Manufactured Narratives: On the Potential of Manipulating Social Media to Politicize World Events

Chris Tsoukaladelis
Stony Brook University

Nick Nikiforakis
Stony Brook University

***Abstract*—Social media platforms are as popular as ever. Billions of users worldwide log in daily to one or more of them to connect with friends, share their day-to-day experiences, keep up with the news, and scout potential employers/employees. News organizations, either directly or through the help of citizen journalists, regularly cite posts and comments made on social media for news stories. A subset of these stories is focused on individuals responsible for radical, sometimes violent actions. It is not uncommon for stories to appear on news platforms after a citizen has committed a crime, attempting to link their posts to their actions, in order to ascertain how to best establish a baseline of radical behavior online, and prevent violent extremism in the future. Things, however, are not always as simple.**

In this paper, we look into the previously ignored possibility that the posts cited are not genuine, or that the profile linked to an individual does not necessarily belong to them altogether. We select 11 social media platforms, and show that by performing seemingly innocuous actions on them, we can craft a fake narrative meant to disinform. We propose two distinct threat models, each tied to a different type of adversarial behavior, and measure the susceptibility of each social media platform across 18 dimensions, through custom experiments mapping on- and off-platform behavior to these threat models.

Through these experiments, we find that 10 out of the 11 social media platforms are susceptible to our threat models, and that the majority of them (8/11) embed links in posts in a way that allows attackers to change the final destinations of these links, reframing posts and ultimately manufacturing narratives.

1. INTRODUCTION

Since their inception, social media have been an integral part of day-to-day life [20]. Billions of people depend on them for keeping up with their family and acquaintances, as well as for the latest developments in politics, sports, and anything else they deem important. It is therefore no surprise that news publishers also look at social media for information to include in their stories.

Let us take a recent example: David DePape, a 42 year old, attacked Paul Pelosi with a hammer in October 2022. San Francisco police arrested him, and once the story was made public, multiple news sources [3], [4], [17] attempted to understand the motivation behind the attack and link his actions to a political ideology. This led to various blog posts being unearthed that linked DePape to various views, from far-right conspiracy theories like Gamergate [13] and Pizzagate [9], to fairies and the occult [18].

While it is entirely possible that an individual struggling with mental health issues could come across a multitude of fringe ideologies and become attached to them, we find that there is yet another possibility: that of a manufactured narrative. In this scenario, a third party, potentially politically motivated, created a blog and multiple *backdated* blog posts related to various far-right ideologies. Then, all they would have to do is register a website in the name of the individual that perpetrated the attack and leak that information to news publishers. Suddenly, there is an online footprint going back months or even years, of a website espousing extremist views linking the perpetrator to fringe political ideologies.

To further highlight an online footprint's temporal focus, we note the following: when visiting DePape's blog through the Internet Archive, we notice a prompt put forth by the Internet Archive: "Context about this archived resource can be found here." Following this prompt takes one to an AP article [12] which refers to the claim that this narrative might have been manufactured as follows: "THE FACTS: The websites existed before the attack, and had entries dating back years." We therefore observe the importance placed on the length of time the websites were up, as well as the age of the posts. Our focus is to show how politically-motivated actors can abuse these seemingly objective metrics.

Existing archiving technologies (such as the Internet Archive) are highly unlikely to detect this kind of abuse, since the frequency with which they crawl and index sites is highly irregular and subject to "login walls." On the research front, we can view this issue as one of integrity (where information is modified in "unauthorized" ways). There, while there has been considerable activity measuring how attackers abuse a lack of integrity to attack web users [31], [35], [38], [40], [46], [51], [52], [58]–[60], [65], [67], this activity has been in the context of traditional computer security. That is, we know little about how a lack of integrity on high-impact websites (like social media) can be abused for misinformation and disinformation.

In this paper, we set out to understand the feasibility of these types of integrity abuses where attackers can create seemingly backdated content to manufacture desired narratives for real-world events. Since a self-hosted web application provides infinite degrees of freedom to attackers

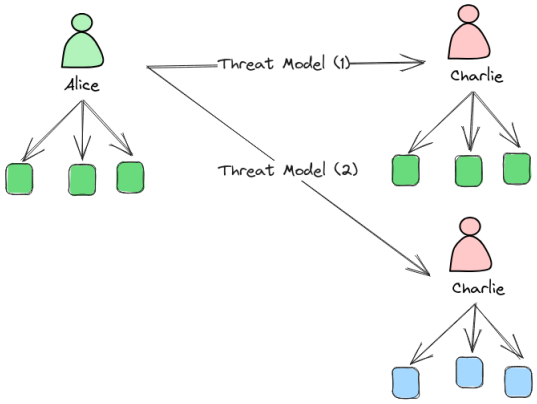

Figure 1: **Our two main threat models. On (1), the information on the user changes but the posts remain the same. On (2), both the user information and the posts change.**

regarding fake and backdated content, we focus our attention on popular social media websites and their resistance to these kinds of attacks. To this end, we create fake profiles on popular social media websites and experiment with the available options and settings regarding post visibility, profile pictures, and even usernames and handles. We review eleven social media platforms across 18 dimensions that could be abused to create online radicalized "puppet accounts." We present two different threat models and present our analysis regarding the susceptibility of the evaluated social-media platforms to our described attacks.

The primary contributions of this work are as follows:

- **Threat model:** We selected 11 social media platforms, in which we created fake profiles and evaluated them across 18 dimensions related to integrity attacks. We define two threat models that allow attackers to weaponize different aspects of each platform for the creation of manufactured narratives.
- **Social media crawler data:** As part of our analysis, we seek to understand how social media engages with off-platform content and whether it can detect changes to that content that are made after the fact.

2. MOTIVATION AND THREAT MODELS

News organizations depend on how quickly they can report a story after it breaks. The first news organization to report on a story tends to get its link shared more across social media, and as a result of that gets the lion's share of clicks and ad revenue. There is therefore a financial incentive for news organizations to react to a breaking story as fast as possible. On top of this, news organizations rely on social-media content to augment their stories, either discovering it themselves or having citizen journalists share relevant content with journalists. As a result, news organizations commonly include social-media posts, comments, and personal blogs in their reporting.

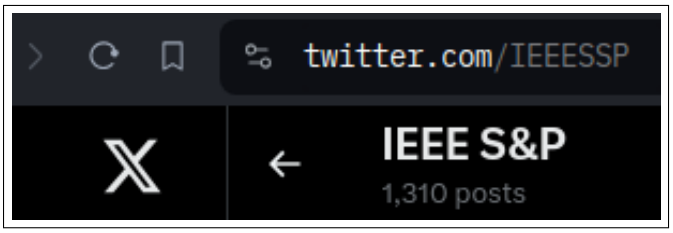

Figure 2: **An X/Twitter handle is part of the URL leading to that account's profile page.**

In the context of this paper, attackers can abuse this reliance on social media by attributing social-media profiles and posts to specific individuals and sharing their "findings" with news organizations. A far-left/far-right profile that can be linked to an unfolding real-world event can immediately color peoples' perceptions, even if the information is eventually discovered to be inaccurate.

The considered adversaries can include individuals that want to promote their own ideology by making the "other side" seem more prone to extremism, as well as nation states attempting to sow discord in a foreign country by pitting different political groups against one another. We define two possible attacks depending on the controls that social media platforms give to their users (visualized in Figure 1).

2.1. Changing the identity of a radicalized account

In our first threat model, attackers can create fake profiles that are already radicalized, i.e., they regularly post explicit far-left/far-right content. Whenever a real-world event unfolds which the attacker wants to use, they can merely change the identity of the existing account to the identity of a real-world person. Depending on what a social-media platform allows, this change could involve changing one's name, their profile picture, and potentially the visibility of some extreme older posts that could have led to a fake account's premature deletion. Prior work has shown that some social media even allow users to change their account names [43] which typically appear in the URLs of social-media platforms (such as IEEE S&P's X account shown in Figure 2) further increasing the believability of this fake account and defeating any history-based forensics (e.g. attempting to use archive sites [6], [15] to establish what a given profile used to link to).

2.2. Changing the content of a previously innocuous account

In this second threat model, attackers can not only change the identity associated with an account (Threat model #1) but can also change the content of existing posts. This could include the editing of existing posts (from benign posts to radicalized ones), the backdating of new extreme content, and even the deletion of content that previously balanced the account (e.g. deleting just the far-right posts from an account, swinging the resulting profile to far-left).

Moreover, in this threat model, attackers may be able to make off-platform changes with on-platform side effects. For example, attackers could link to innocuous articles via a link shortening service and then switch the destination of the short links, without ever modifying the on-platform posts. This would result not just in different destinations when the links are clicked but potentially to new link previews [61] in the fake user's profile.

3. METHODOLOGY

In this section, we describe the methodology we followed to assess how vulnerable social media websites are, against our presented threat models.

3.1. Social Media Platform Selection

Before we perform any integrity-related experiments, we must first compile the set of social media platforms to evaluate. Our goal here is to include some of the most popular social media platforms that billions of people use on a day-to-day basis. For our analysis, we selected 11 distinct social media platforms. We selected these platforms not just based on their popularity, but also their characteristics. Specifically, we aimed for a variety of ways for users to interact with one another (e.g. via video posts, text and images) as well as their deployment models (e.g. centralized vs. decentralized). We included all the "household" names in social platforms (Facebook, Twitter/X, LinkedIn etc.) and further augmented that list with the ones that could have been used in prior real-world incidents (such as DePape hosting a blog on wordpress.com)

Next, we created two accounts on each social media platform. Since most social media platforms are effectively "black boxes" with proprietary code and logic, we do not know the exact set of criteria that could be used to flag our accounts as fake. We therefore took reasonable precautions and created realistic accounts that would avoid many potential such "checks". First, we selected profile pictures from a website that creates realistic, computer-generated images of people who do not actually exist [21]. We then created a persona to fit that image using a fake name generator [11] which gives us a randomly generated name, birthday and occupation. We also registered a phone number for each "person" to use for signing up to the various social media platforms and further minimize the risk of being flagged as a "suspicious" account. We also "aged" our accounts before we began our experimenting. As mentioned, it is impossible for us to know what each platform might consider as "too young" regarding account age, so we made the decision of starting our experiments one month after the accounts were created in the manner described above, to minimize such risks.

Finally, we performed a set of experiments to gauge the degrees of freedom the platform allowed us to operate under, such as editing/deleting posts, sharing posts and backdating, editing user information.

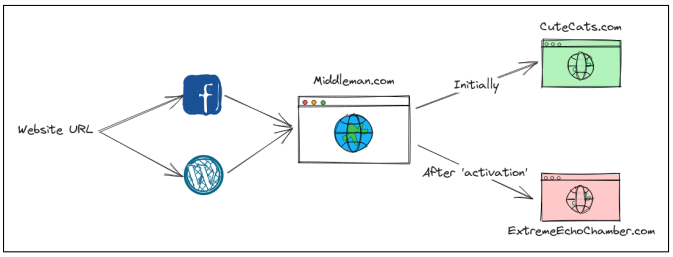

Figure 3: **Diagram of our redirect infrastructure. A URL is posted to various social media platforms pointing to our website. Then, our website redirects the traffic to a benign website. After a period of time, we instruct our website to start redirecting traffic to an extremist website instead.**

Ethical considerations. We took a number of precautions to ensure that our experiments were conducted ethically and did not have any negative effects, neither on the platforms themselves, nor on other users on these platforms. First, by limiting our fake accounts to just 2 per platform, we ensured that we did not place any undue strain on platforms that already have millions (if not billions) of user accounts. We did not attempt to connect with any real users on these platforms and we never posted any malicious content that could negatively affect other users or the platform itself. We also decided against posting extreme content on the studied platforms directly, in order to minimize potential damage in case someone randomly stumbled across our posts. We note that extreme content is **not** required for these attacks, as a concentration of moderately right or left leaning content would likely be sufficient in the context of politicizing a real-world event.

3.2. URL redirection

In addition to the experiments we performed on the social media platforms, we also registered a website that we had control over, and posted links pointing to that website on the evaluated social-media platforms. The server hosting our website acted like a URL redirection/URL shortening service (such as `bit.ly`) redirecting incoming HTTP requests to other websites. Our links initially pointed to benign Wikipedia articles and, as part of our experiments, eventually pointed to politically-fringe websites namely 4chan [1] and wsws [23]. This experiment aimed to simulate an attacker who can change off-platform content (the final destination site) and have a social-media platform reflect those changes without any new on-platform activity. Figure 3 shows the operation of our redirection infrastructure.

4. ANALYSIS

In this section we break down the experiments we performed regarding the possibility to manufacture after-the-fact narratives on the evaluated social-media platforms. Next to presenting individual results, we also provide examples of how features of each platform can be combined to per-

form attacks that follow the two threat models presented in Section 3.

4.1. Experiments and threat model

Table 1 shows the experiments that we performed against the evaluated social-media platforms that essentially constitute 18 different dimensions that attackers could weaponize in the context of integrity abuse. We are mostly interested in manipulating individual posts (e.g. changing a post's text), as well as changing the profile itself (e.g. via the changing of the account's username).

We can observe that parts of our threat models do not apply to some social media platforms. Specifically, social media websites that focus on multimedia as a primary way of making posts (such as Vimeo and TikTok) do not allow the editing of existing posts. Users could only simulate edits by deleting a post and making a new one. Similarly, to "share" a post in a number of platforms, one would have to manually copy the URL of that post, as opposed to embedding the post they are referring to (available on Facebook and X/Twitter). We therefore cannot evaluate the effect of modifying embedded posts on these platforms.

It is important to note that the majority of our tests should not be interpreted through a pass/fail lens. Allowing users to, for example, change their usernames does not in-and-of-itself constitute a weakness of the platform. It is only through the combination of some of these features with a lack of transparency that they become problematic allowing adversaries to weaponize them.

At the same time, there are some experiments that *can* be failed. For instance, there should be no reason for an account's creation date to be manipulated - something that indeed none of the platforms that we studied allows for. Similarly, when a post has been edited, there should be some way for an observer to see this. Furthermore, some combinations of experiments can constitute problematic behavior, for instance if one can backdate a post and there is no indication that this backdating occurred.

Having taken the above into account, through our experiments we can conclude that the platform that can be most abused by an adversary is Wordpress (the non-self-hosted blogging platform), as it allows users to manipulate their posts (i.e. via backdating and editing) without giving any indication that this occurred. Another platform that can be susceptible to adversaries is Facebook, as some of the visual cues that should be in focus when one edits a post, for instance, are hidden behind sub-menus. Contrastingly, a platform like Mastodon that gives users freedom to perform various actions like edits, also ensures that this information is prominently visible to other users by adding a timestamp for the last time an edit was performed. When clicking on that timestamp, a user can see the entire history of edits, along with the latest version of a post.

Overall, across all 11 platforms tested, 100% of them allow users to change their profile picture, while 91% (10 out of 11) of them allow users to change their username. As this combination is the basis of threat model (1), this means that virtually all social media profiles are vulnerable to attacks of this type. That is, attackers can construct profiles with evident political biases that are "keyed" to arbitrary names and, when a real-world event unfolds, merely change the name and profile picture of one of these profiles. Journalists who are scouring social media for posts will then discover these profiles, and include their findings in the resulting articles, thereby implicitly adopting the narrative that attackers manufactured.

Similarly, 72% (8 out of 11) of all platforms allow users to edit their posts, with the exceptions all being platforms that focus on media-based posts, like images or videos. Of these 8 platforms, 6 indicate directly that the post has been edited, 1 has this information hidden under a sub-menu and 1 does not report on it at all. Considering this is the basis of threat model (2), these 2 platforms are susceptible to it, with one being completely susceptible, while the other only partially.

Some platforms attempt to strike a balance between giving users the ability to modify posts without allowing adversaries to abuse that feature, via a temporary-editing feature. This effectively defeats the threat actors operating under group (2) of our threat model.

Below we present some examples of interesting behaviors, cross-referencing them against our threat models. As mentioned earlier, most platforms allow an adversary to casually change their username and profile picture. In many cases, it is also further possible to change a profile "handle", adding an extra layer of deception for the attacker. Specifically, it would nullify the effects of web archiving services, as a new "handle" would usually also mean a new profile URL for that account, such as in the case of X/Twitter. This effect is boosted when considering the ease with which someone can become "verified" in platforms such as X/Twitter [2], since some users may be perceiving verified accounts to be more official than non-verified ones [63].

Next, we revisit our introductory example regarding the perpetrator of the October 2022 attack. One of the websites DePape was using in our example is a Wordpress site, hosted on the `wordpress.com` domain. Since Wordpress allows someone to backdate posts (even earlier than the creation date of the website), all of the posts on that website could have been made at any point up to the moment the website was discovered by a news organization. We therefore see that the claim made in the AP news article [12] cited by the Internet Archive, i.e. that the "age" of the posts implies this profile was active for a long time, is not necessarily valid. In this specific case, considering the website first appeared on archive services as the news about it broke [5], the only way for the public to retrieve information regarding the website's creation would be through DNS records.

Another interesting observation revolves around the decision to show original vs. last-edited timestamps under posts. When coupled with a lack of other visual cues indicating an edited posted, users will not be able to discern an original post from one that has been edited. In the case of Facebook, while it is technically possible to discover when

TABLE 1: *Table of the experiments performed on each social media platform. If the experiment is possible, denoted with a ✓, or ✗ if not. In the case the experiment does not apply for that social media platform, marked with* N/A.

Experiment	Facebook	Twitter/X	Mastodon	Reddit	Vimeo	Instagram	TikTok	Youtube	Linkedin	Wordpress .com	Disqus
Delete post	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Share of post also deleted	✗[1]	✗[1]	✓	✗[1]	✓	N/A	N/A	N/A	N/A	N/A	N/A
Edit post text	✓	✓[2]	✓	✓	✗	✗[3]	✗	✓	✓	✓	✓[2]
Edit post media	✓	✓	✓	✓	✗	✗	N/A	N/A	✗[3]	✓	✗
Visual indication of edit	✗	✓	✓	✓	N/A	N/A	N/A	✓	✓	✗	✓
Timestamp of edit or original	Original	Edit	Both	Original	N/A	N/A	N/A	Original	Original	N/A	Original
Original post available on edit	✓[4]	✓	✓	✗	N/A	N/A	N/A	✗	✗	✗	✗
If shared, go to original post	✓	✓	✓	✓	✓	N/A	N/A	✓	✓	✓	✓
If shared, updated on edit	✓	✓	✓	✓	N/A	N/A	N/A	N/A	✓	N/A	N/A
If shared, indication of edit	✗	✓	✓	✗	N/A	N/A	N/A	N/A	✓	N/A	N/A
If shared, timestamp of edit	✗	✓	✓	✗	N/A	N/A	N/A	N/A	✓	N/A	N/A
Edit username	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
Edit handle	✗	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓
Edit profile picture	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓[5]	✓
Edit account creation date	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Schedule posts	✓	✓	✓[5]	✓[5]	✓	✓[5]	✓[5]	✓	✓	✓	✗
Backdate posts	✓[6]	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
URL redirect cache updated	✗	N/A	N/A	✓	N/A	N/A	N/A	N/A	✗	N/A	N/A

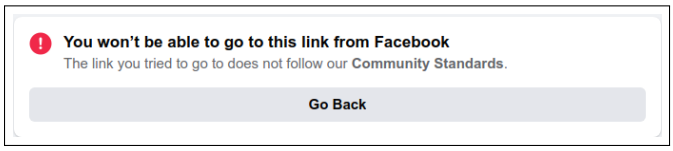

Figure 4: **Facebook link takedown.**

the post has been edited as well as the original post itself, this option is hidden inside a sub-menu and is not intuitive to a casual observer.

4.2. Cached redirects

A category that warrants its own discussion is that of URL redirects. There have been a number of research papers investigating URL shorteners over the years [26], [42], [50] and they all reach the same conclusion; URL shortening and redirections are often used maliciously in various security contexts, such as for spreading malware.

In this paper, we find that they are also a potential threat vector in the context of integrity, when crafting fake narratives. As described earlier, under a "traditional" content-editing scenario, an adversary would need to directly edit or somehow manipulate a post on their social media profile to reframe its content and/or context. When coupled with transparency regarding editing posts and other similar actions (such as changes in visibility) performed on a post, the plans of an adversary could potentially be interrupted.

1. The "shared" post remains, however the post being shared shows up as "deleted" or "removed".
2. Only during a period of time from the original post
3. Can only edit alt text/caption
4. Yes, but option is visually hidden behind menu
5. Yes, but through third-party service
6. Yes, but also indicates that the post has been backdated

However, when the adversary has control over a website they can use to forward traffic to other sides, they need only post the URL to that website and at some point "switch" the redirect from pointing to a benign website, to pointing to an extremist one. In our case, we set the "switch" point to be one week after the link has been posted. On top of this, the adversary's "intermediate" website does not contain any extremist content itself, therefore avoiding being flagged by the crawler of the social media website it is posted on. This does not apply to all the platforms we tested, as some platforms are intuitively less suited to sharing URLs. Such an example is TikTok, which is built around sharing short videos, or Instagram, which focuses on image sharing. Furthermore, some platforms like disqus that are more decentralized may disallow any links that appear suspicious, at the discretion of the community moderator(s).

An issue that complicates our evaluation of the described off-platform content switching is that of the different content policies of closed-source social-media platforms and how exactly a given platform decides to remove a posted link. While most platforms have public content policies [10], [14], [16], [19], [22], [24], [25], it is difficult to know what exactly would trigger a social media platform to "flag" or remove an account or post. For instance, when we shared the aforementioned URL via Facebook, it was initially posted properly. After a period of time however, *before* we "activated" the redirect to point to an extremist website, Facebook removed the post, as shown in Figure 4. This indicates that Facebook was concerned about the quality of the outgoing link itself, as opposed to detecting that an off-platform, content switch happened. It is reasonable to assume that some of the considered adversaries (such as in the case of nation states) will have enough time, budget, and sophistication to methodically build up profiles and external links in ways that they are not flagged as suspicious by the automated systems used by these platforms. This can include

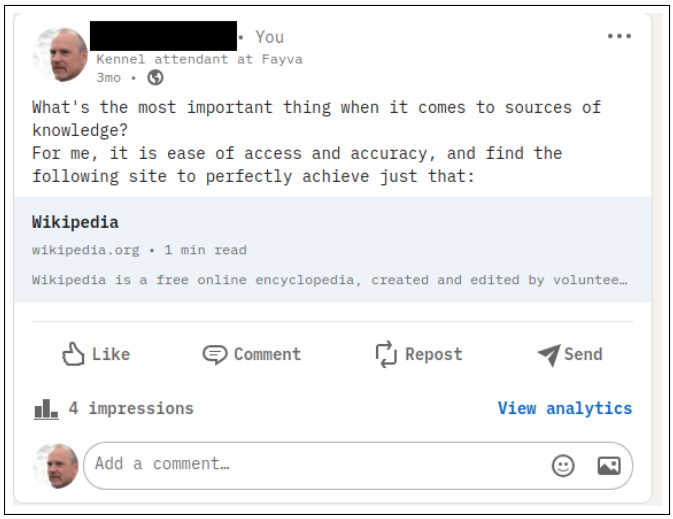

Figure 5: **LinkedIn URL redirection. In this case, we have already performed the "switch" on our website, and this URL when clicked now points to an extremist website, however LinkedIn is unaware of this and embeds a preview of Wikipedia.**

purchasing domain names that recently expired [41], [46], [48] (as opposed to registering new ones), purchasing backlinks to these domains to elevate PageRank-like metrics [64], deploying realistic and professional-looking content on the dummy websites, as well as distancing the registration date of the website with the first time that it used in a link on the platform.

With that issue into consideration, we present our findings regarding URL redirecting. Most websites allow users to post URLs in either the form of a post, or as a comment. Some websites even go as far as to show a preview of the webpage the URL points to (in our case, it would initially be the "benign" website our URL redirects to). Through testing, we identified that in one of our other profiles, Facebook did not delete the link and the post-switch preview remains the same. Similarly, LinkedIn did not update the embed after the link was switched, meaning that even after making the "switch", Facebook and LinkedIn still advertise that the post URL points to the "benign" website. An example of what this looks like can be seen on Figure 5.

Out of 11 platforms tested, 3 of them had URL previews. This effectively means that 72% of the platforms tested did not include any kind of automatically-extracted metadata, allowing attackers to surreptitiously change the destination of their links. Combined with sufficiently vague posts (e.g. "Check this out") we argue that these switched URLs could be clearly weaponized to manufacture after-the-fact political narratives and associate them with a targeted individual. Of the 3 platforms that did include the metadata, 2 retained the stale metadata after the "switch" was made to the extremist website, giving a preview of the original innocuous website.

To understand whether the evaluated platforms visited our off-platform website *after* our links were first posted, we analyzed the web-server logs on our hosts. We used unique links for each evaluated site in order to connect log entries to site crawlers, and also filtered out any requests that did not carry special HTTP parameters (such as a site-related user-agent), thereby removing all background bot and crawler activity from our analysis. TikTok was excluded from this analysis since links are only available in business accounts with sufficiently large sets of followers.

From the remaining social media platforms, we only observed re-crawls from Reddit and Disqus. Reddit performed a re-crawl after a period of approximately two weeks, which also refreshed the on-platform preview to correctly also point to the new site we made the "switch" to. As there was no traffic other than that stemming from the bot at that time, this re-crawl was not caused by a user or bot interacting with the post. Disqus does not appear to have its own web bots checking URLs posted through it, but we did get some traffic on its URL a day after we posted it. Based on the observed user-agent and timing, we attribute that visit to a community moderator who was evaluating our pre-switch URL (redirecting to Wikipedia) before flagging our post as spam. The rest of the social media platforms did not re-crawl our website after their initial visit.

5. DISCUSSION AND LIMITATIONS

In this section we discuss our findings and describe possible steps that can be taken by social media websites in order to combat dis/misinformation in the future. We also discuss the limitations of our experiments, and how they can be addressed in future work.

5.1. Defending against disinformation

The common thread between all these social media websites is that they are all susceptible to disinformation campaigns. Considering the ever-decreasing attention span of the average observer, it is important that social media platforms combat these worrying trends.

Visual cues. As shown by our threat model, an important way of dealing with the novel types of disinformation that we described in this paper is to visually show the user the changes made to a post, a claim supported by relevant literature [30], [66], [68]. For instance, if a post has been edited, other users should be able to straightforwardly discover that it was edited, the timing of the edit, and what the original content was. In this sense, we appreciated the intuitive UI of Mastodon, that gives a user both an original post timestamp, as well as an "edit" timestamp. Clicking on the "edit" timestamp, allows the observer to look at a history of edits, as shown on Figure 7. Furthermore, clicking on any of the "edit" timestamps on the drop-down, shows the corresponding post at that point in time to an observer. Platforms can make additional decisions regarding eventually expiring the original content (e.g. keep the original version of a post for one month, before deleting it) to strike a balance between transparency and storage demands on their infrastructure.

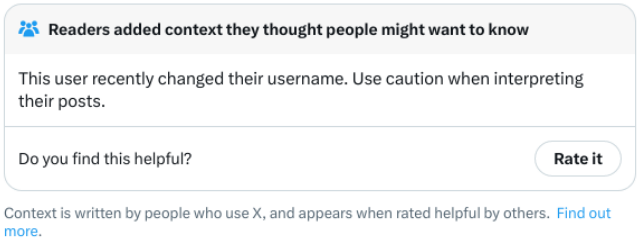

Figure 6: **Mockup of how a community-notes-like mechanism could alert users of important profile changes in the context of the integrity attacks described in this paper.**

Figure 7: **Mastodon shows both the "original" as well as "edit" timestamp on a post. Clicking on the "edit" timestamp shows the history of edits.**

Information on profile changes. Perhaps the most important way of stopping these potential attacks, at least in centralized social media, is by sharing information regarding when a profile has last been altered, i.e. through a username change. Social media platforms already collect this type of information, so making both the original as well as the new version available should be straightforward to implement. Using prominent visual indicators of these types of changes could protect journalists and other users from attackers who modify profile names in order to capitalize on real-world events.

At the same time, we realize that some individuals may want to disassociate from names and profile names that they once used on any given platform. As with our suggestion regarding post edits, this can be tackled by making both the new and the old profile information temporarily available, ensuring that the profile-change mechanism is not abused by attackers yet eventually allowing the platform to "forget" a user's old account names. Alternatively, social-media companies could merely have an indicator showing that a critical piece of profile information has changed, yet only release that information to trusted governmental and non-governmental bodies.

Community input. Another way to battle disinformation is to draw from a social-media platform's community. X/Twitter has taken some steps towards this via the "Community Notes" mechanism [7], providing a framework that other platforms can follow. The high-level idea is straightforward and tangent to our previous point regarding open sourcing parts of a platform's logic; by allowing users to write and vote on notes adding context to post, attempts to misinform or disinform by editing posts, or manipulating a profile, can be exposed by the community members themselves. Figure 6 shows a mockup of how such a notice could be shown to users of these platforms. Taking into account the different potential biases of community contributors as well as how they interact and rate prior notes, the pitfalls of biased individuals adding notes can be minimized.

5.2. Pre-existing accounts

It bears noting that a given user might have pre-existing social media accounts. In that case, the argument could be made that our threat models would not be as effective, as that user's profile would "shield" them from any potential manipulation from a malicious actor's part. We argue, however, firstly that it is extremely unlikely for someone to be on **every** social media platform, especially so when self-hosted blogs can be created with very small amounts of effort. Second, even if someone has innocuous posts on their real accounts, a secondary account can be crafted and planted with convenient ties to the real account, waiting to be discovered at the right time. In that case, we consider it a certainty that the secondary account, being more "caustic" and headline-worthy, would be considered a reflection of the person's "real" thoughts and ideology.

5.3. Limitations

In the undertaking of this research effort, we faced a number of obstacles which we discuss below:

Black box nature of studied platforms. The main difficulty we faced had to do with getting around the seemingly arbitrary rules of each social media platform. As mentioned in Section 4.2, for instance, Facebook removed a post we made containing a URL to a website we controlled, which then in turn pointed to Wikipedia. Facebook cited its content policy for the removal, but — to the best of our understanding — we had not violated any specific clause of that policy. The removal could be related to our unranked, recently registered domain which could have been deemed suspicious to the automated systems that Facebook operates. As such, particularly for the off-platform dimension of this work, different social-media companies must evaluate (using their in-house tools and knowledge of deployed systems) how our described integrity attacks interact with their existing anti-spam systems and whether these systems would be able to automatically capture them.

Account and website age. Another aspect that we considered was the age of all the social media accounts we created for the experiments described in this paper. Namely, soon after creating these accounts, we start experimenting

with different settings and posting content with off-platform links. It is therefore highly possible that the young age of our accounts was an additional signal that made our behavior suspicious to the existing anti-spam mechanisms that these platforms operatore. As such, we consider entirely reasonable that our uncovered set of issues are lower bounds of the true set of issues given the additional features and degrees of freedom given to more seasoned accounts. An interesting future direction would therefore involve evaluating the same attacks using older accounts as well as established off-platform websites.

Language barrier. Our focus was on social media platforms available primarily for English-speaking audiences. We therefore recognize that there may be a bias in our results that would only be uncovered when additional regional social networks and platforms are included. Of particular interest could be comparisons such as the one between TikTok and Douyin [8], its Chinese counterpart, for differences in policy and dealing with disinformation.

6. RELATED WORK

In the era of social media, false information, or fake news are prevalent more than ever before, with far-reaching implications [27], [39], [54], [55]. For this purpose, various researchers have made forays in detecting fake news, using a number of approaches such as deep learning [36], [47], or through investigating linguistic features [34], [44]. Singhal et al. [57] developed two frameworks to detect inaccurate phishing claims on Twitter and misinformation on social media platforms in general. The first framework gave them the ability to label 9% of URLs and 22% of tweets about phishing websites as misinformation. The second framework used supervised classifiers to identify posts discussing the security and privacy of Zoom, and was able to detect 3% to 18% of all posts about the subject, depending on the social media platform, were meant to disinform. Such studies give an added layer to existing misinformation, showing that it is not necessarily limited to the news or health issues, but can also be applied to subjects like technology.

There has also been extensive work in detecting fake users in social media [33], [45], [56], as well as the effects of fake news and misinformation on end users [28], [29]. Moravec et al's study on whether social media users could detect fake news online concluded that users' beliefs remained unshaken regardless of potential false news flags being present [49]. In short, users were more likely to believe news headlines that align with their political opinions and feed their confirmation bias, while those that challenge their opinions are ignored or less likely to be believed.

Guo et al. [32] looked at the subject of post-publication edits to news article titles, by collecting a dataset of over 400K articles and measuring the changes in the titles. Through this process, they are able to showcase the risks posed by news publishers inadvertently creating fake narratives themselves. They also studied social media (namely Twitter) influence over a period of time after the publication of a story, finding that stories achieve maximum propagation in the first few hours, thereby allowing news publishers to use a "clickbait" title on an article to get more traffic, before later changing it silently to a more factual one.

Tsoukaladelis et al. [62] expanded on this work by analyzing changes made in the article bodies as well, including changes in sentiment and meaning introduced by the edits to the content. They also studied the prevalence of "silent" changes to news articles, showcasing the need for informing users of any changes made.

Tangent to our research is the work done by Nikiforakis et al. [50] focusing on third-party link shortening services exposing users to unexpected threats, including malware and the exfiltration of private data. They also make the argument against "linkrot", i.e. when users share a shortened URL on a website they do not control, they unconsciously trust these URLs will work in the future as well, not expecting the shortening service to cease to be operational for instance, thereby causing the links to "rot". We expand on this reasoning in our own work, by expecting a URL shortening service could be controlled by an adversary altogether, or worse, that an adversary registers an expired domain belonging to a prior URL shortening service, and then having full control over the destinations of shortened URLs that were once legitimate.

Khaldarova et al. [37] discusses the role fake news played in the months leading up to, as well as the aftermath of, the annexation of Crimea by the Russian Federation in 2014. They focus on alleged fake news stories planted on Channel One, a Russian TV station, finding that an overwhelming majority of users on Twitter distrusted these stories, thereby making the case for community reviews of dubious stories. They further discuss the potential destructive force disinformation and fake news narratives can have over the real world, in the form of information warfare. A "real" entity such as Channel One can be argued to have been overtaken by an adversary, in this case the Russian government, seeking to leverage the residual trust that the Russian people have in its institution to spread disinformation.

Looking now at an example of the impact that a "fake" entity can have, Shao et al. [53] focus on misinformation campaigns ran by fake users, such as bots. They analyze 14 million messages spreading 400 thousand claims on Twitter during and following the 2016 U.S. presidential election, published by websites that routinely publish false or misleading news. Through their analysis, they found that the accounts that are the most active at sharing fake news have a statistically significant ($p < 10^{-4}$) difference in score from that of randomly selected users that also post at least one link to one such claim. The case is therefore made that the accounts that spread disinformation on social media platforms the most, are fake. In this case, an adversary can use the sheer volume of "fake", bot accounts to spread disinformation by making various subjects "trending". It is a different model compared to the one studied by Khaldarova et al, but both can be lucrative to an adversary depending on their goals and resources.

7. CONCLUSION

Social-media platforms remain as popular as ever, with new platforms regularly surfacing that tackle the perceived shortcomings of existing ones. Next to staying in touch with friends and family, these platforms are increasingly used to propagate breaking news, stay informed on different topics of interest, interact with like-minded individuals, and even investigate people in the context of hiring decisions. News organizations are also increasingly relying on social media when writing their articles, to find content that they can attribute to specific individuals. This is particularly true in the case of breaking events involving perpetrators, where (citizen) journalists search for posts and clues in social-media platforms to understand the character of specific people and find the motivation behind their actions.

In this paper we draw attention to the lack of content integrity in social-media platforms and how this can potentially be abused by attackers to manufacture narratives. Specifically, we show that almost all platforms allow their users to perform seemingly innocuous actions, like change their profile names, profile pictures, and edit their posts. Yet, in the context of disinformation, these actions can be abused by attackers to create fake accounts and link them — after the fact — to individuals involved in real-world events. In this way, attackers can trick journalists (and by extension their readers) in interpreting an event using a specific lens, such as, portraying the perpetrator of a crime as left wing or right wing. To this end, we analyze 11 social-media platforms across 18 dimensions, noting not just the on-platform features that attackers can abuse but how each platform interacts with off-platform changes (such as the URLs of existing posts suddenly redirecting to different destinations). We found that nearly all platforms allow users to change their profile names and pictures, giving them varying degrees of freedom regarding the editing of posts and how prominent they make the fact that a post was edited. Regarding off-platform changes and their effects on on-platform content we find that 72% of platforms do not include previews for externally-linked content and rarely (if ever) revisit the posted links, thereby allowing attackers to switch the final destinations of URLs and completely overhaul the perceived bias of an account.

Recognizing the possible abuses of these platform mechanisms, we propose sensible steps for platforms, from prominently surfacing the original pre-edit information, to allowing community-notes-style notes to all content on their platform. We hope that this study will encourage not just additional research in the space of artificial/manufactured narratives but that it will be used by the platforms themselves to decide on the best tradeoff between features that give to their users, and the integrity of online content.

Acknowledgments

We thank the reviewers for their helpful feedback. This work was supported by the National Science Foundation (NSF) under grants CNS-1941617 and CNS-2126654.

References

[1] 4chan. https://4chan.org/. Online.

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

[3] Alleged assailant filled blog with delusional thoughts in days before pelosi attack. https://www.washingtonpost.com/investigations/2022/10/29/david-depape-blog-pelosi-fairies/. Online.

[4] Alleged paul pelosi attacker posted multiple conspiracy theories. https://www.cnn.com/2022/10/28/politics/pelosi-attack-suspect-conspiracy-theories-invs. Online.

[5] Archive of depape's website, godisloving.wordpress.com. https://web.archive.org/web/20220000000000*/https://godisloving.wordpress.com/. Online.

[6] archive.today. https://archive.is/. Online.

[7] Community notes on x. https://help.twitter.com/en/using-x/community-notes. Online.

[8] Douyin. https://www.douyin.com/. Online.

[9] Elite human trafficking. https://web.archive.org/web/20221028181625/https://www.frenlyfrens.com/post/elite-human-trafficking. Online.

[10] Facebook community standards. https://transparency.fb.com/policies/community-standards. Online.

[11] Fake name generator. https://fakenamegenerator.com/. Online.

[12] False, unfounded claims distort attack on paul pelosi. https://apnews.com/article/fact-check-pelosi-attack-conspiracy-theories-misinformation-028336957768. Online.

[13] Gamer gate. https://web.archive.org/web/20221028174512/https://www.frenlyfrens.com/post/gamer-gate. Online.

[14] Instagram community guidelines. https://help.instagram.com/477434105621119. Online.

[15] Internet archive. https://www.archive.org/. Online.

[16] Linkedin community guidelines. https://www.linkedin.com/legal/professional-community-policies. Online.

[17] Politicized rantings on two blogs by a 'daviddepape' draw scrutiny. https://www.nytimes.com/2022/10/28/us/politics/social-media-david-depape-pelosi-attack.html. Online.

[18] Primacy of consciousness. https://web.archive.org/web/20221028183820/https://www.frenlyfrens.com/post/primacy-of-consciousness. Online.

[19] Reddit content policy. https://www.redditinc.com/policies/content-policy. Online.

[20] The rise of social media. https://ourworldindata.org/rise-of-social-media. Online.

[21] This person does not exist. https://thispersondoesnotexist.com/. Online.

[22] Tiktok community guidelines. https://www.tiktok.com/community-guidelines/en/. Online.

[23] Wsws. https://www.wsws.org/. Online.

[24] The x rules. https://help.twitter.com/en/rules-and-policies/x-rules. Online.

[25] Youtube community guidelines. https://www.youtube.com/howyoutubeworks/policies/community-guidelines/. Online.

[26] Sara Albakry, Kami Vaniea, and Maria K Wolters. What is this url's destination? empirical evaluation of users' url reading. In *Proceedings of the CHI conference on human factors in computing systems*, pages 1–12, 2020.

[27] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.

[28] Oberiri Destiny Apuke and Bahiyah Omar. Fake news and covid-19: modelling the predictors of fake news sharing among social media users. *Telematics and Informatics*, 56:101475, 2021.

[29] Teresa Borges-Tiago, Flavio Tiago, Osvaldo Silva, José Manuel Guaita Martínez, and Dolores Botella-Carrubi. Online users' attitudes toward fake news: Implications for brand management. *Psychology & Marketing*, 37(9):1171–1184, 2020.

[30] Chiara Patricia Drolsbach and Nicolas Pröllochs. Diffusion of community fact-checked misinformation on twitter. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–22, 2023.

[31] Kun Du, Hao Yang, Zhou Li, Hai-Xin Duan, and Kehuan Zhang. The ever-changing labyrinth: A large-scale analysis of wildcard dns powered blackhat seo. In *USENIX Security Symposium*, pages 245–262, 2016.

[32] Xingzhi Guo, Brian Kondracki, Nick Nikiforakis, and Steven Skiena. Verba volant, scripta volant: Understanding post-publication title changes in news outlets. In *Proceedings of the ACM Web Conference 2022*, pages 588–598, 2022.

[33] Aditi Gupta and Rishabh Kaushal. Towards detecting fake user accounts in facebook. In *2017 ISEA Asia Security and Privacy (ISEASP)*, pages 1–6. IEEE, 2017.

[34] Seyedmehdi Hosseinimotlagh and Evangelos E Papalexakis. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. In *Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*, 2018.

[35] Luca Invernizzi, Kurt Thomas, Alexandros Kapravelos, Oxana Comanescu, Jean-Michel Picod, and Elie Bursztein. Cloak of visibility: Detecting when machines browse a different web. In *IEEE Symposium on Security and Privacy (IEEE S&P)*, pages 743–758, 2016.

[36] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788, 2021.

[37] Irina Khaldarova and Mervi Pantti. Fake news: The narrative battle over the ukrainian conflict. *Journalism practice*, 10(7):891–901, 2016.

[38] Deepak Kumar, Zane Ma, Zakir Durumeric, Ariana Mirian, Joshua Mason, J Alex Halderman, and Michael Bailey. Security challenges in an increasingly tangled web. In *Proceedings of the 26th International Conference on World Wide Web*, pages 677–684, 2017.

[39] Srijan Kumar and Neil Shah. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*, 2018.

[40] Nektarios Leontiadis, Tyler Moore, and Nicolas Christin. A nearly four-year longitudinal study of search-engine poisoning. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 930–941, 2014.

[41] Chaz Lever, Robert Walls, Yacin Nadji, David Dagon, Patrick McDaniel, and Manos Antonakakis. Domain-z: 28 registrations later measuring the exploitation of residual trust in domains. In *2016 IEEE symposium on security and privacy (SP)*, pages 691–706. IEEE, 2016.

[42] Federico Maggi, Alessandro Frossi, Stefano Zanero, Gianluca Stringhini, Brett Stone-Gross, Christopher Kruegel, and Giovanni Vigna. Two years of short urls internet measurement: security threats and countermeasures. In *proceedings of the 22nd international conference on World Wide Web*, pages 861–872, 2013.

[43] Enrico Mariconti, Jeremiah Onaolapo, Syed Sharique Ahmad, Nicolas Nikiforou, Manuel Egele, Nick Nikiforakis, and Gianluca Stringhini. What's in a Name? Understanding Profile Name Reuse on Twitter. In *Proceedings of the 26th International World Wide Web Conference (WWW)*, 2017.

[44] David M Markowitz and Jeffrey T Hancock. Linguistic traces of a scientific fraud: The case of diederik stapel. *PloS one*, 9(8):e105937, 2014.

[45] Faiza Masood, Ahmad Almogren, Assad Abbas, Hasan Ali Khattak, Ikram Ud Din, Mohsen Guizani, and Mansour Zuair. Spammer detection and fake user identification on social networks. *IEEE Access*, 7:68140–68152, 2019.

[46] Najmeh Miramirkhani, Timothy Barron, Michael Ferdman, and Nick Nikiforakis. Panning for gold. com: Understanding the dynamics of domain dropcatching. In *Proceedings of the World Wide Web Conference*, pages 257–266, 2018.

[47] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*, 2019.

[48] Tyler Moore and Richard Clayton. The ghosts of banking past: Empirical analysis of closed bank websites. In *Financial Cryptography and Data Security: 18th International Conference*, pages 33–48, 2014.

[49] Patricia Moravec, Randall Minas, and Alan R Dennis. Fake news on social media: People believe what they want to believe when it makes no sense at all. *Kelley School of Business research paper*, (18-87), 2018.

[50] Nick Nikiforakis, Federico Maggi, Gianluca Stringhini, M Zubair Rafique, Wouter Joosen, Christopher Kruegel, Frank Piessens, Giovanni Vigna, and Stefano Zanero. Stranger danger: exploring the ecosystem of ad-based url shortening services. In *Proceedings of the 23rd international conference on World wide web*, pages 51–62, 2014.

[51] Adam Oest, Yeganeh Safaei, Adam Doupé, Gail-Joon Ahn, Brad Wardman, and Kevin Tyers. Phishfarm: A scalable framework for measuring the effectiveness of evasion techniques against browser phishing blacklists. In *IEEE Symposium on Security and Privacy (IEEE S&P)*, pages 1344–1361, 2019.

[52] Adam Oest, Yeganeh Safei, Adam Doupé, Gail-Joon Ahn, Brad Wardman, and Gary Warner. Inside a phisher's mind: Understanding the anti-phishing ecosystem through phishing kit analysis. In *APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–12, 2018.

[53] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, 96:104, 2017.

[54] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[55] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 312–320, 2019.

[56] Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 436–439, 2019.

[57] Mohit Singhal, Nihal Kumarswamy, Shreyasi Kinhekar, and Shirin Nilizadeh. Cybersecurity misinformation detection on social media: Case studies on phishing reports and zoom's threat. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 796–807, 2023.

[58] Johnny So, Michael Ferdman, and Nick Nikiforakis. The more things change, the more they stay the same: Integrity of modern javascript. In *Proceedings of the Web Conference*, 2023.

[59] Johnny So, Najmeh Miramirkhani, Michael Ferdman, and Nick Nikiforakis. Domains do change their spots: Quantifying potential abuse of residual trust. In *IEEE Symposium on Security and Privacy (IEEE S&P)*, pages 2130–2144, 2022.

[60] Marius Steffens, Marius Musch, Martin Johns, and Ben Stock. Who's hosting the block party? studying third-party blockage of csp and sri. In *Network and Distributed Systems Security (NDSS) Symposium*, 2021.

[61] Giada Stivala and Giancarlo Pellegrino. Deceptive previews: A study of the link preview trustworthiness in social platforms. In *27th Annual Network and Distributed System Security symposium, February 2020*, NDSS. Internet Society, 2020.

[62] Chris Tsoukaladelis, Brian Kondracki, Niranjan Balasubramanian, and Nick Nikiforakis. The times they are a-changin': Characterizing post-publication changes to online news. In *IEEE Symposium on Security and Privacy (IEEE S&P)*, 2024.

[63] Tavish Vaidya, Daniel Votipka, Michelle L Mazurek, and Micah Sherr. Does being verified make you more credible? account verification's effect on tweet credibility. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[64] Tom Van Goethem, Najmeh Miramirkhani, Wouter Joosen, and Nick Nikiforakis. Purchased fame: Exploring the ecosystem of private blog networks. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, pages 366–378, 2019.

[65] Thomas Vissers, Timothy Barron, Tom Van Goethem, Wouter Joosen, and Nick Nikiforakis. The wolf of name street: Hijacking domains through their nameservers. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 957–970, 2017.

[66] Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, MB Hunzaker, Keith Coleman, and Jay Baxter. Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *arXiv preprint arXiv:2210.15723*, 2022.

[67] Ronghai Yang, Xianbo Wang, Cheng Chi, Dawei Wang, Jiawei He, Siming Pang, and Wing Cheong Lau. Scalable detection of promotional website defacements in black hat seo campaigns. In *USENIX Security Symposium*, pages 3703–3720, 2021.

[68] Günce Su Yılmaz, Fiona Gasaway, Blase Ur, and Mainack Mondal. Perceptions of retrospective edits, changes, and deletion on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 841–852, 2021.