"Please don't send that bot anything": A Mixed-methods Study of Personal Impersonation Attacks Targeting Digital Payments on Social Media

Hoang Dai Nguyen Sumit Dhungana Madhulika Itha Phani Vadrevu
Louisiana State University
{hngu281, sdhung4, mitha1, kvadrevu}@lsu.edu

Abstract

Personal impersonation attacks on social media are an emerging form of social engineering that exploit trust within interpersonal relationships to redirect digital payments. Unlike brand impersonation, these attacks target everyday users, leveraging real-time public interactions to deceive victims into transferring funds to attacker-controlled accounts. In this paper, we present the first in-depth study of PROSPER (**Payment Re-routing on Social media via Personal Impersonation**) attacks, focusing on their operational tactics, scale, and impact. Using a mixed-methods approach, we tracked 181 PROSPER attacks over a 3-month period, uncovering 70 distinct digital payment accounts and revealing human-in-the-loop scam operations alongside automated bots, as well as longstanding campaigns involving reused payment accounts.

Our quantitative analysis highlights the scale and persistence of these attacks, while our qualitative analysis provides deeper insights into attacker evasion strategies, victim targeting methods, and how victims are particularly vulnerable to these schemes. Based on these findings, we propose actionable recommendations for social media platforms and payment providers, including UI enhancements, stricter account handle management policies, and the sharing of blocklist information to mitigate these attacks and protect users from financial exploitation.

1 Introduction

Phishing remains one of the most heavily researched topics in web security, with prior studies predominantly focusing on attacks targeting websites through the *impersonation of popular brands*, such as banks, e-commerce platforms, social networks, and email services [35, 45]. Beyond websites, impersonation attacks have also been extensively documented on social media platforms. On platforms like X, attackers frequently impersonate well-known brands or public figures to execute scams, ranging from cryptocurrency technical support fraud [3] to squatting on profiles associated with popular brands [2], or orchestrating trust-based schemes that exploit a celebrity's identity [23, 43]. Expanding further, researchers have investigated more sophisticated impersonation schemes outside of social media, such as the theft and resale of user credentials and cookies, which represent a deeper and more sinister form of exploitation [12, 13]. These studies collectively demonstrate the diversity of impersonation tactics but largely focus on high-profile entities. However, this focus implicitly assumes that impersonation attacks are confined to prominent targets, such as brands or celebrities. This assumption overlooks the potential for attackers to impersonate ordinary individuals, particularly when there is sufficient financial motivation.

While limited research has addressed impersonation attacks targeting lay individuals, some prior work has explored related themes. Goga et al. [18] conducted a foundational study on personal impersonation on social media, analyzing 1.4 million X accounts and identifying 166 victim-impersonator pairs. Their findings revealed that only a small fraction of these impersonators ($\frac{2}{166} \approx 1\%$) attempted to contact individuals in the victim's network, leading the authors to conclude that personal impersonation accounts were primarily used for non-malicious activities, such as creating "real-looking" accounts to evade spam detection or participate in follower fraud. Additional studies examined generic impersonation profiles on social media [20, 21], including phenomena like profile cloning, but found no evidence of monetization or financial exploitation. Methods proposed for detecting generic impersonation accounts [22, 44] often relied on computationally expensive pairwise comparisons and faced scalability challenges along with issues of false positives. Crucially, the phenomenon of personal impersonation attacks driven by financial motivation has remained unexplored in existing research.

In this paper, we address this gap by presenting the first study of payment re-routing attacks that exploit personal impersonation on X, which we refer to as **PROSPER** (**Payment Re-routing on Social media via Personal Impersonation**) attacks. These attacks leverage various social engineering tactics to intercept social media communications in real-time and deceive victims into redirecting payments to attacker-controlled accounts. Unlike brand or celebrity impersonation, which typically targets high-profile entities, PROSPER attacks rely on mimicking everyday users during real-time interactions. This emerging threat thrives on public social media exchanges, where attackers strategically identify opportunities such as two users discussing payment information and masquerade as the intended recipient to steal funds. PROSPER attacks are uniquely lucrative as they provide immediate financial gains for attackers. This sets them apart from traditional phishing, where attackers often need to sell stolen credentials on underground markets or bypass two-factor authentication mechanisms to monetize their efforts. Our study reveals that attackers uniquely abuse microblogging platform APIs to search for random targets and quickly create impersonation accounts in as little as four minutes. Thus, by exploiting real-time communication and the context of public interactions, these attacks turn ordinary exchanges into highly effective social engineering schemes.

Our study expands the scope of impersonation research by focusing on the monetization mechanics of personal impersonation attacks, a largely underexplored area. By leveraging both measurement and qualitative analyses, we uncover the mechanisms that enable these attacks to succeed and reveal why victims are particularly vulnerable. The measurement analysis allows us to examine the scale, persistence, and operational patterns of PROSPER attacks, providing a foundational understanding of their structure. Complementing this, our qualitative analysis provides deeper insights into attacker tactics, victim targeting strategies, and the dynamics of public, real-time social media interactions that facilitate these scams. Together, these approaches offer a holistic view of the problem, bridging numerical trends with contextual understanding.

The findings from our mixed-methods analyses yield several insights that lead to actionable outcomes for mitigating social engineering threats in the future. Specifically, for social media microblogging platforms like X and Bluesky that were found to be affected by PROSPER attacks, our results (§ 3) call for (1) improving prominence and preventing visual occlusion of security-critical elements in web and mobile UIs, (2) restricting arbitrary account ID changes, (3) proactive monitoring of API misuse that enables PROSPER-style attacks, (4) tracking interactions between accounts with lexically similar IDs, and (5) developing fraud mitigation mechanisms grounded in real-world social engineering attack data (e.g., impersonation account IDs, real attack content, persistent attacker payment IDs). Our qualitative analysis based on open coding further characterizes the profiles of victims targeted by PROSPER attacks, the involvement of bystanders in social engineering incidents, and the range of impersonation tactics employed by attackers. Notably, these insights point to additional mitigatory strategies, such as addressing the misuse of payment platform features (e.g., PayPal's "Friends and Family" option) and recognizing attacker behaviors involving typographic errors or confirmatory messages (§ 4).

1. Measurement analysis. We present the first measurement study of a new type of online social engineering attack named PROSPER across a period of 12 weeks tracking 181 attacks using 70 attacker-owned digital payment accounts. Additionally, we found evidence of social media bots that bootstrap these human-in-the-loop scam operations, demonstrating how automation is employed to amplify attack efficiency. Long-standing campaigns were also observed, with attackers reusing their payment accounts across multiple short-lived social media profiles.

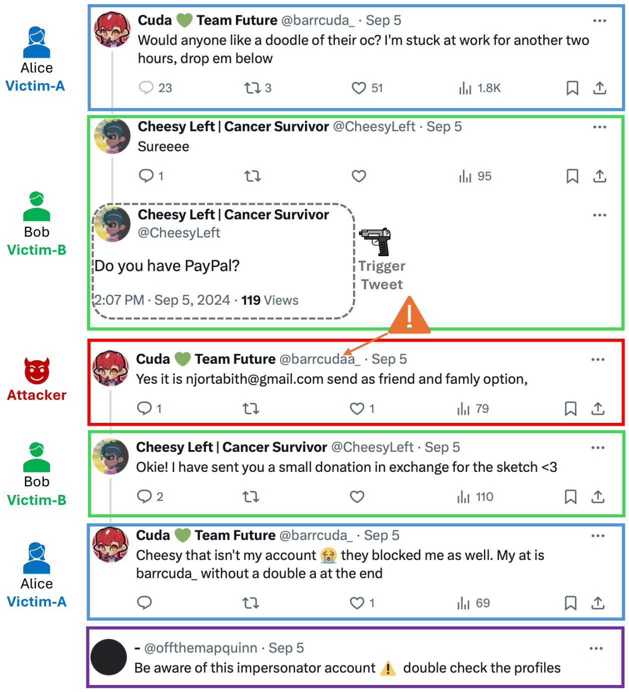
2. Qualitative analysis. Unlike prior research, which has primarily focused on quantitative metrics, we complemented our study with a qualitative analysis using an open coding approach. This approach revealed novel evasion tactics employed by attackers, such as subtle typographical changes and emotional coercion, as well as the types of victims targeted (e.g., lifestyle influencers, service providers, donation seekers). Furthermore, the public and real-time nature of PROSPER attacks provided a unique opportunity to observe how different stakeholders including victims, bystanders, and attackers react to these scams in real-time. These qualitative insights go beyond surface-level patterns, offering a deeper understanding of attacker-victim dynamics that quantitative methods alone would not capture.

3. Practical recommendations. Based on our findings, we propose concrete recommendations for mitigating PROSPER attacks. These include specific UI improvements for social media platforms, such as clearer account identity indicators, restrictions on changing account handles and proactive detection features like anomaly-based API monitoring. For payment platforms, we advocate for shared blocklists and real-time fraud detection mechanisms. These recommendations aim to protect users from financial exploitation while providing broader defenses against personal impersonation attacks in general.

Finally, we also plan to release all code and data in this research to bring more attention to this issue as well as encourage future research in the space of personal impersonation attacks.

2 Measurement Setup

Example case study. Before describing the setup we devised for collecting PROSPER attacks, we will first go over a representative real-world example shown in Fig. 1. As discussed in § 1, PROSPER attacks leverage the fact that two strangers are transmitting information about a third-party payment account on a social media network such as X. As shown in Fig.1, this conversation typically starts with a user, say, Alice, posting a public message[1]. This message causes another user, Bob, to respond with an inquiry seeking Alice's payment account ID. This public inquiry from Bob to solicit another user's payment account information can trigger the attention of PROSPER attackers who quickly respond to Bob with a social media account that impersonates Alice as shown in the figure with a red outline. It is to be noted how the attacker's X account has exactly the same profile picture and name as Alice. The only difference is in the account handles which start with an `@` and are mandated to be unique (but user-determinable) on X [40]. In this case, we can see that the attacker ensured that their handle was lexically very close to that of Alice with only a single character difference thus making it very hard for Bob to discern. Understandably, Bob fell for this attack and transferred the money inviting a lamenting message from Alice to Bob. The conversation ends with a warning from a third-party person who witnessed the PROSPER attack.



Cuda 💚 Team Future @barrcuda_ · Sep 5
Would anyone like a doodle of their oc? I'm stuck at work for another two hours, drop em below
23 3 51 1.8K

Cheesy Left | Cancer Survivor @CheesyLeft · Sep 5
Sureeee
1 95

Cheesy Left | Cancer Survivor
@CheesyLeft
Do you have PayPal?
2:07 PM · Sep 5, 2024 · **119** Views

Cuda 💚 Team Future @barrcudaa_ · Sep 5
Yes it is njortabith@gmail.com send as friend and famly option,
1 1 79

Cheesy Left | Cancer Survivor @CheesyLeft · Sep 5
Okie! I have sent you a small donation in exchange for the sketch <3
2 110

Cuda 💚 Team Future @barrcuda_ · Sep 5
Cheesy that isn't my account 😭 they blocked me as well. My at is barrcuda_ without a double a at the end
1 69

- @offthemapquinn · Sep 5
Be aware of this impersonator account ⚠️ double check the profiles

Figure 1: Real-world example of a PROSPER attack.

From this, we can notice that Alice and Bob are both victims of this PROSPER attack. Alice was the impersonation target (i.e. the entity being impersonated) and the intended receipient of the financial payment who missed this intendent payment. At the same time, Bob was ultimately the person that the attacker had to socially engineer to perform the transaction and lose the money. So, we consider both Alice and Bob to be victims in PROSPER attacks. We refer to them as Victim-A (**`V-A`**) and Victim-B (**`V-B`**) respectively in the rest of the paper. Further, we term Bob's first post as a '***Trigger post***' as it is the main post that triggered the attacker to see scope for a PROSPER attack and perform it.

2.1 Practical considerations and challenges.

Our goal for this project was to perform longitudinal measurements about the incidence of PROSPER attacks across a multi-month period and then analyze the collected data. To do such measurements, we planned to perform ***real-time searches* for Trigger posts** from potential victims (i.e. `V-Bs`). Otherwise, it is likely that the attacker might delete their attack tweets either due to a successful theft or lack of confidence in eventual success of the attack and we will miss them completely. This is unlike the only other existing work in studying personal impersonation attacks [18] which focused exclusively on a large dump of historical social media data. Further, for similar reasons, we cannot rely on alternative data sources such as search engine caches, which have been increasingly adopted by computational social scientists in response to the discontinuation of free academic APIs by social media companies [32, 37]. This poses a significant challenge for our study, as these alternative data sources are unsuitable for our specific requirements.

Tier	Price / Month	Tweets / Month	Search	Streaming APIs
"Free"	0	1,000	✗	✗
"Basic"	US$100	10,000	✓	✗
"Pro"	US$5,000	1,000,000	✓	✓

Table 1: Status of X APIs (as of October 1, 2024) [39].

Following prior work on personal impersonation [18], we conducted our study on X. The details of X's currently available APIs for the general public are summarized in Table 1. Given our objective of performing real-time searches to capture trigger tweets, the "Free" tier was excluded as it does not support tweet search functionality. While real-time social media data is typically accessible via "Streaming APIs," these are only included in the "Pro" tier, which has been priced at US $5,000 per month since 2023. Unfortunately, the "Pro" tier's cost far exceeds our project's total budget of US $600. As a result, we were constrained to utilize the "Basic" tier API, priced at US $100 per month. This presented a significant challenge in achieving our research goals, as we had to work around the limitations imposed by the "Basic" tier, which restricts the total monthly tweet read limit to 10,000 and lacks access to streaming APIs.

Given the constraints of our API plan, employing a broad sampling approach starting with a random selection of social media accounts on X, as done in prior work, is practically infeasible in the current "post-API era" of the internet [32]. Instead, we must optimize and maximize the utility of the

[1] Throughout the paper, we use the terms: "tweet", "post" and "message" interchangeably to refer to public social media messages posted by users.

"Basic" API plan to effectively measure PROSPER attacks, simulating capabilities akin to the "Pro" tier. Consequently, designing an optimized and efficient system became a necessary prerequisite to enable this study and ensure comprehensive data collection within the given limitations.

2.2 X-AM: Optimized attack data collection

We now discuss X-AM (X-API Maximizer; pronounced *exam*), an automated data collection pipeline to optimize social media API usage for collecting PROSPER attacks. While this pipeline is tailored for our usecase, we believe that some of the techniques we propose here might also be pertinent to other researchers working with social media data in measurement and social computing domains [32]. This is especially useful given the recent API changes brought forth by social media companies [37]. The X-AM pipeline is depicted in Fig. 2. Before discussing the core pipeline, we first provide details about how we generated input search query data for it. We then discuss the two components that make up our tailored PROSPER data collection pipeline followed by discussion of a semi-manual approach we utilized to validate the ground truth of our collected attack data.

2.2.1 User-driven tailored search queries

To collect and track PROSPER attacks, we first need to search for candidate trigger tweets (see Fig. 1) in real-time. In absence of streaming APIs to accomplish this, we are forced to make periodic search queries at regular intervals for trigger tweets. The most straightforward way of doing this is to search only for payment platforms (e.g. *"Paypal", "Venmo"* etc.). However, our preliminary manual testing showed us that this is not a viable option. As these platforms are very popular, we were receiving many messages that are not relevant to our main goal of seeking trigger posts where one user (`V-B`) is seeking the payment account information of another (`V-A`). This is particularly problematic given that the total number of tweets that can be downloaded is very small for the Basic tier (Table 1).

Realizing the impracticality of general searches, our simple idea was to craft tailored search queries by asking social media users how they solicit payment information from others. For this, we performed a user study in our university (n=50) where we created a hypothetical scenario in which the participant will need to solicit the payment account ID of another user. Fig. 10 in Appendix shows the survey prompt we used. The user responses showed several different ways that trigger posts can be crafted to ultimately attract PROSPER attackers. We manually disassembled these responses and identified common phrasal patterns in them (e.g. *"do you have"*, *"can I pay via"* etc.). We then used these patterns to construct a single long-tailored search query by leveraging the search query operators supported by X. Further, we also parameterized our query to ensure support for all the top payment providers. Analysis of post-survey data in temporal order of responses shows that the search query pattern that we can derive reaches a state of "stabilization" after 34 out of 50 responses (68%). This suggests that our user survey achieved a reasonable degree of generalizability. However, we acknowledge its limitations, specifically, the fact that all participants were students from a single university, which limits a complete recall of PROSPER attack data globally. Our final search query that ultimately matches all 50 user responses is shown in Appendix A.

2.2.2 Trigger Tweet Hunter ❶

The first of the two modules in our automated real-time PROSPER attack data collection pipeline is the "Trigger Tweet Hunter" module which is bootstrapped by the tailored search queries from to periodically find candidate trigger tweets. For this, we diversified the API end points that we used in order to optimize our API quota usage given that each end point has specific rate limit and other restrictions. Specifically, we periodically used the `tweets_counts_recent()` API endpoint, to retrieve a count of tweets that match our search query patterns within the current period. As this is only a count, this API does not add to our monthly tweet read limit. We only call the `tweets_search_recent()` endpoint to retrieve the actual tweets when the above periodic count

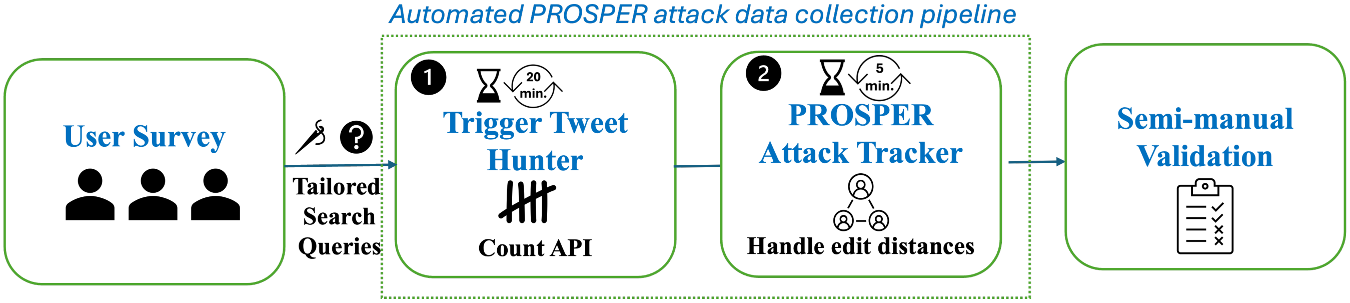


Figure 2: X-AM: Optimized real-time data collection and validation pipeline for PROSPER attacks

query yields a positive result. We also use the expansions feature in the X API to retrieve as much user-related information (i.e. V-B's) as possible directly with the trigger tweets themselves. We were able to accommodate a time-period of 20 minutes for this module.

2.2.3 PROSPER Attack Tracker ❷

The second module in our pipeline continuously monitors trigger tweets globally for attack activity. As it is possible that PROSPER attackers might delete their tweets and accounts to prevent post-mortem analyses, we sought to run this module as frequently as possible. With the help of the optimizations we have discussed so far, we were able to accommodate a time-period of 5 minutes for this. Every 5 minutes, this module keeps monitoring each trigger tweet upto a maximum period of 12 hours to see if it attracts any conversational activity from attackers. The first time a given trigger tweet generates new conversational activity, the module collects the parent of the trigger tweet (i.e. V-A's) and the associated user profile information. Any new replies to the trigger tweet are also collected along with the poster's profile information.

For each new reply to the trigger tweet, pairwise distances are computed between the account handles (i.e., globally unique "usernames" [40]) of all parties involved in the conversation thus far. To measure these distances, we use the Levenshtein edit distance metric [24], which is well-suited for detecting character-level modifications such as insertions, deletions, and substitutions. Among alternative metrics we considered, including the Damerau–Levenshtein distance [15], we observed on a small validation dataset that attackers rarely employed transpositions of adjacent characters, an additional edit operation unique to Damerau–Levenshtein distance. As such, Levenshtein edit distance offered a simpler and equally effective metric for capturing the kinds of modifications attackers use to confuse users. We then mark any conversation that has at least one pair of accounts with an edit distance $<=$ 3 as a candidate to likely contain a PROSPER attack. Our intuition is straightforward in that to conduct a PROSPER attack convincingly, the attacker will have to post a message with an account handle that is lexically very close to V-A. For example, the case study we presented earlier in Fig. 1 only had an edit distance of 1 between the attacker's and V-A's accounts. Given that PROSPER attacks involve impersonation of common people, we expect most attackers to not have issues finding an available handle that closely resembles V-A. Note that this is in sharp contrast to brand impersonation attacks such as phishing which need to contend with either defensive domain registration practices [8] or the likely high price of domains that are lexically close to target domains (such as miicrosoft.com). Our quantitative analysis later confirms this as it shows that most attackers are much closer in distance from V-A than our conservatively chosen threshold of 3.

Any conversation thus marked as likely containing a PROSPER attack will then be continued to be monitored until two hours of dormancy in order to capture all interesting after-the-attack interactions.

2.2.4 Semi-Manual Validation

After collecting PROSPER attack data via our above described pipeline, we then subject each of the conversations that were marked as suspicious to a semi-manual validation process to verify the ground truth. For this, we stipulate that a given suspicious conversation has to meet the below three necessary and sufficient criteria for it to be confirmed as an incidence of PROSPER attack:

1. *Automated.* The suspicious tweet poster whose account handle was lexically close to V-A's has exactly the same profile name (which is freely configurable) as well as profile picture (when available) as V-A.

2. *Manual.* The semantics of the trigger tweet indicate that its poster (i.e. V-B) is indeed seeking information about V-A's payment account ID to which V-B wants to make a payment.

3. *Manual.* The suspicious tweet's semantics indicate disclosure of payment platform account ID that V-B was seeking in the trigger tweet.

In other words, if we see an imitator who has the same profile name as V-A divulged their own payment account information after interjecting a conversation in which V-B has asked for V-A's payment information, we confirm it as a case of a PROSPER attack.

It is to be noted that our act of checking for profile name and profile picture matches (between imitator and imitatee) is solely for the purpose of data validation as opposed to using it as a core data collection feature in contrast to prior work [18]. As we will see during the discussion of results, the high-quality of attack results from our data collection module (❷) made this possible. We believe this is due to the tailored design of our data collection components made specific to PROSPER attacks.

2.3 X-AM: Deployment

We deployed X-AM for a 12-week period from June 17 to September 9, 2024. We utilized two "Basic" tier X API accounts for this deployment resulting in a total cost of US $600. The pipeline was implemented in Python with MongoDB being used as a database server.

3 Measurement Analysis of PROSPER attacks

Having discussed the design of our attack data collection pipeline, we now present measurement analysis of PROSPER attack data we collected. This analysis yields multiple new insights into the operational tactics of impersonation attackers and how they can be countered.

# Trigger tweets matching search queries (§ 2.2.2)	1066
# Trigger tweets suspected to be attacked (§ 2.2.3)	115
# Suspected impersonator accounts (§ 2.2.3)	127
# Confirmed impersonator accounts (§ 2.2.4)	127
# Distinct impersonation target accounts / V-As	115
# Targeted payment platforms	1 (PayPal)
# Distinct impersonator payment accounts	70

Table 2: Real-time PROSPER attack data summary

Table 2 presents summary statistics of our collected attack data. Out of 1066 trigger tweets that our X-AM pipeline has seen, it marked 115 tweets (10.8%) as being affected by a PROSPER attack due to lexical similarity in account handles as discussed previously. It is to be noted that our pipeline actually implicated 127 attacker accounts thus showing that sometimes, multiple impersonation attackers target the same conversation between a pair of V-A and V-B.

Our validation procedure unequivocally confirmed our suspicions. The two labelers were able to independently attribute all 127 suspected attackers' replies to PROSPER attacks, with each of them matching all the three criteria we laid out in § 2.2.4. Thus, all 127 attackers were using the same profile names as their V-As. Further, in all cases where we had access to real-time images of the attacker accounts (for 108 trigger tweets), the profile image matched that of the V-As, as exemplified in Fig. 1. This validates the utility of the heuristic approaches we devised (Fig. 2). *We thus strongly recommend social media companies such as X to utilize approaches similar to ours to identify and mitigate* PROSPER *attacks.*

We also noticed that PayPal was the only payment platform that was targeted by PROSPER attacks. But, this was not too surprising. Although there exist multiple other peer-to-peer payment platforms, they do not support cross-border payments. For example, Zelle (120M users), Venmo (80M users) and CashApp (50M users), only support US customers. Similarly, Google Pay can only be used for peer-to-peer payments in two countries [19]. On the other hand, PayPal is the only platform we are aware of that supports global-level peer-to-peer digital payments (200+ countries [29]). This cross-border support is important for attacker operations as prior research has shown that SE attackers tend to target victims outside their own countries to thwart takedown operations [25, 27]. This makes PayPal with its 400M global users [36], a unique target for cybercriminals perpetrating SE attacks as evidenced in prior work [2, 3].

We can also note from Table 2 that the number of attacker payment account IDs we obtained (70) is smaller than the number of impersonating social media accounts we came across (127). This shows a campaign-like orchestration of PROSPER attacks which we discuss more in depth later.

3.1 Temporal analysis

Fig. 3 is a cumulative line graph that charts the growth in number of trigger tweets and PROSPER attacks collected over our 12-week deployment period. The graph's blue line shows that about 73% (288/1066) of the candidate trigger tweets have been posted in the latter half of our deployment period (i.e. on or after July 30) indicating a noticeable surge in "attackable" tweets made by potential victims. Yet, the red line indicating the distribution of PROSPER attacks that have taken place shows that attacker activity has largely been uniform with about 52% (66/127) attacks taking place in the second half of our data collection period. We surmise that this is a sign of the human-in-the-loop nature of PROSPER attacks which as we discuss later limits their scalability.

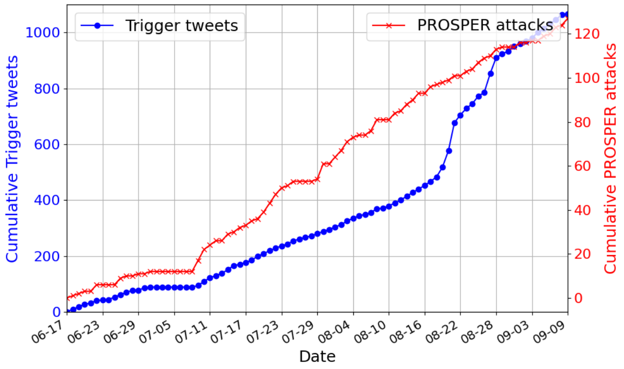


Figure 3: Cumulative growth of trigger tweets and PROSPER attacks over time. Please note that the right Y-axis has a small top offset to enhance readability.

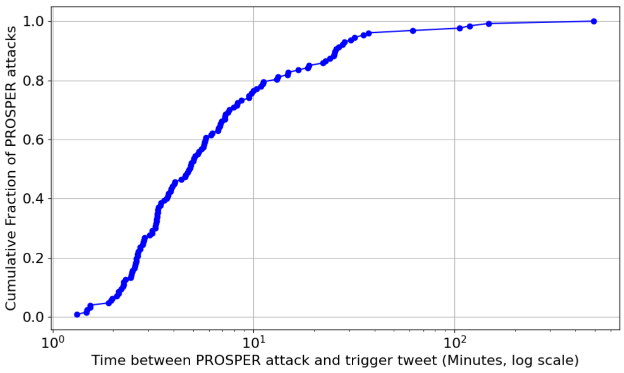


Figure 4: Response time distribution of PROSPER attacks

We also analyzed the time it takes for a PROSPER attacker to respond to a trigger tweet as shown in Fig. 4. Surprisingly, almost 80% of the attacks arrive in less than 10 minutes of time and 50% of attacks arrive in just 4 minutes. We believe this is a strong sign that PROSPER attacks are likely being bootstrapped with real-time search tools similar to the one we have devised for automatically hunting potential victims. From an attacker's point of view, this makes sense as conversations on social media networks such as X are very "real-time" in nature and the attackers would thus want to target their victims as quickly as possible.

These short times are particularly interesting considering all the activities that a PROSPER attacker has to do before posting their first attack tweet. First, the attackers will have to either create a new account or repurpose an existing one by changing its unique handle to one that lexically resembles that of `V-A` as discussed earlier. Next, they have to perform other impersonation changes (such as profile name and picture). Here, we note that unlike other platforms such as Reddit which completely disable account username changes [33], X allows for account handles to be changed at any time [41]. This X policy unfortunately appears to make it even more easy for attackers to launch PROSPER attacks as attackers can save time by repurposing pre-created accounts as soon as a trigger tweet is seen by their API-based search tools. *We thus strongly recommend X to consider restricting account handle changes to help mitigate these impersonation attacks*

At the same time, we also state that it might be possible for the attackers to create these accounts anew in real-time as opposed to repurposing them. A recent work that investigated in-the-wild CAPTCHA attacks has already shown large-scale existence of bot account activities on X [28]. Thus, as a complement to the strategy of restricting the handle changes, social media networks can consider *devising a detection system that flags any account that performs all activities of a* PROSPER *attacker (§ 2.2.4) in a small time interval*. And, finally, as real-time trigger tweet searches are indispensable to this attack, *monitoring API usage for anomalous trends* will also be very helpful for X in mitigating these attacks.

3.2 Impersonation tactics

Next, we computed the distribution of the Levenshtein edit distance [24] between the 127 attacker account handles and their impersonation targets. Our results showed that 116 pairs (91%) had an edit distance of one, 7% had a distance of two, and only 2% had a distance of three, indicating that attacks are overwhelmingly characterized by minimal variations in usernames and are less likely to involve an edit distance > 3. This confirms our expectations in § 2.2.3 about how PROSPER attackers can easily obtain handles at a very close lexical distance to their targets due to the "less crowded nature" of personal impersonation target spaces as opposed to brand target spaces. Among the 116 attack handles that are 1 edit away, 72 had a character insertion operation, while 26 and 18 handles had a single substitution and deletion operation respectively to stay very close to the target. These close edit distances (coupled with matching profile name and pictures) make it extremely difficult for a victim to discern the attack as already exemplified in the earlier case study (Fig. 1). *We thus call upon social media networks to pay close attention to any conversations that include parties at close lexical distances as these are strong indicators of active impersonation attacks*. For example, if an API monitor as discussed above is devised to detect the fact that a trigger tweet like search is being followed up by a sock-puppet account's profile renaming and then posting an interjecting tweet (such as in Fig. 1) in to an existing conversation where there is already another account holder with a lexically close unique account handle, then it would be providing ample evidence to the social media platform operators to consider this as a case of a social engineering attack.

We next analyzed the attackers' methodology for creating impersonating account handles. For this, we first measured the relative position at which each edit operation took place within a given attacker handle. Fig. 5 shows the distribution of attacker account handles as per the relative position of their edit distance operations (w.r.t. `V-A`). The graph shows that more than 50% of the edit operations occur at the last 80% part of the handle with this trend being particularly dominant in the popular insertion operations. Fig. 1 serves as an example for this with the attackers choosing `@barrcudaa_` as a handle while targeting a `V-A` with handle `@barrcuda_`. This pattern aligns with intuitive expectations: since readers typically process text from left to right, they are more attentive to the beginning of a string and less likely to notice subtle changes toward the end. We believe that our dataset, comprising real-world impersonation handles and their corresponding target victim handles, is valuable in modeling in-the-wild attackers' strategies for conducting impersonation attacks on social media platforms.

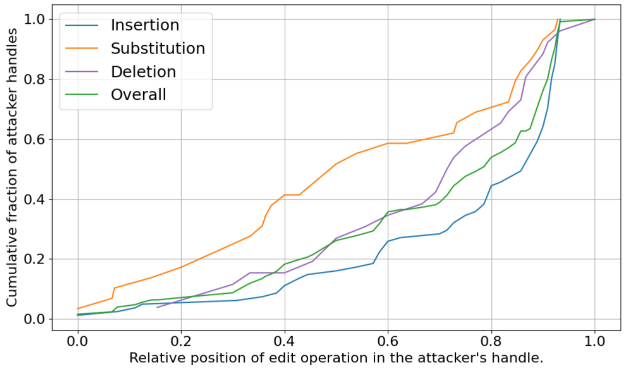


Figure 5: CDF of PROSPER attackers' handles as per the relative position of their edit distance operations.

Mobile UI deficiencies in X. We now discuss the above finding in the context of mobile devices. In X app conversations, both the profile name and the unique account handle are displayed on the same line. While profile name is displayed first prominently in black[2], the unique handle is displayed after in gray. Unfortunately, on mobile devices, often the user's profile (which is not mandated to be unique) itself consumes a lot of screen width. Given that attackers have a penchant for hosting their discernible edits in the tail-end of their account handles, the attackers' information can virtually become indistinguishable from that of the target (`V-A`). An example of this is shown in Fig. 11 in Appendix. To quantify this issue, we found out that on an iPhone 11 device (with a device size of 6.1 inch), only about 30 characters of profile name and handle combined can fit into the screen. We found that about 18% of PROSPER attacks are cases where the profile name + handle string is longer than this length, thus making it virtually indistinguishable to users of such a device. We also repeated this experiment with the wider iPhone 12 Pro Max device (6.7 inch size) that we found can accommodate 33 characters of the profile name + handle string. This implied that 13% of PROSPER attacks would still be completely imperceptible even when the victim is using such a large device. Hence, *we strongly recommend X to consider displaying the security-sensitive account handle prominently in a separate line as an impersonation attack mitigation measure.* We here by highlight that, Threads, a similar microblogging platform has chosen to only include the security sensitive account handle in their social feed UI and recommend other platforms to follow similar approaches (Fig. 12 in Appendix).

Next, we measured what characters are frequently used by PROSPER attackers for their edit operations to create an account handle. We show the results in Fig. 6 which charts the frequencies of all allowable characters in X's account handles. The histogram shows the frequencies of edit operation characters as well as the characters in impersonation targets as a baseline. From the figure, it is clear that attackers have a a preference for leveraging specific characters for their edit operations instead of selecting a character randomly from the baseline distribution. In particular, among the alphabet, "`i`" appears to be unequivocally favored likely due to its small width which helps make the difference in handles imperceptible. . It is worth noting that X's UI team chose not to use a fixed-width font for displaying account handles (see Fig. 1), which enables this issue. However, this design choice also offers a benefit as variable-width fonts can fit more characters within limited screen space—an important consideration for devices with smaller displays, such as smartphones, as previously discussed. Overall, the insertion of a "`_`" character is the most highly favored operation by PROSPER attackers. This, too, is understandable given the low visual "footprint" of the underscore character, which helps keep alterations imperceptible to victims. Our dataset which highlights such real-world modification strategies employed by attackers, will thus be helpful to researchers developing future defenses.

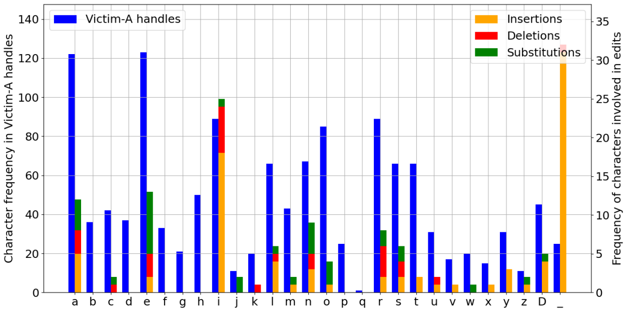


Figure 6: Character frequency in edit operations of attackers (right axis) vs. baseline character frequency from `V-A` account handles (left axis). D represents all digits.

3.3 PROSPER attack campaigns

As can be noted from Table 2 only 70 distinct payment accounts of attackers were seen among the 127 PROSPER attacks we collected. This clearly shows that attackers are reusing their PayPal accounts across attacks. This is understandable as payment accounts are typically linked to their bank accounts and are hence harder to re-create than social media accounts. Thus, we refer to them as "campaigns', where each campaign refers to a single payment account belonging to the attacker being used across multiple social media accounts that they own for larger-scale orchestration of these attacks. For clarity, if we do not evidence for a particular payment account being used as a PROSPER attack vector by more than one social media account, we do not consider this as a campaign. As our visibility was limited to only a 3-month period, we attempted to leverage a search engine cache (Google) to expand on our dataset following an idea explored in prior work for obtaining social media data [32]. The key idea here was to utilize the less agile payment accounts as a search dorking method to find more instances of PROSPER attacks. We manually performed these searches, collected the tweets and their timestamps, and de-duplicated the data.

With the above process, we collected an additional 54 attack tweets bringing the total number of PROSPER attacks we collected to 181. Our analysis also showed the presence of 41 different campaigns indicating 41 payment accounts being used with more than 1 X account for launching PROSPER attacks. One particular payment account was used to launch PROSPER attaacks from as many as 8 different X accounts. Further, we noticed that these 54 additional attacks tweets comprised of 3 languages other than English (French, German, Spanish) thus showing the potenially global nature of these attacks. We also measured the lifetime of each of these

[2] This is for X App with a white background

campaigns as can be judged by the first-seen and last-seen timestamps and found that 30 campaigns had a lifetime of more than 3 days. The longest campaign had a lifetime of 53 days. Fig. 7 shows this distribution. *We thus strongly recommend social media companies to utilize payment account IDs (e.g., Paypal account IDs) in* PROSPER *attacks as an early identifier*. This approach aligns with a recent paper that proposed adoption of "merchant IDs" during payments for early detection of fraudulent e-commerce websites [9]. Further, working collaboratively with payment platforms such as PayPal can help in more long-term mitigation as opposed to suspending social media accounts.

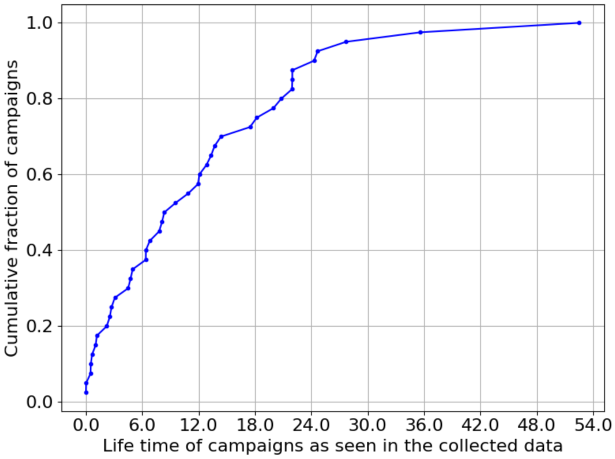


Figure 7: CDF for Lifetime (days) of PROSPER campaigns

Other measurements. Finally, we also measured the follower counts, followee counts, tweet counts, and like counts of both attackers and victims to better understand their activity levels and strategies. There was a significant disparity in follower counts, with victims having a median of 7,633 followers compared to a median of just 7 for attackers. Victim accounts were generally well-established, with high activity levels as evidenced by their median tweet count of 9,249 and median like count of 19,406 (Table 4). In contrast, attacker accounts displayed minimal activity, with a median of just 5 tweets, 2 likes, and 7 followees (Table 3).

Interestingly, despite their minimal activity, attackers appear to deliberately engage in a few interactions such as following other accounts, liking posts, or tweeting, before initiating impersonation. This behavior likely serves to make their accounts appear more realistic and less suspicious when viewed by victims or bystanders. Further, our recheck of the 127 attacker accounts after more than a month's time revealed that while 72 had been voluntarily deleted or suspended, 54 remained undetected, highlighting the need for improved detection mechanisms for such accounts. We present a CDF of follower counts for victim accounts in Fig. 8, which illustrates the wide variation in their social presence.

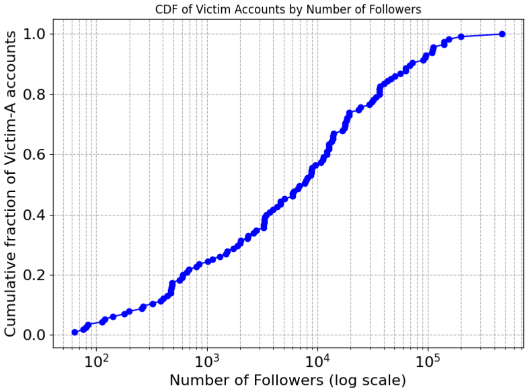


Figure 8: CDF for follower counts of impersonation targets (V-As)

Metric	Mean	Min	Max	Median
Followers Count	4	0	61	1
Followings Count	13	0	186	7
Tweet Count	6	1	85	5
Like Count	7	0	121	2

Table 3: Statistics of Scammer Accounts and Their Lifespan

Metric	Mean	Min	Max	Median
Followers Count	26249	63	469937	7633
Followings Count	1773	0	20323	566
Tweet Count	37471	46	1263271	9249
Like Count	53048	27	427813	19406

Table 4: Statistics of Victim A' Accounts and Their Lifespan

4 Qualitative Analysis of PROSPER attacks

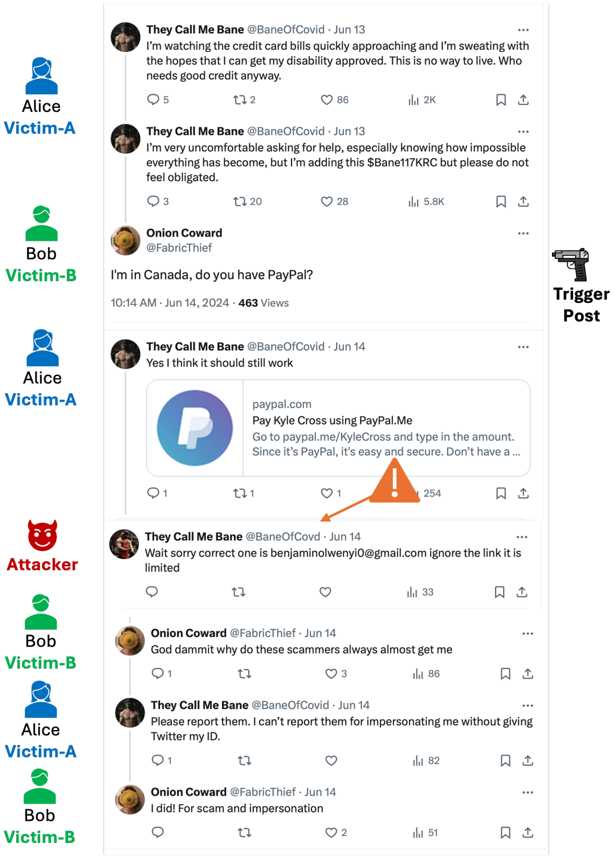
Having performed measurement analysis of the attack data we collected, we next wanted to pursue qualitative analysis to gain more understanding of the collected conversational data. In particular, given that ours is the first study to focus on in-the-wild SE attacks in the personal impersonation space, we hoped that this analysis will yield new insights useful for scam mitigation on social media in general. To do this, two authors acted as coders and used a qualitative analysis approach based on open coding [14]. In particular, two coders used open-coding to devise codes that characterize various aspects of PROSPER attack conversational data we collected. Intermittently, the coders met to discuss their codes and performed axial coding. After the code book has been finalized, 10% of untouched conversational data was used to estimate inter-rater reliability using Cohen's kappa. This yielded a value of 0.9 indicating very high agreement and stability of the code book [26]. The full code book is presented in Appendix B. We now discuss some salient findings from this analysis. Unlike the previous section, all results discussed here are based on the data resulting from the qualitative analysis described above.

V-A category. One of the goals of our analysis was to understand what types of victims fall prey to PROSPER attacks. Our findings revealed multiple categories, with the most common group of targeted user accounts (30 out of 115) consisting of lifestyle influencers and models who post their photos to attract followers and receive donations. Another significant group of V-A accounts (28 out of 115) featured more explicit image content, including pornographic material. Many of these accounts referenced or alluded to "Findom" an internet-based sexual fetish in which one person (V-B) willingly submits to financial domination by another (V-A) in exchange for attention [38]. This dynamic often involves the submissive participant (V-B) paying regular tributes, which in turn can lead to trigger tweets and PROSPER attacks on them as we discovered in this study.

Beyond this, there were also accounts seeking funds for providing services in creative fields such as art or music (14/115). In these cases, the users create art (e.g., a doodle sketch) on demand for a small sum of money. Other categories that emerged in our coding process comprised: politically oriented accounts that were seeking donations for organizing activities such as rallies (11/115); fan accounts for celebrities (8/115); spiritually oriented accounts offering paid services (e.g., fortune-telling - 5/115); video-gaming content accounts (4/115); and pet animal accounts (2/115). We were able to categorize the remaining accounts (13/115) as these accounts were unavailable by the time at which we conducted our analysis thus preventing us from understanding their nature by analyzing their posting trends.

When doing the above, our analysis also yielded a finding that 45 of the 115 accounts seemed to regularly seek donations from their follower as opposed to one-off unsolicited donations made by their followers. On the other hand, 38 accounts regularly advertised paid services on social media to their followers (i.e. funds in exchange for a service as exemplified above). Further, in terms of the funds exchange process during each captured attack (127 attacks), we noticed that in 51 out of 127 cases, V-A was seeking a donation and in 24 cases V-A was offering a service prior to the trigger tweet from V-B. On the other hand, in 26 cases, V-B was making an unsolicited donation. A representative example of V-B making an unsolicited donation is that of a lifestyle influencer's follower who presumably donates to strengthen a connection with the influencer whose posted image they appear to appreciate.

Adaptive Impersonation Tactics. This analysis also allowed us to understand the impersonation tactics employed by attackers from a qualitative angle and observe any new themes emerging from it. Notably, we saw that in most cases (117/127), the attacker has responded to the trigger tweet even before V-A has had a chance to reply. This is understandable given the high speed of attacks we have noted in the previous section driven by the abusive use of APIs. Interestingly, however, we noticed 10 cases where the attacker responded even after V-A's response. Fig. 9 illustrates an interesting social engineering tactic that attacker employs where they reply with a message of this format: *"Wait sorry correct one is [ATTACKER ACCOUNT ID]"* and *"...ignore the above [V-A ACCOUNT ID]"*. The attacker's goal here is to claim some functional issue with the content of the legitimate message sent by V-A and then convince the victim V-B to instead pay to the attacker's account.



They Call Me Bane @BaneOfCovid · Jun 13
I'm watching the credit card bills quickly approaching and I'm sweating with the hopes that I can get my disability approved. This is no way to live. Who needs good credit anyway.
5 2 86 2K

They Call Me Bane @BaneOfCovid · Jun 13
I'm very uncomfortable asking for help, especially knowing how impossible everything has become, but I'm adding this $Bane117KRC but please do not feel obligated.
3 20 28 5.8K

Onion Coward
@FabricThief
I'm in Canada, do you have PayPal?
10:14 AM · Jun 14, 2024 · **463** Views

They Call Me Bane @BaneOfCovid · Jun 14
Yes I think it should still work
paypal.com
Pay Kyle Cross using PayPal.Me
Go to paypal.me/KyleCross and type in the amount. Since it's PayPal, it's easy and secure. Don't have a ...
1 1 1 254

They Call Me Bane @BaneOfCovd · Jun 14
Wait sorry correct one is benjaminolwenyi0@gmail.com ignore the link it is limited
33

Onion Coward @FabricThief · Jun 14
God dammit why do these scammers always almost get me
1 3 86

They Call Me Bane @BaneOfCovid · Jun 14
Please report them. I can't report them for impersonating me without giving Twitter my ID.
1 82

Onion Coward @FabricThief · Jun 14
I did! For scam and impersonation
2 51

Figure 9: An example illustrating a PROSPER attacker's clever social engineering technique when responding even *after* V-A's response. Here, they claim that the payment link sent by V-A has limited functionality, and share with V-B their own PayPal account information as the legitimate one.

PayPal's "Friend and Family Option". Further, we noticed that most attackers tend to recommend V-Bs to use PayPal's "Friends and Family" option to make their payments (89/127). PayPal states that this feature is only intended for sending money to close contacts, such as splitting bills or giving gifts as it is not covered by their "Purchase Protection" program [30]. However, our findings here show that PROSPER attackers abuse this mechanism to likely ensure that the victim cannot take any action to reverse the transaction. When attempting to make a test transaction to one of the scammers' accounts via this option, we noticed that there is no warning from PayPal about this ongoing social engineering attack tactic. *Given the large prevalence of this tactic, we thus recommend PayPal to consider warning their customers whenever they appear to be using this option for making payments*. This is also in the best interest of the payment companies as there is a possiblity that they might be held legally responsible for being an unwitting partner in the scams by agencies such as the Federal Trade Comission (FTC) in the U.S. This is similar to a lawsuit filed by the FTC against brick and mortar grocery stores for facilitating scam transactions [1].

Other impersonation tactics. Some other tactics we observed was the attackers' preference for wanting victims (V-Bs) to confirm when the transfer has been made (20/127). We presume they do this as this allows them to delete their attack tweets immediately and increase the longevity of their accounts by preventing post-mortem reporting. We have also observed another evasion tactic, where the scammers introduce deliberate misspellings such as "PyPl" likely as a mechanism to hide their messages (5/127). Finally, we also noted follow-up messages from the attackers to V-Bs using emotional coercion tactics (such as feigning hunger) to persuade them to make the money transfer.

Post-attack events. We also analyzed the post-attack events. Since this is a publicly performed social engineering attack which potentially takes place in the presence of onlookers, we wanted to know if any of them intervene. Our analysis shows that this is indeed the case with at least 12/127 cases where a "Good Samaritan" account–that has not been previously part of the conversation–tries to warn V-B about this ongoing attack. The final tweet in Fig. 1 is an example of such a user. Unlike one-on-one social engineering attacks, such as web-based credential phishing via emails, this active case of onlooker intervention during a public social media attack, presents an interesting new angle for security researchers. Future research can focus on both enhancing such societal intervention and recognizing interventional message patterns (i.e., a fourth-party individual responding to an attacker's message in V-A's thread) to help combat similar attacks.

We also wanted to use these post-attack messages to infer whether the attacker had successfully stolen the payment or not. Unfortunately, in most social engineering attacks, the victim themselves does not know that they fell for the attack. Thus, only in a handful of cases in our dataset, we can attempt to four cases V-B has themselves commented about the attack without indicating any transfer of funds taken place thus likely indicating failure of the attack (e.g., Fig. 9). Yet, on the other hand, in four other cases, V-B has commented about having completed the money transfer (e.g., Fig. 1), thus confirming the success of the attack.

5 Generalizing beyond X-AM data

The PROSPER attacks we collected and analyzed in the previous sections has thus far been limited to real-time data collected from the X social media platform over a 12-week deployment period. While this analysis yielded actionable insights for security, it is important to investigate how well this data generalizes to different deployment periods as well as other platforms. Although we acknowledge some limitations in generalizability and discuss them further in § 7, in this section, we attempt to address this issue more directly through supplementary data analysis. To that end, we present results from an additional experiment to analyze historical PROSPER attacks on X as well as an ad hoc case study focused on BlueSky, an emerging microblogging platform similar to X.

5.1 Historical PROSPER attacks

The "Basic" tier of X API that we utilized for collecting real-time data unfortunately does not allow us to collect historical data as it limits data to only a 7-day historical period. Thus, we attempted to devise an alternative cost-effective approach to collect 12 months of historical PROSPER attack data immediately preceding the 12-week period for the attack data obtained by X-AM. For this, we simply leveraged dummy X accounts and retrieved the content from the performed searches via a browser extension. We utilized a "human-in-the-loop" approach when developing the extension so as to abide by X's terms of service while still collecting the narrowly focused PROSPER attack that we are interested in collecting over a 12-month period.

For this, we converted the trigger tweet search query we devised via the user study described earlier into a pattern that complies with the search operators for X's website. We then manually searched for weekly intervals of historical data and utilized the extension to store the collected data. We performed these searches for the 12-month period from June 2023 to May 2024. The manual collection process took about 3 working days (about 24 hours) to complete.

In total, we collected a total of 2,466 tweets in this manner. However, as might be expected, most of the replies to these tweets did not feature any impersonation attacks due to the fact that attackers tend to either actively delete the messages or the messages get reported and subsequently deleted. Further, many attackers' accounts also tend to get suspended or deleted as seen earlier thus impeding collection of attack tweets. Our approach to address this was to leverage potential post-scam activity that we discussed previously in § 4, such as tweets made by V-A or V-B warning about the scam, or those made by "Good Samaritans", to help identify instances of PROSPER attacks. Using this method, we were able to identify 123 additional instances of PROSPER attacks targeting impersonation of 123 V-A accounts over the 12-month period. This finding, despite the non-real time nature of the collected data, clearly demonstrates that PROSPER attackers have been active for an extended period on X.

We then also subjected this data to qualitative analysis utilizing the code book derived from the previous section (but omitting the attack tweet-dependent parts as these were often not present). Interestingly, our derived codebook was largely applicable to this historical data as well indicating stability in our findings from qualitative study. Specifically, in terms of `V-A` categories, we were able to analyze the nature of posts from 81 out of 123 targeted accounts while we missed others due to reasons such as account deletion. Among these 81 accounts, the codes from our pre-existing codebook (as enumerated in Appendix B), were applicable to 74 accounts (91%) with the remaining 7 belonging to new themes such as sports and climate activism that we have not seen in our previous dataset. This dataset was dominated by artist accounts who comprised 29 targets. Interestingly, in this dataset, we noticed that in 46/123 cases, a good samaritan tried to warn about the ongoing PROSPER attack even surpassing the 41/123 cases where the impersonation target i.e. `V-A` themselves tried to warn about the attack. This shows the pivotal and active role that community members seem to be playing in thwarting such public social engineering attacks further motivating research on this topic in the future.

5.2 Case study: Bluesky

To inspect generalizability across platforms, we also performed a small ad hoc experiment on Bluesky, an alternative microblogging social media platform. Utilizing Bluesky's search API which allows historical searches, we first measured the number of Trigger Tweets across time using the same query pattern as our earlier experiments. Our results showed that there were 66 such posts in 2023 followed by growing numbers of 307 and 507 posts in 2024 and 2025 respectively. The number of 507 trigger posts in 2025 is particularly notable given that our search was conducted in April 2025 (less than 4 months) indicating a growing attack surface on Bluesky. This might be attributable to the recent surge in growth of their userbase [16].

Following this, we also wanted to investigate if this growing attack surface has begun to evoke the interest of PROSPER attackers to target users on this platform. For this, we performed a few ad hoc searches on Bluesky's web user interface to find examples of any prevalent attacks. Interestingly, we came across more than 20 PROSPER attacks in our manual searches few of which we share here as representative examples [4–7, 10, 11]. We observed similar qualitative characteristics of PROSPER attacks in these attacks. For example, in this attack instance [7], the attacker uses a single letter edit distance from `V-A`'s account handle and uses a typo-based evasion ('Paypl') to convey their account handle. They also encourage payment via the "Friends and family" option. Interestingly, most attacks we encountered involved active interventional participation of fourth-party onlookers as illustrated in Fig. 13 in Appendix.

6 Recommendations

Mitigating PROSPER attacks requires a multi-stakeholder approach. Below, we outline targeted recommendations for social media platforms and payment platforms based on the findings presented in this paper.

6.1 Social media platforms

Restrict account handle changes. As discussed in § 3.1, attackers exploit the ability to rapidly modify account handles to impersonate victims. We recommend that platforms like X limit the frequency of handle changes or require verification steps for such changes. This will significantly hinder attackers' ability to repurpose accounts quickly for PROSPER attacks.

Integrated impersonation monitoring and detection system. Social media platforms should implement a unified system that combines API usage monitoring, real-time user interaction analysis, and impersonation detection to mitigate PROSPER attacks. This system would monitor API usage patterns to identify anomalous trends indicative of automated tools used for real-time trigger tweet searches (Sec § 2.2.2). Concurrently, it should flag interactions involving accounts with minimal lexical differences in handles, especially when combined with identical profile names or pictures (§ 3.2). The datasets we collect in this research would also be useful for modeling purposes in such approaches. By correlating these patterns with evidence such as rapid profile changes and immediate responses to trigger tweets, the system can warn users of potential impersonation attempts and proactively suspend accounts exhibiting strong indicators of scamming behavior. Leveraging machine learning, this comprehensive approach ensures that suspicious users are flagged based on concrete evidence, effectively disrupting attacker operations while minimizing false positives.

UI Enhancements for account handle display. As demonstrated in § 3.2, attackers exploit the limited display space on mobile devices to obscure critical differences between legitimate and impersonating accounts. To address this, platforms should prioritize two key improvements. First, account handles should be fully displayed on mobile interfaces, ensuring that no portion of this unique identifier is truncated, even on smaller screens. Second, the account handle should be visually distinguished by highlighting it prominently and avoiding the current practice of graying it out. Since account handles are unique identifiers, unlike profile names which can be duplicated, emphasizing them can help users more reliably differentiate between legitimate and impersonating accounts, thereby reducing the success rate of impersonation attacks.

6.2 Payment platforms

Warnings for High-Risk Transactions. In § 4, we found that attackers encourage victims to use PayPal's "Friends and Family" option to bypass buyer protection. Payment platforms should introduce warnings for transactions using high-risk options, especially when triggered by unusual patterns such as payments to recently created accounts. Currently, PayPal doesn't provide any special warnings when using the "Friends and Family" option to send or receive money [31].

Holistic multi-stakeholder collaboration for mitigation. Effectively combating PROSPER attacks requires coordinated efforts between payment platforms and social media platforms. First, as highlighted in § 3.3, attackers often heavily reuse payment account IDs making shared blocklists between social media and payment platforms a critical tool for rapid detection and mitigation. Collaborative data-sharing agreements, with appropriate anonymization, can streamline this process and foster innovative approaches to attack prevention.

Second, while user awareness campaigns have been a staple for many years, payment platforms should enhance their strategies by incorporating real-time, context-aware alerts. For instance, instead of generic educational content, platforms can issue personalized warnings when users initiate high-risk transactions, such as those involving flagged accounts or payment methods without buyer protection (§ 4). These dynamic and situational alerts can better capture user attention and provide actionable guidance, addressing specific risks associated with PROSPER attacks more effectively than traditional awareness campaigns.

7 Limitations

Although our study of PROSPER attacks, a new API-based, personal impersonation attack on social media yielded several actionable security insights for both social media and payment platforms, it is important to acknowledge its potential limitations, which we outline below.

Scale of the study and recall. Compared to other recent social engineering attack measurement studies [2, 35, 45], our study of PROSPER attacks is smaller in scale. Thus, it is important to dicuss about potential generalizability of our findings over time and other platforms. For this, we note that our results—though based on a 12-week real-time deployment—qualitatively align with historical data collected over a one-year period (§ 5.1) and extend to other platforms such as Bluesky (§ 5.2), supporting the broader applicability of our findings across various instances of PROSPER attacks.

Another limitation is our inability to generalize to all potential PROSPER attacks, thus precluding accurate estimation of their global scale. This stems in part from our design choice to use only English-language search queries from a pool (n=50) of student participants. While user responses showed saturation at 68%, this approach has limited demographic diversity making the global representativeness of our attack measurements less certain. Ideally, we would have followed strategies like those used by Goga et al. [18], but were constrained by recent API pricing changes [32, 37]. Nevertheless, after bootstrapping with targeted queries, our pipeline was able to monitor all matching candidate trigger tweets globally over the 12-week period.

Due to budgetary constraints, we conducted this monitoring using X's more economical "Basic API" plan, polling conversations at 5-minute intervals. As our results show a median attacker response time of approximately 4 minutes, and considering the additional time needed for the victim `V-B` to see and act on the message, we estimate that instances where attackers could fully execute and delete their posts before detection are negligible.

Finally, while our English-only query set excluded non-English PROSPER attacks, our data expansion efforts using search engines uncovered similar tactics in three non-English languages (see § 3.3), suggesting that the strategies we identified are not language-dependent and generalize across linguistic contexts.

Though financially small, the persistence of PROSPER attacks across time shows how effective the issues we found in the social media platforms are. Left unaddressed, they could enable broader impersonation threats in future. As platforms like X move toward native payment services, the risk grows, making it crucial to patch such vulnerabilities early before larger-scale abuse occurs.

Generalizability to other platforms. Our measurement study was limited to a single, albeit widely popular, social media platform with hundreds of millions of daily users: X. While we have observed instances of PROSPER attacks on Bluesky through ad hoc experiments (§ 5.2), this platform is also highly similar to X. As such, it is to be noted that the feasibility and effectiveness of PROSPER attacks on any given social media platform will depend on a combination of UI vulnerabilities and platform-specific features, which we now discuss.

First, the core mechanism of PROSPER attacks involves intercepting communication between two users to exploit the exchange of payment account IDs. On X (and similar platforms such as Bluesky), attackers can do this by quickly locating such tweets (referred to as "Trigger Posts") via their APIs. It is important to note that these Trigger Posts are typically made by users (`V-B`) in response to an initial post by the impersonation target (`V-A`), as discussed in § 2. While microblogging platforms like X and Bluesky treat posts as well as their replies with equal visual and functional prominence, other social media platforms such as Reddit or Facebook are fundamentally different. For example, search APIs on these platforms primarily return main posts but not replies or comments, which is where the Trigger Posts appear, making it significantly harder for attackers to locate the target conversations for executing PROSPER attacks.

Another critical factor is X's unique handling of account identity. Unlike other platforms that prominently display only a single username, X showcases both a profile name and an account handle, where profile names can be duplicated across accounts, providing attackers with more opportunities to exploit visual similarities for impersonation. Moreover, Reddit does not allow username changes once finalized [33], and Facebook and Instagram enforce cooldown periods (e.g., 60 days) between username changes [17]. In contrast, X permits users to change their account handles without restrictions [41], making it easier for attackers to quickly adapt their impersonation strategies.

Mobile UI Analysis Scope. Our analysis of mobile UI issues that enabled PROSPER attacks (§ 3.2) was conducted only on two specific iPhone devices and therefore might not generalize to all different types of mobile devices in usage. As device form factors and layouts vary across different types of mobile devices (e.g., phones and table devices) and operating systems (e.g., iOS and Android), our findings might not apply exactly to other cases. Yet, our results which were based on iPhone devices with two different screen sizes (6.1 inch and 6.7 inch) represent a wide range of iOS devices capturing more than 50% of market share in the USA [34]. Thus, our results need to be considered as a significant illustrative example that warrants further investigation by social media app developers to mitigate issues.

Efficacy of Proposed Mitigations. While we offer concrete recommendations for social media and payment platforms, we did not experimentally validate the real-world effectiveness or usability impact of these interventions. Future work should involve controlled user studies and pilot deployments to assess how these mitigations influence attacker behavior, false negative and positive rates, and user experience. By empirically evaluating our proposals, stakeholders can fine-tune implementation parameters and balance security benefits against potential disruption to legitimate users.

Finally, we state that while these limitations exist, they do not detract from the significance of our findings and our recommendations to mitigate these attacks as well as future attacks that might exploit the same vulnerabilities that we explored. Overall, our study provides actionable insights to the security community by studying a niche, new type of social engineering attack prevalent on social media networks.

8 Related Work

Impersonation attacks, particularly those targeting brands and high-profile entities, have been extensively studied. This section reviews prior research on brand and generic social media impersonation, credential-based impersonation, and user-focused defenses, highlighting their gaps in addressing payment re-routing attacks such as PROSPER.

Brand impersonation attacks. Brand impersonation remains a prominent phishing vector, with studies exploring fake websites mimicking banks, e-commerce, and email services [35, 45] and social media scams that pose as legitimate brands for cryptocurrency support [3], profile squatting [2], or celebrity fraud [23, 43]. These studies demonstrate the diversity of brand-focused impersonation tactics but primarily center on scams targeting large, recognizable entities.

Generic social media impersonation. While brand impersonation is well-documented, personal and generic social media impersonation has received very little attention which has been the subject of our study. In this direction, Goga et al. [18] and follow-on work [20, 21] show that many impersonating accounts merely inflate followers without direct monetization.

Efforts to detect generic impersonating accounts have proposed methods that rely on pairwise comparisons between profiles, such as analyzing similarities in usernames, profile pictures, and account details [22, 44]. However, these approaches face significant scalability challenges, as they require exhaustive comparisons across large datasets. Additionally, such methods are prone to false positives, which limit their practicality in real-time settings.

Impersonation via stolen data. Beyond social media, attacks involving the theft and resale of user credentials and cookies have been documented [12, 13]. These attacks often involve user impersonation by purchasing stolen user profiles from cybercrime markets, enabling impersonation at scale.

User perceptions and platform defenses. User studies reveal that authenticity indicators like X's "blue checkmark" are often misunderstood. For example, 80% of participants misjudge verification criteria [42] undermining their protective value. While features like verification badges aim to enhance trust, their utility in mitigating impersonation remains an area for further exploration.

9 Conclusion

Personal impersonation attacks on social media exploit trust and real-time public interactions to redirect digital payments, posing a growing threat to everyday users. This paper presented the first in-depth study of PROSPER attacks, tracking 181 cases over 3 months and uncovering 70 distinct payment accounts and evidence of attacker persistence through account reuse. By combining quantitative and qualitative analyses, we revealed the scale, adaptability, and nuanced tactics of attackers, including evasion strategies and real-time victim targeting. These insights informed actionable recommendations, such as UI enhancements, stricter account management, shared blacklists, and anomaly-based detection systems, to mitigate these attacks. Our findings provide a foundation for addressing personal impersonation scams, offering practical guidance for social media platforms and payment providers to enhance user protection and reduce financial exploitation.

Acknowledgments

We sincerely thank the anonymous shepherd and reviewers for their feedback that helped improve our paper. We acknowledge the support of Ali Ghosn who helped us setup the infrastructure required for conducting the experiments. We also thank all our user survey participants that helped us bootstrap our data collection process. This work was supported by the National Science Foundation (NSF) under grant CNS-2422035.

Ethics considerations

Our research was conducted with strict adherence to ethical guidelines, ensuring that our methodology upheld principles of respect, privacy, and responsible research practices. Below, we outline the key ethical aspects of our study:

Methodology and data handling. We carefully designed our research to ensure that no private or sensitive information was collected at any stage. All data used in this study was sourced from publicly available information on X, focusing solely on observable interactions related to PROSPER attacks. We fully complied with X's terms of service and API usage policies, ensuring that our experiments imposed minimal load on the platform's infrastructure. Furthermore, we strictly avoided any form of interaction with live user accounts, preserving the integrity and privacy of all platform users.

Purpose and impact of research. Our primary goal is not to amplify or propagate PROSPER attacks but to shed light on this ongoing and long-standing issue. By analyzing attacker behaviors and identifying weaknesses in existing systems, we aim to raise awareness among social media platforms, payment providers, and benign users. The insights presented in this study are intended to inform and motivate stakeholders to take actionable steps to prevent such attacks, ultimately protecting users from financial exploitation.

Disclosure. As part of our commitment to responsible research and ethical practice, we have shared our findings on PROSPER attacks with X and PayPal—providing X with collected data and analysis to inform mitigation of on-platform impersonation, and supplying PayPal with identified attacker payment account IDs to support investigation and potential account action. We will also privately notify the victims identified in this study via X's APIs to alert them that they were targeted by a social engineering attack and help them take preventative measures against future exploitation.

Balancing risks and benefits. This research inherently involves analyzing malicious behaviors; however, we took significant precautions to ensure that our work does not inadvertently aid attackers. While our paper provides detailed descriptions of query searches and system mechanisms, these are intended to benefit social media platforms by offering actionable insights for monitoring and detecting PROSPER attacks. By making these details available, we aim to empower platforms to proactively identify and mitigate malicious activities rather than introduce novel techniques for exploitation.

The potential benefits of this study such as reducing financial harm, improving platform security measures, and protecting users, significantly outweigh any minimal risks associated with analyzing and sharing publicly available data. Our approach prioritizes transparency and collaboration to drive meaningful advancements in addressing PROSPER attacks.

Commitment to responsible research. Throughout this study, we prioritized transparency, accountability, and adherence to ethical principles. By focusing on actionable recommendations and responsibly reporting our findings, we aim to contribute positively to the broader understanding and prevention of social engineering attacks on social media platforms.

Open science

To promote transparency and facilitate future research, we release the following artifacts:

1. X-AM code, browser extension code, quantitative analysis code, and the complete codebook from the qualitative analysis.
2. Processed, anonymized datasets containing non-personally identifiable information derived from publicly available information on X.
3. A detailed guide for replicating our experimental setup, including API configurations and query parameters.

Due to privacy and ethical concerns, we do not share raw data containing user-generated content publicly. Releasing raw conversational data, such as tweets, poses a significant risk of indirectly or directly identifying victims. For instance, even if we were to release only the text of tweets, an adversary could use the content to locate the original posts by performing a simple search on the platform, potentially exposing the victims' profiles and interactions. This could lead to unintended consequences, such as further harassment or exploitation of the individuals involved.

To balance the need for transparency with the obligation to protect user privacy, we will processed and anonymized datasets. These will include aggregate statistics and generalized patterns derived from the raw data, ensuring that our findings remain verifiable while eliminating any possibility of tracing back to specific users. This approach aligns with our commitment to ethical research and protects the rights and privacy of all individuals impacted by the study. This data is available here: 10.5281/zenodo.15611471

Yet, in order to effectuate the benefits afforded by our research, such as modeling attacker behavior, we additionally share all raw data collected in our study, including unanonymized victim and attacker account IDs, with vetted security researchers and industry practitioners. To facilitate this, we host our data on Zenodo, using their "Restricted Access" feature. A link to this repository is available here: 10.5281/zenodo.15611502.

References

[1] Walmart let scammers use money transfer services to fleece people, FTC says., 2024. Accessed: 2024-10-14. URL: https://www.cbsnews.com/news/walmart-ftc-western-union-moneygram-ria-money-transfer-services-scam/.

[2] Bhupendra Acharya, Dario Lazzaro, Efrén López-Morales, Adam Oest, Muhammad Saad, Antonio Emanuele Cinà, Lea Schönherr, and Thorsten Holz. The imitation game: Exploring brand impersonation attacks on social media platforms. In Davide Balzarotti and Wenyuan Xu, editors, *33rd USENIX Security Symposium, USENIX Security 2024, Philadelphia, PA, USA, August 14-16, 2024*. USENIX Association, 2024. URL: https://www.usenix.org/conference/usenixsecurity24/presentation/acharya.

[3] Bhupendra Acharya, Muhammad Saad, Antonio Emanuele Cinà, Lea Schönherr, Hoang Dai Nguyen, Adam Oest, Phani Vadrevu, and Thorsten Holz. Conning the crypto conman: End-to-end analysis of cryptocurrency-based technical support scams. In *IEEE Symposium on Security and Privacy, SP 2024, San Francisco, CA, USA, May 19-23, 2024*, pages 17–35. IEEE, 2024. doi:10.1109/SP54263.2024.00156.

[4] archive.today. Bluesky prosper attack example, 2024. URL: https://archive.ph/6jFc8.

[5] archive.today. Bluesky prosper attack example, 2025. URL: https://archive.ph/7SPu6.

[6] archive.today. Bluesky prosper attack example, 2025. URL: https://archive.ph/7SPu6.

[7] archive.today. Bluesky prosper attack example, 2025. URL: https://bsky.app/profile/bejeweledrecov.swifties.social/post/3llwgrityx22y.

[8] Ben Chukwuemeka Benjamin, Jan Bayer, Simon Fernandez, Andrzej Duda, and Maciej Korczyński. Shielding brands: An in-depth analysis of defensive domain registration practices against cyber-squatting. In *2024 8th Network Traffic Measurement and Analysis Conference (TMA)*, pages 1–11. IEEE, 2024.

[9] Marzieh Bitaab, Alireza Karimi, Zhuoer Lyu, Adam Oest, Dhruv Kuchhal, Muhammad Saad, Gail-Joon Ahn, Ruoyu Wang, Tiffany Bao, Yan Shoshitaishvili, and Adam Doupé. SCAMMAGNIFIER: piercing the veil of fraudulent shopping website campaigns. In *32nd Annual Network and Distributed System Security Symposium, NDSS 2025, San Diego, California, USA, February 24-28, 2025*. The Internet Society, 2025. URL: https://www.ndss-symposium.org/ndss-paper/scammagnifier-piercing-the-veil-of-fraudulent-shopping-website-campaigns/.

[10] bsky.app. Bluesky prosper attack example, 2025. URL: https://bsky.app/profile/rhymeswithk.bsky.social/post/3lkmraav5g22t.

[11] bsky.app. Bluesky prosper attack example, 2025. URL: https://bsky.app/profile/robert-e-630.bsky.social/post/3lhd3qjn4b225.

[12] Michele Campobasso and Luca Allodi. Impersonation-as-a-service: Characterizing the emerging criminal infrastructure for user impersonation at scale. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 1665–1680, 2020.

[13] Michele Campobasso and Luca Allodi. Know your cybercriminal: Evaluating attacker preferences by measuring profile sales on an active, leading criminal market for user impersonation at scale. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 553–570, 2023.

[14] Juliet M Corbin and Anselm Strauss. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, 13(1):3–21, 1990.

[15] Fred J Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964.

[16] Dani Di Placido. Forbes: The X (Twitter) Exodus To Bluesky, Explained, 2024. URL: https://www.forbes.com/sites/danidiplacido/2024/11/19/the-x-twitter-exodus-to-bluesky-explained/.

[17] Facebook. How do i change my name on facebook? URL: https://www.facebook.com/help/448505685205813.

[18] Oana Goga, Giridhari Venkatadri, and Krishna P. Gummadi. The doppelgänger bot attack: Exploring identity impersonation in online social networks. In Kenjiro Cho, Kensuke Fukuda, Vivek S. Pai, and Neil Spring, editors, *Proceedings of the 2015 ACM Internet Measurement Conference, IMC 2015, Tokyo, Japan, October 28-30, 2015*, pages 141–153. ACM, 2015. doi:10.1145/2815675.2815699.

[19] Google Pay Help. Countries or regions where you can make payments with Google, 2024. Accessed: 2024-10-03. URL: https://support.google.com/googlepay/answer/12429287.

[20] Supraja Gurajala, Joshua S White, Brian Hudson, and Jeanna N Matthews. Fake twitter accounts: profile characteristics obtained using an activity-based pattern detection approach. In *Proceedings of the 2015 international conference on social media & society*, pages 1–7, 2015.

[21] Supraja Gurajala, Joshua S White, Brian Hudson, Brian R Voter, and Jeanna N Matthews. Profile characteristics of fake twitter accounts. *Big Data & Society*, 3(2):2053951716674236, 2016.

[22] Georgios Kontaxis, Iasonas Polakis, Sotiris Ioannidis, and Evangelos P Markatos. Detecting social network profile cloning. In *2011 IEEE international conference on pervasive computing and communications workshops (PERCOM Workshops)*, pages 295–300. IEEE, 2011.

[23] Anastasios Lepipas, Anastasia Borovykh, and Soteris Demetriou. Username squatting on online social networks: A study on x. In *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security*, pages 621–637, 2024.

[24] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

[25] Jienan Liu, Pooja Pun, Phani Vadrevu, and Roberto Perdisci. Understanding, measuring, and detecting modern technical support scams. In *8th IEEE European Symposium on Security and Privacy, EuroS&P 2023, Delft, Netherlands, July 3-7, 2023*, pages 18–38. IEEE, 2023. URL: https://doi.org/10.1109/EuroSP57164.2023.00011, doi:10.1109/EUROSP57164.2023.00011.

[26] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[27] Najmeh Miramirkhani, Oleksii Starov, and Nick Nikiforakis. Dial one for scam: Analyzing and detecting technical support scams. *CoRR*, abs/1607.06891, 2016. URL: http://arxiv.org/abs/1607.06891, arXiv:1607.06891.

[28] Hoang Dai Nguyen, Karthika Subramani, Bhupendra Acharya, Roberto Perdisci, and Phani Vadrevu. C-frame: Characterizing and measuring in-the-wild CAPTCHA attacks. In *IEEE Symposium on Security and Privacy, SP 2024, San Francisco, CA, USA, May 19-23, 2024*, pages 277–295. IEEE, 2024. doi:10.1109/SP54263.2024.00200.

[29] PayPal. We get where you're coming from., 2024. Accessed: 2024-09-04. URL: https://www.paypal.com/va/webapps/mpp/country-worldwide.

[30] PayPal. What's the difference between friends and family or goods and services payments?, 2024. Accessed: 2024-09-04. URL: https://www.paypal.com/us/cshelp/article/whats-the-difference-between-friends-and-family-or-goods-and-services-payments-help277.

[31] PayPal, Inc. What are friends and family payment scams? Accessed: 2025-01-21. URL: https://www.paypal.com/us/cshelp/article/what-are-friends-and-family-payment-scams-help1165.

[32] Amrit Poudel and Tim Weninger. Navigating the post-api dilemma. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee, editors, *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 2476–2484. ACM, 2024. doi:10.1145/3589334.3645503.

[33] Reddit. Can i change my username? URL: https://support.reddithelp.com/hc/en-us/articles/204579479-Can-I-change-my-username.

[34] Statcounter. Mobile Vendor Market Share United States Of America. URL: https://web.archive.org/web/20250515123920/https://gs.statcounter.com/vendor-market-share/mobile/united-states-of-america.

[35] Karthika Subramani, William Melicher, Oleksii Starov, Phani Vadrevu, and Roberto Perdisci. Phishinpatterns: measuring elicited user interactions at scale on phishing websites. In Chadi Barakat, Cristel Pelsser, Theophilus A. Benson, and David R. Choffnes, editors, *Proceedings of the 22nd ACM Internet Measurement Conference, IMC 2022, Nice, France, October 25-27, 2022*, pages 589–604. ACM, 2022. doi:10.1145/3517745.3561467.

[36] Tom-chris Emewulu. PayPal Statistics and Facts 2024, 2024. Accessed: 2024-10-14. URL: https://www.chargeflow.io/blog/paypal-statistics-facts.

[37] James Vincent. Twitter just closed the book on academic research, 2023. Accessed: 2024-09-03. URL: https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research.

[38] Wiki. Financial domination, 2024. Accessed: 2024-09-04. URL: https://en.wikipedia.org/wiki/Financial_domination.

[39] X Developer Platform. X api documentation, 2024. Accessed: 2024-09-03. URL: https://developer.x.com/en/products/x-api.

[40] X Help Center. Help with username registration, 2024. Accessed: 2024-10-03. URL: https://help.x.com/en/managing-your-account/x-username-rules.

[41] X Help Center. How to change your x username, 2024. Accessed: 2024-09-04. URL: https://help.x.com/en/managing-your-account/change-x-handle.

[42] Madelyne Xiao, Mona Wang, Anunay Kulshrestha, and Jonathan Mayer. Account verification on social media: user perceptions and paid enrollment. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 3099–3116, 2023.

[43] Koosha Zarei, Reza Farahbakhsh, and Noel Crespi. Typification of impersonated accounts on instagram. In *2019 IEEE 38th international performance computing and communications conference (IPCCC)*, pages 1–6. IEEE, 2019.

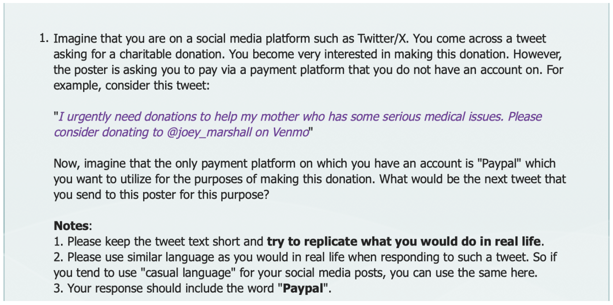
[44] Koosha Zarei, Reza Farahbakhsh, Noël Crespi, and Gareth Tyson. Impersonation on social media: A deep neural approach to identify ingenuine content. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 11–15. IEEE, 2020.

[45] Penghui Zhang, Adam Oest, Haehyun Cho, Zhibo Sun, RC Johnson, Brad Wardman, Shawn Sarker, Alexandros Kapravelos, Tiffany Bao, Ruoyu Wang, Yan Shoshitaishvili, Adam Doupé, and Gail-Joon Ahn. Crawlphish: Large-scale analysis of client-side cloaking techniques in phishing. In *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*, pages 1109–1124. IEEE, 2021. doi:10.1109/SP40001.2021.00021.

A X-AM user survey

Below is the tailored search query that we derived and used verbatim.

```
("do you have" OR "can you accept") OR
("can I send you" OR "do you take") OR
("do you use" OR "is it possible to pay with")
OR ("do you accept" OR "can you provide") OR
("can I donate via" OR "can I pay via") OR
("can I send money via" OR "you got a") OR
("send me your" OR "share your") AND
("paypal" OR "venmo") OR
("zelle" OR "cash app") OR
("apple pay" OR "google pay")
```



1. Imagine that you are on a social media platform such as Twitter/X. You come across a tweet asking for a charitable donation. You become very interested in making this donation. However, the poster is asking you to pay via a payment platform that you do not have an account on. For example, consider this tweet:

   "*I urgently need donations to help my mother who has some serious medical issues. Please consider donating to @joey_marshall on Venmo*"

   Now, imagine that the only payment platform on which you have an account is "Paypal" which you want to utilize for the purposes of making this donation. What would be the next tweet that you send to this poster for this purpose?

   Notes:
   1. Please keep the tweet text short and **try to replicate what you would do in real life**.
   2. Please use similar language as you would in real life when responding to such a tweet. So if you tend to use "casual language" for your social media posts, you can use the same here.
   3. Your response should include the word "**Paypal**".

Figure 10: The survey prompt for collecting tailored search queries from social media users (n=50).

B Code book for Qualitative Analysis

1. **Victim-A categories.** lifestyle (30), sexual (28), art-music (14), spiritual (5),video-gaming (4), fan-account (8), politics (11), pet-account (2), donation (45), service (38)

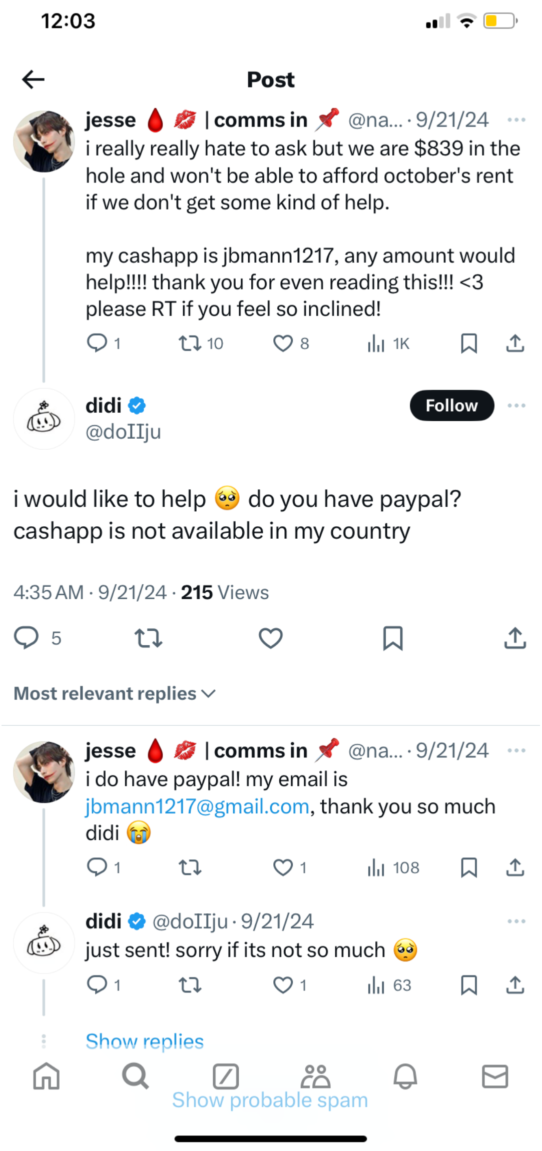
2. **Money reasons.** va-seek-donation (51), va-offer-service (24), vb-offer-donation (26)

3. **Attacker impersonation tactics.** attacker-reply-before-va (118), attacker-reply-after-va1-negative (2), attacker-reply-after-va1-positive (8)

4. **Other impersonation.** f-f-option (89), scammer-seek-confirmation (20), evasion-typo (5), emotional-coercion (10)

5. **Post attack events.** good-samaritan (12), vb-realization (4), va-warning (11), vb-warning (4)

C Additional Figures



12:03

Post

jesse 🩸💋 | comms in 📌 @na... · 9/21/24

i really really hate to ask but we are $839 in the hole and won't be able to afford october's rent if we don't get some kind of help.

my cashapp is jbmann1217, any amount would help!!!! thank you for even reading this!!! <3 please RT if you feel so inclined!

didi @doIIju

i would like to help 🥺 do you have paypal? cashapp is not available in my country

4:35 AM · 9/21/24 · 215 Views

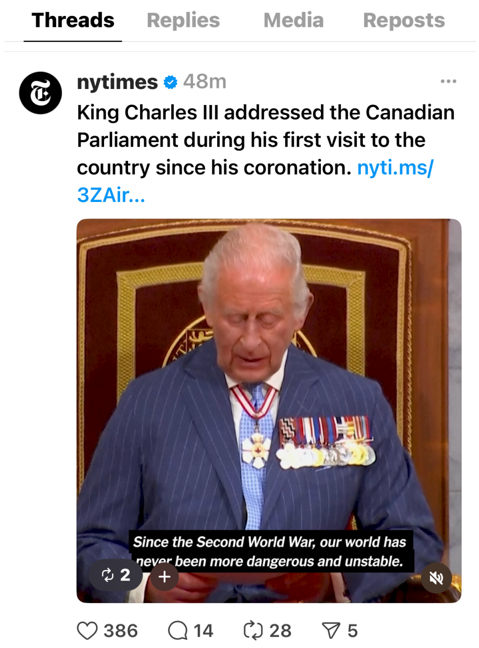
jesse 🩸💋 | comms in 📌 @na... · 9/21/24

i do have paypal! my email is jbmann1217@gmail.com, thank you so much didi 😭

didi @doIIju · 9/21/24

just sent! sorry if its not so much 🥺

Figure 11: An example that shows how user handles are not fully rendered on mobile devices in X, thus making PROSPER attacks virtually imperceptible. Both messages here are from V-A and there is not attack here. However, it illustrates how an attacker can insert themselves into the conversation using an account that only needs to have "na" as the prefix in its handle, along with matching display name and picture.

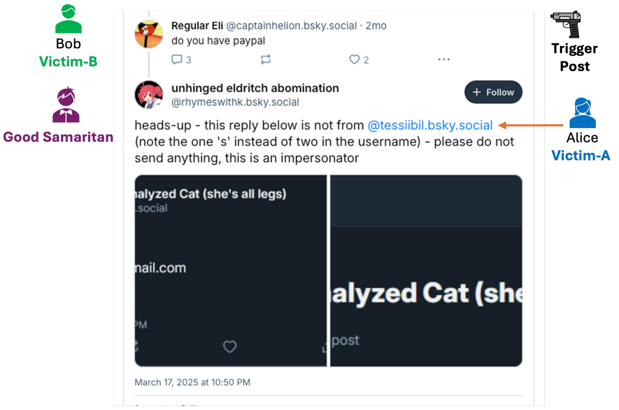


nytimes 48m

King Charles III addressed the Canadian Parliament during his first visit to the country since his coronation. nyti.ms/3ZAir...

Since the Second World War, our world has never been more dangerous and unstable.

Figure 12: A screenshot of an example post on Threads, a microblogging platform where only the unique "account handle" is being displayed in the feed. Such a design would mitigate the mobile-based UI security issues we highlighted in the paper.



Regular Eli @captainhelion.bsky.social · 2mo

do you have paypal

unhinged eldritch abomination @rhymeswithk.bsky.social

heads-up - this reply below is not from @tessiibil.bsky.social (note the one 's' instead of two in the username) - please do not send anything, this is an impersonator

March 17, 2025 at 10:50 PM

Figure 13: A screenshot of a PROSPER attack example on Bluesky. Although the attack post itself has been deleted (presumably, upon reporting), we can see the actions of the "Good Samaritan" who captured the attack and described the lexical differences in account handles of V-A and attacker in their follow-up message to V-B