#DM-Me: Susceptibility to Direct Messaging-Based Scams

Raj Vardhan
Texas A&M University
raj_vardhan@tamu.edu

Alok Chandrawal
Texas A&M University
alokchandrawal94@tamu.edu

Phakpoom Chinprutthiwong
Sisaket Rajabhat University
phakpoom.c@sskru.ac.th

Yangyong Zhang
Texas A&M University
yangyong@tamu.edu

Guofei Gu
Texas A&M University
guofei@cse.tamu.edu

ABSTRACT

In an emerging scam on social media platforms, cyber-miscreants are luring users into sending them a direct-message (DM) and are subsequently exploiting the messaging channel. We term this attack approach as the DM-Me scam. We report on a survey of 214 MTurk participants, in which we make the first effort to systematically study the susceptibility of users in falling victim to DM-Me scams. We find that most participants chose to send a direct message to at least one scammer, and made such choices more than half the time. This susceptibility can be attributed to the misplaced trust in scammers and the lack of negative consequences foreseen by participants in messaging accounts that they do not fully trust. Interestingly, our results also suggest that women mostly from the 31-40 age-group and who predominantly use Instagram a few times a week are less susceptible than men to financial DM-Me scams as they appear to face more discomfort in initiating a conversation with unfamiliar accounts for such services. We conclude with future research directions in mitigating the risks posed by DM-Me scammers, specifically by developing reliable indicators to aid users in assessing the trustworthiness of an account.

CCS CONCEPTS

• **Security and privacy → Human and societal aspects of security and privacy**.

KEYWORDS

Scams, Social media, Survey

ACM Reference Format:
Raj Vardhan, Alok Chandrawal, Phakpoom Chinprutthiwong, Yangyong Zhang, and Guofei Gu. 2023. #DM-Me: Susceptibility to Direct Messaging-Based Scams. In *ACM ASIA Conference on Computer and Communications Security (ASIA CCS '23), July 10–14, 2023, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3579856.3582815

1 INTRODUCTION

Today, many cyber miscreants are abusing social media platforms to operate scams. One platform that is particularly plagued by such scams is Instagram. A recent report [2] demonstrates how young people are being targeted through Bitcoin fraud on Instagram. Another study [1] showed how financial scams, such as money-flipping scams, lured Instagram users into sharing financial information with a false promise of huge profits. The study found 4754 unique scam posts and 1386 unique scammer accounts on Instagram. In light of such activities, Instagram has released guidance on how to report and avoid such scams on its platform [3].

In one kind of scam, scammers post click-bait images to lure users into visiting their profile page. Conceptually, this can be seen as an evolved form of a traditional *phishing* attack. Similar to a phishing victim, a user should click a link embedded in a scammer's post to reach their profile page. The profile page includes more captivating tactics to influence the user to send them a direct message. The messaging channel is subsequently exploited to steal money or private information under the disguise of providing a legitimate service. As such scammers have an account-description containing a hashtag similar to "#DMMe" ("Direct Message me"), we term this emerging attack approach as the DM-Me scam.

Previous research has extensively studied factors that lead to successful phishing attacks [7], the population which is most vulnerable [26], and methods that can be used as defense [13, 17]. However, compared to phishing, a key difference in the DM-Me scam is that a user needs to take an additional step of sending a direct message to the scammer before an exploit can be made. Due to this differentiating characteristic, it is unclear to what extent lessons from existing research can be applied to combat the DM-Me scam. A natural question emerges: *how likely are users to send a direct message to a potential scammer?*

In this paper, we take the first step in systematically studying the susceptibility of users in falling victim to the DM-Me scam. To this end, we surveyed 214 US MTurk participants to address the following research questions: **(1) RQ1:** How likely are users to fall victim to the DM-Me scam i.e. what is the *susceptibility* of users to the DM-Me scam? (Section 5.1.1), **(2) RQ2:** How does the susceptibility to DM-Me scam differ based on demographic factors? (Section 5.1.2), and **(3) RQ3:** Why does the DM-Me scam work in practice? (Sections 5.2, 5.3, and 5.4). To conduct this survey, informed by previous studies on phishing, we created a corpus of benign and scammer accounts found in the wild. Then, we asked participants to perform a roleplay exercise designed to assess how likely they are to send a direct message to different kinds of accounts and to shed light on their decision strategies.

We found that a majority of participants demonstrated high susceptibility to this scam as 84% of participants chose to send a DM to at least one scammer and made such choices more than half

ASIA CCS '23, July 10–14, 2023, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0098-9/23/07…$15.00
https://doi.org/10.1145/3579856.3582815

(54%) the time. The strategies used by participants highlighted that they frequently confuse scammers as being qualified or successful, which added to their trustworthiness. Novice users, with the least security knowledge, were most affected in this manner. Moreover, even when participants were unsure if an account is trustworthy, they often chose to send a DM because they did not foresee any adverse consequences of starting a conversation. Notably, for financial services-based DM-Me scams, we found female participants, among whom 46% belonged to the 31-40 age-group and 54% use Instagram a few times a week, were less susceptible than men as more women expressed discomfort in initiating a conversation for such services with unfamiliar accounts.

Our results demonstrate that users frequently misplace their trust in scammers despite their unrealistic characteristics, and provides insights on why they do so. We conclude with an analysis of our findings and outline potential future work. We hope this work highlights the susceptibility of users to such scams, and inspires future solutions.

2 RELATED WORK

Detecting online miscreants. As social networks grow in popularity, they become a prime target for cyber-miscreants, who abuse such platforms to victimize unsuspecting users. Numerous studies have proposed ways to detect miscreants and prevent scams or phishing attacks on various platforms, such as Twitter [4, 11, 19], Facebook [14], Instant Messaging [5], email [13, 15], or websites in general [20, 23, 33]. As the DM-Me scam has not been studied before, we start by conducting a preliminary investigation of this scam on Instagram (Section 3). Thereafter, we focus on studying the susceptibility of users to the DM-Me scam which is the core contribution of this work.

Demographic factors and susceptibility to scams. Sheng et al. [26] conducted a roleplay survey with 1001 participants towards studying the relationship between demographic factors and phishing susceptibility. Participants were told to assume the identity of a fictitious user and shown 14 images of emails along with some relevant context. Participants were then asked how they would handle the emails themselves. The authors found women to be more susceptible than men. In this work, we show that for DM-Me scams, the susceptibility varies for certain narratives, such as scams that offer financial services (to which we found women to be less vulnerable than men).

Why people fall for phishing-based scams? Many early works in phishing research studied why people fall for phishing-based attacks [6, 10, 12, 21, 28, 29] as well as how to educate people to not fall for such attacks [17, 18, 27, 30]. Dhamija et al. showed twenty web sites to twenty-two participants and asked them to determine which ones were fraudulent [7]. The results of this study showed that 23% of the participants ignored important browser-based cues, such as the address bar, status bar, and security indicators. Consequently, participants made mistakes 40% of the time. Downs et al. reported on interviews and role-playing study aimed at shedding light on the decision strategies used by users, who are relatively naive about security, in dealing with potentially suspicious emails [9]. The authors find that general awareness of phishing or security indicators is not enough for such non-experts to protect themselves against scams, especially against unfamiliar risks. Most of the strategies participants used in determining the trustworthiness of an email were centered around interpreting the text of the email instead of more reliable cues, such as URLs associated with the links. However, these findings cannot be directly used to understand the strategies people use or the indicators they pay attention to towards deciding whether or not to interact with a social media account. Our study seeks to bridge this gap.

3 PRELIMINARY INVESTIGATION

The susceptibility of users to DM-Me scams is naturally tied to the types of narratives and attack-strategies used by such scammers, and their ability to persist on a platform to carry out such time-consuming scams. Therefore, to motivate our work, we first performed a manual investigation on Instagram to understand these basic characteristics of DM-Me scammers in the wild. As Instagram's Terms of Use prohibits automated crawling of information, we designed our investigation to conform to their terms to the best of our ability.

Identifying scammers. We performed the following steps to identify a set of potential scammers. In the first iteration, we manually queried Instagram using well-known hashtags (e.g., "#DMMe") associated with posts of DM-Me scams. To minimize selection bias, we augmented this list with hashtags derived from various combination of keywords associated with popular scams such as romance scams [16, 32], marketing scams [31], and financial scams [1, 2]. From the resulting posts, we shortlisted suspicious accounts and marked them as potential scammers. Our decision was based on properties such as (1) abnormally high posts/followers when the first available post was only few days old, (2) description/-posts include unrealistic or fraudulent claims (e.g., "DM-Me to earn $250,000 weekly"). For these accounts, we recorded the key account attributes and also captured screenshots of their profile-page for use in our survey (as discussed in Section 4.1). One of these attributes include the date of their first available post which we later use to estimate the lifetime of an account. Next, to discover more diverse narratives, we recorded other suspicious hashtags used in an account's description, and in the comments and captions of their three most recent posts. For instance, we found scams (unknown to us before) offering to get people verified with the hashtag "#getverified" and those offering help in gaining followers with "#growfollowers". We utilized such additional hashtags in querying Instagram in subsequent iterations. The full list of hashtags we utilized include #DMME, #DMMeForMore, #getverified, #verificationbadge, #getlikes, #growfollowers, #trader, #investors, #forextrader, #financialfreedom, #financialcoach, #marketing, #digitalmarketing, #marketingexpert, #socialmediamarketing, #creditrepair, #giveaway, #freegiveaway, #datingcoach, #lovecoach, #findlove, #relationshipexpert, #lifecoach, #makemoney, #debtfree, and #getrich. We performed ten iterations of this activity from 16 June 2020 to 30 July 2020 and collected information on 487 potential scammer accounts.

Liveness Checking. Our intuition in distinguishing between benign service providers and scammers is that the accounts of the latter are more likely to get shut down as a consequence of them being reported. Therefore, we asked five of our graduate-student

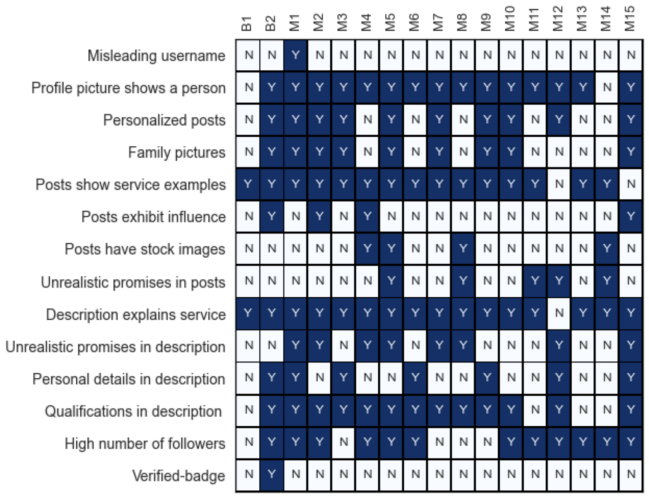

Figure 1: Key characteristics of survey accounts.

colleagues to help verify the liveness of accounts. We assigned nearly 25% of these accounts to each student and requested them to visit all their designated accounts every five days. When an account is found inactive (Instagram display "Sorry, this page isn't available" for such accounts), our colleagues recorded the current date as the date of the account getting suspended. This activity was performed from 16 June 2020 to 3 September 2020. Overall, we estimate this process took us 22 human-hours.

Findings. We made the following key findings. First, we find that scammers use a wide range of narratives from offering help in improving credit-score to providing dating advice. Second, scammers used various strategies to appear qualified and trustworthy, such as using "official" in their username, posting family pictures (not likely to be their own), and claiming to be certified for the service being offered. Some even used strategies that come across as clear red-flags (e.g., unrealistic promises of instant wealth) possibly to only lure the most naive users who would be easier to exploit. We further detail these narratives and strategies in Section 4.1. Third, we found the platform to be active in mitigating the risks posed by potential scammers. In July 2020, Instagram disabled the use of hashtag "#DMMe" for searching posts on its platform. The corresponding page[1] has stopped fetching results since then. However, scammers were quick to adapt, and started using alternate hashtags (e.g., "#DMMeForMore"). Furthermore, we found that 32% (157/487) of the accounts were suspended by Instagram in our monitoring period of 80 days. The average estimated lifetime (date of first post to date of suspension) for scammer accounts was found to be 165 days with a median value of 38 days. This indicates that many scammer accounts could be getting a sufficient time to execute their attacks. Overall, these findings further motivates us to study how users may react to encountering the profile of such diverse scammer accounts, and whether they will be likely to send them a direct message.

Table 1: Scenarios associated with survey accounts.

Service	Scenario	Accounts
Get customized sketches	You would like to get a customized drawing created as a gift for your friend's birthday.	B1
Debt management	You have been reading articles online to learn about money management and financial planning.	B2
Financial trading	You are considering making new investments such as investing money in stocks or cryptocurrency	M1, M2, M4, M5, M7, M8, M10
Credit repair	You have been reading articles online exploring ways to improve your credit score.	M6, M9
Digital marketing	You are exploring ways to promote your company's online presence.	M3, M13
Get verification badge	You wish to get a verified badge for your account on Instagram.	M11, M14
Dating coach	You are interested in privately seeking dating advice.	M12, M15

4 METHODOLOGY

4.1 Survey design

Our survey (see Appendix A) consisted of three sets of questions. In the first set (Appendix A.1), we asked demographic and background questions. We collected information on age, gender, education level, technical background, primary technology platform, and frequency of using Instagram.

In the second set (Appendix A.2), the behavior of participants was measured through a roleplay section. Participants were told to assume the role of Pat Jones, a fictitious Instagram user. For each question, participants were shown an image of an Instagram account that Pat came across under a particular scenario, and were asked to respond as if they were Pat. This exercise is based on an established roleplay methodology that has been shown to have good internal and external validity [8, 26]. The roleplay format also enables us to study users' behavior without conducting an actual scam.

Informed by phishing-based roleplay studies [7, 9, 26] and findings from our preliminary investigation, we curated a corpus of images of 17 Instagram accounts, consisting of 2 benign (B1-B2) and 15 scammer accounts (M1-M15). Each image showed an account's profile page in high-resolution, including three most recent posts, as it would appear in a Chrome browser on a PC. Accounts M1-15 were suspended during the monitoring phase. However, for benign accounts, as there is no assurance of being truly trustworthy, we did our best effort in selecting active accounts that appeared legitimate. Nevertheless, given the lack of ground-truth, our study does not focus on testing the negative effects of suspicion on accounts that we considered benign. We selected these accounts as they offered diverse characteristics (Figure 1) and narratives (Table 1). This would allow us to study users' susceptibility towards different kinds of accounts and scam narratives. Furthermore, as shown in Table 1,

[1] https://www.instagram.com/explore/tags/dmme/

we tailored the scenario associated with each account to be complementary to its primary narrative. The purpose of any scenario was to provide participants an artificial need while showing them one way of getting that need fulfilled.

The roleplay exercise comprised of three groups of scenario-based questions. In each group, participants were shown 2 benign and 4 scammer accounts without being informed of the account's classification (benign/scammer). Groups one, two, and three helped evaluate (1) which accounts are participants likely to send a direct message to and their reasons for doing so, (2) which accounts are they likely to trust and why, and (3) what indicators do they consider important in determining the trustworthiness of an account, respectively. We keep group-2 after group-1 to ensure that participants are not primed to think about trustworthiness when they are providing open-ended reasons on why they would DM (or not DM) an account. Similarly, we keep group-3 after group-2 so that participants can explain why they trust (or distrust) an account without being primed on any relevant indicators of trust (that are shown as a list of choices in group 3).

In the third set (Appendix A.3), participants were assessed on their practical ability and security knowledge using two groups of questions. The first group was driven by our intuition that the *perceived age* of an account (how long a user believes the account has existed on Instagram) can influence the user's decision to trust or interact with it. Note that Instagram does not show the true age of an account (unlike other platforms like Twitter). Not surprisingly, we observed in the preliminary investigation that many scammers had thousands of posts and followers while their first post was only one day old (something that serves as a lower-bound estimate of the account's true age). As these numbers can be easily spoofed by making several posts in a day and using bot-followers, this could potentially be a strategy to inflate their perceived age. Therefore, the first group of questions assessed if participants can estimate the age of an account by seeing its profile page. Here, we showed two scammer accounts whose first post was 1 day old but had many posts and followers. We also asked an open-ended response on how they would estimate this if they are given the freedom to navigate the account's page on Instagram. In the second group of this set, we used the evaluation approach by Sheng et al. [26] for gauging participants' security knowledge by asking them to choose the correct definitions of four concepts related to computer-security: *cookie, phishing, spyware, and virus* [26].

Two researchers coded all open-ended responses to identify the strategies that participants used to DM or trust an account, and to determine an account's age. One of the researchers took the role of the *codemaster* and performed the initial coding, while the second researcher iteratively provided feedback to the codemaster. In the last iteration, both the researchers coded all the responses, with the codemaster resolving any remaining conflicts. Cohen's K, a measure of inter-rater reliability, was 0.978 which indicated strong agreement between coders.

4.2 Recruitment

We recruited adult participants from Amazon MTurk in September 2020. For a participant to be eligible, they were required to be above 18 years of age, be in the United States, and have a 95% previous task approval rating on MTurk. Prior work has shown that these criteria provide reliable participants from MTurk in the context of security surveys and exercises [22, 24]. We paid $4.1 for each finished task, which took approximately 30 minutes to complete.

In total, 274 participants responded to our study. From these, we discarded 42 participants who failed any one of the two attention checks we placed, and an additional 18 who provided duplicate or vacuous answers across multiple questions. Ultimately, we had 214 valid participants. This number is comparable to the number of valid participants in several similar studies [8, 21, 30]. Table 2 shows the demographic information of our participants.

Table 2: Demographic Information

MTurk Participants		
Frequency Of Use		
Several Times A Day	117	54.67%
Once a day	41	19.16%
Few times a week	45	21.03%
Used in the past, no longer use it	9	4.21%
Never used Instagram	2	0.93%
Age		
18-30	78	36.45%
31-40	93	43.46%
41-50	26	12.15%
51-60	11	5.14%
61+	6	2.8%
Gender		
Male	129	60.28%
Female	81	37.85%
Other, Decline to answer	4	1.86%
Education		
Some high school credit	0	0.00%
High school graduate	7	5.18%
Some college credit	23	17.04%
Trade/technical/vocational training	12	8.89%
Bachelor's degree	65	48.15%
Master's degree	24	17.78%
Professional degree	1	0.74%
Doctorate's degree	3	2.22%
IT Degree		
Yes	71	33.18%
No	137	64.02%
Decline to answer	6	2.8%
Computer Security Knowledge		
Expert	56	26.16%
Moderate	131	61.21%
Novice	27	12.61%
Phishing correct	138	64.48%
Phishing wrong	76	35.51%
Technology Use		
Desktop	52	38.51%
Laptop	90	66.67%
Tablet	12	8.89%
Mobile or Smartphone	66	48.89%

5 RESULTS

In this section, we report the findings from our survey. First, we report how susceptible are participants to the DM-Me scam, and

which demographics are more susceptible. Next, we identify the strategies that participants use in deciding to send a DM. Then, we highlight the indicators that were deemed important and the strategies used for determining the trustworthiness of an account. Finally, we show how participants estimate the age of an account.

5.1 Susceptibility to DM-Me Scam

5.1.1 How likely are users to fall victim to DM-Me scam? We consider sending a direct message to a scammer as falling for DM-Me scam. In previous works on measuring susceptibility to phishing, some studies [17] considered clicking on a phishing link as falling for phishing, whereas other studies [26] determined it based on whether users submit information to phishing websites. However, unlike phishing websites that request for user information through a static web form, a DM-Me scammer has the opportunity to persuade the user over a certain time-period to make the exploit. The success rate of such persuasion is difficult to measure through an online survey and we leave investigation into making this estimation as future work. Nevertheless, we believe that sending a direct message to a scammer in itself is a good indicator for a user providing information once the interaction begins. In fact, prior interview and roleplay-based studies on phishing found that about 90% of the participants who click on a phishing link would go on to provide information to phishing websites [17, 26]. Therefore, the susceptibility rates we report are indicative of an upper-bound estimate of victimization.

For the question *How likely are you to DM this person*, asked in group-1 of the roleplay section, participants who responded with "Very Likely" or "Moderately Likely" are deemed as those who will send a DM to that account. As each participant was shown 6 accounts, a total of 1284 cases were shown across 214 users. Overall, 84% (180/214) participants indicated they would DM at least one scammer account. On average, a participant made the mistake of sending a DM to a scammer account 52% (448/856) of the times, while deciding to DM a benign account 59% (254/428) of the times.

5.1.2 How does the susceptibility to DM-Me scam differ based on demographic factors? **Gender.** For scammer accounts, male participants (61%) said that they would send a DM 58% of the times they encountered such accounts in the survey, whereas female participants (38%) said that they would send a DM 47% of the times. Across all scam narratives, we find the groups to not be significantly different from one another in their susceptibility using a two-tailed t-test ($t=1.35$, $p<0.05$). However, on investigating responses to different kinds of narratives, we find that females are less susceptible than males ($t=4.46$, $p< 0.001$) to financial DM-Me scams (accounts corresponding to financial trading and credit repair in Table 1). We shed light on the factors behind the observed lower susceptibility of female participants to financial DM-Me scams in Sections 5.2 and 5.4.

Age. We found that all the age groups were different from one another in their likelihood to send a DM to a scammer account. An analysis of variance (ANOVA) comparing age groups found a significant overall effect ($F(4,70)=3.92$, $p<0.01$) concluding that the groups were significantly different from one another. Moreover, the post-hoc test indicated that no single age group was significantly more susceptible to DM-ME scam than other groups.

Frequency Of Use. For scammer accounts, participants who use Instagram several times a day said they would send a DM more times (60%) than those who use it once a day or less (45%). We found that frequent Instagram users are more susceptible to DM-ME Scam than those who use it less frequently using a two-tailed t-test ($t=2.85$, $p<0.01$). This result seems counter-intuitive at first since frequent users can be expected to have better exposure to the platform. However, many such participants did not distinguish well between benign and scammer accounts and showed a higher inclination to initiate a conversation.

Education. We found that people with a bachelor or higher degree were more likely to DM a scammer. An analysis of variance (ANOVA) comparing educational level found a significant overall effect ($F(1,28)=5.96$, $p<0.05$) with participants having a bachelor or higher degrees being more likely to DM a scammer. Post-hoc tests comparing users with a bachelor or higher degree to other groups were significant at $p<0.05$; however, people with a bachelor or higher degrees (masters, doctorate) were not significantly different from each other.

IT Degree. Participants who had an IT degree (33%) said that they would a send a DM to scammer account 73% of the times, whereas participants without an IT degree (64%) said that they would send a DM to a scammer account 44% of the times. A Mann Whitney U test revealed that people with an IT degree were more likely to send a DM to a scammer account than people without an IT degree ($Z=4.27$, $p<0.001$).

Computer Security Knowledge. We labeled participants who gave the correct definitions for all four terms, i.e., phishing, cookie, spyware and virus, as security *experts*, whereas those who didn't answer any of the definitions correctly as security *novices*. Others were assigned to the security *moderates* group. For scammer accounts, we found that security experts (26%), moderates (61%), and novices (13%) would send a DM 36%, 55%, and 79% of the times, respectively. We find that experts are less susceptible to DM-ME Scam than novices using a two-tailed t-test ($t=-6.03$, $p<0.001$).

5.2 User Strategies for Direct Messaging

By coding the open-ended reasons that participants provided for their answer to *How likely are you to DM this person*, we identified a set of strategies that participants use in deciding whether or not to DM someone. Table 3 shows how often these strategies were used across different accounts.

5.2.1 Reasons to DM. We find that participants primarily use the following two strategies towards being likely to DM someone.

Trustworthiness. Many participants (57%) indicated that they will DM an account because it appears legitimate, e.g., *"The woman looks genuine. I'd like some advice for stocks during COVID."* (P-33 on M1), or trustworthy, e.g., *"I would DM them just to ask for advice on investing because they seem trustworthy."* (P-28 on M2). Ironically, the perceived trustworthiness of an account was the most used strategy to send a DM to scammer accounts (27%), but not the most used for benign accounts (18%).

Participants used various strategies to determine an account's trustworthiness. We observed that novice users did not mention any indicators other than the content of posts, e.g., *"This user is making some analyses reports on his page and looks like I can trust*

Table 3: User strategies towards deciding whether or not to send someone a direct message (DM).

	Benign			Scammer															
	B1	B2	**Total**	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	**Total**
Number of cases	214	214	428	78	87	79	44	54	54	54	43	77	43	58	58	57	34	34	613
Reasons to DM																			
Trustworthiness	9%	**28%**	18%	**29%**	**38%**	20%	**36%**	**26%**	26%	**26%**	**30%**	17%	28%	19%	19%	14%	0%	15%	**27%**
Fulfilment of need	**57%**	18%	**38%**	27%	21%	**25%**	18%	20%	**31%**	15%	28%	**29%**	**30%**	**31%**	**36%**	**39%**	21%	**29%**	24%
Reasons not to DM																			
Need won't be met	24%	16%	20%	9%	9%	22%	2%	9%	17%	15%	7%	14%	19%	9%	17%	16%	15%	12%	13%
They won't reply	0%	11%	5%	3%	2%	1%	2%	2%	0%	4%	0%	0%	0%	0%	0%	2%	0%	0%	1%
Lack of familiarity	1%	2%	2%	4%	1%	0%	0%	4%	0%	0%	5%	1%	2%	0%	0%	0%	0%	6%	2%
Looks untrustworthy	2%	8%	5%	19%	21%	8%	27%	24%	2%	24%	26%	21%	16%	22%	7%	5%	**50%**	21%	18%
Won't DM to fulfil need	1%	9%	5%	6%	5%	4%	14%	2%	4%	2%	2%	9%	0%	7%	7%	5%	6%	12%	5%
Can't decide	2%	4%	3%	1%	2%	8%	0%	4%	13%	7%	0%	8%	5%	3%	5%	11%	0%	3%	5%
Others	3%	4%	4%	1%	1%	3%	0%	9%	7%	7%	2%	1%	0%	9%	9%	9%	9%	3%	3%

him" (P-112, novice, on M4). On the other hand, moderate and expert users used various other cues (e.g., number of followers) towards making this decision, e.g., *"Has a lot of followers and even provides a phone number, making him look credible"* (P-45, moderate, on M4), and *"Good number of followers and posts make look the page genuine"* (P-6, an expert, on M4).

On average, novice users used this strategy towards trusting 1.15 of the four scammer accounts, whereas moderate users and experts used it for trusting 0.95 and 0.70 scammer accounts, respectively. As we hypothesized, trust was indeed an important dimension along which participants decided to DM an account. We further discuss how participants determine trustworthiness in sections 5.3 and 5.4.

Fulfilment of need. Many (76%) of the participants indicated they would send a DM because the account can fulfill their need. Participants used various cues, such as description, posts, and followers, in inferring that the person is qualified to serve their needs, e.g., *"This person has a lot of followers which means he must help lots of people so it would be in my best interest to dm them."* (P-44 on M4). Some participants were also motivated by their interest in the service, e.g., *"As I am very much interested in Cryptocurrency I would like to know more about it to invest in cryto or stock market to get relieved from my financial burdens."* (P-23 on M2). Notably, even many (54%) of the security experts were convinced by at least one scammer account that they are qualified to meet their need, e.g., *I'd be curious about whether he has videos up online somewhere or if he maybe has a discord I could join. He's also expressed a willingness to be contacted and he has experience* (P-106, an expert, on M8).

5.2.2 Reasons not to DM. We find that participants used the following six types of strategies towards not being likely to DM someone.

My need won't be met. Several (53%) participants indicated that they will not DM someone who is not qualified enough to meet their needs. There were two main cues that participants used in making this decision. First, some participants looked at the content of posts to infer if the account owner is a professional. Interestingly, the use of family pictures (possibly to earn trust), by accounts such as M10, made some participants feel that the account owner is not too professional, e.g., *"The personal pictures are a nice touch and make me feel like I would possibly contact them because they might be real, but I wouldn't want to take financial advice from someone who I can't verify as a certified expert."* (P-105 on M10). Second, many participants considered less followers as a sign of the account owner not being successful, e.g., *"She doesn't have as many followers as I would expect a successful business marketer to have."* (P-69 on M3).

I won't get a reply. In our monitoring phase, we observed that scammer accounts use various tactics to show their success and qualification for the offered service. This is achieved through a combination of having high followers, an account description mentioning certifications, and posts that reflect upon their influence. Interestingly, such techniques were counterproductive with some (13%) participants who indicated that they will not DM an account because their perception of the account made them realize that they won't get a reply.

Specifically, there were two factors that convinced participants of this. First, some participants noticed an account having too many followers and felt that their message may go unnoticed, e.g., *"They seem reliable and helpful, but they are a large account and despite them making commitments, it would be tough for them to handle so many requests all by themselves."* (P-18 on M2). Second, some participants were intimidated by an account's page because the account owner appeared too successful/qualified, which added to their hesitation in reaching out, e.g., *"I would not want to appear dumb by not knowing what to ask and how to word the question."* (P-43 on M10). Overall, this was the least-used strategy for scammer accounts.

Lack of familiarity. Some of the participants (6%) indicated that they will not DM someone if they aren't familiar with them, especially for financial services, e.g., *"I don't trust someone online who I don't know when it comes to money issues and potential investments. I would have no way to know if this person is legitimate or if they might try to commit fraud with my investment."* (P-22 on M10). In terms of the people who used this strategy, there are two things to note. First, none of the participants in the novices group used this strategy. Most (78%) of the participants who used this strategy were moderates. Second, we observed that 7% of all females used this strategy as compared to 4% of males, e.g., *"I'm not usually one to send DM's to others unless I know them personally."* (P-28, a female, on B2).

Looks untrustworthy. Many (46%) of the participants indicated that they will not DM a person because the account seems

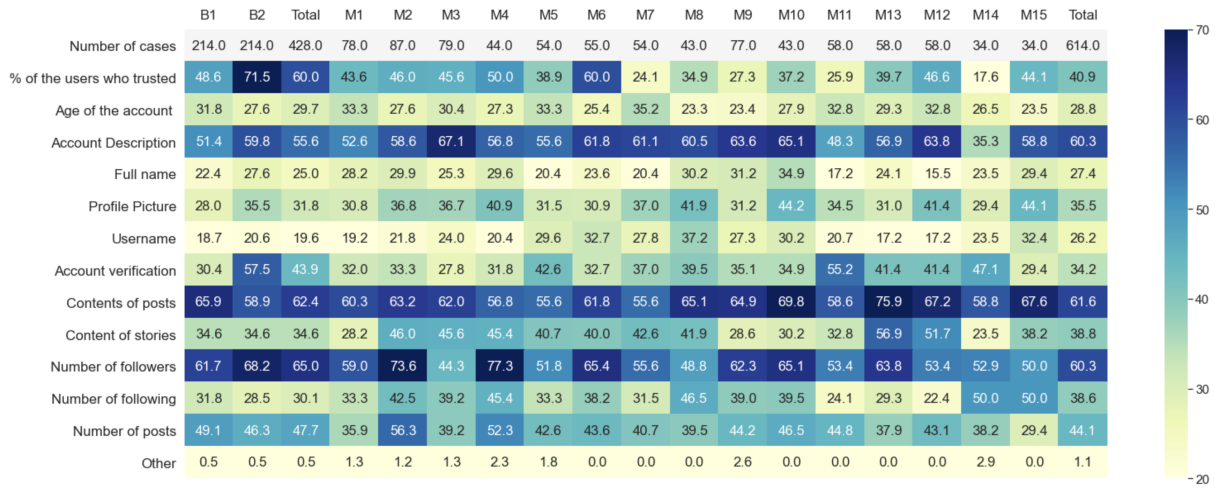

Figure 2: Indicators deemed important for determining trustworthiness. Row1: total participants who saw that account. Row2: % of participants who trusted that account. Rows 3-14: % of times an indicator was selected as a reason to trust that account.

untrustworthy. This strategy was used 18% of the times for scammer accounts and 5% of the times for benign accounts. This strategy was used only by few (15%) of the novices. The novice users did not mention any specific indicators that raised their suspicion. Instead, they considered an account untrustworthy if they did not trust the service, e.g., *"I don't trust binary trades because I don't know about it.. I won't send any message to him."* (P-108, a novice user, on M10), or if the overall account appeared suspicious, e.g., *"I think this is a fake Instagram account or his editor needs to be fired!"* (P-98, a novice user, on M5). Only in rare instances did novice users display cognizance of a service being a potential scam, e.g., *"I feel like this person will charge me a lot of money for the service. I feel like this could be a scam."* (P-34, novice, on M-13).

On the other hand, a higher number of moderates (44%) and experts (68%) used this strategy. These users provided more nuanced explanations that described various indicators, such as suspicious-looking usernames, e.g., *"I chose not to send a DM because their username reminds me of a bot. I think this is a scam."* (P-109, an expert, on M8). Some mentioned the unrealistic claims made in an account's description for finding it untrustworthy, e.g., *"I don't believe the person's claims that they make 250,000 a week and therefore wouldn't trust them, especially with finances."* (P-2, a moderate user, on M4). We further discuss ways in which participants conclude if an account is untrustworthy in sections 5.3 and 5.4.

I won't DM to fulfill such needs. One of the goals of the roleplay section was to test if users would send a DM to fulfill an artificial need given to them. We find that only 20% (42/214) of the participants indicated that they will not DM someone to fulfill their need. There were two kinds of motivations in using this strategy. First, 55% (23/42) of these participants indicated that they do not prefer to send direct messages to enquire about sensitive services, e.g., *"I tend to steer clear of financial and job advice online as a rule, regardless of whatever qualifications the person has. They are most likely just trying to get a commission."* (P-2 on B2). Some indicated that they will use an alternate medium to contact, e.g., *"She provides several other ways to be in touch with her and access her services, I would follow those methods instead of messaging."* (P-159 on M12). Some further explained that there are more reliable alternatives available to fulfill such needs, e.g., *"I won't message a random account for investment advice. there are companies for that"* (P-21 on M1). Both P-159 and P-21 were in the experts group. Second, 33% (14/42) of these participants emphasized their discomfort in sending a DM to fulfill a need. Notably, 93% (13/14) of these participants were females, wherein nearly half of them (6/13) belonged to the 31-40 age-group, and 54% (7/13) of them use Instagram a few times a week. Many of them explicitly indicated being female as part of their explanation for being uncomfortable, e.g., *As a female, I feel that contacting (any males) for this kind of advice on instagram is unwarranted. I don't feel that these people would care about my financial situation in the slightest.* (P-23 on M2). Indeed, P-23 (a female in the moderate group), was one of the few (7%) participants who chose not to DM any account.

I can't decide with the given information. For both benign and scammer accounts, some (16%) participants indicated at least once that they can't say if they will DM the account based on the available information. Some participants indicated they will learn more about the account first. For scammer M13, which claims to provide a verification badge but isn't verified themselves, P-222 (an expert) pointed out the irony, *"They're not even verified? How successful can they be? I'd have to know more before thinking about DMing them, maybe by checking out the testimonials or Googling them."*. Others said they would research about the service before reaching out, e.g., *"I think I would do more research to see if I can improve my credit on my own."* (P-94 on M6). Notably, these participants chose this strategy only a median of 1 time and for 1.47 accounts on average. Across their other answers, most (91%) of this group of participants indicated they will DM at least one scammer account.

5.3 Indicators that Influence Trust

Figure 2 shows how often an indicator was deemed important towards determining the trustworthiness of an account. Across all accounts, the top three indicators consist of the number of followers, content of posts, and account description. The number of followers is a particularly dominant factor for many scammer accounts. For instance, 74% of participants indicated number of followers as a reason for M2, 77% for M4, and 65% for M6.

It is worth noting that participants may pay attention to other prominent indicators even if an account has high followers. Consider M5, which has over 10K followers, but makes several unrealistic claims in its posts and account description. In this case, content of posts and account description were the dominant indicators mentioned 56% of the times. Overall, all accounts included a description that explained their service and posts that showed examples of that service (e.g., screenshots of apparent interactions with customers). Most users paid attention to these indicators to determine the trustworthiness of an account.

We further observe that, on average, verified-badge was not deemed an important indicator. For B2, which has a verified-badge, verification is selected only 57% of the time. Surprisingly, more participants (68%) selected number of followers as a factor for B2. This indicates that participants may choose less reliable indicators and trust their own judgment and instincts in deciding whether an account is trustworthy. Furthermore, for scammer accounts, the lack of account verification was never deemed important by half (56%) of the participants.

5.4 User Strategies for Determining Trust

We find that the strategies participants use for determining the trustworthiness of an account can be categorized as the following. Table 4 shows a summary of these strategies and their usage across different accounts.

5.4.1 Person appears to be legitimate (/suspicious). 6% of the times that participants saw a scammer account, they found it trustworthy because the account owner seemed legitimate. This strategy was used by 16% of participants who primarily formed this perception based on the profile picture and content of posts. We observed that such trust was easily amplified if the posts included family pictures, such as in the case of M1 and M3, e.g., *"she posts the family, posts intimate moments, I believe it is trustworthy for showing that she is a family person, with principles."* (P-10 on M3).

On the other hand, 13% of the times participants saw a scammer account, they exercised caution by using a contrasting strategy ("Person/page looks suspicious" in Table 4) and did not find such posts credible, e.g., *"Leaning towards not trustworthy. Image in profile doesn't match image of woman in posts, other posts look like google image rips."* (P-145 on M10).

5.4.2 Person is qualified (/not qualified). Some (12%) of the times participants saw a scammer account, they considered it to be trustworthy because the account owner appeared qualified. Scammer accounts generally use various techniques to come across as highly qualified. For instance, M2 showed their influence with posts of a person speaking on a stage. This was occasionally successful in instilling trust as 16% of participants mentioned M2 being qualified, e.g., *"He seems trustworthy because he has a picture of him leading a seminar"* (P-28 on M2).

Many participants also paid attention to the account description in finding the account owner qualified. M1 had mentioned that they are a certified trader in their description. Surprisingly, this was sufficient to convince some users of M1's qualifications, e.g., *"She is certified, appears to be a reliable person."* (P-10 on M1). Similarly, many participants simply believed a person is an expert because they said so in their description, e.g., *"The owner's posts and description shows that they are expert of stocks and cryptocurrency."* (P-140 on M8).

On the other hand, 12% of participants, in at least one of their answers, approached with caution and found a scammer account untrustworthy because they did not seem qualified. Many of the strategies used by scammer accounts to earn user's trust (which worked for many naive participants) were the factors that led to such participants becoming suspicious, e.g., *"This person definitely isn't trustworthy. She basically just listed every little profession people try to capitalize on when trying to earn an income online. I highly doubt she is skilled in those subjects."* (P-144 on M10 which had offered numerous services in their description).

5.4.3 Page looks normal (/abnormal). Participants trusted a scammer's account 20% of the times because their page looked normal. Majority of such participants formed an overall positive impression of the page and did not mention any specific indicators that aided them in forming that opinion, e.g., *"She just has an authentic air about her page."* (P-3 on M3). Nevertheless, some participants mentioned indicators such as account description in concluding that the page looks normal, e.g., *"I think this account is offering a reasonable goal of just learning about improving your credit, it is not promising something spectacular and unlikely."* (P-149 on M9).

Again, in a contrasting strategy, many participants found an account to be untrustworthy when the account's page looked abnormal, e.g., *"The account just looks NOT trustworthy due to the mismatched font styles and the way that the account owner displays his various bit coin earnings."* (P-125 on M8). Furthermore, some of the participants were able to spot various kinds of bait that made them suspicious, e.g., *I would not trust this person given the profile info of making $22,500 weekly. That kind of statement screams false or scammy.* (P-144, an expert, on M8). Another user stated, *There is no way he has 24/7 support for stocks. I doubt this is a legit program.* (P-33, a moderate user, on M2).

In our monitoring phase, we had observed that some of the suspended scammer accounts re-spawned after a short time period with a slightly different username. Many such accounts, such as M2, state in their account description that their previous account was hacked, and therefore they have created another one with a slightly modified username. However, such claims in the description made a fraction of the participants suspicious of the account, e.g., *"it says his last account was hacked, but many bad people use this tactic to steal other people's identities and pretend they are the real owners of the account"* (P-1 on M2). Overall, this strategy was used at least once by 30% (8/27) of all novice users, 36% (47/131) of moderates, and 54% (30/56) of experts.

5.4.4 I don't trust such services. Few (10%) of the participants indicated for at least one of the scammer accounts that they do not

Table 4: User strategies to determine the trustworthiness of an account.

	Benign			Scammer															
	B1	B2	**Total**	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	**Total**
Number of cases	214	214	428	78	87	79	44	54	54	54	43	77	43	58	58	57	34	34	613
Account is trustworthy	49%	71%	60%	44%	46%	46%	50%	39%	60%	24%	35%	27%	37%	26%	40%	47%	18%	44%	55%
Verification	1%	**26%**	13%	0%	1%	0%	0%	0%	4%	2%	0%	0%	0%	5%	0%	2%	3%	0%	1%
Number of followers are high	3%	14%	8%	8%	11%	0%	18%	15%	9%	4%	2%	1%	0%	10%	12%	12%	3%	0%	10%
Number of posts, followings	2%	3%	2%	0%	2%	1%	2%	4%	0%	0%	0%	0%	0%	2%	5%	3%	0%	3%	2%
Person is qualified	12%	13%	12%	10%	**16%**	3%	14%	6%	15%	2%	12%	9%	14%	0%	10%	7%	0%	9%	12%
Person looks harmless	3%	2%	3%	6%	1%	16%	0%	2%	11%	2%	7%	0%	2%	0%	2%	9%	0%	6%	6%
Page looks normal	**26%**	12%	**19%**	17%	13%	**24%**	**16%**	11%	16%	13%	**14%**	**14%**	**21%**	5%	9%	12%	6%	24%	20%
Other	2%	1%	2%	3%	0%	1%	0%	0%	4%	0%	0%	3%	0%	2%	2%	2%	6%	3%	2%
Not trustworthy	5%	12%	8%	13%	28%	15%	18%	37%	11%	33%	40%	40%	28%	43%	19%	16%	44%	9%	36%
Lack of verification	0%	0%	0%	1%	0%	1%	0%	4%	2%	0%	2%	5%	0%	**24%**	0%	2%	18%	0%	5%
Number of followers are low	0%	0%	0%	0%	1%	3%	2%	0%	0%	6%	7%	12%	2%	0%	2%	0%	9%	0%	4%
Number of posts, followings	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	5%	3%	2%	2%	2%	0%	3%	0%	1%
Person is not qualified	0%	3%	2%	1%	2%	5%	0%	9%	2%	6%	5%	5%	14%	0%	0%	2%	0%	3%	5%
Page/person looks suspicious	0%	3%	2%	8%	**16%**	4%	11%	**20%**	4%	15%	9%	8%	9%	14%	7%	5%	12%	0%	13%
I don't trust such services	1%	2%	2%	0%	1%	0%	5%	2%	0%	0%	7%	3%	0%	0%	3%	0%	3%	6%	2%
Lack of familiarity	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%	0%	0%
Other	0%	0%	0%	0%	5%	1%	0%	0%	4%	4%	5%	1%	0%	2%	3%	3%	0%	0%	3%
Don't Know/ Not Possible	47%	16%	32%	44%	26%	39%	32%	24%	29%	43%	26%	32%	35%	31%	41%	38%	38%	47%	49%
Verification	6%	2%	4%	8%	6%	0%	9%	2%	2%	4%	2%	3%	5%	12%	5%	10%	12%	9%	8%
Number of followers are low	9%	1%	5%	4%	1%	8%	0%	0%	0%	4%	5%	4%	0%	3%	5%	0%	3%	3%	4%
Number of posts, followings	1%	0%	1%	3%	1%	1%	0%	0%	0%	0%	2%	0%	5%	0%	0%	0%	0%	0%	1%
Page/person looks suspicious	0%	0%	0%	4%	6%	0%	5%	2%	2%	6%	0%	5%	2%	0%	3%	0%	9%	3%	4%
I don't trust such services	2%	1%	2%	0%	0%	3%	2%	0%	4%	0%	0%	1%	0%	0%	2%	2%	0%	3%	1%
Lack of familiarity	1%	0%	0%	1%	2%	1%	2%	0%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%
Insufficient information	24%	9%	16%	**23%**	9%	22%	11%	9%	**20%**	**20%**	12%	14%	19%	16%	**26%**	**26%**	**15%**	**26%**	**25%**
Other	1%	1%	1%	0%	1%	3%	0%	9%	0%	6%	2%	0%	5%	0%	0%	0%	0%	0%	2%

trust such services. Some participants displayed an awareness that the information shown by unverified service providers may not necessarily be accurate, e.g., *"Anyone can create a public persona that is completely different from the person that they actually are."* (P-23 on M3). Furthermore, some participants mentioned their own lack of trust in the services being offered, e.g., *"They appear to be trustworthy on the surface but I simply do not trust the cryptocurrency market so by default, they are not trustworthy."* (P-126 on M8).

5.4.5 Number of followers are high (/low). Excluding M3 and M7-9, other scammer accounts had over 1K followers. For such accounts, 21% of participants attributed the trustworthiness to a high number of followers, e.g., *"This account is trustworthy account as there are more followers which seems to be a genuine account of a celebrity or person known by many people."* (P-24 on M4). These participants did not realize the followers could have been spoofed.

On the other hand, a lack of followers made 10% of participants suspect the bold claims in a scammer account's description, e.g., *"He has made over a million trades and has less than 1,000 followers? That doesn't feel very real or trustworthy."* (P-111 on M8). Furthermore, 4% of the times participants mentioned low followers as a reason for stating that either they don't know if the account is trustworthy or it is not possible to determine that. Some of these users even noticed that the follower-to-following ratio was suspicious, e.g., *"Following to follower ratio is a little suspicious, but I can't outright say he's not trustworthy. Somewhere in the middle."* (P-146, an expert, on M8).

Notably, this strategy was never used by any user in the novices group. A possible explanation is that such users paid less attention to platform indicators, such as the number of followers and primarily made their decision to trust an account based on the content of posts and account description.

5.4.6 Number of posts and followings. Some participants (10%) determined the trustworthiness of an account based on the number of posts or followings. A few (1%) times that participants saw a scammer account, this strategy was mentioned and the answer choice for the account's trustworthiness was either *don't know* or *not possible*, e.g., *"While there are many people following, this person hasn't posted a lot and I would be leary of trusting, but there really isn't a red flag either."* (P-30 on M2).

5.4.7 Lack of familiarity. Few (3%) of the participants took a conservative approach and indicated that they cannot say an account is trustworthy unless they are familiar with the account owner, e.g., *"I don't know this person personally. I would not reach out to anyone that I didn't know on Instagram."* (P-23 on M2). Some clarified further, e.g., *"I just don't know about this account, they seem to be making a living off of teaching about bitcoin but how do I know they aren't just living off of the money they make educating people as opposed to using the strategy they are trying to sell."* (P-149 on M2). Interestingly, 83% (5/6) of these participants belong to the moderate group, with one remaining person being an expert. Moreover, 67% (4/6) of these participants were females. Both of these are consistent with our observations in Section 5.2 where more women and moderates used lack of familiarity as a strategy towards deciding to not send a DM.

5.4.8 Checking for account verification. An official verified-badge is a useful indicator of determining an account's trustworthiness. In our survey, B2 was the trusted by highest number (71%) of participants. However, only some (26%) of the participants indicated account verification as a reason for trusting B2. Interestingly, while verification was chosen as one of the indicators of trust at least once by many (71%) of the participants, it was explicitly mentioned as a strategy to determine an account's trustworthiness by only 36% of these participants. This indicates that many participants

never used verification as a strategy to trust (or distrust) an account even though they later (in part 3 of the roleplay section) displayed knowledge of the verified-badge being a relevant indicator of trust.

In the cases where participants saw a scammer account, the lack of account verification was mentioned as a reason for the account being untrustworthy 5% of the times, e.g., *"The owner is not verified and no information about him is provided, he also follows more people than those who follow him."* (P-136 on M8). We find that only 14% of participants chose a lack of verification as a strategy for finding an account untrustworthy. Notably, no participant in the novices group chose this strategy.

5.4.9 Insufficient information. It is worth noting that by asking users to determine an account's trustworthiness, we did not assume that this decision can (or should) be made solely from the limited and easy-to-spoof account-information presented to participants. Instead, we carefully designed this question to study whether participants would find this information sufficient to make this decision. However, it was only 25% of the times participants saw a scammer account that they indicated its trustworthiness cannot be determined based on the available information. Some participants indeed noted that they can't trust the account without validating its claims, e.g., *"It's hard to tell if this account is trustworthy. There's no way to validate the qualifications or claims of the individual so I can't tell."* (P-21 on M10). Some were rightly skeptical about the credibility of the information, e.g., *"I just can't tell, could be real, or it could just be someone using someone else's pictures and someone else's name."* (P-36 on M2). Indeed, participant P-36 did not trust any account they saw (including benign accounts) primarily due to this reason. Overall, 9% of the participants did not trust any account in the survey and used insufficient information as a strategy nearly half (49%) the times.

5.5 Estimating the Age of an Account

We find that most participants (86%) inferred the age of at least one of the two scammer accounts (both having their first post only one day old) to be several weeks or more. Majority (79%) of the participants based this judgement on the number of posts and followers. Some (8%) participants indicated that they cannot tell the age with the information available, and 4% said that they don't know how to determine the age.

On assuming the freedom to navigate the account's page on Instagram, some (17%) of the participants said that they don't know how to utilize this in estimating the age. Furthermore, several (45%) participants indicated they will just use the number of followers and posts, while 31% (68/177) gave unclear explanations. Finally, only 20% (35/177) of the participants indicated that they will look at the date of the first post. Notably, there was no novice user among these 35 participants.

In Figure 2, we see that age of an account was deemed important for trusting a scammer account several (29%) times. However, the above results show that most participants either do not know how to estimate an account's age or they do it using indicators that can easily be manipulated (e.g., number of posts or followers). Therefore, many people may form an inaccurate notion of trust for an account. We discuss the implications of these results on future work in Section 6.4.

6 DISCUSSION

6.1 Key Takeaways and Analysis

We made the following key findings in this work.

First, 84% of the participants indicated that they would send a DM to at least one scammer and made such choices over half (52%) of the times, indicating a high susceptibility to DM-Me scam. Most participants (91%) trusted at least one scammer and made this mistake more than half (55%) of the time on average. There was a significant association between trust and sending a DM to a scammer account (χ^2=139.77, p<0.001) and therefore, we found that people who trust an account are likely to send a DM to them (r=0.41, p<0.001). Naturally, participants who found an account untrustworthy or couldn't determine its trustworthiness were less likely to DM them (r=0.37, p<0.001). Nevertheless, it is worth noting that over half the participants (53%) chose to DM at least one account that they did not trust, primarily because they do not consider sending a DM as a high-risk activity, e.g., *"Because even though I'm a little bit sketched out by the "_fx" on her username because, again, this may be some Forex scam, I am kind of intrigued by the fact that her and her partner paid off so much debt in 9 months like one of her pictures says. I don't see any harm in at least messaging her."* (P-120 on M10). Although sending a DM is a crucial step in the DM-Me scam, it appears that this doesn't seem to be a hurdle for many participants due to the perceived likelihood of any immediate negative consequences being low. Notably, participants said they were likely to DM an account 31% of the the time when they found account was not trustworthy, and 43% of the time when they couldn't determine its trustworthiness.

Second, we find that women are relatively less susceptible to financial DM-Me scams than men. In part, this is because more women appear to be using a conservative strategy of neither trusting nor sending a DM to unfamiliar accounts. Another reason is that more women appear to be uncomfortable in sending a DM to fulfill a need, and this discomfort is heightened when the service involved is financial. Notably, nearly half (46%) of such female participants belonged to the 31-40 age-group, and just over half of them (54%) use Instagram a few times a week. Overall, this makes such users less susceptible to financial scams in our study which did not include any accounts that participants would easily be familiar with and provided needs that can be fulfilled through alternate and potentially more reliable means (other than sending a DM).

Third, we found that novice users are more susceptible to the DM-Me Scam than experts. Majority (93%) of novices said they would DM at least one scammer as compared to 86% of moderates and 77% of experts. On average, novice users made such choices more often (76%) than moderates (55%) and experts (36%). Examining their reasons revealed that novice users predominantly used weaker strategies, such as trusting someone because they look harmless or qualified. In doing so, such participants naively assume the information presented to them as credible, while not appropriately interpreting any conspicuous red-flags (e.g., unrealistic claims). Notably, novices never used a conservative strategy like familiarity or a robust strategy like lack of verification to not trust someone. Overall, this leaves them more vulnerable.

Fourth, we examined the behavior of 16% (34/214) participants who did not DM any scammer account to understand

why they are least susceptible. This group comprises of 16% (21/129) of all males, 12% (10/81) of all females and 75% (3/4) who chose other. From the perspective of security knowledge, we found that this group comprised of only 2 (out of 27) novice users, with the rest being 23% (13/56) of all experts and 15% (19/131) of all moderates. On exploring their strategies, we find that female participants who are in the moderate group predominantly avoid sending a DM due to their discomfort in sending a DM or trusting unfamiliar accounts. On the other hand, male participants in the moderate group did not send a DM primarily because they felt uncertain about the trustworthiness of the account, and indicated that they will do more research before making a decision. Finally, both male and female experts in this group indicated that they will not send a DM because of specific cues that led them to conclude that the account is untrustworthy. Moreover, for well-crafted scammer accounts that the other 77% of the experts fell for, these 23% (13/56) of the experts generally found them trustworthy as well; however, instead of sending a DM, they indicated they that will use an alternate medium, e.g., "I wouldn't use DM to contact them since it doesn't sound like that's the main way they bring in business - I'd look for a website or email?" (P-49, an expert, on M7).

6.2 Limitations

Our study has several limitations that limit the scope of our results. First, our sample size was relatively small and was drawn only from Amazon MTurk users in the United States. This is not expected to be representative of other populations, such as non-U.S. users.

Second, our study follows a roleplay format, which is an approximation of a real-world setting. In such studies, participants might indicate a higher willingness to perform a risky task if they consider their actions would not adversely affect them [26]. Similarly, some participants might be more conservative in their behavior as they are not risking any real-world opportunity costs. Nevertheless, by examining the open-ended reasons, we find that our participants took the roleplay task seriously and responded in ways close to how they would if they were to encounter such accounts in the real-world. This is in line with prior research that has shown that whereas people's roleplay behavior tends to be slightly less cautious than real-world settings [25], their overall pattern of behavior and use of strategies is very similar [8, 26].

Third, our study uses Instagram as a platform and utilizes a manual approach to identify scammer accounts. Note that the primary goal of this paper is to study the susceptibility of users in falling victim to DM-Me scams, instead of studying the entire DM-Me scam ecosystem across platforms and covering every possible narrative. Nevertheless, we conducted our preliminary investigation to the extent deemed necessary for motivating our user study. We selected Instagram because it is the most representative platform where the DM-Me scam is predominantly prevalent [1–3] and one that offers a diverse set of accounts and scam narratives. To expand on our methodology in future work, we would like to enable performing data collection on a larger scale so that researchers can better understand the prevalence of the DM-Me scam in the wild and systematically study the characteristics, strategies, and various other narratives of such scammers. To this end, we have implemented an automated crawler that can fetch information on potential scammer accounts in Instagram, once the necessary consent is taken from the platform. The reference code for our automated crawler can be found on Github[2]. Furthermore, our work can also be extended to other platforms such as Twitter. Figure 3 in Appendix B shows sample screenshots of two Instagram accounts that we used in our survey (top-row), and examples of suspicious accounts offering similar services on Twitter (bottom-row). We leave further research on the following as separate future work (1) studying the prevalence of DM-Me scams across platforms, (2) comparing attack-strategies between platforms, (3) identifying the difference in susceptibility of users to DM-Me scams based on the target platform, and (4) studying the effectiveness of potential defenses across platforms.

6.3 Ethical Considerations

The following ethical considerations informed our work. For the preliminary investigation, we reached out to Instagram for consent to crawl their platform to automatically retrieve accounts and posts that are returned as search results for DM-Me hashtags. However, we did not receive a positive response. Therefore, we conform to their Terms of Use and conduct the measurement manually on a smaller scale. This limitation naturally reduces the sample space of our study. Nevertheless, we believe that our findings from this limited sample sufficiently represent a collective whole, and further shows that the risk posed by DM-Me scammers is becoming worrisome. Our overall research protocol, recruitment process for the survey, and related materials (e.g., survey questions, corresponding images of scammer and benign accounts, and answer choices) were reviewed and approved by our Institutional Review Board (IRB). We followed IRB guidelines to ensure that (1) participants can review an information sheet to determine if they would like to participate in this study and are allowed to leave at any time, and (2) the identity of the subjects is anonymous and no personally identifiable information is collected.

6.4 Implications for Future Work

Our study revealed that many participants were convinced of a scammer being trustworthy, which made them more likely to send a DM. We believe it is imperative for platforms to better aid users in assessing an account's trustworthiness, given that most of the user-controlled information can easily be fabricated. Currently, many platforms show a badge for verified accounts. When this indicator is present, we find that it is deemed an important factor. Unfortunately, though, the lack of this verified-badge was not deemed equally important, and users instead utilized other visible indicators. Moreover, a high number of posts and followers often led participants to believe that the age of the account is high, which may not necessarily be true. We feel that explicitly showing the age of an account may contribute towards preventing some users from forming an inaccurate notion of its trustworthiness, especially when the age is shown to be low. For instance, a one-day-old account with thousands of followers and hundreds of posts may help raise users' suspicion. However, introducing this indicator alone may not be enough as a scammer could abuse it by purchasing aged-accounts. Nevertheless, this will increase the cost incurred by a scammer. We leave more research on the usability, reliability, and effectiveness

[2]https://github.com/dmmeresearch/InstagramCrawler

of this strategy for future work. Another research direction could be in studying ways to educate users or in developing tools that help in spotting bait-like content and other red-flags in an account.

7 CONCLUSION

In this paper, we studied the susceptibility of users in falling victim to an emerging scam which we termed as the DM-Me scam. Although this scam presents an additional hurdle by requiring users to send a DM, our roleplay-based study revealed that this step does not deter most users. We found that a majority (84%) of the participants would send a DM to at least one scammer and made this choice more than half (52%) the time, indicating a high susceptibility to this scam. Two keys reasons emerged as to why this scam works in practice. First, participants often had misplaced trust in scammer accounts, which made them likely to send a DM. Second, even when participants did not fully trust an account, they were often likely to send a DM because it appeared a low-risk activity. We also found female participants mostly from the 31-40 age-group and who predominantly use Instagram a few times a week to be less susceptible than men for financial services-based DM-Me scams, as more women appeared to be uncomfortable in trusting or initiating a conversation with unfamiliar accounts for sensitive services. Novice users, with the least security knowledge, were also more vulnerable than experts due to them naively finding information presented by scammers to be credible. In conclusion, we believe it is imperative to better assist users in assessing the trustworthiness of an account. We hope this work will inspire future research, such as in exploring the effectiveness of showing the age of an account or developing educational tools that help in spotting red-flags.

ACKNOWLEDGMENTS

This material is based upon work supported in part by the National Science Foundation (NSF) under Grant no. 1816497. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

REFERENCES

[1] 2016. https://www.zerofox.com/blog/zerofox-research-publishes-instagram-scam-whitepaper/.
[2] 2022. https://www.forexfraud.com/news/young-people-targeted-by-instagram-scams-that-means-you-too/.
[3] 2022. https://help.instagram.com/514187739359208/.
[4] Amit A Amleshwaram, AL Narasimha Reddy, Sandeep Yadav, Guofei Gu, and Chao Yang. 2013. CATS: Characterizing automation of Twitter spammers.. In *COMSNETS*. 1–10.
[5] Spiros Antonatos, Iasonas Polakis, Thanasis Petsas, and Evangelos P Markatos. 2010. A systematic characterization of IM threats using honeypots. In *ISOC Network and Distributed System Security Symposium (NDSS)*.
[6] Joshua JS Chang. 2008. An analysis of advance fee fraud on the internet. *Journal of Financial Crime* (2008).
[7] Rachna Dhamija, J Doug Tygar, and Marti Hearst. 2006. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 581–590.
[8] Julie S Downs, Mandy Holbrook, and Lorrie Faith Cranor. 2007. Behavioral response to phishing risk. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*. 37–44.
[9] Julie S Downs, Mandy B Holbrook, and Lorrie Faith Cranor. 2006. Decision strategies and susceptibility to phishing. In *Proceedings of the second symposium on Usable privacy and security*. 79–90.
[10] J Erkkila. 2011. Why we fall for phishing. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems CHI 2011*. ACM, 7–12.
[11] Shehroze Farooqi and Zubair Shafiq. 2019. Measurement and Early Detection of Third-Party Application Abuse on Twitter. In *The World Wide Web Conference*. 448–458.
[12] Ana Ferreira, Lynne Coventry, and Gabriele Lenzini. 2015. Principles of persuasion in social engineering and their use in phishing. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*. Springer, 36–47.
[13] Ian Fette, Norman Sadeh, and Anthony Tomasic. 2007. Learning to detect phishing emails. In *Proceedings of the 16th international conference on World Wide Web*. 649–656.
[14] Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y Zhao. 2010. Detecting and characterizing social spam campaigns. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. 35–47.
[15] Grant Ho, Aashish Sharma, Mobin Javed, Vern Paxson, and David Wagner. 2017. Detecting credential spearphishing in enterprise settings. In *26th {USENIX} Security Symposium ({USENIX} Security 17)*. 469–485.
[16] JingMin Huang, Gianluca Stringhini, and Peng Yong. 2015. Quit playing games with my heart: Understanding online dating scams. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 216–236.
[17] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007. Protecting people from phishing: the design and evaluation of an embedded training email system. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 905–914.
[18] Alexandra Kunz, Melanie Volkamer, Simon Stockhardt, Sven Palberg, Tessa Lottermann, and Eric Piegert. 2016. Nophish: evaluation of a web application that teaches people being aware of phishing attacks. *Informatik 2016* (2016).
[19] Mehrnoosh Mirtaheri, Sami Abu-El-Haija, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. 2019. Identifying and analyzing cryptocurrency manipulations in social media. *arXiv preprint arXiv:1902.03110* (2019).
[20] Tyler Moore, Nektarios Leontiadis, and Nicolas Christin. 2011. Fashion crimes: trending-term exploitation on the web. In *Proceedings of the 18th ACM conference on Computer and communications security*. 455–466.
[21] Kaan Onarlioglu, Utku Ozan Yilmaz, Engin Kirda, and Davide Balzarotti. 2012. Insights into User Behavior in Dealing with Internet Attacks.. In *NDSS*.
[22] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods* 46, 4 (2014), 1023–1031.
[23] Vaibhav Rastogi, Rui Shao, Yan Chen, Xiang Pan, Shihong Zou, and Ryan Riley. 2016. Are these Ads Safe: Detecting Hidden Attacks through the Mobile App-Web Interfaces.. In *NDSS*.
[24] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. 2019. How well do my results generalize? comparing security and privacy survey results from mturk, web, and telephone samples. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1326–1343.
[25] Stuart E Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. 2007. The emperor's new security indicators. In *2007 IEEE Symposium on Security and Privacy (SP'07)*. IEEE, 51–65.
[26] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. 2010. Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 373–382.
[27] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd symposium on Usable privacy and security*. 88–99.
[28] Saniat Javid Sohrawardi, Akash Chintha, Bao Thai, Sovantharith Seng, Andrea Hickerson, Raymond Ptucha, and Matthew Wright. 2019. Poster: Towards robust open-world detection of deepfakes. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 2613–2615.
[29] Huahong Tu, Adam Doupé, Ziming Zhao, and Gail-Joon Ahn. 2019. Users really do answer telephone scams. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 1327–1340.
[30] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, Philipp Rack, Marco Ghiglieri, Peter Mayer, Alexandra Kunz, and Nina Gerber. 2018. Developing and evaluating a five minute phishing awareness video. In *International Conference on Trust and Privacy in Digital Business*. Springer, 119–134.
[31] Monica T Whitty. 2015. Mass-marketing fraud: a growing concern. *IEEE Security & Privacy* 13, 4 (2015), 84–87.
[32] Monica T Whitty and Tom Buchanan. 2012. The online romance scam: A serious cybercrime. *CyberPsychology, Behavior, and Social Networking* 15, 3 (2012), 181–183.
[33] Pengcheng Xia, Haoyu Wang, Bowen Zhang, Ru Ji, Bingyu Gao, Lei Wu, Xiapu Luo, and Guoai Xu. 2020. Characterizing cryptocurrency exchange scams. *Computers & Security* 98 (2020), 101993.

A SURVEY QUESTIONS

A.1 Part 1: Preliminaries.

(1) What is Instagram?
- ○ Instagram is a free email service developed by Google
- ○ Instagram is a photo and video-sharing social networking service owned by Facebook Inc.
- ○ Instagram is primarily a shopping website
- ○ Instagram is a food-delivery application

(2) How frequently do you use Instagram?
- ○ Several times a day
- ○ Once a day
- ○ Few times a week
- ○ I have used it in the past, but I no longer use it
- ○ I have never used Instagram
- ○ Other (please specify)

(3) Please identify your age range.
- ○ 18 - 30
- ○ 31 - 40
- ○ 41 - 50
- ○ 51 - 60
- ○ 61+

(4) To which gender identity do you most identify?
- ○ Male
- ○ Female
- ○ Other, or prefer not to say.

(5) Please specify the highest degree or level of school you have completed.
- ○ Some high school credit, no diploma or equivalent
- ○ High school graduate, diploma or the equivalent
- ○ Some college credit, no degree
- ○ Trade/technical/vocational training
- ○ Bachelor's degree
- ○ Master's degree
- ○ Professional degree
- ○ Doctorate degree
- ○ Other (please specify)

(6) Do you have an Information Technology related degree?
- ○ Yes
- ○ No
- ○ Decline to answer

(7) Which of the following devices do you use primarily?
- □ Desktop
- □ Laptop
- □ Tablet
- □ Mobile or Smartphone

A.2 Part 2: Scenario-based questions

NOTE: In this section, we want you to roleplay an Instagram user named Pat Jones. For each question, we will provide a scenario. Please respond to the question according to that scenario as if you were Pat Jones.

(8) Consider the following scenario. **You would like to get a customized drawing created as a gift for your friend's birthday.** While browsing Instagram, you come across the account shown in the image below. How likely are you to send a direct message to this person on Instagram, such as for requesting a customized doodle or making an inquiry?
[Screenshot of account's profile page shown here]
- ○ Very likely
- ○ Moderately likely
- ○ I won't send a direct message (DM) to this person
- ○ I don't know
- ○ Other (please specify)*[free text]*

Please provide the reason behind the choice you made. *[free text]*

(9) For the account shown below, what is the service being offered that is mentioned in the account's description?
[Screenshot of account's profile page shown here]
- ○ Gives guidance on stock investment upon request
- ○ Creates drawings, doodles, and other artwork upon request
- ○ Provides discounts on travel bookings
- ○ Provides tips on increasing followers on Instagram
- ○ I don't know

(10) Consider the following scenario. **You have been reading articles online to learn about money management and financial planning.** While browsing Instagram, you come across the account shown in the image below. How likely are you to send a direct message to this person on Instagram, such as for making an inquiry? (same options as Q8)

(11) Consider the following scenario. **You wish to get a verified badge for your account on Instagram.** While browsing Instagram, you come across the account shown in the image below.How likely are you to send a direct message to this person on Instagram, such as to make an inquiry? (same options as Q8)

(12) Consider the following scenario. **You are considering making new investments, such as investing money in stocks or cryptocurrency.** While browsing Instagram, you come across the account shown in the image below. How likely are you to send a direct message to this person on Instagram, such as to make an inquiry? (same options as Q8)

(13) Consider the following scenario. **You are interested in privately seeking dating advice.** While browsing Instagram, you come across the account shown in the image below. How likely are you to send a direct message to this person on Instagram, such as to make an inquiry? (same options as Q8)

(14) Consider the following scenario. **You are exploring ways to promote your company's presence on Instagram.** While browsing Instagram, you come across the account shown in the image below. How likely are you to send a direct message to this person on Instagram, such as to make an inquiry?(same options as Q8)

(15) Consider the following scenario. **You would like to get a customized drawing created as a gift for your friend's birthday.** While browsing Instagram, you come across the account shown in the image below. Do you think this account is trustworthy? Why or why not?
[Screenshot of account's profile page shown here]

- ○ I can tell this account is trustworthy
- ○ I can tell this account is NOT trustworthy
- ○ It is not possible to determine trustworthiness of this account with the given information
- ○ I don't know how to determine if this account is trustworthy

Please provide the reason behind the choice you made. *[free text]*

(16) Consider the following scenario. **You have been reading articles online to learn about money management and financial planning.** While browsing Instagram, you come across the account shown in the image below. Do you think this account is trustworthy? Why or why not (same options as Q15)?

(17) Consider the following scenario. **You wish to get a verified badge for your account on Instagram.** While browsing Instagram, you come across the account shown in the image below. Do you think this account is trustworthy? Why or why not (same options as Q15)?

(18) Consider the following scenario. **You are considering making new investments, such as investing money in stocks or cryptocurrency.** While browsing Instagram, you come across the account shown in the image below. Do you think this account is trustworthy? Why or why not (same options as Q15-17)?

(19) Consider the following scenario. **You are interested in privately seeking dating advice.** While browsing Instagram, you come across the account shown in the image below. Do you think this account is trustworthy? Why or why not (same options as Q15-17)?

(20) Consider the following scenario. **You are exploring ways to promote your company's presence on Instagram.** While browsing Instagram, you come across the account shown in the image below. Do you think this account is trustworthy? Why or why not (same options as Q15-17)?

(21) For the account shown below, what is the work title that is mentioned in the account's description?
[Screenshot of account's profile page shown here]
- ○ Lawyer
- ○ Designer
- ○ Dating Coach
- ○ Artist
- ○ I don't know

A.2.3 What indicators influence your decision to trust this account? [questions 22-27]. Note: some of the indicators given in the options may not be explicitly visible in the account shown. You can still choose that option if it would matter to you had the information been available. Also, you can choose multiple options.

(22) Consider the following scenario. **You would like to get a customized drawing created as a gift for your friend's birthday.** While browsing Instagram, you come across the account shown in the image below. What indicators would influence your decision in deciding if the below account is trustworthy?
[Screenshot of account's profile page with various indicators labeled shown here]
- □ Number of followers of this account
- □ Number of followings of this account
- □ Number of posts by this account
- □ Age of this account, i.e., how long this account has existed on Instagram
- □ Profile picture of this account
- □ Account description
- □ Full name of this account
- □ Username of this account
- □ Content of posts made. by this account
- □ Content of stories created by this account
- □ Verification of the account by Instagram, i.e., if the account has a verified-badge or not.
- □ Other (please specify)*[free text]*

(23) Consider the following scenario. **You have been reading articles online to learn about money management and financial planning.** While browsing Instagram, you come across the account shown in the image below. What indicators would influence your decision in deciding if the below account is trustworthy?

(24) Consider the following scenario. **You wish to get a verified badge for your account on Instagram.** While browsing Instagram, you come across the account shown in the image below. What indicators would influence your decision in deciding if the below account is trustworthy?

(25) Consider the following scenario. **You are considering making new investments, such as investing money in stocks or cryptocurrency.** While browsing Instagram, you come across the account shown in the image below. What indicators would influence your decision in deciding if the below account is trustworthy?

(26) Consider the following scenario. **You are interested in privately seeking dating advice.** While browsing Instagram, you come across the account shown in the image below. What indicators would influence your decision in deciding if the below account is trustworthy?

(27) Consider the following scenario. **You are exploring ways to promote your company's presence on Instagram.** While browsing Instagram, you come across the account shown in the image below. What indicators would influence your decision in deciding if the below account is trustworthy?

A.3 Part 3: Decision-making and computer knowledge

(28) For the Instagram account shown below, what can you say about the age of the account on Instagram? That is, how long do you think this account has existed on Instagram?
[Screenshot of account's profile page shown here]
- ○ I think this account has been on Instagram for less than a week
- ○ I think this account has been on Instagram for several weeks
- ○ I think this account has been on Instagram for several months

- I think this account has been on Instagram for several years
- I cannot say anything about the age of this account from the given information
- I don't know

Please provide the reason behind the choice you made. *[free text]*

(29) Consider the following scenario. You visit the Instagram account shown in the image below. You can assume that you are free to navigate the page as you like. How will you estimate the age of this account on Instagram? That is, how will you estimate how long this account has existed on Instagram?
[Screenshot of account's profile page shown here]
- I will estimate the age by: *[free text]*
- I don't know how I will estimate the age of this account

(30) For the Instagram account shown below, what can you say about the age of the account on Instagram? That is, how long do you think this account has existed on Instagram (options same as Q28)? Please provide the reason behind the choice you made. *[free text]*

(31) Please choose the best definition of the term **phishing** from the below options:
- Something that protects your computer from unauthorized communication outside the network
- Something that watches your computer and send that information over the Internet
- Something websites put on your computer so you don't have to type in the same information the next time you visit
- Something put on your computer without your permission, that changes the way your computer works
- Email trying to trick you into giving your sensitive information to thieves
- Email trying to sell you something
- Other software that can protect your computer
- Other software that can hurt your computer
- I have seen this word before but I don't know what it means for computers
- I have never seen this word before
- Decline to answer
- Other (please specify)*[free text]*

(32) Please choose the best definition of the term **virus** from the below options (same as Q31).

(33) Please choose the best definition of the term **spyware** from the below options (same as Q31).

(34) Please choose the best definition of the term **cookie** from the below options (same as Q31).

B EXAMPLE ACCOUNTS

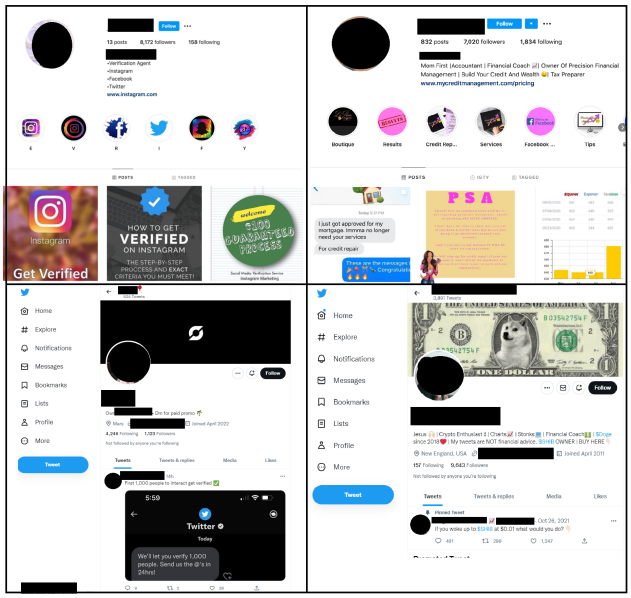

Figure 3: Example accounts. Top-row (left to right) shows images for accounts M11 and M6, respectively, that were used in the survey. Bottom-row shows corresponding examples of similar suspicious accounts on Twitter.