Can Social Media Privacy and Safety Features Protect Targets of Interpersonal Attacks? A Systematic Analysis

Majed Almansoori
University of Wisconsin–Madison
and United Arab Emirates University (UAEU)
Madison, WI, USA
malmansoori2@wisc.edu

Rahul Chatterjee
University of Wisconsin–Madison
Madison, WI, USA
chatterjee@cs.wisc.edu

ABSTRACT

Social media applications have benefited users in several ways, including ease of communication and quick access to information. However, they have also introduced several privacy and safety risks. These risks are particularly concerning in the context of interpersonal attacks, which are carried out by abusive friends, family members, intimate partners, co-workers, or even strangers. Evidence shows interpersonal attackers regularly exploit social media platforms to harass and spy on their targets. To help protect targets from such attacks, social media platforms have introduced several privacy and safety features. However, it is unclear how effective they are against interpersonal threats. In this work, we analyzed ten popular social media applications, identifying 100 unique privacy and safety features that provide controls across eight categories: discoverability, visibility, saving and sharing, interaction, self-censorship, content moderation, transparency, and reporting. We simulated 59 different attack actions by a persistent attacker — aimed at account discovery, information gathering, non-consensual sharing, and harassment — and found many were successful. Based on our findings, we proposed improvements to mitigate these risks.

KEYWORDS

privacy features, social media apps, online interpersonal attacks

1 INTRODUCTION

Social media applications, such as Facebook and TikTok, facilitate social interaction, content sharing, and networking among billions of users today [23]. Users create profiles, share updates, and engage with content through likes, comments, and shares on these apps. Social networks help people stay connected, promote businesses, raise awareness of social issues, and build communities with shared interests, revolutionizing communication and information access.

However, social media platforms have also become enablers of various types of online abuse and harm to individuals, groups, and even society at large [46, 57, 66]. These platforms allow attackers to interact with targets remotely and anonymously.

Such attacks are particularly harmful on the interpersonal level, where an adversary targets people they know, such as a family member, friend, or colleague [62], public figures such as content creators [55, 66], and even strangers [8, 16]. These interpersonal attacks on social media users include harassment, surveillance, bullying, and impersonation [66, 67]. While these attacks could theoretically be avoided by disconnecting from social media apps and going completely "offline" [51], doing so is often impractical. Many targets of abuse need social media platforms to seek support from others and access essential resources, such as shelter services [39].

With the increasing number of social media apps and users, it is important to understand what privacy and safety features these apps provide. Despite numerous studies on various types of online attacks, to our knowledge, no existing work has closely and comprehensively examined social media apps and their privacy and safety features. Thus, we ask:

(1) *What privacy and safety features do popular social media apps provide?*
(2) *How effective are these features in protecting users from online interpersonal attacks?*

In this work, we analyzed the *privacy and safety features* of ten popular social media apps: Facebook, Instagram, LinkedIn, Pinterest, Snapchat, Telegram, TikTok, WeChat, WhatsApp, and X (formerly Twitter). Each app has more than 300 million monthly active users [23] globally, with a collective total of more than 5 billion accounts. Through active information gathering [31, 36], we identified 100 different privacy- and safety-related features, which can be grouped into eight categories: discoverability, visibility, saving and sharing, interaction, self-censorship, content moderation, transparency, and reporting.

We observed that privacy features are not standardized, even among the most popular apps. These apps have disparate sets of privacy features, different names for features with the same effect across different apps, and even different effects for features with the same (or similar) names. For example, among the apps we analyzed, WeChat is the only app that allows users to control who can discover their accounts using usernames. Similarly, we found that only Telegram and WhatsApp allow protecting images from screenshots and allow group chat admins to moderate messages sent within their groups.

We also observed differences in the implementation of features across social media apps. For example, although most apps (except Pinterest and X) allow users to delete messages they have sent, these apps impose varying time restrictions. LinkedIn permits messages to be deleted within an hour of sending, while TikTok allows message deletion only within the first two minutes; Telegram does not impose any time restrictions on message deletion. These subtle differences in features can confuse users and add to the challenges they already face in managing their privacy settings [47–49].

Proceedings on Privacy Enhancing Technologies YYYY(X), 1–18

https://doi.org/XXXXXXX.XXXXXXX

We then assessed the effectiveness of these 100 features against a persistent UI-bound attacker [40], who attempts to spy and harass their targets.[1] To achieve their goals, the attacker needs to conduct one or more of the following *tasks*: (a) *discover the target's social media account(s)*, (b) *collect information about the target through these accounts*, (c) *share information about the target without their permission*, and (d) *make unwanted communication with the target.*

From prior research and the privacy features offered by the apps, we curated a list of 59 potential attack actions to conduct these attack tasks. We then tested the efficacy of these actions under two privacy settings: (a) when the target's account is in its default privacy configuration, and (b) when the target's account is in its most secure configuration. For each action, we considered three scenarios: (1) the attacker is a friend of the target on the app, (2) the attacker is not a friend, and (3) the attacker has been explicitly blocked by the target, possibly due to prior abusive behavior.

We show that even in the most secure configurations, many attacks remain effective. For example, the target's account can be discovered in six different ways in at least 7 apps despite the accounts being in the most secure configurations. If the attacker is a friend, the risk of privacy and safety attacks is higher. For example, under default configurations, a friend will always succeed in sharing information about the target on all apps, but these attacks will fail if the attacker is not a friend. Even under secure configurations, friends can carry out more attacks on five apps compared to non-friends. Finally, while blocking protects targets from many attacks, attackers can easily circumvent it by creating new accounts on these platforms.

Our results highlight that privacy and safety features on social media platforms lack fine-grained control and fail to protect against many types of attacks. We then outlined the necessary changes to existing privacy features and proposed solutions to mitigate these interpersonal attacks.

Contributions of this study.

(1) Through a systematic analysis of privacy and safety features across ten popular social media apps, we identified 100 unique features aimed at enhancing user privacy and safety. We categorized these features into eight groups: discoverability, visibility, saving and sharing, interaction, self-censorship, content moderation, transparency, and reporting.
(2) We found that social media apps lack some essential features within each category, leaving users unable to fully control their personal data and ensure their safety.
(3) We found that privacy and safety features often fail to protect targets from various attacks, including information collection and sustained harassment, underscoring the need for more robust protections.

2 RELATED WORK

Technology-facilitated interpersonal attacks. Interpersonal attacks have become a widespread global issue, impacting over a billion individuals across the world [52]. Recent work has shown that technology plays a significant role in interpersonal attacks. Attackers abuse technology to harass and spy on their targets [40, 41, 50, 54]. Unfortunately, thousands of apps are available for attackers, including spyware apps [30], dual-use apps [27, 30], and creepware apps [54]. Also, attackers can easily find online resources that teach them how to abuse technology to spy on others [26, 68]. Recent studies have examined Internet of Things (IoT) devices [61, 62] and found that cameras, and other IoT technologies are being exploited to spy on and harass targets.

Privacy concerns with social media apps. When using social media apps, it is essential to be mindful of the information shared. Unfortunately, many users share personal details online without considering the associated risks [38]. Careless sharing of information leaves a digital footprint that can be exploited by others, including platform owners, governments, and even other users. Social media has been shown to be used for cyberbullying [24, 29, 44, 71], harassment [65, 72], surveillance [58], spamming [73], and other forms of toxic content [66]. This raises privacy concerns for many users, prompting some to limit their engagement with social media platforms [60]. Indeed, the most effective way to avoid these privacy and safety attacks is by going offline and refraining from using these apps [51]. However, this solution is neither ideal nor feasible, as social media apps offer many benefits, such as staying in contact with family and friends. As a result, users must rely on in-app privacy settings to protect themselves. Unfortunately, many users rarely update or review their privacy settings unless prompted by a concerning incident [56].

To understand what information is accessible to people via social media, McHatton and Ghazinour [51] analyzed the information revealed on Facebook, Instagram, and Twitter under three different configurations, including the default settings applied during account creation. They found that advanced configurations provide adequate security and placed the blame on users for not adjusting their settings. However, users should not be blamed, as there are many challenges involved in managing privacy on social media apps. In fact, research has shown that privacy settings often do not align with users' expectations [47]. Additionally, many users may not be aware of privacy settings, as was the case for a significant minority of Facebook users [25].

This work. We build on the work of McHatton and Ghazinour [51] by exploring ten popular social media apps globally, including Facebook, Instagram, and X. First, we analyze the privacy and safety features available on these apps, organizing them into eight categories. Then, we conduct experiments involving a set of attacks under six different threat scenarios. Our attacks go beyond information collection to include other threats, such as harassment. Our goal is to assess how well social media platforms protect users from interpersonal attacks — rather than blaming users for encountering such risks — which has yet to be fully explored.

3 PRIVACY AND SAFETY THREAT MODEL ON SOCIAL MEDIA

We aim to understand the attacker's capability to surveil or harass targets using social media apps. To do this, we first identify the threat scenarios for social media-based attacks. Then, based on our

[1] We use "target" to denote peron who is receiving the abuse, instead of "victim" or "survivor", following the prior work [66].

survey of prior work (Section 2) and the privacy features we collect in Section 4, we develop a set of potential attack tasks that could be used to compromise the privacy and safety of social media users.

3.1 Threat model and attack scenarios

We consider the threat of an interpersonal adversary [8, 16, 52], who may be someone the target knows (such as an abusive co-worker, friend, or family member) or a stranger. This adversary is willing to invest time and resources to cause harm. In our scenario, the adversary seeks to surveil or harass their target through the social media applications the target uses.

Interpersonal adversaries are persistent threats to their targets and can exploit any means provided by a social media application. They may also use other apps available online that have capabilities to surveil or harass others over social media [27]. The adversary can create multiple accounts on social media platforms and may use multiple devices to carry out their attacks. In our model, the attacker is restricted to the user interface (UI-bound adversary [40]) of social media apps and publicly available tools. Therefore, we do not consider attacks that exploit potential vulnerabilities in social media applications. Additionally, the adversary does not have physical access to the target's devices or know the target's password for any account.[2] However, the adversary may know some of the target's personal information, such as their phone number or username on a different social media app.

We excluded from our threat model attacks carried out by a group of people against one or multiple targets (e.g., group harassment and abuse) [66]. We also excluded attackers who target individuals based on their identity, such as race, religion, age, or gender. While these threats intersect with our model, they require a different analytical approach. Instead, we focus on individual and interpersonal attacks. However, we believe many aspects of our analysis may still be relevant in other contexts, such as group attackers, identity-based attacks, and group-of-victims scenarios, which we leave for future studies to explore in more depth.

In the context of interpersonal attacks, the adversary may or may not be connected to the target on the app, or may have been blocked by the target due to prior abusive behavior. Thus, we consider three possible statuses of the attacker in relation to the target: non-friend, friend, and blocked. Also, targets can have their social media privacy and safety settings configured in various states. We consider two states for the target's account configurations: "default" (no changes made since account creation) and "secure" (where all settings are configured to maximize privacy).

Based on these criteria, we created six different attack scenarios: the attacker is a *friend* on social media with the account set to *default* settings (fr-default), a friend with secure settings (fr-secure), a non-friend with default settings (nf-default), a non-friend with secure settings (nf-secure), a blocked user with default settings (bl-default), and a blocked user with secure settings (bl-secure).

[2]In many interpersonal attacks [40], the adversary may know the target's password and have full access to their social media accounts. However, in this work, we focus on cases where the attacker does not have full access to the target's accounts.

3.2 Attack goals, tasks, and actions

To understand the potential privacy and safety risks for social media users, we considered abuse on online platforms systematized by Thomas et al. [66]. We focused on five *attack goals*: *surveillance*, *toxic content*, *impersonations*, *content leakage*, and *overloading*, as these are relevant in interpersonal threat model we considered. We excluded categories irrelevant to our study, such as *lockout and control*, since the attacker in our model does not tamper with the target's account. We also excluded *false reporting*, as testing such attacks could impact the services provided by these platforms and might involve human analysts. Additionally, under the categories we considered, we excluded attacks that do not fit our threat model, such as dogpiling, which involves multiple attackers, and distributed denial of service (DDoS), which requires a more sophisticated attacker.

We then devise four primary *attack tasks* that are required to achieve the attack goals for an interpersonal attacker:

(1) *Discovering the target's account*, where the attacker wants to find the target's account on a social media app using various identifiers of the target that the attacker knows, such as their name, phone number, email address, or even through a mutual friend. Discovering the account is often the first step towards more severe abuse. This attack goal is irrelevant if the attacker is already connected with the target on that app, for example in fr-default and fr-secure settings.

(2) *Collecting and monitoring information* about the target, such as their connections, profile photo, shared content, and location.

(3) *Sharing information about the target without their consent*, such as their posts and chats, both within and outside the app. Although the target may have shared some information intentionally, they may not have consented to it being shared beyond their connections.

(4) *Making unwanted communication* with the target, for example, by adding them to unwanted groups, sending abusive messages, and impersonating them by creating fake accounts.

We created a series of *attack actions* — sets of steps through which an attacker could complete a given task — for each attack task and tested them against our experimental accounts on each social media app. We created 59 such attack actions across these four attack goals shown on Fig. 6 and Fig. 9 (Appendix B).

4 IDENTIFYING PRIVACY AND SAFETY FEATURES ON SOCIAL MEDIA

We aim to evaluate the protection offered by privacy features in social media apps against various interpersonal threat models. An interpersonal attacker is a knowledgeable, persistent, and resourceful adversary seeking to spy on or harass the target user. To achieve this, we thoroughly examine the user interfaces of popular social media apps to identify and document the privacy features they provide. We then simulate several interpersonal attack scenarios to assess the effectiveness of these features in protecting users.

Selecting social media applications. There are more than 5.04 billion social media users globally [23]. In this study, we focused on ten social media apps based on their global user base and availabil-

Name	Icons	# Users (in millions)
Facebook		3,049
WhatsApp		2,000
Instagram		2,000
TikTok		1,562
WeChat		1,336
Telegram		800
Snapchat		750
X (Twitter)		619
Pinterest		482
LinkedIn		310

Figure 1: Social media applications we consider in this work. The numbers of monthly active users in millions [37, 63] globally is shown in the last column. The data recorded on Feb 29, 2024. The number of users for TikTok does not include the users from Douyin (its Chinese equivalent).

ity in English. We began with the top 15 social media apps with the highest monthly active users, as listed by Statista [37]. From this list, we excluded apps not available in English, namely Kuaishou, Sina Weibo, QQ, and Douyin (the Chinese version of TikTok). We excluded YouTube, as its primary purpose is video hosting, with user interaction limited to the comments section. Instead, we included LinkedIn, which has over 310 million monthly active users worldwide [63]. Additionally, we treated Facebook Messenger and the main Facebook app as a single application for our analysis, while considering WhatsApp and Instagram as separate apps. Although Meta owns all four applications, WhatsApp and Instagram have distinct privacy and safety settings compared to Facebook and Messenger. Our final list of the ten social media apps we analyzed is shown in Fig. 1, along with the number of monthly active users (MAU) in millions for each app globally.

Social media terminologies. There are several terms used in the context of social media that some readers may not be familiar with. Therefore, we provide brief descriptions of these terms in Appendix A and emphasize each *term* upon its first use.

4.1 Aggregating privacy and safety controls

We define a control or feature (used interchangeably in this work) as relevant to privacy or safety if it fulfills one or more of the following criteria: (a) manages information the user directly provides to the platform, such as hiding their profile photo or managing who can view their *stories*; (b) manages information about the user's actions on the platform, such as showing when the user is online or when a message has been read; (c) controls how others can interact with the user, such as limiting who can reply to their posts or send them messages; (d) provides support against abuse, such as reporting a user. We use this definition to classify all controls and features.

We excluded features specifically designed to enhance account security. For example, features like two-factor authentication (2FA), password changes, and app screen locks were excluded because our threat model does not involve attackers attempting to compromise the account or gain physical or remote access to the target's device. Our threat model is described in Section 3.1.

To gather information about available privacy controls and features, we relied on active information gathering [31, 36], a process that involves collecting information about the target system through direct interaction with it. We collected controls provided by the selected social media apps by navigating through the user interface and identifying all user-level controls relevant to privacy and safety. This process involved testing all features — identified using the heuristics mentioned later in the section — regardless of their relevance, documenting the effects of these features, and then eliminating controls irrelevant to privacy and safety. We also reviewed the frequently asked questions (FAQs) of these apps (where available) to identify default privacy features, such as Instagram's notification to users when a recipient takes a screenshot of a chat in vanish mode [15]. Thoroughly examining the user interface, testing all available features — regardless of their relevance to the study — and reviewing FAQs ensures that we capture all relevant features of these social media apps in our analysis.

Identifying privacy- and safety-related controls in social media apps is quite challenging as they are located in different UI paths in different apps, and even within an app. We used a set of heuristics to identify these controls.

(1) *Settings*: We examined the settings of each app, focusing primarily on the privacy tab but also reviewing other tabs for relevant features. The majority of features are found in the settings, including controls over what information about the user is shared and how others can interact with the user.

(2) *Options*: Several privacy and safety features appear as "options" across various sections of the platform, including messages, chats, posts, comments, groups, and profiles. However, there is little standardization in how these options are accessed; some can be found by swiping left or right and tapping, while others are revealed by tapping and holding. For example, tapping and holding a message on WhatsApp displays its options, while tapping the three dots on X posts brings up a list of actions. Additional features are shown when users perform certain actions, such as adding a friend, following an account, posting an image, or liking and reposting content. One example is the option to control the visibility of a post on Facebook, which is available while writing the post.

(3) *Action-triggered features*: Some privacy and safety features are only triggered when specific actions are taken within the app, without an explicit privacy control for them. For example, if someone takes a screenshot of a chat, the other user is notified. We refer to these as *action-triggered privacy features* and include them in our analysis.

We tested eight actions for each social media app and recorded whether any privacy features were triggered: (1) posting or sending a message, (2) editing or deleting a previously posted or sent message, (3) typing a message or post, (4) downloading media content from a post or message, (5) opening and viewing content, (6) taking a screenshot or recording the screen while the app is open, (7) changing certain options, such as adjusting the chat auto-deletion time in Snapchat, which triggers a notification, and (8) previewing the app through the app switcher screen, an action we observed during app testing. We repeated this step each time we added a new trigger condition to our list. An example of an action-triggered privacy feature is sending a notification to users when a screenshot is taken of a chat.

Category	Description	# features
Discoverability	Controls how someone can find the user's account	6
Visibility	Controls who views the user's account information such as posts, comments, and personal information.	23
Saving & Sharing	Limits what information about the user people can share	7
Interaction	Limits who can interacts with the user by commenting on their posts, send messages, and follow the user account.	17
Self-censorship	Controls user's actions, discourse of information and presence, which includes deleting posts, editing comments, and exiting groups.	19
Content moderation	Allows the user to remove and hide content and interactions by others, such as muting posts and deleting comments.	5
Transparency	Logs of interaction done by others, such as who recorded the user's story and who visited their profile.	18
Reporting	Allows user to report other users.	5

Figure 2: We report the categories of privacy features, their descriptions, the number of features under each category.

(4) *Paid features*: LinkedIn [20], Snapchat [17], Telegram [19], and X [2] offer additional privacy and safety features for paid subscribers. We subscribed to these services, repeated the previous steps, and documented any new features or controls relevant to user privacy and safety provided through these subscriptions. For instance, Telegram Premium [19] allows subscribers to hide that they have viewed others' stories.

4.2 Experiment Setup

We used two iPhone devices (running iOS 16 and iOS 17, respectively) and a MacBook laptop (running macOS Sonoma 14.2) to interact with the social media applications. We evaluated the latest versions of the iPhone apps available on the Apple App Store as of February 20, 2024. To ensure accuracy, we cross-checked our findings using the web versions of the apps on the MacBook with the Safari (v17.2) browser. User accounts were created on each platform specifically for this research study, and no interaction with real users took place. Throughout the experiment, our attacks (see Section 5) were demonstrated on these lab accounts. All accounts were permanently deleted after the completion of the study.

We conducted the survey of privacy- and safety features over the course of more than a month of active interaction with these platforms and using the heuristics mentioned before, allowing us to explore the different UI paths available for features. Through this process, we collected 187 unique features from the ten social media apps. We then consolidated features with similar effects (e.g., disappearing messages, images, and videos were grouped as disappearing messages), resulting in 100 features for further analysis. Full list of features and their groups can be found in Appendix B.

4.3 Categories of Privacy and Safety Features

After collecting 100 unique features, we applied thematic analysis to categorize them based on the types of privacy and safety threats they protect against. One researcher reviewed all the features and assigned them to categories. Then, using an iterative approach, the researcher discussed these categories with the other researcher during regular meetings as part of a "peer debriefing" process, refining the categories as needed. Peer debriefing [32, 33] is a method used to ensure the credibility and trustworthiness of results by providing an external perspective and revealing potential biases. In peer debriefing, the researcher discusses findings with peers who are not directly involved in the analysis. We ended up with a total of eight categories of privacy features: *Discoverability*, *Visibility*, *Saving and Sharing*, *Interaction*, *Self-censorship*, *Content moderation*, and *Reporting*. We also conducted another round of peer debriefing, discussing these categories with researchers outside the team. We present all categories, their definitions, and the number of features in Fig. 2 (refer to Appendix B for the full list of features). In Section 5, we describe these features in detail and discuss their limitations.

5 ANALYSIS OF PRIVACY AND SAFETY FEATURES

We analyzed the privacy and safety features provided by the social media apps. In this section, we present our findings, discussing the goals and limitations of these features. We then evaluate the efficacy of these features against interpersonal attacks in Section 6.

5.1 Discoverability

Discoverability allows users to find acquaintances on social media, but it can also serve as a means for initiating abuse by finding the target's account. We identified different methods through which an account can be discovered, and users are sometimes provided with controls to limit who can find their accounts using these methods.

Discoverability by username. Most apps (except WhatsApp) use usernames as identifiers and as a way to find users on the platform. Instead of sharing phone numbers or emails, users can share their usernames to expand their *connections*. Users have limited control over discoverability through usernames. Only WeChat offers an option for users to limit discoverability via their usernames. While Telegram does not provide such a feature, the app does not require having a username, so discoverability by username can be prevented by simply deleting it. However, Telegram does not explicitly state this, and we discovered it through black-box testing. We assume many users are unaware of this hidden feature. Thus, with the knowledge of the username — possibly obtained from other apps — an attacker can find user's account in most social media apps.

Discoverability by email and phone number if contact syncing is enabled. None of the apps allow direct discoverability via email, though many offer a privacy feature to limit discoverability through email. Regardless of whether this feature is enabled, users cannot be found by searching for their email addresses; instead, this feature controls *contact syncing*. Contact syncing allows users to upload their contacts' emails and phone numbers, which the platform uses to match people who may know each other. If enabled, the platform will automatically suggest connections based on synced emails. Similarly, phone numbers cannot be used for direct account searches, but the platform will suggest connections if contact syncing is enabled. After syncing, the apps will try to match

users with mutual contacts. The only exceptions are (a) WhatsApp, where users can always be found if their phone number is known, and (b) Telegram and WeChat, where users can choose whether to allow discoverability via phone numbers.

Discoverability via QR codes. We found that nine apps (except X) provide QR codes for sharing account information. Among these, only WeChat allows users to disable discoverability via QR code, and only WeChat and WhatsApp support resetting the code. Also, QR codes are typically not located under the settings tab, making it difficult for users to easily find and manage them. If a QR code is publicly shared, it can lead to spamming through friend and message requests, even if the account is non-discoverable through other means, such as a username. While we are not aware of any QR-based attacks on social media, they are feasible and could be exploited by attackers. Allowing users to reset QR codes and restrict who can add them via QR code is essential to prevent harassment.

Discoverability via suggestions. Aside from WeChat and WhatsApp, all other social media apps rely on suggestions to help users expand their connections. We found that users have limited control over these suggestions, raising privacy concerns for targets of interpersonal attacks. As previously mentioned, many social media apps use contact syncing to match people who may know each other. However, contact syncing is not the only method for generating suggestions; many apps rely on factors such as name, location, and personal information like job titles on LinkedIn or schools on Facebook to suggest contacts.

We found that only Snapchat and Telegram provide users with clear options to opt out of suggestions. However, even in these apps, the option to disable discoverability via suggestions is not available during account creation, requiring users to navigate the settings to turn it off afterward. In contrast, users cannot fully hide from suggestions on Facebook, Instagram, LinkedIn, Pinterest, TikTok, and X. During our experiments, we found that these apps can identify people known to the user even if the user does not link their phone number or uses a fake email (implications discussed in Section 7). The lack of transparency and control over these suggestion algorithms can lead to confusion and undermine users' agency, while also allowing attackers to find their targets' accounts.

5.2 Visibility

Visibility features allow users to control who can view information related to their accounts, such as posts, comments, and profile details. These features help users manage and limit what others can "see" about them. Visibility is the second layer of privacy to address after discoverability, as it controls the information displayed to users who have already found the account. We identified three distinct visibility features for accounts.

Visibility of account: public vs private. Most social media platforms, excluding instant messaging apps like WeChat, WhatsApp, and Telegram, allow users to switch between public and private accounts. In a public account, profile information and public posts are visible to anyone, whereas in a private account, the account is only visible to approved followers. When creating a new account, Facebook, Instagram, LinkedIn, Pinterest, TikTok, and X automatically set it to public by default, requiring users to manually change the settings to make their account private. Users may not realize that their accounts are public by default, making much of their information visible to others, which gives attackers a greater opportunity to collect data about their target. Facebook does not offer an option to make an account completely private; however, users can adjust various visibility settings, such as hiding their name and current city, to effectively reduce account visibility. While this provides fine-grained control, it can be challenging to use, as users must navigate multiple settings to fully make their accounts private. In contrast, all new Snapchat accounts are private by default, and users can choose to make them public after registration, prioritizing user privacy and offering better protection against interpersonal attackers. However, Snapchat allows users to share their location indefinitely with all friends or some of them on the platform. Prior work [45, 68, 69] has identified several instances of Snapchat location sharing being abused in the context of interpersonal attacks.

Visibility of stories and posts. Users can share text, images, and videos on social media platforms, commonly referred to as *posts*. In 2013, Snapchat introduced the concept of *"stories"* (also called *"status"* by WeChat and WhatsApp), which allows users to upload temporary posts to their profiles. Since then, all but LinkedIn, Pinterest, and X have adopted the story feature. In fact, both LinkedIn and X (formerly Twitter) previously offered stories, but these features have been discontinued [53, 70].

We found that controlling the visibility of stories is significantly different from controlling the visibility of regular posts. All apps offering stories, except TikTok, allow users to exclude specific individuals from viewing their stories. TikTok users can only choose between sharing with everyone, friends, or only themselves. In contrast, apps provide less control over the visibility of posts. Among the seven apps with posts, Facebook, LinkedIn, TikTok, WeChat, and X allow users to limit who can view a specific post. The only exception is WeChat, which allows users to exclude specific individuals from seeing posts. For Instagram and Pinterest, the only way to control post visibility is by switching to a private account, which only hides posts from non-followers. The lack of fine-grained control over post visibility can be particularly challenging for targets of interpersonal attacks, especially when the abuser is a "friend" on the platform and cannot be removed for fear of escalation — a common scenario in cases of intimate partner violence (IPV) [41]. In such situations, the target cannot exclude the attacker from viewing their posts without also impacting other followers.

Visibility of connections. Users can control who can see their *connections* (also known as contact list, friend list and follower list) on social media apps. We found that Snapchat, Telegram, WeChat, and WhatsApp do not display friends publicly. Among the remaining six apps, only Facebook, LinkedIn, and TikTok allow users to hide their connections from non-friends. For Instagram, Pinterest, and X, the only way to hide connections is by switching to a private account, which only hides the list from non-followers. Thus, even if the attacker is not on the target's friend list, they may still be able to monitor the target's connections.

5.3 Saving and Sharing

Nearly all social media apps allow users to share content from other users, either on their own profiles, through private messages,

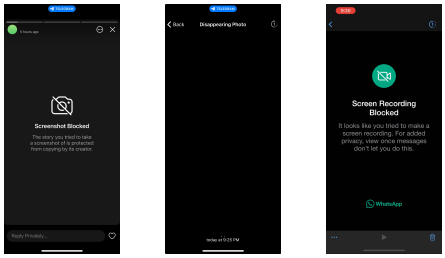
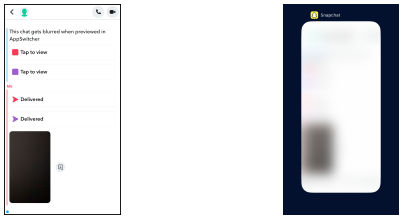


Figure 3: Screenshots of features implemented by Telegram and WhatsApp to prevent screen capture. Telegram prevents taking screenshots and recording stories, secret chats, and disappearing media, while WhatsApp only prevents taking screenshots and recording disappearing media.



(a) Actual chat **(b) Screen in App Switcher**

Figure 4: This figure shows the automatic blurring feature found in Snapchat and Telegram (Section Section 5.2). Screenshot (a) shows the chat within Snapchat itself, and (b) shows the same chat blurred when previewed in iOS's App Switcher.

or outside the platform by downloading or screen capturing the content. While this is generally desirable, it can be harmful in interpersonal attack scenarios.

Sharing posts and stories. Facebook and TikTok users can restrict others from sharing their posts to stories, while Instagram users can prevent others from sharing their stories through messages. Apps like WeChat and WhatsApp do not allow story sharing at all. On Telegram, users can share others' stories in messages, repost them, and share them outside the app via a link. Story sharing can be prevented by disabling screenshots within Telegram, though the option does not explicitly state that disabling screenshots also prevents story sharing.

We observed message forwarding is not allowed in Telegram's *secret chats* [12], which are encrypted chats designed to preserve privacy, and in Instagram's *vanish mode* [15], a mode where messages are automatically deleted after leaving the chat. Facebook also offers an encrypted chat called *secret conversation* [18], which restricts forwarding if disappearing messages are enabled (this feature is disabled by default). LinkedIn, Snapchat, TikTok, WeChat, and WhatsApp do not prevent message forwarding, while Pinterest and X do not allow forwarding by default.

Providing users with full control over what can be shared and with whom is a crucial aspect of privacy. Currently, platforms allow attackers to share information from targets' accounts, including private messages, posts, and stories, without fine-grained restrictions or meaningful notifications for the targets.

Sharing and saving media. Users can generally download images and videos, but we found that some apps limit this feature, primarily for videos. Facebook, LinkedIn, and Pinterest prevent users from downloading videos shared on timelines, while Instagram, TikTok, and X allow users to disable video downloads. All apps allow video downloads in chats, except for LinkedIn and Pinterest. We found that Instagram disables downloading images and videos shared in vanish mode, while users can download media in Facebook's secret conversations and Telegram's secret chats. Telegram users can further protect their images and videos by sending them as *self-destruct* messages, which are limited by time or number of views. Similarly, Facebook, WhatsApp, and Snapchat protect disappearing images and videos from being downloaded. However, Snapchat users can choose to allow others to save their *Snaps* (the term for disappearing images and videos in Snapchat), although this feature is disabled by default.

While many users may not see this as a risk, it is essential to offer the option to protect all media from being saved, not just disappearing media. Enforcing this feature by default is crucial for ensuring better privacy, especially for users who may be unaware of potential attackers. Additionally, targets may wish to control older media shared with an attacker; however, most apps currently do not provide a way to protect older media from being downloaded.

Preventing screen capture. We found that Telegram and WhatsApp have features that prevent people from capturing information about others. Telegram prevents screen capture for secret chats, protected stories, and disappearing messages (self-destruct). When a user attempts to screenshot a secret chat or disappearing image, the app hides all messages and displays a blank chat or background. For stories, the app shows a message indicating that the user has been blocked from capturing. For disappearing videos, Telegram freezes the video when someone tries to record the screen, preventing it from being captured. WhatsApp, on the other hand, only prevents screen capture for disappearing images and videos and does not offer protection for chats or stories. Illustrative screenshots are shown in Fig. 3. Facebook, Instagram, and Snapchat allow users to capture disappearing messages, which defeats the purpose of these features. Users have no way to prevent others from recording and sharing the information.

App screens can also be captured in app switcher, which is the screen that allows users to view all the currently running apps and switch between them. Snapchat and Telegram prevents screen capture even in app switcher by automatically blurring the screen, but they implement it differently (illustrated in Fig. 4). Snapchat blurs all chats, images, videos, and stories (excluding profiles and Snap Map), and this cannot be disabled by users. Telegram, however, only blurs content in secret chats, where screenshots are also disabled, to protect the information.

This blurring effect is not automatically enforced by operating systems; it is implemented on the desired pages within the app. Such features deter someone from secretly capturing chats or media without triggering any notification (discussed in Section 5.6).

5.4 Self-censorship

Users, especially those experiencing online abuse, may want greater control over what they share to protect their privacy and safety. *Self-censorship* refers to the intentional act of controlling one's actions and the disclosure of information to others. This can also be

exercised by permanently or temporarily deleting accounts [22, 64]. Targets of interpersonal attacks often engage in self-censorship both in general [43] and on social media [34], frequently driven by fear [13, 14, 35, 40, 59]. We discuss three different types of controls for users to self-censor on social media.

Hiding profiles (instead of deleting). All apps, except Telegram and WhatsApp, allow users to temporarily deactivate their accounts and reactivate them later. Users can hide their Snapchat and X accounts for up to 30 days, after which they are deleted if not reactivated. The other six apps allow users to reactivate their accounts at any time. Telegram and WhatsApp only support permanent account deletion. However, Telegram also allows users to set an inactivity period (e.g., 1 year) for automatic account deletion. When experiencing digital abuse [27, 30, 41], users may prefer to temporarily hide their profiles rather than permanently delete their accounts and create a new one. Thus, features that allow users to hide or temporarily delete their accounts are important, as they support users' *"right to delete"* their personal information [4].

Changing personal information. While all apps allow users to change their usernames and full names, some impose restrictions. Facebook allows users to change their full names once every 60 days. Instagram usernames can be changed twice every 14 days, while Snapchat and WeChat usernames can only be changed once per year. Restricting users from changing this information could harm targets by preventing them from hiding their social media presence. Allowing these changes serves as a preventative measure against attackers who already know the target's account. By modifying their names and usernames, targets can make it more difficult for attackers to find their accounts.

Deleting and hiding previous posts, stories, and messages. Apps offer various mechanisms for managing previously published posts (and comments), stories, and private messages. All apps allow users to edit their posts without restriction, except for X, which only allows *premium* users to edit posts within 30 minutes of publishing. In contrast, apps are less flexible with stories; only Telegram and WeChat allow users to edit their stories after posting.

We observed a significant discrepancy in message editing mechanisms across apps. Nearly half of the social media platforms (Pinterest, Snapchat, TikTok, and X) do not allow users to edit sent messages. Some apps permit message editing within a specified time frame: Facebook, Instagram, and WhatsApp allow edits within 15 minutes of sending, while WeChat allows edits only within the first two minutes. The number of times a message can be edited also varies: Facebook and Instagram limit users to 5 edits, while Telegram, WeChat, and WhatsApp allow unlimited edits.

None of the apps impose restrictions on deleting posts and stories; users can delete them at any time. However, some apps limit the deletion of messages. WhatsApp messages can be deleted from the recipient's chat within two days of sending, while LinkedIn allows deletion within 60 minutes (the same limit applies to message editing). Similarly, TikTok messages can only be deleted within the first 3 minutes, and WeChat messages within the first 2 minutes. Interestingly, Telegram offers a unique feature: users can delete an entire chat from the recipient's side, not just individual messages, with no time limit on this feature.

5.5 Interaction

Social media applications offer various forms of interaction between users, such as commenting and following. For our purposes, we define interaction privacy features as those that allow users to control who can communicate with them and restrict specific actions such as commenting and following. In this section, we describe the most notable interaction privacy and safety features we found on the ten platforms.

Controlling who can connect. Although connecting with others is one of the primary purposes of social media applications, in certain situations — such as for targets of interpersonal abuse — users may want to limit these interactions to protect themselves from abuse. Social media apps offer various controls for managing connections. Some apps limit who can send friend requests; for instance, Facebook and LinkedIn allow users to restrict requests to people with a direct connection. Both apps also provide options to limit who can follow a user. WeChat users can choose to hold friend requests and manually approve them, rather than having requests automatically approved. Also, users can control how others are able to add them. For example, they can prevent others from adding them via a *"contact card"*, which refers to the user's contact information when shared by friends. Moreover, WeChat users can block people from adding them to shared groups or through QR codes, a privacy feature not available in any other app.

We also found that some privacy features allow users to prevent others from adding them to groups. This was observed in Instagram, LinkedIn, Telegram, WeChat, and WhatsApp. In contrast, other apps allow anyone, including non-friends, to add users to groups without restrictions. The only exception is TikTok, which only allows friends to add each other to groups, though it does not offer a privacy feature to restrict friends from doing so.

Blocking and restricting users. The most common form of restriction is blocking, which prevents blocked users from sending messages, friend requests, and from viewing, liking, or commenting on posts. All apps have implemented some form of user blocking. Some apps explicitly inform users when they have been blocked. For example, X displays "You're blocked" when a blocked user visits the blocker's profile. In contrast, Pinterest, Telegram, and WhatsApp do not explicitly notify users; instead, they hide the blocker's profile picture and bio. In addition, Telegram displays "last seen a long time ago" under the blocker's name. Other apps — Facebook, Instagram, LinkedIn, Snapchat, and TikTok — hide the blocker from the blocked user, making it appear as if the account was deleted. The blocker also disappears from the contact list, followers, and following lists. WeChat, however, does not hide the blocker or their information. Instead, if the blocked person tries to send a message, they receive a notification stating that the message was "successfully sent but rejected by the receiver".

Facebook and Instagram allow *restricting users* [3, 7, 11]. On Facebook, users can be added to a *restricted list*, which limits them to viewing only public posts [3]. Additionally, users can be restricted on Facebook Messenger, muting all calls and messages from the restricted person. On Instagram, restricting users includes several features, such as hiding their comments so that only the restricted person can see them [7]. These restrictions can be more beneficial than blocking for targets who are vulnerable to an escalation of

abuse, especially in cases of intimate partner violence (IPV) which involves physical abuse [28].

Limiting interactions. Social media apps offer a variety of ways for users to control how others can interact with them. Users can limit who can call, message, send voice and video messages, mention, tag, and comment on their posts. However, these features are not uniformly available across the apps we tested. The most common feature is the ability to limit who can comment on posts, which is available in all apps except WeChat. This feature typically comes in two forms: (a) turning off comments on specific posts entirely, or (b) preventing comments from certain groups of users, such as non-followers. The first option is the most common, implemented by five of the seven apps that allow restrictions on comments.

Instagram allows users to block comments from specific individuals, which is particularly helpful in abusive situations, as it prevents comments from a particular person without restricting others. Snapchat, on the other hand, offers a robust way to manage comments on *Spotlights* [21] (a feature similar to *Reels* on Facebook [10] and Instagram [9], and TikTok videos). Users can choose to automatically approve comments posted on a Spotlight, manually review all comments, or selectively review some of them.

Social media interactions can be exploited by attackers to spam and harass targets. Therefore, having control over these interactions is essential for preventing and combating interpersonal attacks.

5.6 Transparency

We refer to features that focus on the visibility of actions conducted by other users as "transparency". These features provide a log of interactions performed by others on the app that relate to the target user's account. While transparency features often do not have explicit controls, we analyze them as an important privacy aspect of social media apps. We identified three main types of transparency features offered by these platforms.

Record of who viewed stories, posts, and profiles. Many social media platforms inform users about who has viewed their content, such as stories, posts, and profiles. All social media apps, except WeChat, notify users about who viewed their stories, regardless of whether the viewer is on their friend list. Users generally cannot hide the fact that they have viewed someone else's story, except for Telegram users with premium accounts.

Profile-visit transparency is less common among apps, with only LinkedIn and TikTok offering this feature. LinkedIn always logs profile visits; however, users can choose to hide their information, including their name and profile, when viewing someone else's profile. In contrast, TikTok allows users to visit profiles secretly, but when they hide their presence, they lose the ability to see who visited their own profile. Similarly, post views are logged on TikTok and can be hidden from others, but this also prevents the user from seeing who viewed their posts.

These features are useful for targets being stalked or monitored on social media paltforms, as they alert targets and allow them to collect evidence of abusive behavior.

Record of actions. Transparency features extend beyond views to include actions such as reading chats, typing, editing posts, and deleting messages. Some of these actions are logged temporarily; for example, an indicator appears when a contact is typing and disappears once they stop. Other actions are permanently logged. For instance, editing messages leaves a flag in all apps, indicating that a message was edited. Similarly, when a message is deleted, some apps notify the receiver that the message was removed.

Attackers could tamper with chats by editing or deleting their own messages to manipulate the conversation against targets. Similarly, they could tamper with their comments on the target's posts. Permanent logs help ensure that any modifications are traceable, protecting targets from such tampering.

Notification of capturing the screen or saving media content. In Section 5.3, we discussed how Telegram and WhatsApp detect when a user tries to capture the screen by taking a screenshot or recording it, and how they prevent these actions. Other apps use similar detection mechanisms but, instead of preventing the capture, they inform the user that their screen has been captured. Snapchat is well-known for this feature, notifying users when their story, chat, or profile is captured. Snapchat also detects when someone downloads media shared by others and sends a notification to the user. In addition to Snapchat, screen recordings and screenshots are detected and flagged by Instagram's vanish mode and Telegram's secret chat. Facebook Messenger has a similar feature implemented in secret conversations, which is triggered only when disappearing messages are enabled. With such notifications, attackers may be deterred from capturing screens or saving media, especially when combined with the blurring effect discussed in Section 5.3.

5.7 Content moderation

Content moderation is a well-known process that allows users to remove or hide content shared or written by others [5]. To promote healthy interactions among users, social media platforms must provide a range of controls for moderating interactions through posts and comments. Surprisingly, we found that social media apps offer only a limited number of content moderation features.

Hiding and filtering unwanted content. Apps provide two main ways to hide content: (a) explicitly hiding and muting users, and (b) muting specific words chosen by the user. The first method primarily involves disabling notifications from the muted user and hiding their content from the timeline or messages on the main page. In contrast, the second method, available in only five apps, allows users to mute specific words, which hides posts, comments, or messages containing any of the muted words. Additionally, LinkedIn offers the option to hide profile photos of other members. Users can choose to hide all profile pictures, show pictures of connections and people within their network, or display all pictures. These features can help targets prevent harassing comments from attackers.

Removing comments and group messages. Social media apps allow users to delete comments from their posts and messages from group chats. These features are essential for protecting against spammers and harassers. We found that all apps allow users to delete comments, except X, which only allows users to hide comments. While deleting comments is widely supported, we observed less support for group message moderation. Although all apps allow the creation of group chats, only Telegram and WhatsApp give admins control over messages sent in their group chats. Facebook

Report category										X
Harassment	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓
Threats	✗	✗	✓	✓	✓	✓	✗	✗	✗	✓
Spam	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓
Fake account	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓
Data leakage	✗	✗	✗	✓	✓	✓	✓	✗	✗	✓
Impersonation	✓	✓	✓	✓	✓	✗	✓	✗	✗	✓
Detailed report	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗

✓ Report category is included ✗ Report category not included.

Figure 5: The categories in which content can be reported for each app, and whether detailed reports can be submitted.

offers two types of group chats: typical group chats and community chats (linked to Facebook groups). In typical group chats, the admin has no control over the content, whereas in community chats, admins have more control over the shared content. In some cases, the attacker might be in a shared group with the target, and if the attacker has admin permissions, they could harass the target by removing them from the group or deleting their messages.

5.8 Reporting

We identified five reporting features in social media apps: reporting an account, reporting an entire chat, reporting a single message, reporting posts, and reporting stories. All apps offer at least three reporting features, except WeChat, which limits reporting to accounts and chats, and Telegram, which only allows reporting stories (shown in Fig. 8). Only Facebook, Instagram, TikTok, WhatsApp, and X provide reporting options for all applicable cases.

To better understand the reporting features provided by social media applications, we reviewed the options available under these features to identify the types of actions and information that can be reported. We aggregated the relevant categories in Fig. 5. Categories such as misinformation, drugs, and fraud were excluded as they are not relevant to our study of privacy features and interpersonal attacks. We only included categories explicitly mentioned in the apps and did not make assumptions about what they might support.

We observed that harassment and spam are the main categories supported by social media platforms, likely due to their prevalence. Reports of impersonation are also well-supported by most apps, except for Telegram, WeChat, and WhatsApp. Only half of the apps support reports related to threats and data leakage, and even fewer provide options to report fake accounts.

WhatsApp does not support explicit or detailed reporting; it only accepts general reports. More concerning is that most apps do not allow users to provide detailed information with their reports, limiting them to selecting from predefined categories without the option to explain the reasons for their report.

6 EFFICACY OF PRIVACY FEATURES AGAINST INTERPERSONAL ATTACKS

Experiment set up. We created between three and six accounts on each social media platform specifically to evaluate the efficacy of these attacks. One account on each platform was always designated as the target, while the attacker could create one or more accounts on the platform as needed. We used different IP addresses located in the United States and different devices for the target and the attacker to simulate a scenario where the attacker and the target do not live together. This step was necessary to prevent social media suggestion algorithms from using IP-level information. For the target's account, we set up a profile with basic information, created small groups, and uploaded content such as posts and stories as needed for the experiments.

Attack tasks	Attack scenario	# attacks succeeded: Avg. per app	Median per app	On all apps
Discover the target's account (10 attacks: e.g., discover the target through a shared group)	fr-default	-	-	-
	nf-default	7.0	8	4
	fr-secure	-	-	-
	nf-secure	5.6	7	0
Collect and monitor information about the target (30 attacks: e.g., secretly screenshot profile information)	fr-default	17.2	18	16
	nf-default	15.2	16	11
	fr-secure	13.2	14	5
	nf-secure	12.1	13	4
Share information about the target (7 attacks: e.g., share the target's posts outside the app)	fr-default	5.0	5	7
	nf-default	2.9	3	0
	fr-secure	1.9	3	0
	nf-secure	0.5	0	0
Make unwanted communication with the target (12 attacks: e.g., spam the target with mentions)	fr-default	8.5	9	9
	nf-default	7.2	8	4
	fr-secure	6.0	6	3
	nf-secure	5.1	6	3

Figure 6: We report the attacker's goals, the number of attack actions we considered that can help achieve that goal, and an example attack. Then, for the four threats, we report the average and median of the number of attacks that succeeded (rounded up) and the number of attacks that succeeded for all apps. nf represents *non-friend* and fr represents *friend*, and both cases are either under *default* or *secure* configurations.

In our simulation, we assume the attacker begins with a single account to conduct the attacks and creates additional accounts (no more than five in total) as needed. Some of these additional accounts were created to test the efficacy of blocking. We first conducted experiments with the target account set to the default privacy configurations, where the attacker was a friend of the target. Then, we tested the non-friend scenario, followed by the scenario where the attacker was explicitly blocked by the target. Afterward, we repeated the same experiments with the target account set to the secure configuration, following the same sequence of attacker scenarios. For each setting, we simulated each attack by following a series of steps for each social media application and recorded whether the attacks succeeded or failed. To prevent the attacks we demonstrated from being misused in practice, we do not release the specific attack steps anywhere.

Ethical considerations. Our study did not require IRB approval, as no human subjects were involved. All experiments were conducted on lab accounts created specifically for this study, and we refrained from testing interactions that would require the participation of other users. We deleted all accounts upon completion of the study. Also, we did not test any abuse reporting features, as this could involve human reviewers. We manually conducted the attacks, relying solely on the provided user interface, external devices, and tools. None of our methods exploited security vulnera-

bilities, ensuring that our experiments had negligible impact on the apps. Our results cannot be used to compromise, overload, deny, or negatively affect these platforms or their services in any way. While the information we gathered is publicly available, we have deliberately refrained from sharing specific details of our attack methods to prevent potential abusers from replicating them to monitor or harass victims. Our goal is to provide insights into the effectiveness of privacy features and their limitations, helping to improve these features and reduce the range of possible attacks.

We report the highlights of our attacks and refrain from discussing details on how these attacks were conducted for ethical reasons. All experiments were conducted in 2024 from February 1st until February 20th.

Overall attack efficacy. As shown in Fig. 6, a large number of attack actions are successful on social media platforms under default configurations. Even after securing the account, a significant number of attacks still succeeded. For instance, out of 30 attack actions aimed at collecting the target's information, an average of 17.2 attacks succeeded for friends under default settings (when the target is most vulnerable), dropping to 12.1 under the most secure configuration. This suggests that despite securing the account by removing the attacker or adjusting settings, many attacks remain plausible for all attack tasks except information sharing. For sharing attack actions under the most secure configuration, a median of 0 successful attacks was observed across platforms.

Discovering accounts is feasible even with most secure configurations. As described in Section 4, a non-friend adversary must first discover the target's account in order to launch attacks. Under default configurations, an average of 7 attacks per app were successful, as shown in Fig. 6. When the target's account was set to secure, the average number of successful attacks slightly decreased to 5.6 per app, which is still high in the context of interpersonal attacks. Regardless of the account's configurations, the attacker can always discover the target's account on all apps, except for WeChat, if they know the target's username from other platforms. Similarly, attackers can always discover the target via their first and last names on all apps, except for Telegram, WeChat, and WhatsApp.

We found that QR codes are generally insecure because they are directly linked to the target's account. Only WeChat and WhatsApp allow QR code resets, preventing attackers from using an old QR code to find information about the target.

Additionally, suggestion algorithms introduce significant risks to targets by helping attackers find people they know with minimal effort. All apps that support user suggestions fail to protect against discovery attacks, even when the account is secured. Attackers can use known information, such as the target's university or city, and contact syncing to locate their targets, and targets generally cannot prevent being suggested to others.

Variety of information can be collected. Attackers aim to collect data about the target by monitoring their profile and interactions. We tested 30 different attacks as described in Fig. 6; these attacks focus on collecting information generally accessible to many users (e.g., profile picture), rather than private information like chats with other individuals. Once the attacker discovers the target's account, they can access it across all 10 apps, regardless of the configurations, and view additional information about the user. For example, attackers can always see the target's first and last name, and in most apps, they can also view the profile picture and bio. Only WhatsApp and Telegram can prevent the attacker from accessing this information under secure configurations. Information such as connections, posts, and comments is generally visible to attackers; however, it can be hidden if the account is secured.

Under secure configurations, if the attacker is not a friend of the target, they cannot view stories on any app except Instagram, which allows all users to view stories unless the user restricts the audience. Once added as a friend, the attacker will be able to view any story. However, in all apps, targets can secure their stories by restricting the attacker from viewing them. Similarly, attackers can access posts and comments even if they are non-friends. If the account is private, only attackers who are friends can access these posts, and some apps provide the option to exclude attackers from viewing posts, but not comments.

On LinkedIn, attackers can view all of a target's interactions. These interactions are summarized under the "all activity" tab, which cannot be hidden from other LinkedIn users. This tab displays all reactions to posts, comments, and public posts (including images and other information) made by the target. Additionally, under default configurations, an attacker can view all of the target's activity without even logging into the app.

Among the data attackers seek to collect, phone numbers and emails are generally hidden by apps. Only LinkedIn shows emails and Telegram shows phone numbers, but this is limited to friends under default configurations. Attackers can always access the target's phone number on WhatsApp, but this is expected, as WhatsApp relies on phone numbers as unique identifiers. Other information, such as the online status indicator, cannot be viewed in most apps if the account is secured, and it is hidden from non-friend attackers under default configurations in many apps.

Circumventing transparency features can be achieved with simple tools and methods. As discussed in Section 5.6, transparency features can detect when a user reads a chat, captures the screen, downloads media, and more. However, we found that collecting data and monitoring the target is relatively easy due to the lack of robust transparency features across most social media apps. Transparency features are important because they introduce challenges for attackers. For example, if an attacker tries to capture the screen on Snapchat, the target will be notified of this action. However, attackers can circumvent these transparency features, and the easiest way for them to do so is by using a second phone to capture the screen, which can also bypass features shown in Fig. 3 and Fig. 4. Although the quality of captured images and videos is lower than those taken with built-in features, attackers can still secretly save the desired information.

Another method we found is through dual-use apps [27] that allow attackers to download stories secretly. Additionally, attackers can use built-in features or dual-use apps designed for screen mirroring, enabling them to mirror the app's screen to another device, such as a laptop, and then record or take screenshots of the mirrored screen. This approach avoids triggering transparency features. Lastly, attackers could root their device to bypass these features, though this falls outside the scope of our threat model.

Attackers can also hide that they have viewed messages by simply adjusting their privacy settings. For stories, anonymous viewing can be done using built-in features (paid on Telegram and free on TikTok), external apps (for Snapchat and WhatsApp), or airplane mode (for Facebook and Instagram).

Unwanted content sharing cannot be prevented easily. Sharing the target's posts and stories cannot be prevented under default configurations if the attacker is a friend. If the target's posts are visible to the attacker, they can always repost them, as most apps do not offer a feature to limit reposting (except for WeChat, which does not allow reposting). Even if such a feature exists, the attacker can simply capture the screen and then share it with others. Attackers can also share posts within apps as direct messages to other users or as links that can be accessed outside of these apps. Similarly, attackers can share stories within the app using built-in sharing features or by generating links provided by most apps. Additionally, attackers can record the screen and share the target's story. Telegram is the only app that successfully prevents most story-sharing attempts, but this is effective only when screenshots are disabled within the app. However, if the attacker uses a second device or dual-use apps (e.g., screen mirroring apps), they can still capture and share the story (as discussed later in this section). Many of these risks also apply to followers — users with a one-way connection to the target. These attacks can be mitigated by hiding posts and stories from the attacker.

Blocking attackers limit most if their interactions.. Blocking users limits their interactions and restricts the information they can view. Similar to the threat models for friends and non-friends, we explored what information is accessible to an interpersonal attacker logged into their blocked account and how restricted their interactions are under both default and secure configurations. We found that the threat model of a blocked attacker generally remains the same under both default and secure configurations for most apps. On Facebook, Instagram, LinkedIn, Snapchat, and TikTok, blocked attackers will see that the target's profile no longer exists. In contrast, on X, users will be explicitly notified that they have been blocked if they visit the target's profile, and on Pinterest, if they try to follow the target or repost their content. Attackers on Telegram, WeChat, and WhatsApp will not receive an explicit notification or see that the target's profile has been removed. Instead, on Telegram, they will see that the user has been inactive for a long time. On WeChat, they will receive a notification when trying to send a message, indicating that the message failed to send. On WhatsApp, all user information will be hidden from the attacker, but the profile itself will remain accessible.

Once blocked, attackers will not be able to send private messages to targets or view and interact with their content (with the exception of Pinterest, which does not hide the user's posts)[3]. On X and Pinterest, attackers can still view certain information, such as the user's ID, name, and profile picture. If both the target and the attacker are in a shared group, the attacker may still be able to view the target's messages and interact with them. For instance, on Facebook, attackers can join group calls with the target and view their messages, although messages sent by the attacker will be hidden from the target. Other apps, such as WhatsApp, do not restrict the interactions of blocked attackers with targets in shared groups at all. Failing to fully isolate the attacker, even after being blocked, introduces significant risks to targets.

Creating new accounts bypasses blocking. Blocking can prevent many interactions with harassing users, as it stops them from viewing posts, commenting, sending messages, and engaging in other interactions. However, attackers can easily circumvent blocking by creating new accounts. Social media platforms have attempted to address this issue by enforcing verification processes. However, attackers can easily create temporary email addresses online to set up new verified accounts. Additionally, many services provide virtual numbers that attackers can use for account verification. When we tested free virtual numbers, most failed due to high demand. However, when we rented virtual numbers, which can be as inexpensive as $5, they worked for all apps requiring phone verification. In many cases, attackers do not even need phone verification, as most apps allowed attackers to interact with other users once the attacker's email is verified.

On Instagram, when a user blocks someone, all of that person's current accounts, as well as any accounts they may create in the future, are supposed to be blocked, as claimed by the app. However, we found that this feature does not work perfectly. We tested this feature by creating three accounts for the attacker and blocking one using the target's account. As a result, the other two accounts were also blocked since all three were connected to the same device and verified using the same phone number. We then created three new accounts: one using Safari's private mode on the first device (the equivalent of Chrome's incognito mode), one on the Instagram app on the same device after deleting and clearing its data, and one on a second iPhone. For all these accounts, we used new emails that were not linked to the original three[4]. Using these methods, the attacker was still able to contact the target and continue harassing them. We tested linking the new accounts to the old ones to see if they would be automatically blocked, but none of the new accounts were blocked. We observed that other accounts owned by the attacker are automatically blocked only if they are already linked to the blocked profile or if the attacker creates a new profile using an Instagram app or browser session connected to the blocked account. This shows that blocked attackers can simply create new accounts to continue harassing and monitoring the target.

7 DISCUSSION

We found that current social media privacy and safety features lack consistency and fall behind in protecting users from interpersonal attacks (Section 5). Platforms can take several steps to improve the safety and privacy of users, especially those who are experiencing interpersonal violence.

Transparent and controllable user-matching process. Some platforms rely on more than just emails and phone numbers to match users and help them discover people they may know (discussed in Section 5.1). Users are not informed about the specific information used for this matching process, nor are they given control over it, raising concerns about the privacy practices of these

[3] In a new update after our experiment, X started to show posts and comments of the blocker to blocked users.

[4] Instagram allows users to link multiple accounts together under "*Accounts Center*" [1]

platforms. Therefore, it is essential for these platforms to be transparent about the data used for matching users and to provide users with control over this process, allowing them to protect themselves from potential attacks.

Fine grained control over interactions. Apps lack fine-grained control over certain privacy features. For example, many apps do not allow users to exclude specific users from viewing posts; instead, they must choose between making the post public for everyone or private for their connections. Such coarse-grained control can negatively impact the user experience. For example, a content creator may not want to switch their account to private, as it would limit their reach as a content creator.

Additionally, attackers can evade blocking by simply creating a new account. While some apps, such as Instagram, attempt to block all accounts created with the same email or phone number, many social media apps do not. Creating accounts using temporary email addresses or phone numbers is also possible, underscoring the need for better user verification by social media apps. Also, as discussed in Section 6, blocking in some apps fails to eliminate some risks because it does not completely limit all interactions between the attacker and their target. We believe that limiting all interactions of the blocked is needed, but introducing such blocking might be challenging especially in the context of shared groups.

Finally, we believe that features that restricting users, which is implemented by Facebook and Instagram as discussed in Section 5.5, is beneficial in many contexts and should be widely adopted. For example, in a domestic violence context where the attacker and target live together, blocking might result in escalation of violence. Thus, restricting features that does not alarm attackers are crucial.

Streamline privacy management. We observed several limitations of privacy and safety controls across all apps. For example, apps do not allow adjusting discoverability features during account creation, thus leaving a window of time between account creation and changing account discoverability settings when an attacker can discover a user's account. Moreover, even when users wish to keep their account secret, app contact suggestions can inadvertently advertise the target's account to the attacker due to shared email, phone numbers, mutual friends, or location.

Social media users often struggle to manage their privacy settings [47–49], and the lack of standardized labeling across platforms can further complicate this process, as users must learn to navigate each app independently. While privacy checkups implemented by some apps, such as Facebook [6], may assist users, these checkups do not cover all the privacy and safety features we identified.

Thus, we propose an *account privacy management framework* to simplify and standardize the process of configuring user privacy settings. During account creation, users would first be prompted to manage their discoverability settings, as being easily discovered can lead to other privacy risks. Next, they would be asked about their visibility settings, such as whether to make their account private. Then, users would configure what information others can share and how people can interact with them, including who can comment on posts, send direct messages, and add them to groups. We believe this framework will encourage users to manage their privacy settings and guide them efficiently through complex configurations. Additionally, it will increase awareness of various privacy settings, as users are often unaware of the security and privacy features on their devices [42]. Implementing standardized labeling across social media platforms may be challenging given that apps have their unique features. However, pursuing this goal—requiring collaboration among platforms—could improve usability and foster a safer digital environment.

Trade-offs of fine-grained privacy features. Introducing more fine-grained privacy features is essential for enhancing users' control over their privacy and safety. However, it may pose challenges for both users and social media platforms. Breaking down existing features into finer controls would significantly increase the number of privacy controls, potentially complicating settings for users who are already struggling to manage their configurations [47–49]. To address this concern, platforms could offer coarse-grained controls as the default for simplicity (basic settings), while providing fine-grained controls as advanced settings.

Several factors should be considered when implementing more fine-grained features. Platforms must ensure that these features are easy to navigate to minimize the complexity of managing privacy settings. Additionally, these features should be clearly described to align with users' expectations, as many features fall short in this regard [47]. Another challenge is that increasing controls could impact the platform's usability. For example, users might unintentionally restrict their interactions with others if they misconfigure visibility or interaction settings.

Limitations. We designed our feature-collection methodology to be as comprehensive as possible, but we may have missed some features due to potential bugs in the apps or updates introduced after our experiments. Additionally, we could not test the efficacy of reporting features, as this would require submitting false reports. Finally, since we only analyzed ten popular social media platforms, there may be additional features in other apps, and scaling our methodology to cover all platforms is a challenging task.

8 CONCLUSION

We analyzed 100 privacy features implemented in 10 popular social media apps and grouped them into eight categories: discoverability, visibility, saving and sharing, interaction, self-censorship, content moderation, transparency, and reporting. Our analysis revealed that many apps lack easily implementable privacy features. We then measure the effectiveness of the privacy features by designing 59 privacy attacks and simulating them via a set of accounts created on the ten social media applications specifically for this study. We observed that none of the applications initiate new accounts with the most secure settings; instead, new profiles are created with most privacy features turned off. Even when all available privacy features are turned on, we demonstrate that the account remains vulnerable to many interpersonal privacy and safety attacks.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their feedback, which helped improve the paper significantly. We acknowledge the use of ChatGPT-4o to assist with grammar and typo corrections throughout the paper. PI Chatterjee acknowledges funding from the OVC grant #15POVC-23-GK-01414-NONF and NSF #2339679.

REFERENCES

[1] About Accounts Center. URL: https://help.instagram.com/1731078377046291.

[2] About X Premium. URL: https://help.twitter.com/en/using-x/x-premium.

[3] Add or remove someone from your Restricted List on Facebook. URL: https://www.facebook.com/help/206571136073851?cms_platform=iphone-app&helpref=platform_switcher.

[4] California consumer privacy act (ccpa). URL: https://oag.ca.gov/privacy/ccpa.

[5] Content moderation | wikipedia. URL: https://en.wikipedia.org/wiki/Content_moderation.

[6] Facebook Privacy Checkup. URL: https://www.facebook.com/help/443357099140264.

[7] How do I restrict or unrestrict someone on Instagram? URL: https://help.instagram.com/2638385956221960.

[8] Interpersonal Violence - Counseling Center - Missouri State. URL: https://CounselingCenter.MissouriState.edu/InterpersonalViolence.htm.

[9] Introducing Instagram reels. URL: https://about.instagram.com/blog/announcements/introducing-instagram-reels-announcement.

[10] Reels on Facebook. URL: https://www.facebook.com/help/398606435303267.

[11] Restrict or unrestrict someone on Messenger. URL: https://www.facebook.com/help/messenger-app/1021314848608781?cms_platform=iphone-app&helpref=platform_switcher.

[12] Secret chats. URL: https://core.telegram.org/blackberry/secretchats.

[13] Self-censorship | Cambridge Dictionary. URL: https://dictionary.cambridge.org/us/dictionary/english/self-censorship.

[14] Self-censorship | Wikipedia. URL: https://en.wikipedia.org/wiki/Self-censorship.

[15] Send messages in vanish mode on Instagram. URL: https://help.instagram.com/888592124998543/.

[16] Signs and symptoms of interpersonal violence. URL: https://coloradolinkproject.com/about/child-and-elder-maltreatment-research/domestic-violence/interpersonal-violence/.

[17] Snapchat+. URL: https://www.snapchat.com/plus.

[18] Start end-to-end encrypted chats or calls in Messenger. URL: https://www.facebook.com/help/messenger-app/811527538946901/.

[19] Telegram Premium FAQ. URL: https://telegram.org/faq_premium.

[20] Welcome to LinkedIn Premium. URL: https://premium.linkedin.com/.

[21] What is Spotlight? URL: https://help.snapchat.com/hc/en-us/articles/7012271311892-What-is-Spotlight.

[22] Toxic Twitter – The Silencing Effect, March 2018. URL: https://www.amnesty.org/en/latest/news/2018/03/online-violence-against-women-chapter-5-5/.

[23] Internet and social media users in the world 2024, Jan 2024. URL: https://www.statista.com/statistics/617136/digital-population-worldwide/.

[24] Ghada M Abaido. Cyberbullying on social media platforms among university students in the united arab emirates. *International journal of adolescence and youth*, 25(1):407–420, 2020.

[25] 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.

[26] Maied Almansoori, Mazharul Islam, Saptarshi Ghosh, Mainack Mondal, and Rahul Chatterjee. The Web of Abuse: A Comprehensive Analysis of Online Resource in the Context of Technology-Enabled Intimate Partner Surveillance. In *2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P)*, pages 773–789. IEEE, 2024.

[27] Majed Almansoori, Andrea Gallardo, Julio Poveda, Adil Ahmed, and Rahul Chatterjee. A global survey of android dual-use applications used in intimate partner surveillance. *Proceedings on Privacy Enhancing Technologies*, 4:120–139, 2022.

[28] Rosanna Bellini, Emily Tseng, Nora McDonald, Rachel Greenstadt, Damon McCoy, Thomas Ristenpart, and Nicola Dell. " so-called privacy breeds evil" narrative justifications for intimate partner surveillance in online forums. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–27, 2021.

[29] Tommy KH Chan, Christy MK Cheung, and Zach WY Lee. Cyberbullying on social networking sites: A literature review and future research directions. *Information & Management*, 58(2):103411, 2021.

[30] Rahul Chatterjee, Periwinkle Doerfler, Hadas Orgad, Sam Havron, Jackeline Palmer, Diana Freed, Karen Levy, Nicola Dell, Damon McCoy, and Thomas Ristenpart. The spyware used in intimate partner violence. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 441–458. IEEE, 2018.

[31] Glen Clarke. Active information gathering for pentesting, Mar 2021. URL: https://www.dummies.com/article/academics-the-arts/study-skills-test-prep/comptia-pentestplus/active-information-gathering-for-pentesting-275736/.

[32] John W Creswell and J David Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.

[33] John W Creswell and Cheryl N Poth. *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications, 2016.

[34] Sauvik Das and Adam Kramer. Self-censorship on facebook. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 120–127, 2013.

[35] Jonathan Day. What is self-censorship? how does it kill media freedom?, Jun 2021. URL: https://www.liberties.eu/en/stories/self-censorship/43569.

[36] Ivan Dimov. Information gathering [updated 2019], May 2019. URL: https://resources.infosecinstitute.com/topics/penetration-testing/information-gathering/.

[37] Stacy Jo Dixon. Most popular social networks worldwide as of january 2024, ranked by number of monthly active users, Feb 2024. URL: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[38] Rubia Fatima, Affan Yasin, Lin Liu, Jianmin Wang, Wasif Afzal, and Awaid Yasin. Sharing information online rationally: An observation of user privacy concerns and awareness using serious game. *Journal of Information Security and Applications*, 48:102351, 2019.

[39] Thomas Feiter. How domestic violence victims can use social media for help. URL: https://www.fighterlaw.com/how-domestic-violence-victims-can-use-social-media-for-help/.

[40] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. "A stalker's paradise": How intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13, 2018.

[41] Diana Freed, Jackeline Palmer, Diana Elizabeth Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. Digital technologies and intimate partner violence: A qualitative analysis with multiple stakeholders. *Proceedings of the ACM on human-computer interaction*, 1(CSCW):1–22, 2017.

[42] Alisa Frik, Juliann Kim, Joshua Rafael Sanchez, and Joanne Ma. Users' expectations about and use of smartphone privacy and security settings. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2022.

[43] James L Gibson and Joseph L Sutherland. Keeping your mouth shut: Spiraling self-censorship in the united states. *Political Science Quarterly*, 138(3):361–376, 2023.

[44] Gary W Giumetti and Robin M Kowalski. Cyberbullying via social media and well-being. *Current Opinion in Psychology*, 45:101314, 2022.

[45] Sam Havron, Diana Freed, Rahul Chatterjee, Damon McCoy, Nicola Dell, and Thomas Ristenpart. Clinical computer security for victims of intimate partner violence. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 105–122, 2019.

[46] Avery E Holton, Valérie Bélair-Gagnon, Diana Bossio, and Logan Molyneux. "Not their fault, but their problem": Organizational responses to the online harassment of journalists. *Journalism Practice*, 17(4):859–874, 2023.

[47] 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.

[48] 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.

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

[50] Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F Churchill, and Sunny Consolvo. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2189–2201, 2017.

[51] Jeremy McHatton and Kambiz Ghazinour. Mitigating social media privacy concerns-a comprehensive study. In *Proceedings of the 9th ACM International Workshop on Security and Privacy Analytics*, pages 27–32, 2023.

[52] James A Mercy, Susan D Hillis, Alexander Butchart, Mark A Bellis, Catherine L Ward, Xiangming Fang, and Mark L Rosenberg. Interpersonal violence: global impact and paths to prevention. *Injury Prevention and Environmental Health. 3rd edition*, 2017.

[53] Jay Peters. LinkedIn gives up on stories, Aug 2021. URL: https://www.theverge.com/2021/8/31/22650740/linkedin-stories-ephemeral-video-shut-down.

[54] Kevin A Roundy, Paula Barmaimon Mendelberg, Nicola Dell, Damon McCoy, Daniel Nissani, Thomas Ristenpart, and Acar Tamersoy. The many kinds of creepware used for interpersonal attacks. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 626–643. IEEE, 2020.

[55] Patrawat Samermit, Anna Turner, Patrick Gage Kelley, Tara Matthews, Vanessia Wu, Sunny Consolvo, and Kurt Thomas. {"Millions} of people are watching {you"}: Understanding the {Digital-Safety} needs and practices of creators. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5629–5645, 2023.

[56] K Saravanakumar, K Deepa, et al. On privacy and security in social media–a comprehensive study. *Procedia Computer Science*, 78:114–119, 2016.

[57] Anirban Sengupta and Anoshua Chaudhuri. Are social networking sites a source of online harassment for teens? Evidence from survey data. *Children and Youth Services Review*, 33(2):284–290, 2011.

[58] Leslie Regan Shade and Rianka Singh. "Honestly, we're not spying on kids":

School surveillance of young people's social media. *Social Media+ Society*, 2(4):2056305116680005, 2016.
[59] Keren Sharvit, Daniel Bar-Tal, Boaz Hameiri, Anat Zafran, Eldad Shahar, and Amiram Raviv. Self-censorship orientation: Scale development, correlates and outcomes. *Journal of Social and Political Psychology*, 6(2):331–363, 2018.
[60] Jessica Staddon, David Huffaker, Larkin Brown, and Aaron Sedley. Are privacy concerns a turn-off? engagement and privacy in social networks. In *Proceedings of the eighth symposium on usable privacy and security*, pages 1–13, 2012.
[61] Sophie Stephenson, Majed Almansoori, Pardis Emami-Naeini, and Rahul Chatterjee. "it's the equivalent of feeling like you're in jail": Lessons from firsthand and secondhand accounts of iot-enabled intimate partner abuse. In *32nd USENIX Security Symposium (USENIX Security 23)*, 2023.
[62] Sophie Stephenson, Majed Almansoori, Pardis Emami-Naeini, Danny Yuxing Huang, and Rahul Chatterjee. Abuse vectors: A framework for conceptualizing iot-enabled interpersonal abuse. In *32nd USENIX Security Symposium (USENIX Security 23)*, volume 3, 2023.
[63] Robert C. Stern. Linkedin stats looking into 2023, Feb 2023. URL: https://www.linkedin.com/pulse/linkedin-stats-looking-2023-robert-c-stern/.
[64] Rima Tanash, Zhouhan Chen, Dan Wallach, and Melissa Marschall. The decline of social media censorship and the rise of {Self-Censorship} after the 2016 failed turkish coup. In *7th USENIX Workshop on Free and Open Communications on the Internet (FOCI 17)*, 2017.
[65] Mark Taylor, John Haggerty, David Gresty, Natalia Criado Pacheco, Tom Berry, and Peter Almond. Investigating employee harassment via social media. *Journal of Systems and Information Technology*, 17(4):322–335, 2015.
[66] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 247–267. IEEE, 2021.
[67] Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein. "it's common and a part of being a content creator": Understanding how creators experience and cope with hate and harassment online. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2022.
[68] Emily Tseng, Rosanna Bellini, Nora McDonald, Matan Danos, Rachel Greenstadt, Damon McCoy, Nicola Dell, and Thomas Ristenpart. The tools and tactics used in intimate partner surveillance: An analysis of online infidelity forums. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1893–1909, 2020.
[69] Emily Tseng, Diana Freed, Kristen Engel, Thomas Ristenpart, and Nicola Dell. A digital safety dilemma: Analysis of computer-mediated computer security interventions for intimate partner violence during covid-19. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021.
[70] Kurt Wagner and Bloomberg. Twitter ending fleets, its stories clone, due to lack of interest, Jul 2021. URL: https://fortune.com/2021/07/14/twitter-fleets-stories-clone-closing/.
[71] Elizabeth Whittaker and Robin M Kowalski. Cyberbullying via social media. *Journal of school violence*, 14(1):11–29, 2015.
[72] Randy Yee Man Wong, Christy MK Cheung, Bo Xiao, and Jason Bennett Thatcher. Standing up or standing by: Understanding bystanders' proactive reporting responses to social media harassment. *Information Systems Research*, 32(2):561–581, 2021.
[73] Yin Zhu, Xiao Wang, Erheng Zhong, Nathan Liu, He Li, and Qiang Yang. Discovering spammers in social networks. In *proceedings of the AAAI conference on artificial intelligence*, volume 26, pages 171–177, 2012.

A SOCIAL MEDIA TERMINOLOGY

We describe the terminology we used in the paper:

- *Chats*: real-time online conversations between two or more users. Chats are not limited to text-based messages but can also include media such as images, videos, voice messages, and other content. If the chat involves only two parties, we refer to it as a private chat, whereas if it involves at least three parties, we refer to it as a group or group chat.
- *Connections/contacts list*: shows the people with whom the user is connected on the platform.
- *Contact syncing*: allows platforms to import contacts from the user's phone or email, helping users connect with people they may know through suggestions.
- *Disappearing messages*: messages that are automatically deleted after being viewed by the recipient or after a specified period of time. The term "self-destruct" is used by some apps instead of the term "disappearing".
- *Followers list*: shows the people who are following the user's account. Followers can view the content shared by the user.
- *Following list*: shows the people that the user is following.
- *Groups*: online communities (spaces) within the platform created by users. In a group, members can interact and share content with each other, even if they are not directly connected. This differs from group chats, but platforms often use the term for communities and chats interchangeably. We will make the distinction throughout the paper when necessary.
- *Messages*: unlike posts, messages are shared on private conversations or small group chats.
- *Online indicator*: shows whether a user is currently online. Some apps refer to it as "activity status."
- *Posts*: pieces of content shared on the user's timeline. Depending on the app, the shared content could be text, links, images, videos, or audio files.
- *Private account*: an account restricted to approved followers or contacts only. Content shared on private accounts is visible exclusively to users who have been granted access. Non-connections must request permission from the account holder before being able to view their content.
- *Public account*: an account that is accessible by anyone. Content shared on public accounts is visible to all users, regardless of whether they are connected to the account owner or not.
- *Secret chat*: a mode in Telegram that utilizes end-to-end encryption and additional privacy features to ensure secure communication between users
- *Secret conversation*: a mode in Facebook Messenger that uses end-to-end encryption, similar to secret chats, to provide secure communication between users.
- *Snaps*: a term exclusively used by Snapchat for images and videos that disappear after being viewed, or within 24 hours if shared as a story.
- *Snap Map*: a feature exclusive to Snapchat that shows the real-time locations of all friends on a map.
- *Stories*: temporary posts that typically disappear after 24 hours.
- *Timeline*: a chronological listing of posts that users can view on their homepage. These posts include the user's own posts as well as posts from their connections.
- *Typing indicator*: shows when a user is currently typing a message in a chat. Some apps extend this feature to also indicate when a user is sending an image or recording a voice message.
- *Vanish mode*: a feature in Instagram in which messages disappears when closing the chat or turning off the mode. Whether the mode encrypts messages or not is not stated.

B LIST OF FEATURES AND ATTACKS

We report all the 100 privacy features in Fig. 7 and Fig. 8. We also report our attack actions in Fig. 9

Category	Description	# features	Feature	Facebook	Instagram	LinkedIn	Pinterest	Snapchat	Telegram	TikTok	WeChat	WhatsApp	X
Discoverability	Controls how someone can find the user's account	6	Limit discoverability by username	✗	✗	–	✗	✗	✗	✗	✓	–	✗
			Limit discoverability by email	✓	–	✓	–	✓	–	✓	–	–	✓
			Limit discoverability by phone number	✓	–	✓	–	✓	✓	✓	✓	✗	✓
			Limit discoverability by QR code	✗	✗	✗	✗	✗	✗	✗	✓	✓	–
			Limit discoverability in suggestions	✗	✗	✗	✗	✓	✓	✗	✓	–	✗
			Limit discoverability via forwarded messages or reposts	✗	–	✗	–	–	✓	–	–	–	✗
Visibility	Controls who views the user's account information such as posts, comments, and personal information.	23	Make account private	✓	✓	✓	✓	✓	–	✓	–	–	✓
			Limit who can view last name	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
			Limit who can view posts	✓	✗	✓	✗	–	–	✓	✓	–	✓
			Limit who can view stories	✓	✓	–	–	✓	✓	✓	✓	✓	–
			Limit who can view profile photo	✗	✗	✓	✗	–	✓	✗	✗	✓	✗
			Limit who can view phone number	✓	–	–	–	–	✓	–	–	✗	✗
			Limit who can view email address	✓	✓	✓	✓	✓	–	✓	✓	–	✓
			Limit who can view own connections	✓	✗	✓	✗	–	–	✓	–	–	✗
			Limit who can view own live broadcast	✓	✓	–	–	–	–	✓	–	–	✗
			Limit who can view broadcasts currently viewed by you	–	–	–	–	–	–	–	–	–	✓
			Listen to live broadcasts anonymously	–	–	–	–	–	–	–	–	–	✓
			Hide online indicator	✓	✓	✓	–	✓	✗	✓	–	✓	–
			Hide read receipts	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓
			Hide typing indicator	✗	✗	✓	✗	✗	✗	✗	–	✗	✗
			Hide "last seen" status	✓	✓	–	–	–	✓	✓	–	✓	–
			Hide own interactions with others (e.g., likes)	✗	✗	✗	✗	–	–	✓	–	–	✗
			Hide posts you are tagged in from your profile	✓	✗	–	–	–	–	–	–	–	–
			Hide personal information like city, bio, education, etc.	✓	✗	✓	✗	✓	✓	✗	✗	✓	✓
			Post in groups anonymously	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
			Manage location sharing settings	✓	–	–	–	✓	✓	–	✓	✓	✓
			Hide that you viewed people's posts	–	–	–	–	–	–	✓	–	–	–
			Hide that you viewed people's stories	✗	✗	–	–	✗	✓	✗	✗	✗	–
			Hide that you viewed people's profiles	–	–	✓	–	–	–	✓	–	–	–
			Hide that you viewed people's profiles	–	–	✓	–	–	–	✓	–	–	–
Saving & Sharing	Limits what information about the user people can share	7	Limit reposting posts/stories	✓	✓	✗	✗	✗	✓	✓	–	–	✗
			Prevent chats/media from being captured	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗
			Limit sharing posts/stories in private messages	✗	✓	✗	✗	✗	✓	✗	–	–	✗
			Limit sharing posts/stories outside app	✗	✗	✗	✗	✗	✓	✗	–	–	✗
			Limit who can forward messages	✗	✗	✗	✗	✗	✓	✗	–	–	✗
			Protect media from being downloaded	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗
			Blur app when viewed in App Switcher	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗
Interaction	Limits who can interacts with the user by commenting on their posts, send messages, and follow the user account.	17	Block other users	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
			Limit who can comment on posts	✓	✓	✓	✓	✓	–	✓	✗	–	✓
			Limit who can send private messages	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓
			Limit who can add you from a group chat	✗	✓	✓	✗	✗	✓	✗	✓	✓	✗
			Limit others from adding you to groups	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗
			Limit others from mentioning you	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗
			Limit others from tagging you	✓	✓	✓	–	–	–	✓	–	–	✓
			Limit specific users and restrict them	✓	✓	✗	✗	✗	–	✗	–	–	✗
			Limit access to owned group and permissions of members	✓	✓	✓	✓	✗	✓	✗	✓	✓	✗
			Limit friend requests	✓	✗	✓	✗	✗	–	✗	✓	–	✗
			Limit who can follow your account	✓	✗	✓	✗	✗	–	✗	–	–	✗
			Limit who can add you by contact card	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗
			Limit who can send voice and video messages	✗	✗	✗	–	✗	✓	–	✗	✗	✗
			Limit who can call you	✗	✗	–	–	✗	✓	–	✗	✗	✓
			Mute calls of others	✓	✓	–	–	✓	✗	–	✗	✓	✗
			Limit participation on live broadcasts	✓	✓	–	–	–	–	✓	–	–	✓
			Remove friends and followers	✓	✓	✓	✓	✓	–	✓	✓	–	✗

Facebook, Instagram, LinkedIn, Pinterest, Snapchat, Telegram, TikTok, WeChat, WhatsApp, X (Twitter)

Figure 7: We report the list of features under their designated category, and whether the feature exists in the app (✓), if it does not exist (✗), or if it is not applicable to that platform (–). Features related to private messages are bolded.

Category	Description	# features	Feature	f	ig	in	p	sc	tg	tt	wc	wa	X
Self-censorship	Controls user's actions, discourse of information and presence, which includes deleting posts, editing comments, and exiting groups.	19	Delete account permanently	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
			Deactivate account temporarily	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓
			Change phone number	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
			Change email	✓	✓	✓	✓	✓	–	✓	✓	–	✓
			Change username	✓	✓	–	✓	✓	✓	✓	✓	–	✓
			Change first & last name	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
			Delete messages from both sides	✓	✓	✓	✗	✓	✓	✓	✓	✓	✗
			Delete entire chat from both ends	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
			Delete own posts/comments/stories	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
			Auto-delete messages	✓	✓	✗	✗	✓	✓	✗	✗	✓	✗
			Edit messages	✓	✓	✓	✗	✗	✓	✗	✓	✓	✗
			Edit stories	✗	✗	–	–	✗	✓	✗	✓	✗	–
			Edit posts	✓	✓	✓	✓	–	–	✓	✓	–	✓
			Edit comments	✓	✗	✓	✓	✗	–	✗	✗	–	✓
			Unpin posts and messages	✓	✓	✓	–	–	✓	✓	–	–	✓
			Edit/remove personal information such as bio	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
			Edit/remove interactions such as likes	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
			Limit notifications related to profile edits	✓	–	✓	–	–	–	–	–	–	✓
			Exit groups	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Content moderation	Allows removing interactions made by others, such as muting posts and deleting comments.	5	Mute and filter words	✓	✓	–	✓	–	–	✓	–	–	✓
			Hide/mute posts, accounts, chats, and stories	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
			Remove someone's comment from own post	✓	✓	✓	✓	✓	–	✓	✓	–	✓
			Remove someone's message from group	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗
			Control which member's profile photos to see	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
Transparency	Logs of interaction done by others, such as who recorded the user's story and who visited their profile.	18	Logs who has viewed your story	✓	✓	–	–	✓	✓	✓	✗	✓	–
			Logs if a user has edited a message	✓	✓	✓	–	–	✓	✓	✓	✓	–
			Logs if a user has deleted a message	✓	✗	✓	–	✓	✗	✓	✓	✓	–
			Logs if chat auto-delete has been activated	✗	–	–	–	✓	✓	–	–	✗	–
			Logs if a chat has been screenshotted or recorded	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗
			Logs if profile info has been screenshotted or recorded	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
			Logs if story has been screenshotted or recorded	✗	✗	–	–	✓	✗	✗	✗	✗	–
			Logs who has visited your profile	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
			Logs if a user has edited a post	✓	✗	✓	✗	–	–	✗	✗	–	✓
			Logs who has reposted your post	✓	✓	✓	✓	–	–	✗	–	–	✓
			Logs who has interacted with your post or message	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
			Logs posts and messages edit history	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓
			Logs if sent messages has been read	✓	✓	✗	✗	✓	✓	✓	✗	✓	✓
			Logs who has read sent messages in groups	✗	✗	✗	–	✓	✓	✗	✗	✓	✗
			Logs if user is currently typing or sending a message	✓	✓	✗	✗	✓	✓	✗	✗	✓	✗
			Logs if images or videos have been saved from chat	✓	✓	✗	✗	✓	✓	✗	✗	✓	✗
			Logs who viewed a post	✗	✗	✗	✗	–	–	✓	✗	–	✗
			Logs if others have removed or blocked you	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓
Reporting	Allows user to report other users.	5	Report accounts	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
			Report chats	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓
			Report single messages	✓	✓	✗	✗	✓	✗	✓	✗	✓	✓
			Report posts	✓	✓	✓	✓	–	–	✓	✗	–	✓
			Report stories	✓	✓	–	–	✓	✓	✓	✗	✓	–

f: Facebook, ig: Instagram, in: LinkedIn, p: Pinterest, sc: Snapchat, tg: Telegram, tt: TikTok, wc: WeChat, wa: WhatsApp, X: X (Twitter)

Figure 8: We report the list of features under their designated category, and whether the feature exists in the app (✓), if it does not exist (✗), or if it is not applicable to that platform (–). Features related to private messages are bolded.

Attack tasks	Attack actions
Discover the target's account on a target platform:	This task can be done by searching the target's account using the social media search functionality, leveraging: (1) known usernames of the target from other social media platforms; (2) the email linked to the account; (3) the phone number linked to the account; (4) the target's full name. (5) The account can also be identified by monitoring the target's interactions (e.g., likes and reposts) on other users' profiles; (6) through common friends or known followers (e.g., celebrities followed by the target); or (7) by being in shared groups with the target. (8) Additionally, the target's account can be discovered via account information shared in chats (e.g., contact cards) or (9) through a previously known QR code, even if the target has changed their username. (10) Finally, accounts may also be discovered through the app's suggestion features.
Collect information about the target and monitor them:	This task can be performed by first (1) accessing the target's account page after identifying it. The attacker can then collect and monitor changes in the following information: (2) the first and last name associated with the account, (3) phone number, (4) email, (5) biography, (6) profile picture, (7) approximate location, (8) live location, and (9) other personal details such as education and job information found on the account page. The attacker can also monitor the target's activity and interactions on the platform, including: (10) their connections, (11) posts, (12) comments, (13) stories, (14) edit history of posts and comments, (15) interactions with other users (e.g., likes), and (16) live broadcasts they view. Additionally, they can monitor: (17) whether the target is currently online, (18) the last time they went online if not currently connected, (19) the time they viewed a chat, or (20) listened to a voice message. Some methods leave traces and notify the target, but certain actions can be done secretly, such as: (21) viewing account information, (22) chats, (23) stories, and (24) posts; (25) saving media from chats; or (26-29) screenshotting/recording profile information, chats, stories, and posts, all without triggering notifications. Finally, (30) the attacker can monitor whether the target has enabled privacy and safety features, such as hiding their online status or making their account private.
Share information about the target without their consent:	This task can be accomplished using built-in features that allow sharing, primarily through: (1) reposting the target's posts. Also, it can be achieved by (2-3) sharing the target's posts and stories via the attacker's own stories; and (4-5) sharing content through direct messages within the platform. (6-7) Finally, platform-generated links can be used to share the target's posts and stories outside the app.
Make unwanted communication with target:	This task can be done by spamming the target through: (1) calls; (2) messages; (3) comments; (4) tags; (5) mentions; or (6) repeatedly adding them to unwanted groups Unwanted communication also includes faking and impersonating information about the target, such as: (7) images; (8) videos; (9) voice messages; (10) chats; or (11) impersonating their account. (12) Finally, the attacker can leak private information they know about the target without their consent (e.g., information not disclosed by the target on the platform).

Figure 9: The table shows the social media privacy and safety attack objectives and attack actions used in our experiments.