI	p-value
	Tweet Frequency	0.22	[0.21, 0.23]	0.59	[0.54, 0.63]	***<**.001**
	Mentions	61.79	[60.99, 62.59]	64.11	[63.19 , 65.02]	***<**.001**
	Verified Ment.	10.34	[9.63, 11.05]	15.30	[14.25, 16.35]	***<**.001**
	Non-follower Ment.	52.28	[50.34, 54.23]	56.80	[54.66, 58.94]	***<**.001**
	Reply	42.43	[41.48, 43.39]	43.29	[42.23, 44.35]	.022
u_x	RT	22.59	[21.68, 23.51]	22.02	[21.07, 22.96]	.058
	QT	8.15	[7.80, 8.50]	8.39	[7.98, 8.81]	.169
	Urls	7.99	[7.68, 8.30]	8.49	[8.07, 8.92]	.013
	Hashtags	3.70	[3.43, 3.97]	5.60	[5.15, 6.05]	***<**.001**
	Media	15.12	[14.61, 15.63]	15.07	[14.48, 15.65]	.835
	English	17.06	[15.98, 18.14]	17.28	[16.17, 18.39]	.178
	Tweet Frequency	0.28	[0.27, 0.28]	0.50	[0.48, 0.51]	***<**.001**
	Mentions	64.08	[63.77, 64.38]	65.73	[65.45, 66.02]	***<**.001**
	Verified Ment.	11.00	[10.76, 11.25]	11.10	[10.86, 11.34]	.228
	Non-follower Ment.	57.16	[56.62, 57.71]	57.41	[56.89, 57.93]	.097
	Reply	41.31	[40.94, 41.68]	41.73	[41.38, 42.08]	***<**.001**
u_b	RT	25.57	[25.20, 25.93]	25.19	[24.85, 25.53]	***<**.001**
	QT	8.53	[8.40, 8.66]	8.61	[8.48, 8.73]	.065
	Urls	8.36	[8.24, 8.49]	8.69	[8.56, 8.82]	***<**.001**
	Hashtags	4.19	[4.07, 4.31]	5.16	[5.03, 5.29]	***<**.001**
	Media	15.47	[15.27, 15.67]	15.80	[15.61, 15.98]	***<**.001**
	English	19.55	[19.11, 19.99]	19.63	[19.19, 20.07]	.090
	Tweet Frequency	0.47	[0.44, 0.50]	0.54	[0.51, 0.57]	***<**.001**
	Mentions	66.43	[65.41, 67.44]	68.64	[67.86, 69.41]	***<**.001**
	Verified Ment.	15.78	[14.69, 16.86]	13.36	[12.50, 14.23]	***<**.001**
	Non-follower Ment.	57.35	[55.89, 58.80]	55.20	[53.96, 56.45]	***<**.001**
	Reply	40.35	[39.19, 41.52]	42.45	[41.47, 43.44]	***<**.001**
u_o	RT	27.35	[26.20, 28.50]	26.78	[25.80, 27.77]	.093
	QT	8.82	[8.33, 9.31]	8.94	[8.60, 9.28]	.522
	Urls	8.82	[8.32, 9.32]	8.91	[8.57, 9.25]	.673
	Hashtags	4.82	[4.36, 5.28]	5.50	[5.14, 5.87]	***.001**
	Media	16.99	[16.25, 17.72]	16.13	[15.62, 16.64]	.004
	English	23.05	[21.62, 24.48]	23.36	[21.98, 24.74]	.127

Table 5.4: Paired samples t-test comparing number of tweets of u_x, u_b, and u_o when they are protected and public (with Bonferroni correction *$p < (0.05/14)$, **$p < (0.01/14)$, ***$p < (0.001/14)$).

Table 5.4 shows the means of values, 95% confidence intervals, and paired t-test p-values for each feature of the tweets for u_x, u_b, and u_o. Users in u_x tweet more when they are public. They also mention Twitter users in their tweets more when they are public, albeit with a low effect size. The effect sizes are bigger for mentioning verified users and non-followers. Since a user's tweets can only be seen by followers, it is not surprising that they are switching to public to mention a verified account or someone who does not follow them. u_x also share more tweets with hashtags in them when they are public. They share media and links along with their tweets, reply, retweet, and quote tweet in similar rates. The effect sizes for comparisons of tweeting frequency, verified user mentions, and non-followers mentions were small and the remaining comparisons all had very small effect sizes.

Similarly with u_x, users in u_b tweet and mention other Twitter users in their tweets more while in the public setting. They also share media, links, and hashtags in their tweets and reply more when public. On the other hand, they retweet more when they are protected. Surprisingly, they mention verified users and non-followers in similar rates. They also tweet in English and quote tweet similarly when public and protected. Aside from the tweeting frequency, which had a small effect size, effect sizes were all very small for comparisons of u_b.

Users in u_o also tweet more when they are public. They mention users, reply, and share hashtags more when they are public. However, surprisingly, they mention verified users and non-followers relatively more when they are protected. They tweet in English, retweet, quote tweet, share media, and links in similar rates. The effect sizes for comparisons of u_o were all very small.

5.4 User Surveys

The above data collection and analysis provides an interesting view on *what* users did when they switch account visibility, but it does not provide an understanding of *why* they are engaging in these actions. In this section we describe two surveys that we ran to better understand the reasons users change their settings and how common those reasons are. We start with an initial open-ended survey where we elicit reasons why people have switched in the past. We then combine the results of that survey with related work and our own observations during the data collection detailed in the previous section to create a list of common reasons people might change their visibility settings. We use the list in a second closed-ended survey to understand how common

different switching reasons are. We give the full list and the sources of each reason (i.e. Open-ended survey, Twitter data collection) in the Appendix D.

5.4.1 Identifying reasons to switch account visibility

The first survey aims to identify common reasons that lead users to switch between public and protected. The survey is not intended to identify a comprehensive set of reasons for switching, and instead focuses on identifying common reasons which are then combined with information from literature (e.g. [112, 132, 158]) to construct the options used in the next survey.

We recruited 100 Twitter users who have switched their tweet visibility at least once in the prior two months. The study was advertised as a "Short survey about changing your tweets between public and protected on Twitter" to Prolific Academic (PA) [128] users who are fluent English speakers on March, 2021. We conducted 2 pilots surveys with 6 participants each to adjust the wording of the questions as well as to get an accurate estimation of the time required. 100 participants took the final survey with an average completion time of 5.5 minutes and a compensation of £1.15. The study design was approved through our University's ethics process.

Survey Instrument: The survey consisted of: informed consent, a screening question about if they had changed their Twitter account visibility (the ad clearly stated that this was required), four retrospective questions described below, demographics questions which included Twitter usage questions, and an optional free text comments box. For the retrospective questions we asked them to "think back to a recent time when you have changed your Tweets from public to protected or from protected to public." The first question was free text and asked them what had motivated them to make the switch, followed by another free text question asking them what they hoped to achieve by making the switch. We also asked if the switch enabled them "to achieve the effect I was trying for" (Likert) and what direction the switch was in (to public, to protected, or multiple changes). We share this survey in Appendix B.

Analysis: Two researchers reviewed the responses and determined that the two free-text answers were often very related to each other, where the answer for what they hoped to achieve expanding on the motivation for the switch, such that viewing them together provided a more comprehensive understanding of the reason. The answers were therefore analyzed together. We used an affinity diagram [59] type approach to conduct a thematic analysis. The answers were placed in a shared spreadsheet,

one researcher went through and sorted the answers into themes. They then met and the second researcher who read through all the answers and adjusted the themes discussing with the first researcher as they went. The result was a set of themes that both researchers agreed on. The resulting themes can be seen in Table 5.5.

Participants: Participants were 23.6 years old on average with a max age of 39 years. 52 identified as male and 48 female. 65 described a situation where they switched to protected, 11 described switching to public, and 24 described an event that required multiple visibility changes. When asked about the normal visibility of their account, 42% indicated that they kept their account public most of the time, 27% kept it mostly protected, while 31% indicated that they switched more often. 62% had also tweeted within the last week.

5.4.1.1 Reasons to Switch:

Participants most commonly switched to protected to talk about sensitive, political, or controversial topics freely. Reasons in this theme also overlapped with other themes, particularly around avoiding problematic interactions with strangers and proactively limiting the audience. Another common reason was to prevent the people they know, such as friends and family, from seeing their tweets. Some of these users were uncomfortable because a person they knew found their account leading them to switch visibility. Other participants were uncomfortable with strangers or non-followers seeing their tweets. Participants also changed to protected temporarily to share personal information such as private Instagram links and then deleted the tweet before switching back to public. Harassment, having a tweet go viral, and sharing tweets that were not safe for work (NSFW) were some of the other reasons given. Uncommon but interesting reasons included wanting to archive an account and to avoid getting suspended due to complaints.

Our participants primarily changed to public to reach a broader audience, interact with non-followers, as well as participating in giveaways where the account must be public to get selected for the prize. One participant cited all of these reasons when explaining why they changed to public. *"I hoped to achieve what the platform could offer to its fullest, that included retweeting thing in order to comment on them, "taking part" in a trending hashtag or even taking part in giveaways."* Interestingly, one participant discussed going public for a while in order to gain more followers, then switching back to protected. Others said that it was challenging for friends and family to find them when they were protected, so they went public for a while to make them-

Theme	# users	Theme	# users
Sensitive/Political/Controversial	19	Giveaways*	4
To prevent people I know from seeing	16	Harassment	3
Personal Information	15	Avoid suspension	3
To prevent non-followers from seeing	10	Viral Tweet	3
Sense of privacy	10	Archival purposes	2
Broader audience*	7	NSFW	1
Interacting with non-followers*	4		

Table 5.5: Themes of free-text answers about reasons of switching. Most reasons are for switching from public to protected, while (*) indicates reason to switch from protected to public.

selves easier to find. We also asked if changing visibility enabled them to achieve the effect they were trying for and 89% of the participants agreed that it did.

5.4.2 Prevalence of Reasons

The second survey looked at how common different reasons for switching visibility are. We used a combination of the results of the first survey, our observations from the Twitter data collection, and findings from related work to construct lists of reasons people switch to public and to protected. Then used the survey to find how many participants have switched due to these reasons.

We advertised a study entitled "5 min survey about changing your tweets between public and protected on Twitter" on PA to fluent English speakers on April, 2021. The advertisement stated that we were looking for people who had changed their tweet visibility between public and protected two or more times in the last year to ensure respondents had experience with switching. We conducted a pilot survey with 5 participant to ensure accurate time estimates. 324 participants took the survey, requiring an average of 4 minutes to complete. They were compensated £0.75. Ethics approval was given by our University.

Survey Instrument: Participants were first shown a consent page, followed by a screening to see if they had switched two or more times in the last year, followed by

three multiple answer questions, questions about their Twitter experience, demographics, and an optional comment box. The first two multiple answer questions asked users to indicate "which of the following has previously lead you to change your tweet visibility settings to public?" and also for protected. The third multi answer question showed a list of protection activities, like deleting tweets when moving from protect to public, and asked users to select those they had used previously. This second survey for this study is given in Appendix C.

Participants: Participants had an average age of 25 years old ($\sigma = 7.6$). 46% identified as female, 53.4% as male, with the remainder preferring not to say.

Participants were asked what their normal public/protected balance was on their account. 30.5% indicated their account was "Always" or "Mostly" protected (Protected here forth), 28.4% indicated their account was 'Always" or "Mostly" public (Public), and finally 41.0% indicated that their visibility was "Balanced" or "Somewhat" public/protected who are referred as Balanced users for the remainder of this section.

We also gave users a scenario about a person who tweeted while protected and then switched to public and asked if their tweet was now public or protected. The majority (67.9%) of participants accurately said that the tweet would now be publicly visible. 16.4% thought that only users logged into Twitter could see it, but were aware that the tweet would be public to all Twitter users. But 15.7% incorrectly thought that the tweet would remain protected.

We asked participants how often they had changed their visibility in the last three months. 19% had not changed their tweet visibility settings in the last three months. 35.8% had changed once, 27.5% twice, 15.4% three to five times, and 2.2% changed six or more times.

5.4.2.1 Results:

We asked participants to select all of the reasons that have caused them to change their visibility settings to public (Table 5.6) and protected (Table 5.8). On average users selected 3.3 ($\sigma = 2.1$) options from the protected reasons list and 3.4 ($\sigma = 2.0$) options from the public reasons list. The small number of selections and the range of selections shown in the table suggest that respondents were indeed reading the choices and only selecting those that applied to them. We also give the reasons selected by participants divided by sex and age in Table 5.7 and Table 5.9.

The most common reason to turn public for those that were mostly public or balanced was to "reach a broader audience and get more interactions". The finding makes

Reasons to turn **public**	Public	Balanced	Protected
To reach a broader audience and get more interaction with my tweets	65.7%	50.4%	39.1%
To mention/reply to a user who does not follow me	49.5%	42.9%	41.3%
To retweet other users	49.5%	41.4%	31.5%
To gain more followers	43.4%	33.1%	27.2%
To quote tweet other users	33.3%	27.1%	21.7%
To enter to get giveaways or freebies	21.2%	21.8%	20.7%
To share articles or links	20.2%	20.3%	15.2%
To share pictures	23.2%	15%	16.3%
To associate a tweet with hashtags or trends publicly	24.2%	12.8%	14.1%
To boost the visibility, popularity, or ranking of a hashtag or topic	25.3%	13.5%	8.7%
To mention/reply to celebrities, famous people, or other VIPs	17.2%	9.0%	8.7%
To have a professional image	7.1%	14.3%	6.5%
To get customer service	9.1%	9.8%	6.5%
To boost the visibility of another user's tweet	11.1%	6.8%	8.7%
To find potential employment	0	5.3%	3.3%
To sell things or receive donations	4%	1.5%	0
Other	1.0%	1.5%	0
I did not change my tweet visibility settings to public before	2.0%	0.8%	7.6%

Table 5.6: Reasons to turn public. Answers divided based on if their account is more often protected, public, or balanced. Percentages are out of the total number of mostly public, balanced, and mostly protected participants respectively.

Reasons to turn **public** (%)	Female	Male	18-24	25-34	35-44	45-54	55-64
To reach a broader audience and get more interaction with my tweets	57.7	46.8	52	57.6	39.3	37.5	33.3
To mention/reply to a user who does not follow me	53	37.6	44.5	51.8	21.4	37.5	66.7
To retweet other users	38.9	43.4	44.5	38.8	28.6	25	33.3
To gain more followers	38.3	31.8	38.5	31.8	21.4	12.5	33.3
To quote tweet other users	31.5	24.3	30	25.9	14.3	25	33.3
To enter to get giveaways or freebies	19.5	23.1	21	24.7	17.9	12.5	0
To share articles or links	16.8	20.8	14.5	23.5	28.6	50	0
To share pictures	18.1	17.9	19.5	14.1	14.3	37.5	0
To associate a tweet with hashtags or trends publicly	24.8	9.8	16.5	16.5	17.9	25	0
To boost the visibility, popularity, or ranking of a hashtag or topic	19.5	12.7	15.5	17.6	14.3	12.5	0
To mention/reply to celebrities, famous people, or other VIPs	15.4	8.1	14	9.4	0	12.5	0
To have a professional image	8.7	11	5.5	14.1	28.6	12.5	0
To get customer service	8.1	9.2	4.5	15.3	21.4	0	0
To boost the visibility of another user's tweet	9.4	8.1	9.5	8.2	7.1	0	0
To find potential employment	6	0.6	2.5	1.2	14.3	0	0
To sell things or receive donations	1.3	2.3	2	0	3.6	12.5	0
Other	0.7	0.6	0.5	2.4	0	0	0
I did not change my tweet visibility settings to public before	2	4	4.5	0	0	12.5	0

Table 5.7: Reasons to turn public, divided by sex and age.

Reasons to turn **protected**	Public	Balanced	Protected
I wanted to prevent non-followers from seeing tweets with personal content	50.5%	52.6%	63.0%
People I know found my account and that made me uncomfortable	47.5%	51.9%	44.6%
To get a sense of privacy	35.4%	38.3%	46.7%
To prevent people I know, such as friends and family, from seeing my tweets	36.4%	38.3%	42.4%
To prevent interactions from strangers	30.3%	26.3%	38.0%
To avoid harassment	23.2%	27.1%	22.8%
To talk about a sensitive, controversial, or political topics freely	17.2%	15.8%	14.1%
To take a temporary break from interactions with non-followers	19.2%	14.3%	4.3%
I did not want people to retweet me	11.1%	8.3%	14.1%
To archive the account without deleting it	10.1%	9.8%	8.7%
To tweet about someone without them being able to see the tweets	9.1%	9.0%	8.7%
I did not want people to quote tweet me	13.1%	6.0%	7.6%
To share pictures	5.1%	6.8%	10.9%
My tweet unexpectedly went viral	6.1%	2.3%	6.5%
To share content that is not safe for work (NSFW)	5.1%	4.5%	3.3%
To quote tweet other users	3.0%	4.5%	1.1%
To prevent account suspension	2.0%	2.3%	5.4%
To share articles or links	4.0%	0.8%	3.3%
To retweet other users	2.0%	3.0%	0
Other	3.0%	1.5%	1.1%
I did not change my tweet visibility settings to protected before	1.0%	3.0%	0

Table 5.8: Reasons to turn protected. Answers divided based on if their account is more often protected, public, or balanced. Percentages are out of the total number of mostly public, balanced, and mostly protected participants respectively.

Reasons to turn **protected** (%)	Female	Male	18-24	25-34	35-44	45-54	55-64
I wanted to prevent non-followers from seeing tweets with personal content	63.1	48	55	52.9	64.3	50	33.3
People I know found my account and that made me uncomfortable	53	45.1	53	44.7	35.7	25	33.3
To get a sense of privacy	46.3	34.7	41.5	36.5	35.7	37.5	66.7
To prevent people I know, such as friends and family, from seeing my tweets	43	35.8	44.5	35.3	25	0	0
To prevent interactions from strangers	38.3	24.9	30	30.6	39.3	37.5	0
To avoid harassment	27.5	22.5	24	25.9	17.9	50	33.3
To talk about a sensitive, controversial, or political topics freely	16.8	15	15.5	17.6	10.7	12.5	33.3
To take a temporary break from interactions with non-followers	18.1	8.7	13.5	14.1	10.7	0	0
I did not want people to retweet me	8.7	12.7	11.5	11.8	3.6	12.5	0
To archive the account without deleting it	10.7	8.7	9.5	9.4	14.3	0	0
To tweet about someone without them being able to see the tweets	12.8	5.8	11.5	7.1	0	0	0
I did not want people to quote tweet me	8.1	9.2	9.5	9.4	0	0	33.3
To share pictures	9.4	5.8	9	1.2	14.3	12.5	0
My tweet unexpectedly went viral	4.7	4.6	5	3.5	3.6	12.5	0
To share content that is not safe for work (NSFW)	3.4	5.2	3.5	5.9	7.1	0	0
To quote tweet other users	3.4	2.9	4	1.2	0	12.5	0
To prevent account suspension	3.4	2.9	4.5	1.2	0	0	0
To share articles or links	2	2.9	2	2.4	3.6	12.5	0
To retweet other users	1.3	2.3	2	2.4	0	0	0
Other	2	1.2	2.5	1.2	0	0	0
I did not change my tweet visibility settings to protected before	2	1.2	2	1.2	0	0	0

Table 5.9: Reasons to turn protected, divided by sex and age.

sense given that interacting with tweets is a key functionality of Twitter. The most common set of reasons to turn public was to be able to interact with other accounts through mentioning, replying, retweeting, and quote tweeting. For those who were mostly protected, interactions like these seem to be the main driver to change visibility.

The most common reason to turn protected for all groups was "to prevent non-followers from seeing tweets with personal content". In the first survey we saw several people switching to protected because someone they knew had found their account, and in this survey we similarly see that being a large reason to switch, even more so than preventing friends and family from seeing tweets, suggesting that the issue is around specific individuals more than just people they know. Preventing interaction from strangers and harassment were also common answers. "To get a sense of privacy" was also a common answer, though interestingly not the most common. In the first survey, a common vague answer was that the user just wanted more privacy, which is why we added the answer option to the second survey despite its low specificity. Interestingly, sharing content like pictures, links, and not safe for work content were not common reasons to turn protected. Similarly, interacting with other users, which was a common reason to turn public, was not a common reason to turn protected.

We also examined the combinations of answers users gave to why to turn protected and public. The most frequent pair was "To reach a broader audience and get more interaction with my tweets" and "I wanted to prevent non-followers from seeing tweets with personal content" with 30.2% of participants giving both those answers. Similarly, 18.5% both indicated turning public to get more interactions while also indicating that they turned protected to prevent interactions from strangers. Another pair selected by 11.4% participants was "To enter to get giveaways or freebies" and "I wanted to prevent non-followers from seeing tweets with personal content".

We also asked the participants to select actions they have taken to control who can see and interact with their tweets (Table 5.10 and Table 5.11). More than half of participants indicated that they delete some or all of their tweets before changing their visibility to public. The control approaches of soft blocking, muting, and blocking were also used by many. Majority protected participants in particular preferred soft blocking, possibly because it removed followers without creating a notification to the impacted Twitter account and may be seen as less harsh.

Interestingly, 23.1% of participants indicated either temporarily deactivating their accounts to prevents interactions or changing to protected when they are not logged in

Control Visibility & Interactions	Public	Balanced	Protected
Delete some or all tweets when moving from protected to public	49.5%	57.9%	52.2%
Remove a follower without blocking them (soft blocking)	46.5%	43.6%	54.3%
Mute a follower so you can't see their interactions	44.4%	39.8%	39.1%
Block followers to prevent them from interacting even when your account is public	34.3%	36.8%	33.7%
Temporarily deactivate your account to prevent all interaction for a time	17.2%	16.5%	10.9%
Change to protected when not logged in or otherwise unable to respond to interactions	12.1%	10.5%	7.6%
Have a clear list of engagement rules prominently shown or linked to that detail what is acceptable interaction or following behaviour	4.0%	0.8%	1.1%
Other	0	0	1.0%
I have not used any of the above	6.1%	4.5%	3.3%

Table 5.10: Actions taken to control who can see and interact with tweets. Answers divided based on if their account is more often protected, public, or balanced. Percentages are out of the total number of mostly public, balanced, and mostly protected participants respectively.

and cannot respond to interactions. Twitter allows 30-day grace period for accounts that are deactivated, if a user re-activates their account in this time frame all content will be restored. We added these options based on prior research showing that some Facebook users delete their accounts when not online as a way of preventing interactions when they are unable to respond [112]. Based on these results, some Twitter users may be using a similar tactic. Temporarily deactivating might also explain why we could not reach some accounts during the automated Twitter data collection.

5.5 Discussion

5.5.1 Summary of Findings

In this chapter, we focused on our third main research question concerning privacy settings usage and reasons behind the behaviour of frequently changing these settings on

Control Visibility & Interactions (%)	Female	Male	18-24	25-34	35-44	45-54	55-64
Delete some or all tweets when moving from protected to public	57.7	50.9	55	50.6	53.6	50	66.7
Remove a follower without blocking them (soft blocking)	53	43.4	51.5	44.7	32.1	37.5	33.3
Mute a follower so you can't see their interactions	43	39.3	46.5	35.3	28.6	12.5	33.3
Block followers to prevent them from interacting even when your account is public	36.2	34.7	37.5	35.3	28.6	0	33.3
Temporarily deactivate your account to prevent all interaction for a time	18.1	12.7	16	15.3	7.1	25	0
Change to protected when not logged in or otherwise unable to respond to interactions	7.4	12.1	10	9.4	14.3	12.5	0
Have a clear list of engagement rules prominently shown or linked to that detail what is acceptable interaction or following behaviour	0.7	2.9	2	1.2	3.6	0	0
Other	0	0.6	0.5	0	0	0	0
I have not used any of the above	5.4	4	4	3.5	7.1	25	0

Table 5.11: Actions taken to control who can see and interact with tweets, divided by sex and age.

Twitter. Firstly, we quantified privacy settings usage on Twitter to answer our RQ3.1 "How frequently do Twitter users change their tweet visibility settings?". After monitoring a set of 107K protected accounts on Twitter over three months, we found that 40% of these accounts changed their visibility to public at least once over this period. Over a quarter of those who changed, did so 10 or more times. This shows an interesting behaviour of Twitter users when it comes to their use of tweet visibility settings, which motivated our RQ3.2 related to the posting strategies of these users when they are public vs. protected. Our analysis showed that users tweet less frequently when they are protected, even for those users who stayed mostly protected (u_x) during the data collection. It was also interesting to find that all user groups mention others more when they are public. However, u_o mention verified users or non-followers more while protected, u_x mentions them more while public, while there is no change for u_b.

This indicates that users change their usual setting to mention verified users and non-followers. This also applies to hashtags, where all user groups use hashtags more when they are public. Effect sizes for these results vary between small and very small. Aside from tweeting frequency for u_x and u_b, verified mention and non-follower mention for u_x, the effect sizes were very small.

We conducted two user surveys to answer the remaining research questions RQ3.3 and RQ3.4. We investigated the reasons behind the tweet visibility changes between public and protected to answer our RQ3.3. Over 40% of our participants changed their tweet visibility settings to public so they can mention non-followers. Other interaction-based reasons such as retweeting, quote tweeting, and entering giveaways were also popular reasons to turn public. On the other hand, our participants changed their tweet visibility to protected so they can limit the audience and unwanted interactions. Our participants were also wary of people they know, e.g. family and friends, finding their accounts. Avoiding harassment and being able to talk freely about possibly controversial topics were among the most popular reasons to turn protected.

Our RQ3.4 concerns the strategies users utilize to manage their audience and interactions. The most popular strategy our participants employed was deleting tweets before moving the account from protected to public. Hidden preventions such as soft-blocking, blocking someone and unblocking them immediately to remove them from followers, and muting were more popular compared to blocking, which can be noticed by the blocked person. Our participants also changed their account visibility to protected when they are unable to respond to interactions or deactivated their accounts so they can prevent all kinds of interactions including interactions from followers.

5.5.2 Implications

Twitter's relatively simplistic privacy settings imply that accounts are either public, where anyone on the internet can see all their tweets, or protected, where only followers can see the tweets. An observation from this work is that for many users the visibility of their tweets is not a static setting but instead one that changes as their needs and circumstances change. For these users, it is inaccurate to think of them as simply public or protected as they are changing their settings to get a mix of the affordances granted by both those states.

Current setting language also focuses around the protection of the tweets themselves rather than the person: "Unprotecting your Tweets will cause any previously

protected Tweets to be made public" [30]. We find that users made changes to settings to protect themselves (i.e. harassment) or to create safe spaces (i.e. discussing controversial topics), where the focus in both cases was around the users themselves and their ability to have conversations with other users in a private space rather than to necessarily protect the tweets themselves. The overlap here is obviously large since the conversations they are having are happening via tweets so they have to protect the tweets in order to create a safe space to have these conversations. But the conceptual difference is quite important to understand as it contextualizes the different actions users are taking to protect themselves.

Privacy expectations and notions of people are dynamic [10] representing a continual effort to control their presentation of self in a social context [54]. While Twitter's binary setting option is simple, which helps with understanding it, that simplicity also makes it more challenging to conduct fine grain boundary management. Particularly since users are not just moving themselves between a protected and public space, the way they might when they leave the house to go to a coffee shop, they are also bringing all their past tweets with them when they move between audience spaces, closer to holding an open house where anyone can enter the previously private space. The bringing of past events with them when they move between audience spaces makes proper boundary management more challenging since the whole tweet feed has to be curated to be appropriate for the new audience which can be cumbersome [164, 64, 174]. Alternatively users may rely on self-censorship [111, 140] where they refrain from sharing sensitive information even when protected to limit the risks of it becoming public later. Approaches like these, or even lack of awareness, may be why 46% of survey respondents indicated that they don't delete tweets when changing from protected to public. Another reason for not deleting past tweets might be the archival value of historical tweets and protecting meaningful self-representation [164, 64, 174].

Some of our users definitely curated their feeds when switching between protected and public. Deleting tweets was the most common audience control action selected in the survey, and in the tweet dataset we noticed fewer tweets when people were protected than when they were public which *might* be an indication that users are deleting tweets before switching. Participants also mentioned the need to briefly protect a specific tweet or be able to briefly engage with a public account. For example, a participant mentioned being protected to share a private Instagram account with their followers to gain *"a momentary bit of privacy before of deleting the tweet and making my account public again."* Prior work has also shown that users delete their tweets for various rea-

sons including preventing harassment, revealing too much information [141], or simple typos [9]. In line with Mondal et al. [114], we find that users utilize deletion to protect their privacy, control information flow, and manage audiences.

Users both want to "achieve what the platform could offer to its fullest" while also "keep tweeting freely". They are not just looking for a way to create a safe space, they also want to be able to interact with the wider community by doing things like replying to a public tweet or making parts of their tweet stream public. Hence, we see the active management of tweet visibility in our dataset where we find that some users changed their settings more than 200 times in three months. The nature of the Twitter usage also seems to affect this behaviour. As a participant stated: *"how much you change your public settings depends on which side of twitter you are. I have a fan account so I get more harassment than if I had a personal account and I change my setting more often"*.

One solution our participants made use of was to have multiple accounts which had different settings, similar to Stutzman and Hartzog [146]. A participant stated: *"in my twitter bubble it is very common to have two accounts, one public (and anonymous) and one private for the few close and trusted friends made on the platform, talking about private matters and sharing images ect."*. Instagram users use "finstas" for similar purposes [172]. Some users even appear to share their protected accounts on their bio and ask to be "private moots (mutual follow)" where they usually add only protected users to their network to minimize information leak, which can happen if they have conversations with public accounts.

5.5.3 Design Recommendations

The largest design recommendation we have is that Twitter should think about tweet visibility in terms of conversations, time frames, or audiences rather than focusing on the account level only.

Visibility Pinning of Tweets: Provide users with the ability to "pin" a tweet as either public or protected so that it stays that way even when the account changes visibility. Doing so would allow users to more easily share a single tweet publicly or privately without needing to change their whole account visibility. It would also allow predominately protected users to share a public set of tweets that would allow them to attract new followers without opening their entire account up.

Paired Accounts: Directly support the practice of having both public and pro-

tected accounts. Users already use this approach to manage the visibility of their tweets and their audiences, but they are doing it ad hoc. Twitter could more directly support this practice by offering to create two linked accounts and making it easier to switch between them to post to appropriate audiences. Doing so would allow for better boundary management as well as assist users who are looking to gain followers by having some tweets public.

Temporal Settings: Make some settings time based so that they encompass a time period or a conversation stream rather than just a single tweet. A user may change to protected to have a private conversation with their followers and then change back to public, but then they would need to go an delete all the tweets happening in that time frame. It would be helpful to users to have the ability to delete tweets that happened within a frame of time. It may also be helpful to have tweets that disappear after a set time frame or disappear when the account visibility changes. Posts with time-limits that disappear after a set time is already adopted by various social media platforms including Snapchat, Instagram, and WeChat. These features help users to cope with temporal context collapse to a degree and users appreciate these ephemeral posts [64]. However, the time-limit should be controlled by the users instead of the platform. According to Yilmaz et al. [174], users found the automated deletion is beneficial only when the users are in control of the deletion date. Tweet deletion is also a common practice among users who do not necessarily switch their account visibility [114, 9, 141] so this feature may benefit them too.

Limiting Interactions: As there were participants who made their account public to interact, there were also participants who protected their accounts to limit interactions from non-followers. Some of our participants even deactivated their accounts temporarily to prevent all interactions for a while. There were also some participants who chose to limit interactions when they are unable to respond to them. Twitter recently introduced a feature for users to limit the replies to specific tweets. However, other interactions such as retweeting, quote tweeting, and liking are still possible for these tweets. Especially, quote tweets can be used as replies by other users when replies are limited, which also increases the reach of the original tweet more than the replies [52]. Users can stop all interaction by deactivating their accounts up to 30 days without deletion but deactivation will hide users' tweets, even from the user themselves. Enabling users to stop all interactions without changing their account visibility to protected or deactivating can help the participants adopting those strategies to limit/stop interactions. Additionally, prior work has shown that people see an archival

value in their social media accounts [64, 175, 164]. This feature will benefit nearly 10% of our participants who also wanted to archive their accounts without deleting it.

Removing Followers: Allow users to remove followers directly rather than requiring them to soft-block them. Doing so would allow users to more easily curate their followers when they switch between public and protected. It would also be important to consider allowing them to do so silently without raising a notification to the other person since people prefer online sanctions that can create "plausible deniability" rather than visible ones, especially when dealing with followers who are strong-ties [132].[4]

5.5.4 Limitations

We curated our initial set of users by collecting mentioned protected accounts. Hence, our dataset is skewed towards users who stay protected most of the time. However, we observed that the users we collected tweets from represent a wide range of percentage of time spent protected vs private as well as exhibiting a good distribution over that range. Another limitation is the representation of mostly protected users in t_x and t_o. We were only able to collect tweets of accounts when they are public, so in the case where a user deletes their tweets before changing their tweets to public, we would not be able to collect the deleted tweets. Hence, for some users we might be getting only the t_x tweets that users are comfortable sharing publicly.

Our survey participants also represent a younger population, who may have a different view on privacy and have different goals than people of other age groups. In the United States the median Twitter user age is 40 and 62% of worldwide Twitter users are younger than 34 [171, 39]. Which means that our survey participants are indeed younger than a typical Twitter user but not greatly so. We recruited our survey participants from the crowdsourcing platform Prolific Academic and crowdsource participants tend to be more privacy-conscious than the general population [78]. This population source may be the cause of the higher percentage of mostly protected accounts in our participants (28%) compared to their general percentage among Twitter users (5%) [101]. Self reporting is also known to have biases, particularly around memory [117], and our survey relies on answering questions about past events.

The first survey was also intended to create an initial list of reasons that people switch between public and protected rather than create a comprehensive list. The number of participants surveyed, in our view, is high enough to identify common issues, but

[4]Twitter recently provided a feature for Twitter Web users to remove followers. However, this feature was introduced after we conducted our studies.

too few to really understand the scope of all possible reasons for switching. Also, as noted above, our participant sample is younger than an average Twitter user which also likely impacted the types of reasons they provided. We try and balance this issue by also including reasons found in prior work, but it is quite likely that if we were to more aggressively survey people from different age ranges, cultures, physical locations, or other demographics we would likely find a wider set of reasons for switching.

The breakdown analysis with age and sex demographics for the user survey was conducted posthoc.

5.6 Summary

In this study, we investigated Twitter users' privacy setting switching behaviour and the reasons behind the changes (RQ3). To do so, we collected the account visibilities of 107K initially protected users for three months and found that nearly 40% of them changed their settings at least once. We also collected the switchers' tweets to understand whether their sharing behaviour differs when they are protected vs public. We find that users utilize the privacy settings dynamically, sometimes changing as much as daily. They send tweets with mentions and hashtags more when they are public compared to protected. We coupled our Twitter data with two user surveys to get insights into the potential reasons behind the changes. We find that users change their accounts to protected to control their audience and interactions. On the other hand, users prefer to change their account visibility to public to use Twitter features freely, possibly at the expense of their privacy. We also suggest some design implications to protect the privacy of the users while enabling them to experience what the platform offers fully.

Until now, we found that networks of users can disclose information even when an account is protected. In our birthday study, we see that the interactions between a public and a protected account are not clearly understood. There are users who stay in one privacy setting but there are some users utilize privacy settings frequently. However, it is not clear how good these users understand the implications of changing privacy settings or the visibility of interactions. In the next chapter, we study the user understanding of privacy settings especially the visibility of information and interactions with respect to different account settings. We also investigate what factors lead to better understanding of information visibility.

Chapter 6

User Understanding of Privacy Settings and Visibility of Information

6.1 Overview

In the previous chapter, we investigated the privacy settings usage of Twitter users and the reasons behind the setting changes. In this chapter, we focus on answering the last main research question (RQ4): "How well do Twitter users understand the visibility of user information and tweets in relation to different privacy settings?" The binary privacy configuration of Twitter is fairly easy to understand at the surface level so it might be expected that most users understand the implications of the public/protected configuration. However, as seen in the previous chapters, inadvertent disclosure still happens on Twitter. While many factors likely impact these disclosures, one possibility is that information visibility is less clear in cases where public and protected accounts interact with each other, cases where a user changes their account visibility type, and the visibility of different types of account profile information. The simplicity of the configuration options may be leading users to a false sense of confidence where they believe that a post or account information will only be seen by a restricted set of followers, when that is not actually the case. Without this strong understanding it becomes difficult for users to enact their privacy intentions on Twitter. The potential problem is further exacerbated by the fact that any unaware person interacting with a protected user can leak information about them [77, 8].

In this study, we want to gauge the users' awareness regarding the visibility of user information on Twitter, as well as the tweet visibility especially when users interact with each other (RQ4). Specifically, we investigate the following research questions:

RQ4.1 How well do Twitter users understand the visibility of user information and tweets in relation to different privacy settings?

RQ4.2 What factors contribute to users' awareness of the visibility of information and tweets? In particular, what is the role of the user's account type on their awareness?

To answer our research questions, we conducted a user survey with 336 participants who have Twitter accounts with a range of privacy settings, including participants who only use public account, protected account, and some switching between the two settings. Our findings show that the participants are mostly aware of who can see tweets of users when they are tweeting by themselves, i.e. not interacting with other users. They also mostly understand what account information is publicly visible with the exception of topics and lists. Interactions between public and protected accounts was more confusing with only 40% of these questions answered correctly. Surprisingly, the normal audience (public, protected, switching) of the participants did not have any significant impact on their knowledge of the platform functionality around privacy settings. However, the frequency of replying to protected accounts, Twitter usage, and being able to easily see that they are interacting with a protected account have an impact on our participants' Twitter privacy functionality understanding. Our contribution includes comparing awareness of users with different privacy settings, including the ones who change them frequently. We also investigate the understanding of interactions in addition to the single posts. Our findings suggest that the design of Twitter UI might be sub-optimal, especially when it comes to dealing with protected accounts, where users are not fully aware of visibility of some of their activities on the platform. We list possible privacy violations that could happen in the platform and suggest design implications.

6.2 Related Work

There is limited work on understanding privacy settings and information visibility on Twitter. Proferes [127] conducted a user survey with 434 participants to measure their understanding of Twitter in terms of techno-cultural and socioeconomic aspects. They provided participants various statements, in a range of topics including data, users, governance, algorithms, etc., and asked them whether the statement was accurate. They found that the participants did not understand the long term storage of past tweets

as well as the visibility of information to other users. Only 24.7% of the participants correctly answered that number of tweets, followers, followees etc. were public for protected accounts. While the statements covered some topics around account information visibility and interactions between accounts, they did not investigate which of those account information were understood better or what was the participants' expected audience for the interactions.

In Chapter 4, we asked their participants two interaction visibility questions between public and protected accounts. We found that participants commonly thought these interactions could happen but the audience will be limited to the followers of the protected accounts. Compared to the study conducted by Proferes and our previous study, we expand on the account information questions and ask more detailed interaction questions in this chapter. We also divide our results depending on the account type of the users and compare their understanding.

6.3 Methodology

We conducted a user survey to measure Twitter users' awareness of information and tweet visibility. We conducted a prescreening survey asking participants if their Twitter account is public, protected, or switches between the two. We invited a balanced number of participants of each account type to take our survey so that we can measure the possible effect of account type on the user understanding. The main study consisted of questions around information and tweet visibility on Twitter along with demographic questions around their account. We recruited our participants from Prolific Academic (PA) [128].

The main survey was pilot tested with 6 PA participants before launch to estimate time required to complete the survey and get feedback on the clarity of questions. We designed and ran both surveys following the University's ethics protocol and compensated each participant with £2.25 for filling the main survey (£9 per hour).

Prescreen The visibility of a user's account likely impacts their understanding of Twitter's visibility settings so we used a prescreen to ensure that we invited even numbers of people who are protected, public, and switch between. Doing so is especially important given that only 4% of Twitter users have protected accounts [101] so prescreening is necessary to ensure adequate participation from all groups. We conducted the screening survey at the end of June 2021. We used PA and limited the survey visibility to those who can speak English fluently and have Twitter accounts using PA's

filtering feature. We compensated each participant with £0.09 for filling the prescreen. We asked participants what their normal Twitter audience was on a 7-point likert scale from: "Always protected" to "Always public". We then split respondents into three groups: **Public** ("Always public"), **Protected** ("Always protected"), and **Switching** (other answers). Out of 1074 users who had a Twitter account, 179 (16.7%) were protected, 408 (38%) were public, and the remaining 487 (45.3%) switched. Following the prescreen, we invited equal number of participants from each group to take our main survey in July 2021.

6.3.1 Survey Instrument

After informed consent, we asked participants if they had a Twitter account (they all did) followed by four set of questions around Twitter functionality with respect to different account types. The first set of five questions asked about the visibility of individual tweets posted by: public users, protected users, users who were protected but changed to public, users who were public but changed to protected, and lastly users who were public and stayed public.

Next, we asked questions about the visibility of 11 types of account information (e.g. followers, lists) for public and protected accounts. Followed by 11 scenario-based questions about the what would happen if a public and protected account attempted to interact in various ways (e.g. quote tweet, retweeting). Finally we asked four true/false statement questions targeting potential misconceptions involving how Twitter behaves towards different account types (e.g. users with protected accounts cannot be tagged in photos).

The last section covered demographics including the normal audience question from the prescreening, frequency of Twitter usage, the information they have on their profile, number of followers, as well as the number of users they follow. We also asked if they can easily tell the type of account when replying and if they look at account type before engaging (reply, mention, retweet) with a tweet. Finally, we provided an optional free text comment box.

Full survey text is available in the Appendix E.

6.3.2 Participants

In total, 459 participants completed the survey, but 123 were excluded due to failing an attention check question resulting in 336 users. Common participant demograph-

ics (e.g. sex, age, nationality) as well as the demographics related to the filters set by researchers (e.g. social media usage) are provided by PA. According to those demographics, 141 (42%) of our participants were female and 193 (57.4%) were male with 1 preferring not to respond. Our participants had an average age of 25.2 (sd 7.4) and a median of 23. 214 (63.7%) of the participants were between 18-24 years old, 88 (26.2%) were between 25-34 years old, 25 (7.4%) were between 35-44 years old, and the remaining 8 (2.4%) being 45 or older. Most of the participants used Twitter daily (194, 57.7%) followed by weekly (95, 28.3%), monthly (29, 8.6%), and a couple times a year (18, 5.4%).

In regards to account type, 75 (22.3%) of the participants keep their accounts always protected while 131 (39%) keep them always public, the remaining 130 (38.7%) change their privacy settings. Out of those who change their privacy settings, 27 (8%) stays mostly protected, 15 (4.5%) somewhat protected, 14 (4.2%) balanced, 20 (6%) somewhat public, and 54 (16.1%) mostly public. Twitter accounts are public by default and protected users need to change their settings only once, while users who switch indicate that they utilize the privacy settings more often than other users. For simplicity, we use the terms public users, protected users, and switching users in the remainder of the chapter to refer to these account types.

Most participants (62.5%) had less than 100 followers. They themselves mostly followed between 100 and 499 users (50.9%) or less than 100 users (37.8%). 82.4% had a profile picture, 59.5% had a header photo, and 70.5% biographic information. In regards to more sensitive data, 33.6% had their birthdays, 26.2% their location, and 13.1% had a website on their profiles.

Interacting with Protected Accounts 61.9% of participants said they can easily tell when they are replying to a protected account (70.7% of protected users, 63.4% of public users, and 55.4% of switching users). However, only 32.4% said they check account type when they are engaging with a tweet (36% of protected, 32% of public, and 30.8% of switching).

We asked participants how often they interact with protected users via liking tweets, replying to their tweets, or mentioning the account. 36% of our participants “Always” or “Often” like protected accounts’ tweets, 17% reply to them, and 8% mention them.

Question Set	Accuracy
All Questions	71.7%
Individual Tweet Visibility	86.7%
Public Account Information Visibility	85.9%
Protected Account Information Visibility	82.9%
Interaction Visibility	51.2%
Misconceptions	39.1%

Table 6.1: Percentages of correct answers given to each question set.

Question Set	Female	Male	18-24	25-34	35-44	45+
All Questions	72.5%	71%	74%	69.7%	63%	58.6%
Individual Tweet Visibility	84.4%	88.3%	90.2%	85.2%	72.8%	52.5%
Public Account Information Visibility	85.4%	86.2%	86.2%	85.3%	84.7%	87.5%
Protected Account Information Visibility	83.5%	82.6%	84.7%	81.8%	73.8%	72.7%
Interaction Visibility	55.7%	47.8%	55.7%	45.9%	37.8%	33%
Misconceptions	38.7%	39.1%	40.7%	39.2%	30%	18.8%

Table 6.2: Percentages of correct answers given to each question set, as well as to the questions in individual visibility questions and misconceptions. Answers divided based on sex and age of the participants.

6.4 Results

6.4.1 User Awareness of Visibility

We asked about the visibility of information, tweets, and Twitter functionality to our participants. Table 6.1 shows the percentage of correct answers given by our participants to each set of questions described in the survey instrument. In general, users were able to answer questions correctly. Over 80% of questions about individual tweet visibility and account information visibility were answered correctly by our participants. But users had more trouble when answering questions that involved interactions between two accounts of different types. Misconception questions, which focused on less common Twitter actions, were the least understood with 39.1% of the participants answered correctly. Participants answered comparably when we divide them based on sex. However, the accuracy of the participants were usually lower in older age groups compared to younger ones (Table 6.2).

Individual Tweet Visibility We asked five questions about the visibility of tweets

Individual Visibility Questions	Accuracy
Who can see a public account's tweets	86.9%
Who can see a protected account's tweets	95.5%
Change visibility from public to protected	89.3%
Change visibility from protected to public	76.2%
Keep visibility setting public	85.7%

Table 6.3: Percentages of correct answers given to tweet visibility questions for an account.

Individual Tweet Visibility	Female	Male	18-24	25-34	35-44	45+
Who can see a public account's tweets	83.7%	89.1%	92.5%	84.1%	64%	37.5%
Who can see a protected account's tweets	96.5%	94.8%	94.4%	98.9%	92%	100%
Change visibility from public to protected	87.2%	90.7%	89.7%	90.9%	84%	75%
Change visibility from protected to public	72.3%	78.8%	83.2%	69.3%	60%	12.5%
Keep visibility setting public	82.3%	88.1%	91.1%	83%	64%	37.5%

Table 6.4: Percentages of correct answers given to tweet visibility questions for an account, broken by sex and age.

under different account types and in cases where the account tweets and then later changes type, the answer breakdowns are visible in Table 6.3 and Table 6.4. As a simple check, we started this section with two very easy questions asking who can see public account tweets and who can see protected account tweets. 87% of the participants correctly answered that anyone on internet can see tweets posted by public accounts with the remaining 13% incorrectly selected that only logged in Twitter users could see public account tweets, indicating an awareness of wide visibility but not properly understanding just how wide. For protected accounts, 95% indicated that only the followers of a protected account can see the tweets. Answers to the other questions show less understanding of how historical tweets are handled when an account switches type. Though only 7% incorrectly thought that when changing from protected to public, past tweets would stay protected and only 8% thought that when changing from public to protected past tweets would stay widely visible to logged in Twitter users or anyone on the internet.

Account Information Visibility We asked about information visibility of public and protected account information to measure awareness of their visibility. We asked our participants about the visibility of the 11 types of information a Twitter account can

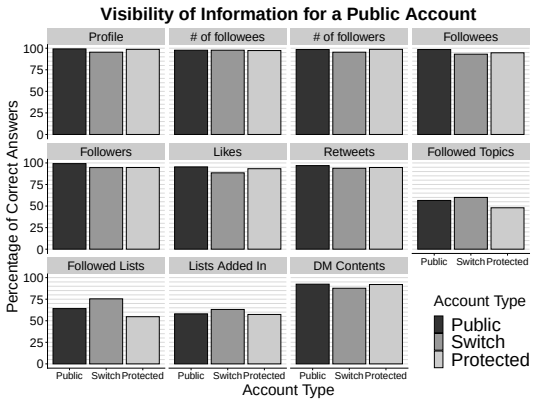

Figure 6.1: Percentage of correct answers given to account information visibility questions regarding a public account.

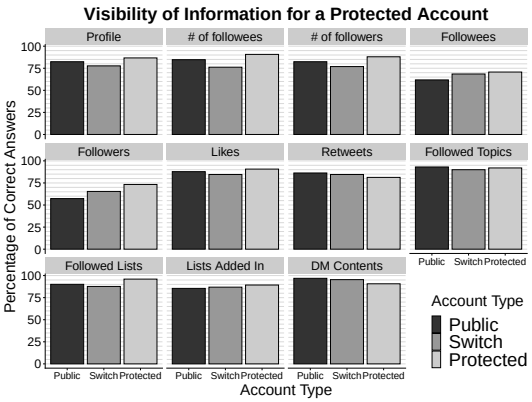

Figure 6.2: Percentage of correct answers given to account information visibility questions regarding a protected account.

have. Figures 6.1 and 6.2 show the percentages of correct answers given by participants broken down by account types, as well as sex (Figures 6.3 and 6.4) and age (Figures 6.5 and 6.6). Currently, all of the information shown in Figure 6.1, aside from the "DM Contents", are publicly visible for a public account. On the other hand, only "Profile", "# of followees", and "# of followers" are publicly visible for a protected account (Figure 6.2). Most of our participants correctly selected whether the information was publicly visible or not for all types of information. However, the participants had lower understanding around topics and lists for public accounts. The visibility of followers and followees of protected accounts were also less understood by participants.

Interaction Visibility We give the percentage of correct answers given by our participants to interaction visibility questions in Figure 6.7 (broken down by account type), Figure 6.8 (sex), and Figure 6.9 (age). We asked participants various scenarios between two accounts that follow each other and asked them what would happen if one of these accounts interacted with the other by replying, mentioning, quote tweeting, and retweeting. All of the questions included interactions with protected accounts. The first question set, which had a public account interacting with a protected one, was challenging for participants with most participants answering incorrectly for replying, quote tweeting and retweeting. A common error involved a public account replying to a protected account's tweet, the public account's reply would be public, but 66% of participants incorrectly thought that only the protected accounts followers could see the reply tweet.

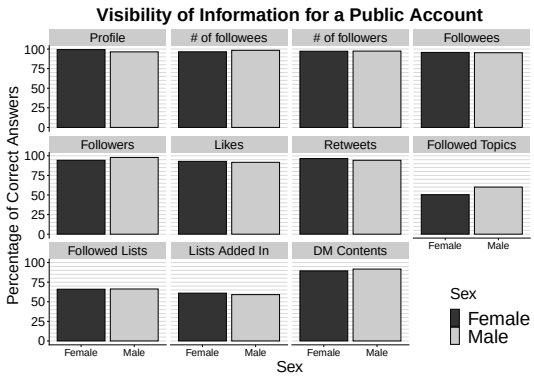

Figure 6.3: Percentage of correct answers given to account information visibility questions regarding a public account, broken by sex.

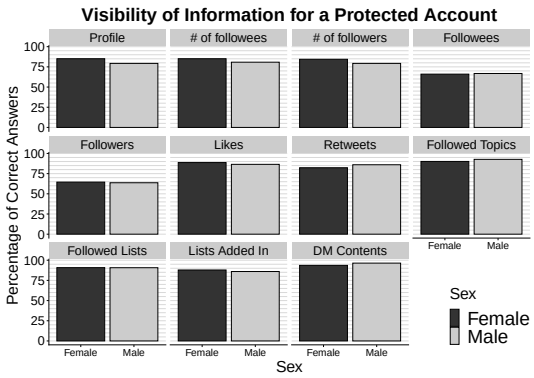

Figure 6.4: Percentage of correct answers given to account information visibility questions regarding a protected account, broken by sex.

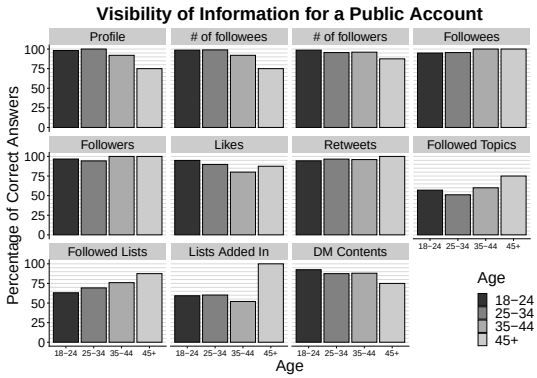

Figure 6.5: Percentage of correct answers given to account information visibility questions regarding a public account, broken by age.

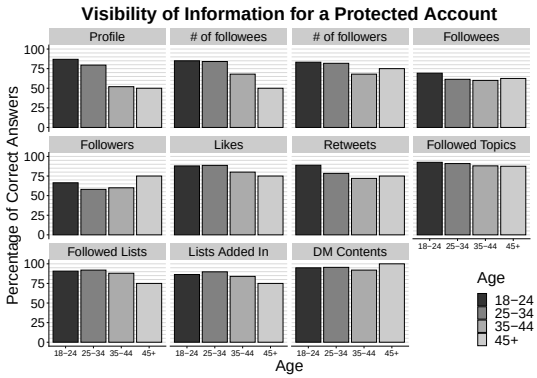

Figure 6.6: Percentage of correct answers given to account information visibility questions regarding a protected account, broken by age.

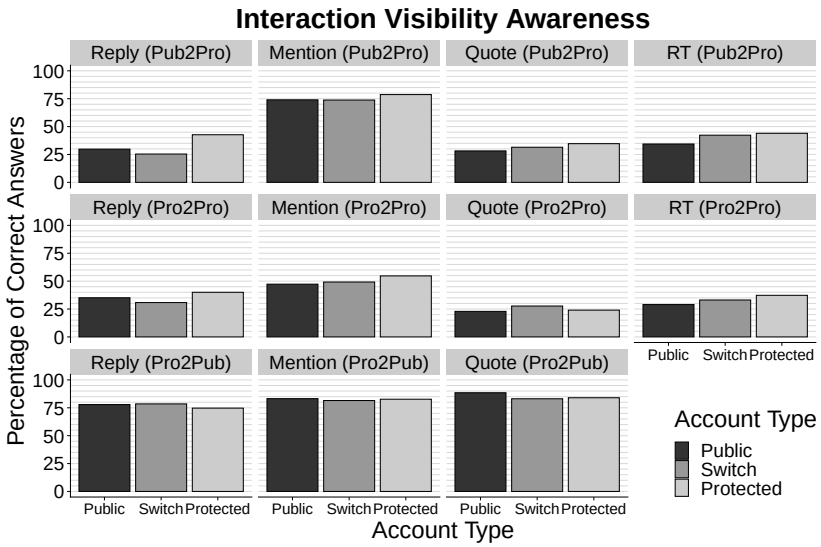

Figure 6.7: Percentages of correct answers given to interaction visibility questions. First row for interactions from public accounts to protected ones, second row for protected to protected interactions, and the last row for protected to public interactions.

In the second scenario we asked what would happen if a protected account interacted with another protected account. For all of the interaction types, the majority of participants incorrectly answered that the interaction could happen but only users who follow both of the protected accounts could see it. Participants performed the best in the third scenario where we asked what would happen if a protected account interacted with a public one with over 75% of participants answering all questions correctly.

Based on their answers to these questions, participants clearly thought that when interacting with a protected account the tweet or other interaction would only be visible to the followers of the protected account. Since tweet visibility is only connected to the poster's account type, the possible interactions initiated by public accounts such as replying and mentioning will be visible publicly. If two protected accounts interact, then the followers of the account who initiated the interaction, i.e. reply and mention, are the only ones who can see those tweets.

Misconceptions Around Account Type and Twitter Functionality In addition to the visibility questions we also asked four questions around Twitter functionality with different account types. Our participants had the lowest performance in this set of questions which somewhat expected for this question group. Direct message (DM) requests and picture tagging are controlled by a different set of settings than the public/protected account type setting which is potentially confusing. Which turned out to be the case, less than half the participants knew that protecting an account will not

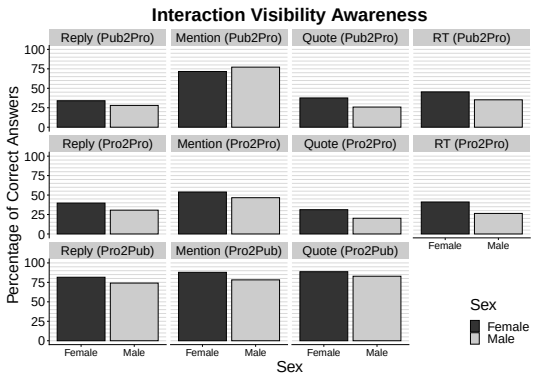

Figure 6.8: Percentage of correct answers given to questions around Twitter functionality for different account types, broken by sex.

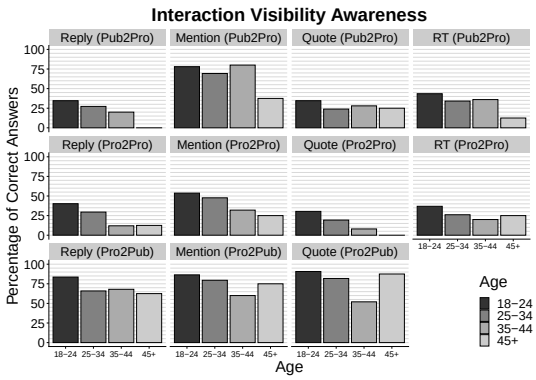

Figure 6.9: Percentage of correct answers given to questions around Twitter functionality for different account types, broken by age.

disable tagging that account in pictures. Worse, only 18.8% of participants knew that protecting an account does not disallow non-followers from sending DMs to that account.

Lastly, we asked two questions around tweet deletion and the residual data [114]. Replies to a deleted tweet stay in the platform even when the deleted tweet was posted by a protected account. Most of our participants (55.1%) correctly said that deleting a public account's tweet will not delete replies to it. However, only 36% correctly answered when the same question was asked regarding a protected account.

6.4.2 Factors That Contribute to User Awareness

We now investigate to what extent user characteristics affect their ability to determine who can see their information. For this purpose, we construct a generalised linear model that contains the main factors which explain variation in the accuracy of answers.

Method Generalised linear models are an extension of linear regression models. Here, we use a type of generalised linear model that is adapted to binary outcome variables, namely logistic regression. The outcome variable we are looking at is *accuracy* - whether the respondent answered a question about tweet or account information visibility correctly.

We use the term "model" since the aim of model building is to construct a generalised linear model, described by a logistic regression equation, that explains the high-

Misconceptions around functionality	Total	Public	Switching	Protected
Protected accounts cannot be tagged in photos	46.4%	48.1%	47.7%	41.3%
Protecting an account will disable DMs from non-followers	18.8%	10.7%	23.1%	25.3%
Deleting a public account's tweet will delete replies to it	55.1%	53.4%	56.9%	54.7%
Deleting a protected account's tweet will delete replies to it	36%	35.1%	35.4%	38.7%

Table 6.5: Percentage of correct answers given to questions around Twitter functionality for different account types. Answers divided based on the account type. Percentages are out of the total number of public, switching, and protected participants respectively.

Misconceptions around functionality	Female	Male	18-24	25-34	35-44	45+
Protected accounts cannot be tagged in photos	41.1%	50.3%	50%	44.3%	24%	37.5%
Protecting an account will disable DMs from non-followers	19.1%	18.7%	13.6%	31.8%	24%	0
Deleting a public account's tweet will delete replies to it	54.6%	54.9%	60.3%	50%	36%	25%
Deleting a protected account's tweet will delete replies to it	39.7%	32.6%	38.8%	30.7%	36%	12.5%

Table 6.6: Percentage of correct answers given to questions around Twitter functionality for different account types, broken by sex and age of the participants.

est amount of variation in the data set in the most parsimonious way possible. Thus, we assess model quality using the well established AIC (Akaike Information Criterion), which rewards models that fit the data well, and penalises models with many parameters. When comparing multiple models of the same data set, a lower AIC is better.

We have already seen that there are some aspects of tweet visibility that users understand well, and others that they struggle with. Therefore, our baseline model is $accuracy = 1 + question$, where $question$ stands for one of the aspect of Twitter functionality and tweet visibility that are covered in the main body of the questionnaire.

We then use a process of greedy stepwise selection to add the most appropriate m user factors to the model. These factors are selected from a fixed set of n candidates.

The procedure used is as implemented in `stepAIC` in R, MASS package.

First, we construct four models that cover different aspects of users, and compare their performance. The models are:

Social Media: whether the user has social media experience on Facebook, Tiktok, Instagram, Snapchat, or Linkedin

Profile Elements: the types of information the user includes in their Twitter profile (bio, website, picture, header, date of birth, location)

Twitter-specific: This includes all twitter-specific features except for audience, such as number of followers (ordinal scale converted to numeric), frequency of Twitter use, whether they can easily tell the type of account when replying, whether they look at the account type before engaging, and whether they like, reply to, or mention protected accounts.

Audience: The typical audience of the user. We use two variables, audience (always public, always protected, switching between public and protected), and the original seven item Likert scale (audience7).

We then compared how well each of the resulting four models explain the amount of variation in the data set using Anova, with the Chi Square test to establish significant differences. Finally, we used greedy stepwise selection to construct a final, full model using all features from the four individual models.

Results Table 6.7 summarises the four models created using different user characteristics. In the model specifications, variables are listed in the order in which they were added to the model in the stepwise selection process.

We see clearly that the model which only takes audience into account performs worst. Use of other social media platforms yields a somewhat better model, just as considering the information disclosed in a person's Twitter biography. However, the model that explains the data best is the one that focuses on people's knowledge of Twitter, the number of followers, and frequency of use. When comparing the amount of variation explained by the four models using analysis of deviance (`anova` function in R), the Twitter-specific model outperforms all others with $p < 0.00001$, and the audience-only models is worse than the model based on profile elements alone ($p < 0.00001$) and the model based on other social media activity alone ($p < 0.00001$).

In other words, the initial four models show that users who use the "protected" switch do not necessarily understand what it does—users who frequently engage with

Model	User-Specific Variables Included In Order	AIC
Audience	Audience7	12829
Profile Elements	header, profile picture, website, location, date of birth	12784
Social Media	on Instagram, on LinkedIn, on Tiktok, on SnapChat, on Facebook	12776
Twitter-specific	can easily spot protected accounts, replies to protected accounts, usage frequency (numeric), number of followers	**12607**

Table 6.7: Performance of Four Models Covering Different Aspects of Users. AIC = Akaike Information Criterion. Variables listed in the order in which they were included in the model during stepwise greedy selection. All models also include an intercept and the variable for Question, which is part of the baseline.

Twitter, know how to spot protected accounts, and have a sizeable number of followers do. Participants who strongly agreed that they can easily tell the account they are replying to is protected answer 76.7% of all questions accurately, while participants who answered strongly disagree perform worse with 64.5%. Participants who reported they always reply to protected accounts get 80.7% of questions right, where this rate is 69.3% for participants who never replies to protected accounts' tweets. Finally, participants who used Twitter daily answer 73.5% of questions correctly, while participants who reported they use Twitter a couple of times a year perform at 62.7%.

The full model, as selected from all variables in the four previous models, is shown in Table 6.8. We do not show the complete logistic regression model with all coefficients since there are more than 60. Instead, we focus on the relative importance of the variables included in explaining the variation in the data set. In Table 6.8, variables are listed in the order in which they were added to the model in the greedy stepwise selection process. The p-value given for each variable indicates whether adding the variable yields a significant improvement over the previous model that did not include the variable. For example, once the model includes the first seven variables in Table 6.8, up until the number of followers, the next best addition is information about whether the user is on Snapchat ($p < 0.05$). Once information on Snapchat use has been integrated, the next most important information is whether the user is on Tiktok ($p < 0.005$)

We see that, again, the most important variables indicate whether users can easily spot protected accounts, and whether they engage with protected accounts. Experience

Variable	df	Deviance	Pr(>Chi)
Question	41	4074.4	$p < 0.00001$
Can easily spot protected accounts	4	174.0	$p < 0.00001$
Replies to protected accounts	4	42.2	$p < 0.00001$
User on Instagram	1	24.0	$p < 0.00001$
Frequency of use	1	17.0	$p < 0.00001$
User on LinkedIn	1	13.8	$p < 0.0005$
Number of Followers	4	19.8	$p < 0.001$
User on Snapchat	1	4.8	$p < 0.05$
User on Tiktok	1	9.1	$p < 0.005$
Has profile picture	1	3.8	$p < 0.1$
Audience	1	4.7	$p < 0.05$
Web site in profile	1	4.2	$p < 0.05$
Location in profile	1	4.3	$p < 0.05$

Table 6.8: Relative Importance of Each Factor in the Order in Which It was Added to the Model. Df: degrees of freedom. Deviance: measure of variation in the data set covered by variable. Pr(¿Chi): probability that the model with variable x_i is an improvement over the model with variables $x_1, \ldots, x_{i-1}$

with Instagram and LinkedIn is also important. While the type of information mentioned in the user's profile and the audience setting (public, protected, different levels of switching) do cover significant additional variation in the data set, their relative contribution, as indicated by the reduction in residual deviance, is small.

6.5 Discussion

In this study, we investigated the last research question around the user understanding of the information and tweet visibility on Twitter. One of the important aspects of protecting privacy in online social networks is understanding who can see the information and the posts that are being shared. Previous work has shown that users struggle to configure privacy settings to reflect their privacy expectations [105, 100] and they are confused with different privacy settings [140]. However, these studies are done on social media platforms that have more granular privacy options than Twitter. The findings in this study show that users confusion with privacy settings occur even with the relatively simplistic binary Twitter privacy settings. These results are in line with Proferes' [127] work on user beliefs about Twitter, though in some cases our participants did show better understanding. In their study, only 24.7% of the participants correctly answered the number of tweets, followers, followees, and so on that would be public for protected accounts. However, 80% of our participants correctly answered for both the number of followers and followees. Considering the seven years between the two studies, user awareness of such settings could have changed. Its also possible that since our sample includes more protected and switching users than a random sample, that is impacting the level of awareness, though given our own results, the impact is likely minimal.

Bartsch and Dienlin [14] found that the time spent on Facebook and the frequency of utilizing the privacy settings on the platform led to better online privacy literacy. Similarly, we find that frequency of Twitter usage is a factor that contributes to the knowledge of the information and tweet visibility. However, the participants' normal audience did not have any significant impact on the awareness. Even the users who report they switch their settings frequently have similar knowledge around the information and tweet visibility of different privacy settings.

Tweets from public accounts can be seen by anyone on internet while tweets from protected accounts can only be seen by approved followers. Individual tweet visibility is well understood by our participants with over 85% of questions correctly answered.

However, who can see the tweets when these accounts interact is not necessarily as clear. In this study, most of the participants could not answer what would happen if an account interacted with a protected one (excluding mentioning). Less than 40% knew that protected account tweets could not be retweeted, and only 31% were aware that anyone could see the reply by a public account to a protected account's tweet.

While our sampling strategy resulted in 22% of our participants having always protected accounts and a further 40% sometimes having protected accounts, the general Twitter population has a much lower percentage of protected accounts [101] with only 13% of the US adults choosing to have their Twitter account protected [171, 135]. Given that, it is somewhat surprising that 81% of the Always Public participants indicated they had liked, mentioned, or replied to a protected account. Crowdworkers on websites like Mechanical Turk are known to have more privacy concerns than average users [78] which may explain some of the finding. But the finding still suggests that despite being uncommon, public accounts do interact with protected ones, more than might be expected based purely on the number of protected accounts. Surprisingly, protected account holders answered the interaction questions in a similar way to public account holders, suggesting that even though they actively chose to have their tweets private with their followers only, they share similar misconceptions and may not be aware that half of their conversations with public accounts are visible to everyone, possibly violating their intended privacy outcomes.

Compared to the general public, Twitter users are younger and more educated [171] as well as being more skilled in using internet [156]. Having better internet skills is shown to have a strong effect on privacy protection [27]. In addition, the 42% of our survey participants were female, who are shown to be more privacy conscious [48], where only 31.9% of all Twitter users are female [40]. Despite these factors our participants had a low understanding of the visibility of information around topics and lists, as well as the visibility of tweets when interacting with protected accounts. A recent article also states that Twitter is aware of the low understanding their users have of privacy settings [166].

6.5.1 Design Implications

Our findings suggest that the two main points of confusion about information visibility on Twitter are: 1) interactions, especially with protected users and 2) topics and lists. We discuss the privacy implications of these points and give design suggestions for

helping user understanding below.

Showing potential audience when drafting tweets: Users have an imagined audience when they share posts [163], which is often smaller than the real audience size [17]. Underestimating the audience of the tweets can lead to privacy problems, especially when interacting with protected accounts. Showing users the potential audience of their tweets can help users to better contextualize the reach of their tweets [97, 157]. For example, when users with public accounts are interacting with protected accounts, it could be stated that the tweets can be seen by everyone, not only the followers of that protected account.

Hiding interactions with protected accounts: Twitter UI prevents retweets/quote tweets of protected tweets, hence the effects of not being informed on the expected behavior is minimal. However, our findings suggest that big chunk of users believe that their tweets can only be seen by the followers of the protected account when they engage with one. This may lead them to disclose information they did not intend to share with the general public. Searching a person's account (e.g. "@username") in Twitter will bring up the tweets sent to them such as mentions and replies, even when they are a protected account. Hiding these tweets from non-followers can help protect the privacy of the protected accounts, which is closer to what users tend to believe currently.

Showing types of the interacted accounts clearly: Even if the users were perfectly aware who could see their tweets when they interact with protected accounts, it is possible that these users are not aware they are interacting with one. Only 29% of our participants strongly agreed that they can easily tell that they are replying to a protected account, while this rate was 10% for checking the account types of the users they engage with. Our analysis showed that the participants who were able to easily tell that they are replying to a protected account had higher awareness of the platform. Its unclear the direction of the relationship though. It could be that those who are more conscious of protected users' privacy are consequently more self-assured in their ability to notice such accounts, or it could be that the ability to notice such accounts causes users to become aware. Design changes to the interface could make protected accounts more visible, especially in cases where users interact with multiple accounts (i.e. replying to a reply).

Informing users about the visibility of topics/lists: Our participants were mostly aware whether different types of account information are publicly visible. The visibility of lists and topics on public lists was confusing, both are publicly visible by anyone

for public accounts along with the lists a user has been added to. Users can follow topics they are interested in and create lists to curate a timeline of users they want to keep track of without necessarily following individual accounts. While there is an option to create private lists, there is no option to hide the topics that a public account follows on Twitter. Public users cannot hide the lists they are following which are created by other users. They also cannot remove themselves from a list without blocking the list creator. Blocking a user may create discomfort and users might not be willing to block another user especially if that person is a close friend or a family member [132]. These public topics and lists can leak information about the user including their interests and personal ties. Users should be clearly notified that this information around topics and lists is public.

6.5.2 Limitations

Many of the more specific limitations of the study are already presented in the discussion in relation to their findings. More generally, this study recruited from Prolific Academic which can draw participants from more privacy-conscious crowd, as shown for Amazon Mechanical Turk [78]. In addition, our sample is younger than global Twitter users [39] which also may translate into having better internet and privacy protection skills [156, 27]. This actually might show that the general public might be less aware of the visibility of their information and tweets. The study also uses a survey approach which means we are able to ask about only issues and answer options we know about in advance. We countered this issue by familiarizing ourselves with Twitter's range of options and ensuring that the full range was presented to users. We also endeavored to be comprehensive and clear in our question and answer presentation. Finally, we included a comment box at the end of the study in case participants noticed anything they strongly felt was missing, reviewing these comments resulted in no serious identified omission. The breakdown analysis with age and sex demographics for the user survey was conducted posthoc.

6.6 Summary

In this chapter, we focused on answering the fourth main research question concerning the user understanding of information and tweet visibility on Twitter. To do so, we conducted a user survey with 336 participants to understand users' awareness of the

visibility of information and tweets shared on Twitter, including the tweet visibility when accounts with different types interact. Our findings suggest that users are mostly aware information shared on accounts depending on the account type. They also understand the the audience of tweets sent by different account types when those tweets are not interacting with others by mentioning or replying. However, the functionality and visibility of interactions between accounts are not clear, especially when the interacted account is protected. Our participants tend to think these interactions will only be shown to the followers of the interacted protected account. Surprisingly, being a protected or switching account holder did not translate into a better understanding of interactions with protected accounts. Frequently using Twitter and being able to easily tell that they are replying to a protected account contributed to better performance in our participants. Informing users on how engagements work between different accounts is essential. Users also should be notified better when they are interacting with protected accounts.

Chapter 7

Conclusion

People share a wide variety of information on social media, including personal and sensitive information, without understanding the size of their audience which may cause privacy complications. Networked nature of the platforms further exacerbates these complications where the information can be shared without the information owner's control. People struggle to achieve their intended audience using the privacy settings provided by the platforms. Hence, they employ various strategies in addition to these settings to protect their privacy while also wanting to gain social capital. In this thesis, I analyzed potential privacy violations caused by social media users and their networks, as well as the usage and understanding of privacy settings. I focused on Twitter which has a rather simplistic privacy settings with binary states. In this chapter, I firstly give summary of findings of the studies detailed in the previous chapters and state thesis contributions. Then I give overall implications about privacy protection on social media platforms through the findings of the studies in this thesis. I conclude the chapter with limitations of my approach and the potential future directions for research.

7.1 Thesis Contributions and Findings

The first two research questions in this thesis were concerning the personal information shared by users' networks and the reactions of the users to these tweets. To answer these research questions, I conducted two studies detailed in Chapter 3 and Chapter 4. **In Chapter 3**, I investigated personal information disclosures by networks using congratulatory messages. To do so, I collected 635K tweets with the phrase "happy for you" over four months. I analyzed these messages and detect 12 types of life events including relationships, illness, familial matters, and birthdays through LDA topic mod-

elling. Eight percent of these life event tweets were directed to protected accounts, possibly exposing the content. The most popular celebration topic in my dataset was directed to users who were having a new baby. Most of the congratulatory tweets around a sensitive topic were replies to existing tweets. More common life events such as graduations and birthdays were sent as stand-alone tweets, without apparent prompt, more than others. The majority of the users interacted with these tweets by liking, retweeting, or replying. The work in this chapter has been published in the 12th ACM Conference on Web Science (WebSci'20) with the title *"Analysing privacy leakage of life events on twitter"* [83].

In Chapter 4, I further focused on birthdays which was one of the most popular life events found in Chapter 3. However, with birthdays, there is a potential date of birth disclosure which has security implications besides the privacy ones. I collected 18 million birthday celebrations in English over 45 days. I filtered these tweets to leave out retweets, tweets towards verified accounts, and tweets with multiple mentions. I found that 2.8 million of these tweets were directed to 724K accounts. Further analysis showed that for 50K accounts, the age was likely mentioned revealing their DOB, and 10% of them were protected accounts. The findings show that the majority of both public and protected accounts seem to be accepting of their birthdays and DOB being revealed online by their friends even when they do not have it listed on their profiles. The user survey showed that giving birthday wishes to others online is considered a celebration and many users are quite comfortable with it. The work in this chapter has been accepted for publication in the 16th International Conference on Web and Social Media (ICWSM'22) with the title *"From an Authentication Question to a Public Social Event: Characterizing Birthday Sharing on Twitter"* [85].

The user survey in the Chapter 4 also showed an interesting way users were identifying their account types. Other than platform-provided types of "public" and "protected", these participants selected that they have "sometimes protected" accounts. **In Chapter 5** of my thesis, which answers the RQ3, I focused on these users who were changing their privacy settings unexpectedly on Twitter. I inspected the privacy setting changes of 100K Twitter users over three months and noticed that 40% of those users changed their privacy settings at least once with 16% changing it over five times. I compared the tweeting behaviour of users when public vs protected and showed that users who switch their privacy settings mention others and share hashtags more when their setting is public. The following user surveys highlighted that users turn protected to share personal content and regulate boundaries, while they turn public to interact

with others in ways prevented by being protected. The work in this chapter has been published in the CHI conference on human factors in computing systems (CHI'22) with the title *"Understanding Privacy Switching Behaviour on Twitter"* [86].

In Chapter 6 of the thesis, I investigated the user awareness of information and tweet visibility of different account types by conducting a user survey (RQ4). Since the publicly visible information changes based on the account type and the visibility of tweets also depends solely on the poster's account type, unintended disclosures can occur especially when users interact with each other. The user survey had 336 participants with valid answers to questions ranging from profile information visibility to tweet visibility in interactions. I showed that the users are aware of the visibility of their profile information and individual tweets. However, the visibility of followed topics, lists, and interactions with protected accounts is confusing. Less than third of the survey participants were aware that a reply by a public account to a protected account's tweet would be publicly visible. Surprisingly, having a protected account did not result in a better understanding of the information or tweet visibility. The work in this chapter is under revision with the title *"Twitter has a Binary Privacy Setting, are Users Aware of How It Works"*.

The main contributions of this thesis are focused around the interactions of users; how do they leak information, how users try to manage their interactions, and whether the users understand the visibility of interactions they have on the platform. From the studies in Chapter 3 and Chapter 4, I show that public replies are enough to infer the content of the original message, even if the event subject hides or deletes the message. These results are in line with Mondal et al. [114], who analyzed deletion behaviour on Twitter and found that protected and deleted tweets leave residual information that could help infer the original tweet.

One of the most interesting findings of this thesis is how users overcome the shortcomings of the platform-given functionalities, shown in Chapter 5. Some users change their privacy settings frequently to circumvent the restrictions of the binary settings provided by Twitter. Prior studies conducted on Twitter check the account types of the users only once to decide whether they are protected or public [34, 96]. This finding shows that it is not safe to assume that an account is always protected or always public.

Rashidi et al. [132] found that users prefer invisible sanctions while regulating interactions. I also show that users employ boundary regulation strategies that are not easily noticeable by their networks such as using soft-blocking to remove followers by using blocking feature. boyd and Marwick [22] found that some teenagers were deac-

tivating their Facebook profiles when they logged out of the platform and reactivating their accounts when they want to log in. Similar behaviour can be seen in our study where the participants agreed that they temporarily deactivate their accounts to prevent all interaction for a while or protect their accounts while they were away and unable to respond to interactions.

Another main contribution of my thesis is the lack of interaction visibility understanding throughout different account types, including the switching account type I introduced in Chapter 5. Compared to Proferes [127], our participants performed better in some individual visibility questions regarding protected accounts. However, I found that interactions with protected accounts are less understood, especially when the interaction is between two protected accounts. When the user understanding of information visibility are compared, interestingly, I found that having a switching or protected account did not translate into having a better visibility understanding. This indicates that while these users actively try and protect their privacy, they are unaware of the implications of interacting with other users.

7.2 Implications

Information that were shared with a small set of people in physical settings is now shared publicly for everyone to see online. People gain social capital by sharing this information [45] and once shared the information becomes co-owned by the shared audience [123]. Engagements on social media posts can increase the visibility of personal information depending on the platforms' algorithms. Other people can also share information about users which is not easily controlled by the information owner. Hence, in line with theories of Altman [10] and Petronio [123], this thesis supports that privacy protection on online social networks is not an individual task. Interactions with a social media post can disclose information about the owner of the post, even when the owner chooses the hide that information. Communication Privacy Management (CPM) Theory [123] defines "boundary turbulence" when there is a conflict between the privacy rules of the information owner and the co-owners. As a result of this conflict, the information is shared with unintended third parties. In my studies around life events and birthday, it is shown that replies to protected tweets can disclose such unintended information to public, creating a boundary turbulence as defined by the CPM. The information owner has limited resources to resolve this unintended information disclosure created by the boundary turbulence. One of the resources provided by the

platforms is privacy settings. However, most platforms only provide privacy settings for the posts created by the users. The posts mentioning or tagging a user can only be modified or deleted by the poster. Users can request others to delete such information which may cause discomfort between the user and their connections. Hence, there is a need for collective privacy understanding and management framework on social media platforms. Post owners and stake-holders, e.g. users tagged in the post, should be able to modify and decide on the privacy settings together at the very least. There is also research that calls for even more elaborate understanding of privacy on social media. For example, networked privacy theory of boyd and Marwick [112] asks for more relationship-based framing of privacy rather than individual or group-based management ones. Platforms should move towards less individualistic models of privacy and future research should focus on how to translate complex privacy protection strategies we use in real life to these platforms.

Petronio's CPM Theory [124] states that people have privacy rules they use to regulate their boundaries. Some of these privacy rules can be seen in Chapter 5 where the reasons to change privacy settings are selected. For some people posts with personal content must be hidden from strangers, and some do not want their students to see their posts. According to CPM, one of the factors determining these rules is the risk-benefit ratio. This is in line with my findings where users navigated the desire to fully experience the platform functionalities while protecting their privacy.

According to Altman's Theory [10], privacy of a person does not have a static state, rather it changes according to the conditions, internal and external. For example, a person has different privacy behaviours and management mechanisms when they are a children vs. when they are an adult. A person's privacy regulation methods might also change depending on the environment they are in. This definition of privacy is supported by the findings of this thesis. It is shown in Chapter 5 that users employ dynamic privacy management strategies, even when they are provided with static settings. They change their privacy settings according to their interactions, the content of their posts, or depending on the social circle they are in, e.g. having different privacy rules for a fan account compared to a professional one.

The dynamic privacy setting usage also has ethical implications for researchers, developers, and companies. The Twitter API only gives access to tweets of public accounts but, as this thesis shows, the premise that accounts can be neatly sorted into public or protected is flawed and that a good number of accounts change state regularly. So simply assuming that all users who currently have a public account are comfortable

with having all their tweets, including historical tweets, available via API at scale may be incorrect. Users may also not be aware of how fast their tweets can be collected or the range of people who might try and record them such as researchers, developers, third-party apps, and companies [17, 49, 147, 4].

As mentioned, the publicly shared information can have audience that may not be obvious to the users such as companies, search indexes, researchers, developers, and third-party apps [17, 49, 147, 4]. OSNs may warn their users about who can see the posts shared on their platform but these warnings are usually buried in help pages instead of the front page of the application. For example, the information that third-party applications installed by the users can reach these users' protected tweets or the links of pictures shared by protected accounts are not actually protected and can be seen by anyone who has the link is only given in the help pages for privacy settings on Twitter. Increasing user understanding of the reach of the information they share is necessary for privacy protection. Platforms should also consider providing users a way to block third-party use via the API. For example, a flag can be put on the profile of the users who choose to opt-out from such data use as Fiesler and Proferes suggested [49].

One of the main findings of this thesis is the role of interactions, therefore networks, in protecting privacy. These interactions leak information about the parties involved and the visibility of the interactions is not clear to people. Even the users who actively utilized privacy settings did not fully understand the reach of their interactions where interaction management was one of the main reasons for them to use the privacy settings. Interacting with others is one of the main purposes of OSNs where these interactions are mostly enjoyed and sought after by users. For example, users share screenshots of their profile pages with the Twitter-provided balloons to let their followers know it is their birthday. Interactions are also valued by algorithms used by OSNs as more platforms start to show the liked posts of a user's network in the main browsing pages, e.g. timelines, feeds. Berstein et al. [17] found that users underestimate the size of the audience of their own posts where I find that my participants reasonably understand the individual tweet visibility. However, the visibility of interactions is not clear to the users, especially when the interaction happens with less common account types. These interactions can help infer the original posts even when those posts are inaccessible, e.g. hidden or deleted [114]. According to Jurgens et al. [77], the interactions can be even more informative than the users' own tweets. Increasing the user understanding of the interactions and implications of these interactions is essential. Platforms should collaborate with researchers and make increased

efforts to inform their users around the functionalities of their products, as well as the protections they have in place.

Interaction management was one of the main drivers of changing tweet visibility settings to protected in the study conducted in Chapter 5. Supporting safe interaction spaces where users are less worried about unintended data disclosures can encourage them to reach out and get the support and help they need. This is especially important when the consequences of the data leak is critical. Vulnerable populations can choose to leave their communities because of these privacy concerns where cutting off communication with their communities might be harmful to them [44]. Sharing experiences can be a stepping stone in the healing process of a person [12]. The risk of information leak might prevent people from reaching out and starting their healing process since people might employ self-censorship [91] to protect their privacy. Sleeper et al. [140] found that some people would share things they self-censor if they could perfectly configure the intended audience.

Another strategy users employ to protect their privacy on social media platforms is deleting posts [74]. I also found that nearly half of the user survey participants in Chapter 5 deleted tweets to protect their privacy. I also encountered tweet deletions in all of my studies' Twitter data. However, as seen in Mondal et al. [114] and my work in this thesis, i.e. Chapter 3 and Chapter 4, deleting a tweet might not be enough to protect privacy since the interactions stay on the platform. In Chapter 6, I show that just over half of the participants correctly thought deleting a public tweet will not delete the replies. Only over a third of the participants knew deleting a protected tweet will not delete the replies, even the protected accounts answered similarly with other accounts. This shows the importance of the problem where some people delete their tweets and do not realize the interactions that are left on the platform.

There is a serious overhead of understanding and configuring privacy settings for users. Every platform has their own set of features for sharing information and these platforms also have different privacy settings. Understanding these settings and configuring them correctly for each platform to manage information flow is not trivial [105]. Prior research shows that these settings are cumbersome for users [155]. It is also common to stay with the default settings provided [48] or configure the settings at the account creation stage [145]. In the information visibility study, detailed in Chapter 6, it is shown that using other social media platforms like Instagram and Snapchat is correlated with the increased awareness of Twitter functionality around visibility. However, people can also incorrectly carry their privacy understandings from one plat-

form to another. This can cause users to share more information than their intentions. OSNs should relay privacy settings and their implications actively rather than putting them into help pages that might not be seen by the users. Even when a particular user does not wish to change their privacy settings, it is necessary for them to understand the visibility of their interactions with users who choose to utilize these settings.

7.3 Ethical Considerations

I followed the University of Edinburgh ethics protocol while designing and running all of the studies, including the social media data collection and following surveys. Even though I access the tweets of public accounts in all of these studies, users may have a more fluid account types as shown in Chapter 5. While Twitter is quite clear about the implications of making an account public and most of the survey participants in the studies evidenced awareness that their tweets can be accessed by non-followers when they changed their tweets to public, it might not be clear to the users that researchers are also included on that set [49]. To mitigate possible risks, only the metadata curated from the social media data are reported. I report on aggregate information and refrain from singling out individuals in quotes, links, or anything else identifying. When using quotes, I carefully select those that are generic and represent common tweet content (i.e. “Happy sweet 16th birthday!”). I also do not collect the survey participants’ Twitter data or link it to their answers. For each study, only the researchers involved have access to the collected raw data.

7.4 Limitations and Future Directions

There are some common limitations of the surveys conducted in this thesis. Firstly, I pre-screen participants relying on their self-report and I decide on their eligibility for the respective surveys depending on their answers. While Prolific [128] has a filtering feature that I use, they also rely on the honesty of the people they recruit. Hence, I do not check whether the participants really have Twitter profiles or if they changed their privacy settings in the past. Another limitation of the surveys is the lack of stratified sampling while recruiting participants. While I compared the participants to existing Twitter demographics, I did not collect a stratified sample of Twitter population in the surveys. Lastly, the breakdown analysis with age and sex demographics for the user surveys were conducted posthoc.

In this thesis, I focused on quantifying the potential privacy violations, privacy settings usage, and the user understanding of privacy settings and information visibility on Twitter. To do so, I collected data from Twitter and user surveys. I found that interactions between accounts are not clearly understood. A possible next step would be increasing the awareness of the users. In my studies, the Twitter data collection and the user surveys are not linked. Combining data collected from profiles with user surveys and intervening using various strategies such as informational boards, privacy nudges, and so on may help to increase awareness of the users [2, 126]. Increase of the awareness can be evaluated by follow-up Twitter data collection and/or user surveys.

Twitter is mainly a text-based social media. While tweets can have pictures and videos, they usually have an accompanying text. Hence, in my studies I also focused on text-based privacy violations. For example, I used LDA on the texts or the focus is on the text-based questions in the surveys. In Chapter 5, I do check whether users share more pictures when public vs protected, and also ask picture related questions in the survey. However, the content and context of these pictures are not investigated in detail. Twitter also recently added a feature for users to share audio with their tweets or create "Spaces" to hold voice-based conversations with other users. A future direction may be investigating the potential privacy violation through these pictures and audio data. While picture based privacy is highly studied [172, 148, 121], voice-based social platforms are emerging and remain understudied. Especially, voice-based conversations that happen over "Spaces" or voice-based social media platforms like Clubhouse [35] can be investigated for unintended disclosures and privacy perceptions of users around these new types of social media communication.

The median age of Twitter user in the United States is 40 [39] which is younger than the general public. However, recently emerged platforms like TikTok [151] has considerably younger demographics than even Twitter. Nearly half of the TikTok users from the United States are younger than 30 and the quarter of the users are in their teens. This brings a whole set of different questions around privacy and data disclosure. Privacy perceptions of teenagers in social media are studied [112]. However, TikTok is different from other platforms in terms of the content-type which is focused on short videos with eye-catching content. Hence, a future direction may be investigating the privacy settings usage and the perception of the TikTok users.

Aside from the study for investigating privacy settings switching behaviour detailed in Chapter 5, all of the tweets collected from Twitter were restrained only to English language. All of the user studies conducted in English and participants were

recruited according to their fluency in English. Considering only around the third of all tweets are in English [159] and around the fifth of the world's population speaks English [120], my studies are limited to a part of the Twitter users. Even more, the analysis is conducted without considering cultural differences in English speaking communities. Privacy perceptions and understandings can change depending on the culture [10, 1] and English speaking community is not a culturally monolithic group. Further nuance is needed to analyze the cultural differences regarding privacy. Future work can also investigate the privacy perceptions of users who use other languages on the platform and from non-English speaking countries. Japanese, Spanish, Korean, and Arabic are some of the popular languages on Twitter [159]. Some of the countries with top Twitter users are Japan, India, Brazil, Indonesia, and Turkey [41]. More inclusive research especially focusing on non-WEIRD (Western, Educated, Industrialized, Rich and Democratic) populations is needed considering they are underrepresented [57, 168].

Bibliography

[1] Norah Abokhodair and Sarah Vieweg. Privacy & social media in the context of the arab gulf. In *Proceedings of the 2016 ACM conference on designing interactive systems*, pages 672–683, 2016.

[2] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, et al. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)*, 50(3):1–41, 2017.

[3] Alessandro Acquisti and Ralph Gross. Imagined communities: Awareness, information sharing, and privacy on the facebook. In *International workshop on privacy enhancing technologies*, pages 36–58. Springer, 2006.

[4] Esma Aïmeur, Gilles Brassard, and Jonathan Rioux. Data privacy: An end-user perspective. *International Journal of Computer networks and communications security*, 1(6):237–250, 2013.

[5] Mohammad Akbari, Xia Hu, Nie Liqiang, and Tat-Seng Chua. From tweets to wellness: Wellness event detection from twitter streams. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[6] Fatma Al Maqbali and Chris J Mitchell. Web password recovery: A necessary evil? In *Proceedings of the Future Technologies Conference*, pages 324–341. Springer, 2018.

[7] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[8] Abeer Aldayel and Walid Magdy. Your stance is exposed! analysing possible factors for stance detection on social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–20, 2019.

[9] Hazim Almuhimedi, Shomir Wilson, Bin Liu, Norman Sadeh, and Alessandro Acquisti. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 897–908, 2013.

[10] Irwin Altman. A conceptual analysis. *Environment and behavior*, 8(1):7–29, 1976.

[11] Mary Jean Amon, Rakibul Hasan, Kurt Hugenberg, Bennett I Bertenthal, and Apu Kapadia. Influencing photo sharing decisions on social media: A case of paradoxical findings. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 79–95, 2020.

[12] Nazanin Andalibi and Andrea Forte. Announcing pregnancy loss on facebook: A decision-making framework for stigmatized disclosures on identified social network sites. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018.

[13] James P Bagrow, Xipei Liu, and Lewis Mitchell. Information flow reveals prediction limits in online social activity. *Nature human behaviour*, 3(2):122–128, 2019.

[14] Miriam Bartsch and Tobias Dienlin. Control your facebook: An analysis of online privacy literacy. *Computers in Human Behavior*, 56:147–154, 2016.

[15] Eric PS Baumer, Phil Adams, Vera D Khovanskaya, Tony C Liao, Madeline E Smith, Victoria Schwanda Sosik, and Kaiton Williams. Limiting, leaving, and (re) lapsing: an exploration of facebook non-use practices and experiences. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3257–3266. ACM, 2013.

[16] Natalya N Bazarova and Yoon Hyung Choi. Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of Communication*, 64(4):635–657, 2014.

[17] Michael S Bernstein, Eytan Bakshy, Moira Burke, and Brian Karrer. Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 21–30. ACM, 2013.

[18] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[19] Joseph Bonneau, Elie Bursztein, Ilan Caron, Rob Jackson, and Mike Williamson. Secrets, lies, and account recovery: Lessons from the use of personal knowledge questions at google. In *Proceedings of the 24th international conference on world wide web*, pages 141–150. International World Wide Web Conferences Steering Committee, 2015.

[20] Joseph Bonneau, Sören Preibusch, and Ross Anderson. A birthday present every eleven wallets? the security of customer-chosen banking pins. In *International Conference on Financial Cryptography and Data Security*, pages 25–40. Springer, 2012.

[21] Danah Boyd. Networked privacy. *Surveillance & society*, 10(3/4):348, 2012.

[22] Danah Boyd and Alice E Marwick. Social privacy in networked publics: Teens' attitudes, practices, and strategies. In *A decade in internet time: Symposium on the dynamics of the internet and society*, 2011.

[23] Danah M Boyd and Nicole B Ellison. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, 13(1):210–230, 2007.

[24] Louis Brandeis and Samuel Warren. The right to privacy. *Harvard law review*, 4(5):193–220, 1890.

[25] Petter Bae Brandtzaeg and Marika Lüders. Time collapse in social media: extending the context collapse. *Social Media+ Society*, 4(1):2056305118763349, 2018.

[26] Alan S Brown, Elisabeth Bracken, Sandy Zoccoli, and King Douglas. Generating and remembering passwords. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 18(6):641–651, 2004.

[27] Moritz Büchi, Natascha Just, and Michael Latzer. Caring is not enough: the importance of internet skills for online privacy protection. *Information, Communication & Society*, 20(8):1261–1278, 2017.

[28] Leland Bybee, Bryan T Kelly, Asaf Manela, and Dacheng Xiu. The structure of economic news. Technical report, National Bureau of Economic Research, 2020.

[29] Ricardo Castro. #hbd: Celebrate your birthday on twitter, July 2015. URL: https://blog.twitter.com/official/en_us/a/2015/hbd-celebrate-your-birthday-on-twitter.html.

[30] Twitter Help Center. About public and protected tweets, 2021. URL: https://help.twitter.com/en/safety-and-security/public-and-protected-tweets.

[31] Jeffrey T Child, Judy C Pearson, and Sandra Petronio. Blogging, communication, and privacy management: Development of the blogging privacy management measure. *Journal of the American Society for Information Science and Technology*, 60(10):2079–2094, 2009.

[32] Jeffrey T Child and David A Westermann. Let's be facebook friends: Exploring parental facebook friend requests from a communication privacy management (cpm) perspective. *Journal of Family Communication*, 13(1):46–59, 2013.

[33] Mina Choi and Catalina L Toma. Social sharing through interpersonal media: Patterns and effects on emotional well-being. *Computers in Human Behavior*, 36:530–541, 2014.

[34] Yoon Hyung Choi and Natalya N Bazarova. Self-disclosure characteristics and motivations in social media: Extending the functional model to multiple social network sites. *Human Communication Research*, 41(4):480–500, 2015.

[35] Clubhouse. Clubhouse: The social audio app. URL: https://www.clubhouse.com/.

[36] European Commission. Gdpr personal data – what information does this cover?, 2020. URL: https://www.gdpreu.org/the-regulation/key-concepts/personal-data/.

[37] Sauvik Das and Adam Kramer. Self-censorship on facebook. In *Seventh international AAAI conference on weblogs and social media*, 2013.

[38] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Major life changes and behavioral markers in social media: case of childbirth. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1431–1442, 2013.

[39] Statista Research Department. Distribution of twitter users worldwide as of july 2021, by age, June 2021. URL: https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/.

[40] Statista Research Department. Distribution of twitter users worldwide as of july 2021, by gender, August 2021. URL: https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/.

[41] Statista Research Department. Leading countries based on number of twitter users as of october 2021, October 2021. URL: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.

[42] Thomas Dickinson, Miriam Fernandez, Lisa A Thomas, Paul Mulholland, Pam Briggs, and Harith Alani. Identifying prominent life events on twitter. In *Proceedings of the 8th International Conference on Knowledge Capture*, pages 1–8, 2015.

[43] Charles Duhigg. How companies learn your secrets, February 2012. URL: https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

[44] Brianna Dym and Casey Fiesler. Vulnerable and online: Fandom's case for stronger privacy norms and tools. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 329–332, 2018.

[45] Nicole B Ellison, Jessica Vitak, Charles Steinfield, Rebecca Gray, and Cliff Lampe. Negotiating privacy concerns and social capital needs in a social media environment. In *Privacy online*, pages 19–32. Springer, 2011.

[46] ENISA. Tips for secure user authentication, 2020. URL: https://www.enisa.europa.eu/news/enisa-news/tips-for-secure-user-authentication.

[47] Facebook. Facebook, 2022. URL: https://www.facebook.com/.

[48] Casey Fiesler, Michaelanne Dye, Jessica L Feuston, Chaya Hiruncharoenvate, Clayton J Hutto, Shannon Morrison, Parisa Khanipour Roshan, Umashanthi Pavalanathan, Amy S Bruckman, Munmun De Choudhury, et al. What (or who) is public?: Privacy settings and social media content sharing. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 567–580. ACM, 2017.

[49] Casey Fiesler and Nicholas Proferes. "participant" perceptions of twitter research ethics. *Social Media+ Society*, 4(1):2056305118763366, 2018.

[50] Glenn Fleishman. How facebook devalued the birthday, 2018. URL: https://www.fastcompany.com/40550725/how-facebook-devalued-the-birthday.

[51] David Garcia, Mansi Goel, Amod Kant Agrawal, and Ponnurangam Kumaraguru. Collective aspects of privacy in the twitter social network. *EPJ Data Science*, 7(1):3, 2018.

[52] Kiran Garimella, Ingmar Weber, and Munmun De Choudhury. Quote rts on twitter: usage of the new feature for political discourse. In *Proceedings of the 8th ACM Conference on Web Science*, pages 200–204, 2016.

[53] Shirley Gaw, Edward W. Felten, and Patricia Fernandez-Kelly. Secrecy, flagging, and paranoia: adoption criteria in encrypted email. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, page 591–600, 2006. doi:10.1145/1124772.1124862.

[54] Erving Goffman. *The presentation of self in everyday life*. Garden City, NY Double Day, 1959.

[55] Sukeshini A Grandhi, Linda Plotnick, and Starr Roxanne Hiltz. Do i stay or do i go? motivations and decision making in social media non-use and reversion. *Proceedings of the ACM on Human-Computer Interaction*, 3(GROUP):1–27, 2019.

[56] Paul Grassi, Michael Garcia, and James Fenton. Digital identity guidelines, 2017-06-22 2017. doi:https://doi.org/10.6028/NIST.SP.800-63-3.

[57] Joseph Henrich, Steven J Heine, and Ara Norenzayan. Most people are not weird. *Nature*, 466(7302):29–29, 2010.

[58] Ki Mae Heussner. Woman loses benefits after posting facebook pics, November 2009. URL: https://abcnews.go.com/Technology/AheadoftheCurve/woman-loses-insurance-benefits-facebook-pics/story?id=9154741.

[59] Karen Holtzblatt and Hugh Beyer. 6 - the affinity diagram. In Karen Holtzblatt and Hugh Beyer, editors, *Contextual Design (Second Edition)*, Interactive Technologies, pages 127–146. Morgan Kaufmann, Boston, 2017. URL: https://www.sciencedirect.com/science/article/pii/B9780128008942000065, doi:https://doi.org/10.1016/B978-0-12-800894-2.00006-5.

[60] Mat Honan. How apple and amazon security flaws led to my epic hacking, August 2012. URL: https://www.wired.com/2012/08/apple-amazon-mat-honan-hacking/.

[61] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[62] Roberto Hoyle, Srijita Das, Apu Kapadia, Adam J. Lee, and Kami Vaniea. Viewing the viewers: Publishers' desires and viewers' privacy concerns in social networks. In *Proceedings of the Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, pages 555–566, 2017.

[63] Hongxin Hu, Gail-Joon Ahn, and Jan Jorgensen. Multiparty access control for online social networks: model and mechanisms. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1614–1627, 2013.

[64] Xiaoyun Huang, Jessica Vitak, and Yla Tausczik. " you don't have to know my past": How wechat moments users manage their evolving self-presentation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[65] Instagram. Instagram, 2022. URL: https://www.instagram.com/.

[66] Danesh Irani, Steve Webb, Kang Li, and Calton Pu. Modeling unintended personal-information leakage from multiple online social networks. *IEEE Internet Computing*, 15(3):13–19, 2011.

[67] Stephen Jackson, Nadia Vanteeva, and Colm Fearon. An investigation of the impact of data breach severity on the readability of mandatory data breach notification letters: Evidence from us firms. *Journal of the Association for Information Science and Technology*, 70(11):1277–1289, 2019.

[68] Prachi Jain, Paridhi Jain, and Ponnurangam Kumaraguru. Call me maybe: Understanding nature and risks of sharing mobile numbers on online social networks. In *Proceedings of the first ACM conference on Online social networks*, pages 101–106. ACM, 2013.

[69] Steve MJ Janssen and David C Rubin. Age effects in cultural life scripts. *Applied Cognitive Psychology*, 25(2):291–298, 2011.

[70] Carter Jernigan and Behram FT Mistree. Gaydar: Facebook friendships expose sexual orientation. *First Monday*, 14(10), 2009.

[71] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2):1–33, 2018.

[72] Haiyan Jia and Heng Xu. Autonomous and interdependent: Collaborative privacy management on social networking sites. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4286–4297. ACM, 2016.

[73] Seung-A Annie Jin. "to disclose or not to disclose, that is the question": A structural equation modeling approach to communication privacy management in e-health. *Computers in human Behavior*, 28(1):69–77, 2012.

[74] Maritza Johnson, Serge Egelman, and Steven M Bellovin. Facebook and privacy: it's complicated. In *Proceedings of the eighth symposium on usable privacy and security*, pages 1–15, 2012.

[75] Yumi Jung and Emilee Rader. The imagined audience and privacy concern on facebook: Differences between producers and consumers. *Social media+ society*, 2(2):2056305116644615, 2016.

[76] David Jurgens. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[77] David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. Writer profiling without the writer's text. In *International Conference on Social Informatics*, pages 537–558. Springer, 2017.

[78] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara Kiesler. Privacy attitudes of Mechanical Turk workers and the U.S. public. In *Symposium On Usable Privacy and Security*, SOUPS '14, pages 37–49, July 2014.

[79] Amir Karami and Aida Elkouri. Political popularity analysis in social media. In *International conference on information*, pages 456–465. Springer, 2019.

[80] Amir Karami, Morgan Lundy, Frank Webb, and Yogesh K Dwivedi. Twitter and research: a systematic literature review through text mining. *IEEE Access*, 8:67698–67717, 2020.

[81] Ravneet Kaur, Ravtej Singh Sandhu, Ayush Gera, Tarlochan Kaur, and Purva Gera. Intelligent voice bots for digital banking. In *Smart Systems and IoT: Innovations in Computing*, pages 401–408. Springer, 2020.

[82] Dilara Keküllüoğlu, Nadin Kokciyan, and Pinar Yolum. Preserving privacy as social responsibility in online social networks. *ACM Transactions on Internet Technology (TOIT)*, 18(4):42, 2018.

[83] Dilara Keküllüoğlu, Walid Magdy, and Kami Vaniea. Analysing privacy leakage of life events on twitter. In *Proceedings of the 12th ACM Conference on Web Science*, 2020.

[84] Dilara Keküllüoğlu, Walid Magdy, and Kami Vaniea. Happy 18th birthday! measuring birth date disclosure on twitter. In *6th International Conference on Computational Social Science*, 2020.

[85] Dilara Keküllüoğlu, Walid Magdy, and Kami Vaniea. From an Authentication Question to a Public Social Event: Characterizing Birthday Sharing on Twitter. In *Proceedings of The 16th International AAAI Conference on Weblogs and Social Media (ICWSM'22)*, June 2022.

[86] Dilara Keküllüoğlu, Kami Vaniea, and Walid Magdy. Understanding Privacy Switching Behaviour on Twitter. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, May 2022.

[87] Jennifer King, Airi Lampinen, and Alex Smolen. Privacy: Is there an app for that? In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, page 12. ACM, 2011.

[88] Nadin Kökciyan, Nefise Yaglikci, and Pinar Yolum. An argumentation approach for resolving privacy disputes in online social networks. *ACM Transactions on Internet Technology (TOIT)*, 17(3):27, 2017.

[89] Katharina Krombholz, Heidelinde Hobel, Markus Huber, and Edgar Weippl. Advanced social engineering attacks. *Journal of Information Security and applications*, 22:113–122, 2015.

[90] Cliff Lampe, Jessica Vitak, and Nicole Ellison. Users and nonusers: Interactions between levels of adoption and social capital. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 809–820, 2013.

[91] Airi Lampinen, Vilma Lehtinen, Asko Lehmuskallio, and Sakari Tamminen. We're in it together: interpersonal management of disclosure in social network services. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3217–3226. ACM, 2011.

[92] Christopher A Langston. Capitalizing on and coping with daily-life events: Expressive responses to positive events. *Journal of personality and social psychology*, 67(6):1112, 1994.

[93] Kevin Lee, Benjamin Kaiser, Jonathan Mayer, and Arvind Narayanan. An empirical study of wireless carrier authentication for {SIM} swaps. In *Sixteenth Symposium on Usable Privacy and Security ({SOUPS} 2020)*, pages 61–79, 2020.

[94] Jiwei Li, Alan Ritter, Claire Cardie, and Eduard Hovy. Major life event extraction from twitter based on congratulations/condolences speech acts. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1997–2007, 2014.

[95] Tianlin Li, Amish Mehta, and Ping Yang. Security analysis of email systems. In *2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud)*, pages 91–96. IEEE, 2017.

[96] Hai Liang, Fei Shen, and King-wa Fu. Privacy protection and self-disclosure across societies: A study of global twitter users. *new media & society*, 19(9):1476–1497, 2017.

[97] E. Lieberman and R.C. Miller. Facemail: showing faces of recipients to prevent misdirected email. In *Proceedings of the 3rd symposium on Usable privacy and security*, page 122–131. ACM, 2007.

[98] LinkedIn. Linkedin. URL: `https://www.linkedin.com/`.

[99] Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1–22, 2016.

[100] Yabing Liu, Krishna P Gummadi, Balachander Krishnamurthy, and Alan Mislove. Analyzing facebook privacy settings: user expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 61–70, 2011.

[101] Yabing Liu, Chloe Kliman-Silver, and Alan Mislove. The tweets they are a-changin': Evolution of twitter users and behavior. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[102] Riemer Hess LLC. Is your disability insurer monitoring your social media? [insider tips]. URL: `https://www.riemerhess.com/wiki/top-5-ways-social-media-can-hurt-your-long-term-disability-claim`.

[103] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, page 63–70, USA, 2002. Association for Computational Linguistics. `doi:10.3115/1118108.1118117`.

[104] Michelle Madejski, Maritza Johnson, and Steven M Bellovin. A study of privacy settings errors in an online social network. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*, pages 340–345. IEEE, 2012.

[105] Michelle Madejski, Maritza Lupe Johnson, and Steven Michael Bellovin. The failure of online social network privacy settings. 2011.

[106] Walid Magdy, Yehia Elkhatib, Gareth Tyson, Sagar Joglekar, and Nishanth Sastry. Fake it till you make it: Fishing for catfishes. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 497–504. ACM, 2017.

[107] Merja Mahrt, Katrin Weller, and Isabella Peters. Twitter in scholarly communication. *Twitter and society*, 89:399–410, 2014.

[108] Huina Mao, Xin Shuai, and Apu Kapadia. Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, pages 1–12. ACM, 2011.

[109] Stephen T Margulis. Privacy as a social issue and behavioral concept. *Journal of social issues*, 59(2):243–261, 2003.

[110] Ereni Markos, George R Milne, and James W Peltier. Information sensitivity and willingness to provide continua: a comparative privacy study of the united states and brazil. *Journal of Public Policy & Marketing*, 36(1):79–96, 2017.

[111] Alice E Marwick and Danah Boyd. I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society*, 13(1):114–133, 2011.

[112] Alice E Marwick and Danah Boyd. Networked privacy: How teenagers negotiate context in social media. *New media & society*, 16(7):1051–1067, 2014.

[113] Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20, 2015.

[114] Mainack Mondal, Johnnatan Messias, Saptarshi Ghosh, Krishna P Gummadi, and Aniket Kate. Forgetting in social media: Understanding and controlling longitudinal exposure of socially shared data. In *Twelfth Symposium on Usable Privacy and Security ({SOUPS} 2016)*, pages 287–299, 2016.

[115] Mainack Mondal, Günce Su Yilmaz, Noah Hirsch, Mohammad Taha Khan, Michael Tang, Christopher Tran, Chris Kanich, Blase Ur, and Elena Zheleva.

Moving beyond set-it-and-forget-it privacy settings on social media. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 991–1008, 2019.

[116] Steven J Murdoch and Ross Anderson. Verified by visa and mastercard securecode: or, how not to design authentication. In *International Conference on Financial Cryptography and Data Security*, pages 336–342. Springer, 2010.

[117] Andreas Möller, Matthias Kranz, Barbara Schmid, Luis Roalter, and Stefan Diewald. Investigating self-reporting behavior in long-term studies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 2931–2940, 2013.

[118] Yee Man Margaret Ng. Re-examining the innovation post-adoption process: the case of twitter discontinuance. *Computers in Human Behavior*, 103:48–56, 2020.

[119] Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.

[120] Ethnologue Languages of the World. English. URL: `https://www.ethnologue.com/language/eng`.

[121] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision*, pages 19–35. Springer, 2016.

[122] OPC. Guidelines for identification and authentication, 2016. URL: `https://www.priv.gc.ca/en/privacy-topics/identities/identification-and-authentication/auth_061013/`.

[123] Sandra Petronio. *Boundaries of privacy: Dialectics of disclosure*. Suny Press, 2002.

[124] Sandra Petronio. Communication privacy management theory. *The International Encyclopedia of Interpersonal Communication*, pages 1–9, 2015.

[125] Richard A Posner. Economic theory of privacy. *Regulation*, 2:19, 1978.

[126] Stefanie Pötzsch. Privacy awareness: A means to solve the privacy paradox? In *IFIP Summer School on the Future of Identity in the Information Society*, pages 226–236. Springer, 2008.

[127] Nicholas Proferes. Information flow solipsism in an exploratory study of beliefs about twitter. *Social Media+ Society*, 3(1):2056305117698493, 2017.

[128] Prolific academic, 2019. URL: https://www.prolific.co/.

[129] Yu Pu and Jens Grossklags. Valuating friends' privacy: Does anonymity of sharing personal data matter? In *Thirteenth Symposium on Usable Privacy and Security ({SOUPS} 2017)*, pages 339–355, 2017.

[130] Qualtrics xm. URL: https://www.qualtrics.com.

[131] Ariel Rabkin. Personal knowledge questions for fallback authentication: Security questions in the era of facebook. In *Proceedings of the 4th symposium on Usable privacy and security*, pages 13–23. ACM, 2008.

[132] Yasmeen Rashidi, Apu Kapadia, Christena Nippert-Eng, and Norman Makoto Su. "it's easier than causing confrontation": Sanctioning strategies to maintain social norms and privacy on social media. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–25, 2020.

[133] Reddit. Reddit, 2022. URL: https://www.reddit.com/.

[134] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[135] Emma Remy. How public and private twitter users in the u.s. compare — and why it might matter for your research, 2019. URL: https://medium.com/pew-research-center-decoded/how-public-and-private-twitter-users-in-the-u-s-d536ce2a41b3.

[136] Data Reportal. Global social media stats, January 2022. URL: https://datareportal.com/social-media-users/.

[137] Elizabeth Rivelli. Your social media could affect your insurance rates, August 2022. URL: https://www.coverage.com/insurance/status-update-you

r-social-media-activity-could-affect-your-insurance-rates-more/.

[138] Kelley Robinson. What i learned about security from calling 35 contact centers, 2019. URL: https://www.twilio.com/blog/learned-about-security-from-calling-35-contact-centers.

[139] Koustuv Saha, Jordyn Seybolt, Stephen M Mattingly, Talayeh Aledavood, Chaitanya Konjeti, Gonzalo J Martinez, Ted Grover, Gloria Mark, and Munmun De Choudhury. What life events are disclosed on social media, how, when, and by whom? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2021.

[140] Manya Sleeper, Rebecca Balebako, Sauvik Das, Amber Lynn McConahy, Jason Wiese, and Lorrie Faith Cranor. The post that wasn't: exploring self-censorship on facebook. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 793–802, 2013.

[141] Manya Sleeper, Justin Cranshaw, Patrick Gage Kelley, Blase Ur, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. " i read my twitter the next morning and was astonished" a conversational perspective on twitter regrets. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3277–3286, 2013.

[142] Ben Smyth. Forgotten your responsibilities? how password recovery threatens banking security. Technical report, Citeseer, 2010.

[143] Statista. Age distribution of twitter users in great britain from q2 2013 to q1 2018, 2021. URL: https://www.statista.com/statistics/278320/age-distribution-of-twitter-users-in-great-britain/.

[144] Statista. Gender breakdown of twitter users in great britain (gb) from may 2014 to february 2018, 2021. URL: https://www.statista.com/statistics/278319/gender-breakdown-of-twitter-users-in-great-britain/.

[145] Katherine Strater and Heather Richter Lipford. Strategies and struggles with privacy in an online social networking community. *People and Computers XXII Culture, Creativity, Interaction 22*, pages 111–119, 2008.

[146] Frederic Stutzman and Woodrow Hartzog. Boundary regulation in social media. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 769–778, 2012.

[147] Frederic D Stutzman, Ralph Gross, and Alessandro Acquisti. Silent listeners: The evolution of privacy and disclosure on facebook. *Journal of privacy and confidentiality*, 4(2):2, 2013.

[148] Jose M Such, Joel Porter, Sören Preibusch, and Adam Joinson. Photo privacy conflicts in social media: A large-scale empirical study. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3821–3832, 2017.

[149] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34, 2000.

[150] Sue Yeon Syn and Sanghee Oh. Why do social network site users share information on facebook and twitter? *Journal of Information Science*, 41(5):553–569, 2015.

[151] TikTok. Tiktok - make your day. URL: `https://www.tiktok.com/`.

[152] Mina Tsay-Vogel, James Shanahan, and Nancy Signorielli. Social media cultivating perceptions of privacy: A 5-year analysis of privacy attitudes and self-disclosure behaviors among facebook users. *new media & society*, 20(1):141–161, 2018.

[153] Twitter. Twitter. URL: `https://twitter.com/`.

[154] Twitter. Twitter api. URL: `https://developer.twitter.com/`.

[155] Brendan Van Alsenoy, Valerie Verdoodt, Rob Heyman, Ellen Wauters, Jef Ausloos, and Günes Acar. From social media service to advertising network: a critical analysis of facebook's revised policies and terms. 2015.

[156] Alexander JAM Van Deursen, Jan AGM Van Dijk, and Oscar Peters. Rethinking internet skills: The contribution of gender, age, education, internet experience, and hours online to medium-and content-related internet skills. *Poetics*, 39(2):125–144, 2011.

[157] Kami Vaniea, Lujo Bauer, Lorrie Faith Cranor, and Michael K. Reiter. Out of sight, out of mind: Effects of displaying access-control information near the item it controls. In *PST 2012: Conference on Privacy, Security, and Trust*, 2012.

[158] George Veletsianos, Shandell Houlden, Jaigris Hodson, and Chandell Gosse. Women scholars' experiences with online harassment and abuse: Self-protection, resistance, acceptance, and self-blame. *New Media & Society*, 20(12):4689–4708, 2018.

[159] Vicinitas. 2018 research on 100 million tweets: What it means for your social media strategy for twitter. URL: `https://www.vicinitas.io/blog/twitter-social-media-strategy-2018-research-100-million-tweets`.

[160] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42, 2009.

[161] Jessica Vitak. The impact of context collapse and privacy on social network site disclosures. *Journal of broadcasting & electronic media*, 56(4):451–470, 2012.

[162] Jessica Vitak. Facebook makes the heart grow fonder: relationship maintenance strategies among geographically dispersed and communication-restricted connections. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 842–853, 2014.

[163] Jessica Vitak, Stacy Blasiola, Sameer Patil, and Eden Litt. Balancing audience and privacy tensions on social network sites: Strategies of highly engaged users. *International Journal of Communication*, 9:20, 2015.

[164] Jessica Vitak and Jinyoung Kim. " you can't block people offline" examining how facebook's affordances shape the disclosure process. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 461–474, 2014.

[165] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[166] Kurt Wagner. Twitter plans new privacy tools to get more people tweeting, September 2021. URL: `https://www.bloomberg.com/news/articles/20`

21-09-02/twitter-plans-new-privacy-features-to-get-more-people-tweeting.

[167] Ding Wang, Ping Wang, Debiao He, and Yuan Tian. Birthday, name and bifacial-security: understanding passwords of chinese web users. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 1537–1555, 2019.

[168] Yang Wang. Inclusive security and privacy. *IEEE Security & Privacy*, 16(4):82–87, 2018.

[169] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. " i regretted the minute i pressed share" a qualitative study of regrets on facebook. In *Proceedings of the seventh symposium on usable privacy and security*, pages 1–16, 2011.

[170] Rick Wash. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, pages 1–16, 2010.

[171] Stefan Wojcik and Adam Hughes. Sizing up twitter users. *Washington, DC: Pew Research Center*, 2019. URL: https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.

[172] Sijia Xiao, Danaë Metaxa, Joon Sung Park, Karrie Karahalios, and Niloufar Salehi. Random, messy, funny, raw: Finstas as intimate reconfigurations of social media. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[173] Diyi Yang, Robert E Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[174] Günce Su Yılmaz, Fiona Gasaway, Blase Ur, and Mainack Mondal. Perceptions of retrospective edits, changes, and deletion on social media. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM'21)*, 2021.

[175] Xuan Zhao, Niloufar Salehi, Sasha Naranjit, Sara Alwaalan, Stephen Voida, and Dan Cosley. The many faces of facebook: Experiencing social media as

performance, exhibition, and personal archive. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1–10, 2013.

[176] Elena Zheleva and Lise Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th international conference on World wide web*, pages 531–540. ACM, 2009.

Appendix A

Birthday Celebrations on Twitter User Survey

A.1 Overview

This survey was conducted as a part of the study detailed in Chapter 4. I recruited participants from Prolific Academic [128] and the survey was hosted by Qualtrics XM Online Survey Software [130]. I firstly showed the participants an information sheet and consent form explaining the study. I then filtered participants according to their Twitter usage and asked about demographic information, Twitter functionality, and various questions around birthday celebrations and the date of birth disclosure.

A.2 Demographics

1. Do you have a Twitter account?

   ○ Yes, I use it often - at least once a month.

   ○ Yes, I use it rarely - less than once a month.

   ○ No, but I had one I used frequently - at least once a month.

   ○ No. *(Screen out participants if selected.)*

2. Is your Twitter account clearly linked to your real identity?

   ○ Yes.

   ○ No.

○ Not sure.

3. Which best describes your primary Twitter account?

   ○ Public - Anyone can read my tweets.

   ○ Protected - Only select people can read my tweets.

   ○ Sometimes protected - I change between public & protected sometimes.

4. (If "Sometimes protected" selected) What are the most common reasons why you change your account from protected to public and vice-versa? (Open-text Question)

5. Is your birthday publicly visible on any of your social media accounts? (Twitter Facebook Instagram etc.)

   ○ Yes.

   ○ No.

   ○ Not sure.

A.3 Twitter Functionality

6. Alice (@alice) has a public Twitter account and Bob (@bob) has a protected account. They follow each other. In your opinion, what would happen if Alice (Public) retweeted one of Bob's (Protected) tweets using the Twitter website?

   ○ Twitter would not allow Alice to tweet it.

   ○ Twitter would warn Alice.

   ○ Twitter would let Alice tweet, but only followers of Bob could see it.

   ○ It would let Alice tweet, and anyone on Twitter could see it.

   ○ I don't know.

7. Alice (@alice) has a public Twitter account and Bob (@bob) has a protected account. They follow each other. In your opinion, what would happen if Alice (Public) tweeted at Bob (Protected) using his handle (@bob) in the tweet?

   ○ Twitter would not allow Alice to tweet it.

- ○ Twitter would warn Alice.
- ○ Twitter would let Alice tweet, but only followers of Bob could see it.
- ○ It would let Alice tweet, and anyone on Twitter could see it.
- ○ I don't know.

8. For each statement choose the option that describes you the best.

   (a) When replying to a tweet I can easily tell if the poster's account is public or protected.

   - ○ Disagree.
   - ○ Neither agree nor disagree.
   - ○ Agree.

   (b) When engaging (reply, mention, retweet) with a tweet, I look to see if the poster's account is protected.

   - ○ Disagree.
   - ○ Neither agree nor disagree.
   - ○ Agree.

A.4 Birthday Questions

9. Imagine it was **your** birthday today. How comfortable would you be with your friends or family Tweeting publicly that it is your birthday today?

   - ○ Very comfortable
   - ○ Comfortable
   - ○ Neither comfortable nor uncomfortable
   - ○ Uncomfortable
   - ○ Very uncomfortable

10. Imagine it was **your** birthday today. How comfortable would you be with your friends or family Tweeting publicly that it is your birthday along with your age (Birth date)?

    - ○ Very comfortable

- ○ Comfortable
- ○ Neither comfortable nor uncomfortable
- ○ Uncomfortable
- ○ Very uncomfortable

11. Imagine if it was **your friend's or family member's** birthday today. How likely would you be to Tweet publicly that it is their birthday today?

    - ○ Very likely
    - ○ Likely
    - ○ Neither likely nor unlikely
    - ○ Unlikely
    - ○ Very unlikely

12. Imagine if it was **your friend's or family member's** birthday today. How likely would you be to Tweet publicly that it is their birthday today along with their age (Birth date)?

    - ○ Very likely
    - ○ Likely
    - ○ Neither likely nor unlikely
    - ○ Unlikely
    - ○ Very unlikely

13. Imagine a friend or family member tweeted about your birthday publicly on Twitter, how would you respond? Select all that apply.

    - ○ Like the tweet
    - ○ Retweet it
    - ○ Thank them by replying to their tweet
    - ○ Thank them by direct message
    - ○ Ask them to remove it

14. Please give at least one example of a good thing that could happen if someone knew your birthday and age. (Open-text Question)

15. Please give at least one example of a bad thing that could happen if someone knew your birthday and age. (Open-text Question)

16. Any other comments or questions you would want to add. (optional) (Open-text Question)

Appendix B

Explaining Privacy Switching Reasons User Survey

B.1 Overview

This survey was conducted as a part of the study detailed in Chapter 5. We recruited participants from Prolific Academic [128] and the survey was hosted by Qualtrics XM Online Survey Software [130]. I firstly showed our participants an information sheet and consent form explaining the study. Participants were filtered according to their privacy settings usage on Twitter. The survey followed with questions about their recent privacy settings change and the motivation behind the mentioned change. The survey concluded with demographic information questions.

B.2 Screening Questions

1. Do you have a Twitter account?

   ○ Yes.

   ○ No. *(Screen out participants if selected.)*

2. How you ever changed the audience of your Tweets from Public to Protected, or from Protected to Public using the setting shown below? (Figure B.1)

   ○ Yes.

   ○ No. *(Screen out participants if selected.)*

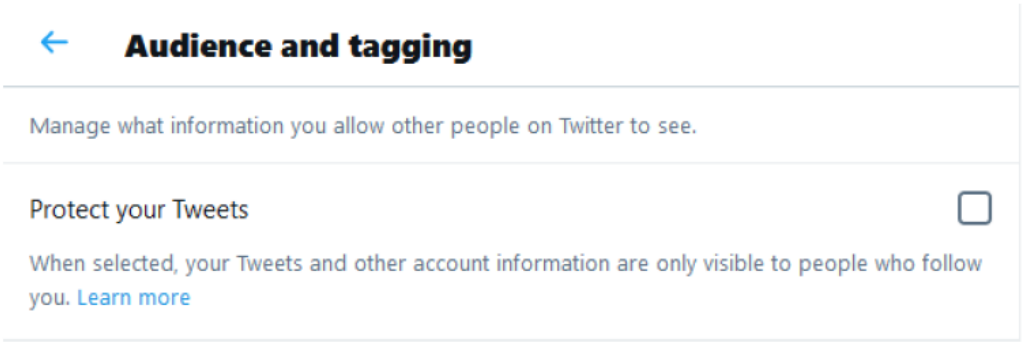

Figure B.1: Settings to change tweet visibility

B.3 Switching Reasons

3. Please think back to a recent time when you have changed your Tweets from public to protected, or from protected to public. Please pick an event you can clearly recall and answer the following four questions in regards to this event. For this event, what motivated you to make the change from public to protected or protected to public? Please be specific in your answer. (Open-text Question)

4. What effect(s) were you hoping to achieve by changing between public and protected? What were you expecting to change? (Open-text Question)

5. Changing between public and protected allowed me to achieve the effect I was trying for.

   ○ Strongly disagree

   ○ Disagree

   ○ Neither agree nor disagree

   ○ Agree

   ○ Strongly agree

6. In the event you described above, which of the following best describes the change you made?

   ○ I made my Tweets **public**.

   ○ I made my Tweets **protected**.

   ○ During this event I changed between public and protected multiple times.

B.4 Demographics

For the following questions, consider the Twitter account used during the change between public and protected you discussed on the previous page.

7. Which of the following best describes your normal Twitter audience?

   ○ Always protected – I keep my Tweets protected all the time.

   ○ Mostly protected – I keep my Tweets protected most of the time.

   ○ Somewhat protected – On average my Tweets are more often protected.

   ○ Balanced – My Tweets are public/protected about half the time.

   ○ Somewhat public – On average my Tweets are more often public.

   ○ Mostly public – I keep my Tweets public most of the time.

   ○ Always public – I keep my Tweets public all the time.

8. When was the last time you Tweeted?

   ○ Today

   ○ This week

   ○ Over a week ago

   ○ Over a month ago

   ○ Over 3 months ago

   ○ I have never sent a Tweet before.

9. Is your Twitter account linked to your real identity?

   ○ Yes, I use my real identity in my profile.

   ○ Yes, I don't include my real identity in my profile, but it is easy to find out from my Tweet content or links.

   ○ Not sure, I avoid sharing personal details, but maybe someone could find my identity if they tried hard.

   ○ No, the content I share is unlikely to be linked to my identity.

   ○ No, I am very careful to avoid sharing content linked to my identity.

10. Imagine you post some Tweets while protected, then you change to public. Would your past Tweets be publicly visible?

    ○ Yes - Everyone could see the past Tweets.

    ○ No - Only my followers could see the past Tweets.

11. Any other comments or questions you would want to add. (optional) (Open-text Question)

Appendix C

Quantifying Privacy Switching Reasons User Survey

C.1 Overview

This is the second survey that was conducted as a part of the study detailed in Chapter 5. We recruited participants from Prolific Academic [128] and the survey was hosted by Qualtrics XM Online Survey Software [130]. I firstly showed our participants an information sheet and consent form explaining the study. Participants were filtered according to their privacy settings usage on Twitter. The survey followed with questions about their recent privacy settings changes and the motivations behind these changes. Unlike the first survey, participants were prompted to answer the questions considering all of the privacy setting changes they did in the past and the questions were multiple choice. The survey concluded with demographic information questions.

C.2 Screening Questions

1. Do you have a Twitter account?

   ○ Yes.

   ○ No. *(Screen out participants if selected.)*

2. In the last year, have you changed the ”Protect your Tweets” setting (see picture above) of your Tweets two or more times? (Figure C.1)

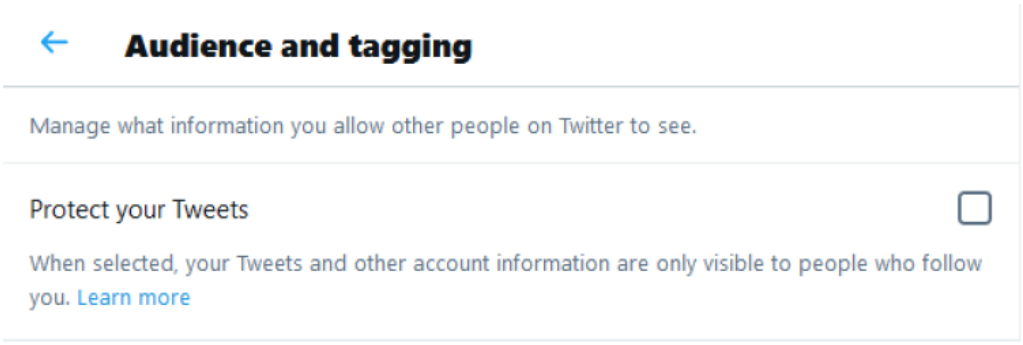

Figure C.1: Settings to change tweet visibility

- ○ Yes.
- ○ No. *(Screen out participants if selected.)*

C.3 Switching Reasons

3. Which of the following has previously lead you to you change your tweet visibility settings to **public**? Please select all the reasons that apply below:

   - ○ To reach a broader audience and get more interaction with my tweets
   - ○ To gain more followers
   - ○ To mention/reply to a user who does not follow me
   - ○ To find potential employment
   - ○ To have a professional image
   - ○ To sell things or receive donations
   - ○ To enter to get giveaways or freebies
   - ○ To retweet other users
   - ○ To quote tweet other users
   - ○ To associate a tweet with hashtags or trends publicly
   - ○ To boost the visibility, popularity, or ranking of a hashtag or topic
   - ○ To boost the visibility of another user's tweet
   - ○ To share articles or links
   - ○ To share pictures

- ○ To get customer service
- ○ To mention/reply to celebrities, famous people, or other VIPs
- ○ Other, please speficy
- ○ I did not change my tweet visibility settings to public before.

4. Which of the following has previously lead you to you change your tweet visibility settings to **protected**? Please select all the reasons that apply below:

   - ○ People I know found my account and that made me uncomfortable
   - ○ My tweet unexpectedly went viral
   - ○ I wanted to prevent non-followers from seeing tweets with personal content
   - ○ To prevent people I know, such as friends and family, from seeing my tweets
   - ○ To archive the account without deleting it
   - ○ To avoid harassment
   - ○ To tweet about someone without them being able to see the tweets
   - ○ To prevent account suspension
   - ○ I did not want people to retweet me
   - ○ I did not want people to quote tweet me
   - ○ To talk about a sensitive, controversial, or political topics freely
   - ○ To prevent interactions from strangers
   - ○ To share pictures
   - ○ To share content that is not safe for work (NSFW)
   - ○ To retweet other users
   - ○ To quote tweet other users
   - ○ To get a sense of privacy
   - ○ To share articles or links
   - ○ To take a temporary break from interactions with non-followers
   - ○ Other, please specify
   - ○ I did not change my tweet visibility settings to protected before.

5. Which of the following actions have you previously taken to control who can see and interact with your tweets?

   ○ Delete some or all tweets when moving from protected to public

   ○ Change to protected when not logged in or otherwise unable to respond to interactions

   ○ Block followers to prevent them from interacting even when your account is public

   ○ Remove a follower without blocking them (soft blocking)

   ○ Mute a follower so you can't see their interactions

   ○ Have a clear list of engagement rules prominently shown or linked to that detail what is acceptable interaction or following behaviour

   ○ Temporarily deactivate your account to prevent all interaction for a time

   ○ Other, please specify

   ○ I have not used any of the above.

C.4 Demographics

6. During the last **three months**, how many times have you changed your tweet visibility settings?

   ○ 0

   ○ 1

   ○ 2

   ○ 3-5

   ○ 6-9

   ○ 10+

7. Imagine that Alex has her tweets set to protected and posts a tweet about her new socks. She then changes her "Protect your Tweets" setting to public. After Alex changes the setting, who can see her tweet about her socks?

   ○ Anyone on the Internet can see Alex's sock tweet.

- ○ Anyone logged into Twitter can see Alex's sock tweet.
- ○ Only Alex's followers can see Alex's sock tweet.

8. Which of the following best describes your normal Twitter audience?

   - ○ Always protected – I keep my Tweets protected all the time.
   - ○ Mostly protected – I keep my Tweets protected most of the time.
   - ○ Somewhat protected – On average my Tweets are more often protected.
   - ○ Balanced – My Tweets are public/protected about half the time.
   - ○ Somewhat public – On average my Tweets are more often public.
   - ○ Mostly public – I keep my Tweets public most of the time.
   - ○ Always public – I keep my Tweets public all the time.

9. Any other comments or questions you would want to add. (optional) (Open-text Question)

Appendix D

Curation of Reasons for Switching Account Visibility

Reasons to turn **public**	Twitter Data	Free-Text Survey
To reach a broader audience and get more interaction with my tweets		+
To gain more followers		+
To mention/reply to a user who does not follow me	+	+
To find potential employment		
To have a professional image		+
To sell things or receive donations		
To enter to get giveaways or freebies		+
To retweet other users	+	+
To quote tweet other users	+	+
To associate a tweet with hashtags or trends publicly	+	+
To boost the visibility, popularity, or ranking of a hashtag or topic		+
To boost the visibility of another user's tweet		
To share articles or links	+	
To share pictures	+	+
To get customer service		
To mention/reply to celebrities, famous people, or other VIPs	+	+

Table D.1: Reasons to turn public - Options given in the survey and their sources

Reasons to turn **protected**	Twitter Data	Free-Text Survey
People I know found my account and that made me uncomfortable		+
My tweet unexpectedly went viral		+
I wanted to prevent non-followers from seeing tweets with personal content		+
To prevent people I know, such as friends and family, from seeing my tweets To archive the account without deleting it		+
To avoid harassment		+
To tweet about someone without them being able to see the tweets		+
To prevent account suspension		+
I did not want people to retweet me	+	
I did not want people to quote tweet me	+	
To talk about a sensitive, controversial, or political topics freely		+
To prevent interactions from strangers		+
To share pictures	+	+
To share content that is not safe for work (NSFW)		+
To retweet other users	+	
To quote tweet other users	+	+
To get a sense of privacy		+
To share articles or links	+	
To take a temporary break from interactions with non-followers		+

Table D.2: Reasons to turn protected - Options given in the survey and their sources

Appendix E

Information and Tweet Visibility User Survey

E.1 Overview

This survey was conducted as a part of the study detailed in Chapter 6. I recruited participants from Prolific Academic [128] and the survey was hosted by Qualtrics XM Online Survey Software [130]. I firstly showed the participants an information sheet and consent form explaining the study. I filtered participants who did not have a Twitter account. The survey has questions around individual information and tweet visibility, interaction visibility, misconceptions around Twitter functionality, and demographic information.

E.2 Individual Tweet Visibility

Please answer the following questions based on your own current understanding of how Twitter works. There is no need to look up the correct answer. The point of the research is to understand how people currently think Twitter works.

1. Imagine Emily has a **public** account. Who can see Emily's tweets?

   ○ Anyone on the Internet can see Emily's tweets.

   ○ Anyone logged into Twitter can see Emily's tweets.

   ○ Only Emily's followers can see Emily's tweets.

   ○ No one but Emily can see Emily's tweets.

2. Imagine Michael has a **protected** account. Who can see Michael's tweets?

   ○ Anyone on the Internet can see Michael's tweets.

   ○ Anyone logged into Twitter can see Michael's tweet.

   ○ Only Michael's followers can see Michael's tweets.

   ○ No one but Michael can see Michael's tweets.

3. Imagine that Alex has her tweets set to **protected** and tweets about her new socks. She then changes her tweet visibility setting to **public**. After Alex changes the setting, who can see her tweet about her socks?

   ○ Anyone on the Internet can see Alex's sock tweet.

   ○ Anyone logged into Twitter can see Alex's sock tweet.

   ○ Only Alex's followers can see Alex's sock tweet.

   ○ No one but Alex can see Alex's sock tweet.

4. Imagine that Blake has his tweets set to **public** and tweets about his potted plant. He then changes his tweet visibility setting to **protected**. After Blake changes the setting, who can see his tweet about his plant?

   ○ Anyone on the Internet can see Blake's plant tweet.

   ○ Anyone logged into Twitter can see Blake's plant tweet.

   ○ Only Blake's followers can see Blake's plant tweet.

   ○ No one but Blake can see Blake's plant tweet.

5. Imagine that Jacob has his tweets set to **public** and tweets about his completed puzzle. He then keeps his tweet visibility setting as **public**. After Jacob keeps his setting, who can see his tweet about his puzzle?

   ○ Anyone on the Internet can see Jacob's puzzle tweet.

   ○ Anyone logged into Twitter can see Jacob's puzzle tweet.

   ○ Only Jacob's followers can see Jacob's puzzle tweet.

   ○ No one but Jacob can see Jacob's puzzle tweet.

E.3 Profile Information Visibility

Twitter Topics: Topics are a way to see more of your interests on Twitter without having to follow individual accounts

Twitter Lists: From your own account, you can create a list of other Twitter accounts by topic or interest (e.g., a list of friends, coworkers, celebrities, athletes)

6. Which of the following information about both types of accounts is publicly visible? Please select all that apply for both the **public** and the **protected** accounts.

	A public account	A protected account
Profile Information (Profile photo, bio, etc.)	○	○
Number of users they follow	○	○
Number of users following the account	○	○
Users the account follows	○	○
Users who follow the account	○	○
Tweets the account liked	○	○
Tweets the account retweeted	○	○
Twitter Topics the account follows	○	○
Twitter Lists the account created	○	○
Twitter Lists the account follows	○	○
Twitter Lists the account is added on	○	○
Contents of their direct messages	○	○
Twitter password	○	○

E.4 Interaction Visibility

Please answer the following questions based on your own current understanding of how Twitter works. There is no need to look up the correct answer. The point of the research is to understand how people currently think Twitter works.

Quote Tweet: You have the option to add your own comments, photos, or a GIF before Retweeting someone's Tweet to your followers.

Mention: Mentioning other accounts in your Tweet by including the @ sign followed directly by their username is called a "mention".

Username: A username (or handle) is how you're identified on Twitter, and is always preceded immediately by the @ symbol.

Reply: A response to another person's Tweet.

E.4.1 Public to Protected

7. Alice (@alice) has a **public** Twitter account and Bob (@bob) has a **protected** account. They follow each other. In your opinion, what would happen if Alice (**Public**) retweeted one of Bob's (**Protected**) tweets using the Twitter website.

   ○ Twitter would not allow Alice to tweet it.

   ○ Twitter would let Alice tweet, but only followers of Bob could see it.

   ○ Twitter would let Alice tweet, and anyone on Twitter could see it.

8. Alice (@alice) has a **public** Twitter account and Bob (@bob) has a **protected** account. They follow each other. In your opinion, what would happen if Alice (**Public**) quote tweeted one of Bob's (**Protected**) tweets using the Twitter website.

   ○ Twitter would not allow Alice to tweet it.

   ○ Twitter would let Alice tweet, but only followers of Bob could see it.

   ○ Twitter would let Alice tweet, and anyone on Twitter could see it.

9. Alice (@alice) has a **public** Twitter account and Bob (@bob) has a **protected** account. They follow each other. In your opinion, what would happen if Alice (**Public**) replied to one of Bob's (**Protected**) tweets using the Twitter website.

   ○ Twitter would not allow Alice to tweet it.

   ○ Twitter would let Alice tweet, but only followers of Bob could see it.

   ○ Twitter would let Alice tweet, and anyone on Twitter could see it.

10. Alice (@alice) has a **public** Twitter account and Bob (@bob) has a **protected** account. They follow each other. In your opinion, what would happen if Alice (**Public**) mentioned Bob (**Protected**) using his username (@bob) in the tweet.

    ○ Twitter would not allow Alice to tweet it.

    ○ Twitter would let Alice tweet, but only followers of Bob could see it.

    ○ Twitter would let Alice tweet, and anyone on Twitter could see it.

E.4.2 Protected to Public

11. Alice (@alice) has a **public** Twitter account and Bob (@bob) has a **protected** account. They follow each other. In your opinion, what would happen if Bob (**Protected**) quote tweeted one of Alice's (**Public**) tweets using the Twitter website.

    ○ Twitter would not allow Bob to tweet it.

    ○ Twitter would let Bob tweet, but only followers of Bob could see it.

    ○ Twitter would let Bob tweet, and anyone on Twitter could see it.

12. Alice (@alice) has a **public** Twitter account and Bob (@bob) has a **protected** account. They follow each other. In your opinion, what would happen if Bob (**Protected**) replied to one of Alice's (**Public**) tweets using the Twitter website.

    ○ Twitter would not allow Bob to tweet it.

    ○ Twitter would let Bob tweet, but only followers of Bob could see it.

    ○ Twitter would let Bob tweet, and anyone on Twitter could see it.

13. Alice (@alice) has a **public** Twitter account and Bob (@bob) has a **protected** account. They follow each other. In your opinion, what would happen if Bob (**Protected**) mentioned Alice (**Public**) using her username (@alice) in the tweet.

    ○ Twitter would not allow Bob to tweet it.

    ○ Twitter would let Bob tweet, but only followers of Bob could see it.

    ○ Twitter would let Bob tweet, and anyone on Twitter could see it.

14. Alice (@alice) has a **public** Twitter account and Bob (@bob) has a **protected** account. They follow each other. In your opinion, what would happen if Bob (**Protected**). Please select Twitter would not allow Bob to tweet it. (Attention Check)

    ○ Twitter would not allow Bob to tweet it.

    ○ Twitter would let Bob tweet, but only followers of Bob could see it.

    ○ Twitter would let Bob tweet, and anyone on Twitter could see it.

E.4.3 Protected to Protected

15. Bob (@bob) has a **protected** Twitter account and Charlie (@charlie) has a **protected** account. They follow each other. In your opinion, what would happen if Charlie (**Protected**) retweeted one of Bob's (**Protected**) tweets using the Twitter website.

    ○ Twitter would not allow Charlie to tweet it.

    ○ Twitter would let Charlie tweet, but only users who follow both Charlie *and* Bob could see it.

    ○ Twitter would let Charlie tweet, but only followers of Charlie could see it.

    ○ Twitter would let them tweet, and anyone on Twitter could see it.

16. Bob (@bob) has a **protected** Twitter account and Charlie (@charlie) has a **protected** account. They follow each other. In your opinion, what would happen if Charlie (**Protected**) quote tweeted one of Bob's (**Protected**) tweets using the Twitter website.

    ○ Twitter would not allow Charlie to tweet it.

    ○ Twitter would let Charlie tweet, but only users who follow both Charlie *and* Bob could see it.

    ○ Twitter would let Charlie tweet, but only followers of Charlie could see it.

    ○ Twitter would let them tweet, and anyone on Twitter could see it.

17. Bob (@bob) has a **protected** Twitter account and Charlie (@charlie) has a **protected** account. They follow each other. In your opinion, what would happen if Charlie(**Protected**) replied to one of Bob's (**Protected**) tweets using the Twitter website.

    ○ Twitter would not allow Charlie to tweet it.

    ○ Twitter would let Charlie tweet, but only users who follow both Charlie *and* Bob could see it.

    ○ Twitter would let Charlie tweet, but only followers of Charlie could see it.

    ○ Twitter would let them tweet, and anyone on Twitter could see it.

18. Bob (@bob) has a **protected** Twitter account and Charlie (@charlie) has a **protected** account. They follow each other. In your opinion, what would happen if Charlie (**Protected**) mentioned Bob (**Protected**) using his username (@bob) in the tweet.

    ○ Twitter would not allow Charlie to tweet it.

    ○ Twitter would let Charlie tweet, but only users who follow both Charlie *and* Bob could see it.

    ○ Twitter would let Charlie tweet, but only followers of Charlie could see it.

    ○ Twitter would let them tweet, and anyone on Twitter could see it.

E.5 Misconceptions around Twitter Functionality

19. For the following statements, indicate if they are true or false.

	True	False	Not sure
Users with **protected** accounts cannot be tagged in photos.	○	○	○
Making an account **protected** will disable DMs from non-followers.	○	○	○
If a **public** user deletes one of their own tweets, the replies to it will be deleted too.	○	○	○
If a **protected** user deletes one of their own tweets, the replies to it will be deleted too.	○	○	○

E.6 Demographics

20. Which of the following best describes your normal Twitter audience?

    ○ Always protected – I keep my Tweets protected all the time.

    ○ Mostly protected – I keep my Tweets protected most of the time.

    ○ Somewhat protected – On average my Tweets are more often protected.

    ○ Balanced – My Tweets are public/protected about half the time.

- ○ Somewhat public – On average my Tweets are more often public.
- ○ Mostly public – I keep my Tweets public most of the time.
- ○ Always public – I keep my Tweets public all the time.

21. How often do you use Twitter?

- ○ Daily
- ○ Weekly
- ○ Monthly
- ○ A couple times a year
- ○ Never

22. What information do you have on your Twitter profile? Please select all that apply.

- ○ Profile photo
- ○ Header photo
- ○ Bio
- ○ Location
- ○ Website
- ○ Birth date
- ○ None of the above

23. How many followers do you have?

- ○ Less than 100
- ○ 100-499
- ○ 500-999
- ○ 1000-4999
- ○ 5000+

24. How many users do you follow?

- ○ Less than 100
- ○ 100-499

- ○ 500-999
- ○ 1000-4999
- ○ 5000+

25. When replying to a tweet I can easily tell if the poster's account is public or protected.

    - ○ Strongly disagree
    - ○ Disagree
    - ○ Neither agree nor disagree
    - ○ Agree
    - ○ Strongly agree

26. When engaging (e.g. reply, mention, retweet) with a tweet, I look to see if the poster's account is protected.

    - ○ Strongly disagree
    - ○ Disagree
    - ○ Neither agree nor disagree
    - ○ Agree
    - ○ Strongly agree

27. How often do you interact with **protected** users through the given options below?

	Never	Rarely	Occasionally	Often	Always
Mention their username in a tweet	○	○	○	○	○
Reply to their tweets	○	○	○	○	○
Like their tweets	○	○	○	○	○

28. Any other comments or questions you would want to add. (optional) (Open-text Question)