A First Look at Zoombombing

Chen Ling[♣*], Utkucan Balcı[♦*], Jeremy Blackburn[♦], and Gianluca Stringhini[♣]
[♣]Boston University, [♦]Binghamton University
ccling@bu.edu, ubalci1@binghamton.edu, jblackbu@binghamton.edu, gian@bu.edu

***Abstract*— Online meeting tools like Zoom and Google Meet have become central to our professional, educational, and personal lives. This has opened up new opportunities for large scale harassment. In particular, a phenomenon known as zoombombing has emerged, in which aggressors join online meetings with the goal of disrupting them and harassing their participants. In this paper, we conduct the first data-driven analysis of calls for zoombombing attacks on social media. We identify ten popular online meeting tools and extract posts containing meeting invitations to these platforms on a mainstream social network, Twitter, and on a fringe community known for organizing coordinated attacks against online users, 4chan. We then perform manual annotation to identify posts that are calling for zoombombing attacks, and apply thematic analysis to develop a codebook to better characterize the discussion surrounding calls for zoombombing. During the first seven months of 2020, we identify over 200 calls for zoombombing between Twitter and 4chan, and analyze these calls both quantitatively and qualitatively. Our findings indicate that the vast majority of calls for zoombombing are not made by attackers stumbling upon meeting invitations or bruteforcing their meeting ID, but rather by insiders who have legitimate access to these meetings, particularly students in high school and college classes. This has important security implications because it makes common protections against zoombombing, e.g., password protection, ineffective. We also find instances of insiders instructing attackers to adopt the names of legitimate participants in the class to avoid detection, making countermeasures like setting up a waiting room and vetting participants less effective. Based on these observations, we argue that the only effective defense against zoombombing is creating unique join links for each participant.**

I. Introduction

One of the earliest promises of the Internet was to enable quick, easy, and real-time communications, not just via text, but also audio and video. While it took some time, there are now numerous online meeting tools like Skype, Zoom, and Google Meet that are used in a variety of contexts, both personal and professional. In 2020, society has found itself increasingly reliant on these online meeting tools due to the COVID-19 pandemic, with many business meetings, online classes, and even social gatherings moving online. Unfortunately, the mass adoption of these services has also enabled a new kind of attack where perpetrators join and deliberately disrupt virtual meetings. This phenomenon has been dubbed *zoombombing*, after one of the most used online meeting platforms [6, 49].

To mitigate the threat of zoombombing, security practitioners have begun discussing best practices to prevent these attacks from happening or limit their effects. These include requiring a password to join online meetings, setting up a waiting room and manually vetting participants before letting them in, and not sharing meeting links publicly [11, 55]. While helpful to keep out casual and unmotivated attackers, there is an inherent tension between tightening the security of online meeting rooms and the need for them to be easily accessible to a number of people, especially in the case of large public events [6]. Most importantly, devising effective security policies requires a good understanding of the capabilities of attackers and their modus operandi. To date, however, the research community lacks a good understanding of how zoombombing attacks are called for and how they are carried out. For example, it remains unclear how attackers obtain meeting links in the first place. This type of knowledge is crucial because, for example, protecting against attackers proactively bruteforcing the ID of meeting rooms is very different (and calls for different countermeasures) than mitigating attacks called for by insiders.

In this paper, we perform the first measurement study of calls for zoombombing attacks on social media. We first select ten popular online meeting services, spanning a wide range of target users, from businesses to individuals. We then analyze the security features that these services offer to their users, with a particular focus on the mechanisms that allow them to restrict and control who can join and participate in the meeting. We next identify posts that contain online meeting information. We decide to focus on two online services for this purpose, a mainstream social network, Twitter, and a fringe Web community, 4chan, which previous work showed is often involved in harassment attacks against online users [23, 33]. Between January and July 2020, we identify 12k tweets and 434 4chan threads discussing online meeting rooms. We then apply thematic qualitative analysis [47] to identify posts that are indeed calling for a zoombombing attack, and to further characterize them. We identify 123 4chan threads discussing such attacks and 95 tweets. We then adopt a mixed methods approach to perform further analysis. We first analyze this dataset quantitatively, looking at temporal properties of posts and apply natural language processing techniques to better understand the topics of discussion. We then dig deeper into our qualitative analysis results to get a more nuanced view of the zoombombing phenomenon. Finally, we discuss our findings in view of existing countermeasures, reasoning about their effectiveness.

In summary, we make the following key findings:

- The majority of the calls for zoombombing in our dataset target online lectures (74% on 4chan and 59% on Twit-

[*]Utkucan Balcı and Chen Ling contributed equally to this work.

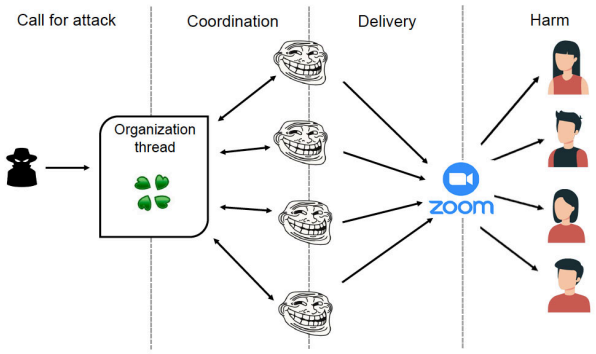

Figure 1: Threat Model for a zoombombing attack. Charlie calls for an attack against a Zoom meeting created by Alice, by creating a thread on an online service (e.g., 4chan). Participants then join the Zoom meeting, report back on the thread about the status of the attack, and harm the legitimate participants to the meeting.

ter). We find evidence of both universities and high schools being targeted.

- Most calls for zoombombing come from insiders who have legitimate access to the meetings (70% on 4chan and 82% on Twitter). This has serious security implications, because it makes passwords ineffective to protect the meeting rooms as attackers can share them with whoever participates in the attack. In some cases we find that the insider shares additional information like names of real students in the class, allowing participants to select those names and make it difficult for teachers and moderators to identify intruders.
- Almost all calls for zoombombing target meetings happening in real time (93% on 4chan and 98% on Twitter), suggesting that these attacks happen in an opportunistic fashion and that defenders cannot prepare for zoombombings by identifying posts ahead of time.

Disclaimer. Due to their nature, zoombombing messages on social media are likely highly offensive. In this paper we do not censor any content, therefore we warn the reader that some of the quotes included in the following sections are likely to be upsetting and offensive.

II. Background

In this section, we first describe the threat model that we assume for this paper. We then describe how we chose the ten meeting services that we study, and describe their features.

A. Threat Model

We consider a zoombombing attack as being composed of four phases (see Figure 1), based on anecdotal evidence of how zoombombing accounts unfold, as well as following empirical evidence reported by previous research that studied coordinated online aggression, trolling, and harassment on other social media platforms (e.g., Reddit, YouTube) [15, 23, 28, 34]. Note that in this paper we focus on calls for attacks that aim at attracting multiple participants; single attackers stumbling upon meeting rooms and disrupting them are out of scope. In the following, we describe the four phases in detail through an example in which Charlie is orchestrating a coordinated attack against a Zoom meeting created by Alice.

i) Call for attack. Charlie obtains information about Alice's Zoom meeting. As we will show later, this is often because Charlie is a legitimate participant of the meeting (e.g., a student in an online lecture). Charlie then posts information about the Zoom meeting on an online service of his choice (starting an *organization thread*), asking other members of the community to participate in a coordinated attack. Previous research showed that attacks like this are often organized on polarized Web communities (e.g., /pol/, 4chan's Politically Incorrect Board), where the person calling for an attack posts a link to content on another service that was created by the victim (e.g., a Zoom meeting), followed by an invite to the person (e.g., through the phrase "you know what to do") [23, 33].

ii) Coordination. The organization thread created by Charlie now becomes an aggregation point for attackers, who will report additional information and coordinate the attack by replying to the thread. For example, attackers will post details like a password to access the meeting or personal information about the host.

iii) Delivery. The attackers will then join the online meeting and harass the participants, for example sending them hateful messages, shouting profanities, or displaying offensive or indecent images through their webcams [6].

iv) Harm. The goal of the attack is to cause harm to the group of people. Depending on its success and intensity, victims could suffer serious psychological [16, 22] or even physical harm [32].

B. Online Meeting Services

To select a representative set of online meeting tools to study in this paper, we ran Google queries for "online meeting services" and manually vetted the results for Web pages that actually advertise a service (excluding, for example, news articles talking about a certain meeting platform). After this process, we obtained the list of the ten highest ranked meeting tools. These services are Zoom, Hangouts, Google Meet, Skype, Jitsi, GotoMeeting, Microsoft Teams, Cisco Webex, Bluejeans, and Starleaf.

In the following, we describe the general characteristics of each of these services (see Table I). We then analyze the security relevant features offered by the various platforms (e.g., whether they allow hosts to set a password for meetings). We are particularly interested in understanding what characteristics of a service might make it a popular target platform for attackers, or might reduce the risk for a successful attack.

Length of operation. Half of our ten services were established after 2010, with the notable exception of Webex which started in the 90s. Major tech companies like Microsoft, Google, and Cisco have their own solution, with Microsoft and Google having two of them (Skype and Teams for Microsoft and Hangouts and Meet for Google). While Google started retiring Hangouts

Platform	Est.	Headquarters	Parent Company	Target Users	User base	Plan
Zoom	2011	US	-	Both individual and business	300M	Free, upgrade available starts from $15/month
Meet	2017	US	Google	Both individual and business	100M	Free, upgrade available starts from $12/month
Webex	1993	US	Cisco	Business	324M	Free, upgrade available starts from $13.5 /month
Jitsi	2017	AU	Atlassian	Both individual and business	-	Free
Skype	2003	US	Microsoft	Both individual and business	100M	Free, charge for phone calls
GotoMeeting	2004	US	LogMeIn	Business	-	Starts from $12/Month
Teams	2017	US	Microsoft	Business	75M	Free, upgrade available starts from $5 per user/month
Hangouts	2013	US	Google	Individual	14M	Free, charge for phone calls
Bluejeans	2009	US	Verizon	Business	-	Starts from $12/Month
Starleaf	2008	UK	-	Business	3,000	Free, upgrade available starts from $14.99 /month

Table I: Overview of the ten online meeting services studied in this paper.

in October 2019, we will later show that this platform is still very much used and many meeting links to it are posted on social media. There are also companies that focus on online communication services, like Zoom and Starleaf. During the COVID-19 pandemic, where millions of people have been forced to work, learn, and socialize remotely, Zoom has risen to the top, with over 300 million daily participants in virtual meetings, and has also become the top target of attack; hence the phrase "zoombombing."

User base. Most of the online meeting services are aimed at business users. While Hangouts is the only service specifically devoted to individuals, five of them are geared towards both business *and* individual users. Based on the most current data [8, 18, 38, 43] (July 2020), four of the online meeting services have a user base of over 100M (Zoom, Meet, Skype, and Webex). We hypothesize that the user base of a service plays a role in which services face the most attacks.

User plan. Most online meeting services provide free accounts for individuals and small companies, with five of them allowing paid plans that provide additional features. GotoMeeting and Bluejeans, however, exclusively target business consumers (charging hosts $12/month) and do not provide free accounts. Teams paid plans are somewhat different, as they are not based on a per-host basis, but on a per-user basis. Google Hangouts and Skype are free, but charge for phone calls to local numbers.

Features. We next analyze the features that are specific to each online meeting platform, with a particular focus on the security measures that they put in place to prevent zoombombing. To this end, we compare the features offered to free accounts. Since GotoMeeting and Bluejeans do not provide free accounts, we list the features that they offered to paid customers in this section. An overview of the features offered by each platform is reported in Table II.

First, we look at the security features offered by the meeting platforms. Eight of the ten services require an account to join a meeting. This is done to prevent attackers from flooding meeting rooms and provide some accountability, e.g., suspending misbehaving accounts. Only Jitsi and Zoom do not require a registration to join meetings, although Zoom hosts allow hosts to require participants to have an account in order to join. Authentication-wise, the security model of online meeting services is the following: anyone with an account on the platform and who knows the meeting ID can join the meeting. This is not dissimilar to other security sensitive services that have been studied by the community in the past, from online document editing [27] to file download platforms [30]. To prevent anyone knowing the meeting ID from joining a room, Zoom, Webex, GotoMeeting, and Bluejeans allow hosts to specify a password participants need to provide upon joining. Only Zoom and Google Meet allow a waiting room for hosts to vet the identity of participants. Google Meet automatically admits participants whose accounts are included in the invitation list into the meeting room and puts others in a waiting room, where the host can let them in manually. Only Zoom and Webex provide a registration system with one-time unique links per registrant, which can help restrict and trace participants. Generally, other meeting services use unique links for each meeting, with Google Hangouts and Google Meet allowing a link to be reused within a 90 day period. Skype does not have a one time unique link function. Due to privacy concerns, Google Meet, Google Hangouts, and Jitsi do not allow hosts to mute all participates [20, 25]. Google Meet only allows educational accounts to mute participants [19].

Second, we look at whether services limit the number of users that can join a meeting, as well as the maximum duration of a meeting for free users. All the services under study have a participant limit in their free version. Zoom, Google Meet, and Webex limit meetings to 100 participants, and Teams only supports four attendees in its free version. When looking at the maximum duration of a meeting, we find that three services (Zoom, Webex, and Starleaf) limit meetings to between 40 and 50 minutes for free users.

III. Datasets

In this section, we describe the datasets use in this paper as well as our data collection process. We first discuss how we identify social media posts containing links to meeting rooms. We then discuss the online services we collect data from.

Identifying posts containing meeting URLs. To find posts that contain meeting URLs on the online services we monitor, we first identify the their DNS domain names. To avoid simple evasion attempts, we use regular expressions that only consider alphanumeric characters and dots. We manually examine posts and find that Zoom meetings are often not shared via a URL, but rather via a message containing the meeting ID, which users can input in the Zoom application to join, like in the following example:

Platform	Requires account to join in	Max particp.	Max time	Allows password	Allows waiting room	one-time unique link	Mute upon entry
Zoom	No	100	40min	Yes	Yes	for each particp.	Yes
Google Meet	Yes	100	Unlimited	No	Yes	No	No
Webex	Yes	100	50min	Yes	No	for each particp.	Yes
Jitsi	No	75	Unlimited	No	No	Yes	No
Skype	Yes	50	Unlimited	No	No	No	No
GotoMeeting*	No	26	Unlimited	Yes	Yes	Yes	Yes
Teams	Yes	4	Unlimited	No	No	Yes	No
Hangouts	Yes	25	Unlimited	No	No	No	No
Bluejeans*	Yes	50	Unlimited	Yes	Yes	Yes	Yes
Starleaf	Yes	20	45min	No	No	Yes	Yes

Table II: Comparison of the features offered by the online meeting services studied in this paper to free accounts. Services marked with * do not provide a free version and are only available to hosts who pay a subscription.

> "Date: 03/24/2020 Time: 12:00PM Meeting ID: [ZOOM ID] Passcode: [ZOOM PASSWORD]"

To account for these posts, after lowercasing and removing non-alphanumeric characters, we search for a pattern with "id" followed by at least nine consecutive digits by using regular expressions. We then further filter these by only including posts with the keyword "zoom" in them.

4chan. 4chan [36] is an imageboard where users start a thread anonymously, and other users comment on it. 4chan is organized in boards that cover different special interests (e.g., Anime & Manga, Sports) or host more generic discussions (e.g., Politically Incorrect, Random). Unlike traditional online services, threads on 4chan boards are *ephemeral*: only a fixed number of threads is alive at any given time. Once a new thread is created, the active thread that has least recently been used is removed from the catalog of live threads. Previous research showed that 4chan is a popular platform used by miscreants to carry out abuse, e.g., organizing coordinated harassment campaigns [23, 33, 36]. We therefore hypothesize that zoombombing is widespread on the platform.

We develop a custom crawler following the same methodology of previous research on 4chan [23, 37], and collect all posts between January 1st, 2020, and July 24th, 2020. We then identify posts containing online meeting links and invitations following the methodology discussed in the previous section. Every time we identify a post containing information about a meeting, we pull the entire thread. In total, we identify 47,221 posts from 434 threads with a URL or an ID for at least one meeting platform room.

Twitter. Twitter [29] is a microblogging social media platform on which registered users can share posts publicly or privately. While private accounts can only reach their followers, public accounts can reach any user on Twitter. The posts are called "tweets" and can be re-shared (retweeted) by other users to share with their followers. Tweets can contain "hashtags" where users can put the "#" symbol at the beginning of a word. By using the same hashtags, people can create trends, which can also be used to look up tweets on the same topic.

Leveraging the Twitter streaming API, a public service that provides a random 1% sample of all tweets posted worldwide, we identify 12,077 tweets containing links or IDs to online meeting rooms. These tweets were posted between January 1st, 2020, and July 18th, 2020. Note that due to limitations in the Twitter API we could not retrieve any replies to tweets containing meeting IDs.

Ethics. Since this work only involved publicly available data and did not require interactions with participants, it is not considered human subjects research by our institution. We however acknowledge that data from social media is sensitive, as it can contain personal information. In this study, we adopted standard best practices to ensure that our study followed ethical principles [2, 39] In particular, we did not try to further de-anonymize any user.

IV. Identifying Zoombombing Threads

While it is relatively straight forward to automatically find posts that include links to meetings, the challenge is in determining the intent behind the link being posted, and in particular whether the post is calling for a zoombombing attack. We expect that most meeting links on social media are posted for benign reasons; therefore, to carry out this study we need a way to separate harmless posts from those that are calls for zoombombing. Since zoombombing is a human driven phenomenon, developing automated techniques to identify posts calling for attacks is challenging and prone to false positives and false negatives. To avoid these issues, we perform manual annotation based on thematic analysis of all posts in our dataset, with the goal of identifying a reliable ground truth dataset.

In this section, we develop a codebook to guide the thematic annotation process for our 4chan and Twitter datasets. We break the development of this codebook in two phases. First, we perform a binary labeling to determine if posts are indeed calls for zoombombing or not. As a second step, we further characterize the posts and threads that contain zoombombing invitations, with the goal of understanding the behavior of attackers and the targets that they choose.

To build our codebook and perform annotation we follow the same methodology described in recent security research [47], in which the authors studied posts from online infidelity forums and their relation with intimate partner surveillance tools and tactics. More precisely, we follow these four steps:

1) The four authors of this paper independently screened our dataset and produced initial codes using thematic coding [5].

2) We then discussed these initial codes and went through multiple iterations, using a portion of the data to build a final codebook. The process continued until the codebook reached stability and additional iterations would not refine it further.
3) To investigate the common agreement on the codebook by multiple annotators, we have them rate a portion of our dataset and discuss disagreements until a good agreement is reached.
4) We split the rest of our dataset and each annotator labels one portion of it.

We next describe our process and our codebook in more detail.

Phase I: labeling zoombombing content

As we mentioned, the first phase of our annotation process deals with identifying social media posts and threads that contain an invitation to zoombombing. We start by labeling 4chan threads. Following the methodology from [47], we first randomly choose 10 threads from the 434 threads that contain a link to a meeting room, and have each author of the paper review and discuss them together to build a shared understanding of what a zoombombing invitation looks like. From this initial dataset, the authors agreed that two threads were "bombing" threads (i.e., they were encouraging/calling for a zoombombing) while the remaining eight were not (i.e., "non-bombing").

We then aim to test each author's ability to independently identify bombing threads. To this end, we chose 20 additional threads (balanced as per the overall distribution of meeting platform links on 4chan), and had each author label them as either bombing or non-bombing. We used the following definition to make a decision: *a zoombombing thread should include an invitation to bomb along with a URL to a meeting room or a meeting ID*. One interesting caveat here is that while discussing the initial set of threads we noticed that the invitation to bomb did not necessarily appear in the same post as the meeting link itself, and thus we added the following additional condition where applicable: *the same user posted the link or meeting ID and the textual invitation to bomb, even if they were not in the same post*. Note that although users on 4chan are anonymous, users are given a unique ID that identifies them within the same thread [23]. It is important to note that invitations to bomb are not necessarily *explicit*. 4chan's users are well known to use coded language and slang [33], and thus we relied on our domain expertise when coding posts that include phrases like "you know what to do" and "do ya thing." Finally, because of the overall uncertainty of things, we decide to be conservative and label any threads we are unsure about as non-bombing. A typical bombing invitation looks as follows:

> "[ZOOMURL] My English class, come in and trolley for a while."

While a typical non-bombing invitation looks as follows:

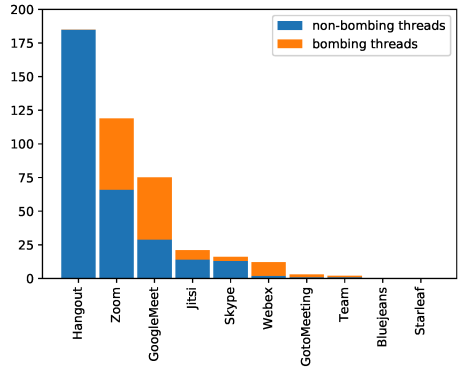

Figure 2: Ratio of bombing and non-bombing threads on 4chan.

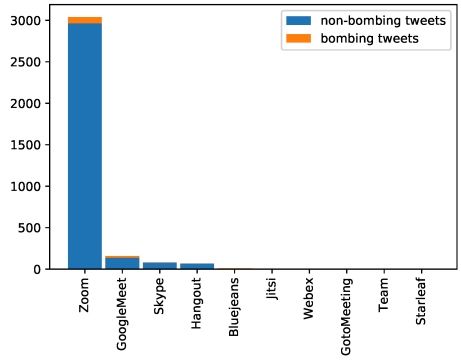

Figure 3: Ratio of bombing and non-bombing tweets on Twitter.

> "Lets chat anon. Send your favorite memes!!!![ZOOMURL]"
> "Be professional anon, please."

The four authors of this paper independently coded each thread to determine whether it was bombing related or not. From this testing phase of 20 threads, we calculated Fleiss' agreement score between the annotators and found perfect agreement (κ = 1.0) [13, 14]. This indicates that all authors were able to reliably identify zoombombing threads. From here, we expand our annotation to the full dataset of 434 threads, split evenly between the four annotators.

In the end, we find that 123 of the 434 threads in our 4chan dataset are bombing threads. As seen in Figure 2, nearly half (43.96%) of the Zoom meeting links in our dataset were determined to belong to a bombing thread, and a majority (59.72%) of Google Meet links appeared in bombing threads. On the other hand, Google Hangouts and Skype links are mostly posted with benign intentions.

We followed the same labeling procedure for Twitter. From our preliminary screening of the tweets, we find that a large portion are non-English. Thus, we use the "language" field provided by the Twitter API and restrict our analysis to only English tweets from our total 12,077 tweets, which left us with 3,510 candidate tweets.

A challenge we faced when labeling tweets is that Twitter is a much different platform than 4chan in its user base and general tone. 4chan is dominated by trolling and irony,

and veiled calls to join meetings can often be interpreted as bombing invitations. Here is an example of a bombing invitation from 4chan:

> "Ok retards, this is an id of a zoom web lessons. Do your worst [ZOOM ID] [ZOOM PASSWORD]."

On the other hand, Twitter is a general audience social network, therefore we expect most meeting invitations to be benign. For example, this is a bombing invitation from Twitter:

> "Raid this class as fast as u can.... #zoomcodes #zoomclasscodes #zoomclass #zoom [ZOOMURL]"

To reflect this difference and avoid potential false positives, we decided to be stricter when determining if a tweet is a zoombombing invitation. More precisely, a bombing tweet needs to meet the following two criteria:

- An invitation to bombing with a link (invitation text usually comes with a link)
- A clear indication of bombing, such as "raid," "bomb," "troll," "discord," "disruptive," and "make fun of it."

As with 4chan, we were generally conservative in our labeling and default to non-bombing in uncertain cases.

From the 3.5K English tweets, we randomly sampled 500 so all services were equally represented (i.e., balanced with respect to services). From these 500, we manually selected 20 tweets, and four coders independently determined whether they were a bombing tweet or not. The inter-rater reliability again shows perfect agreement (Fleiss' κ = 1.0). Because of the high agreement scores on the initial testing set, as well as the agreement on the 4chan ratings, we had a single annotator label the remaining 3,490 tweets in this dataset. Note that this is a much quicker process than on 4chan, since the coder had to look at single tweets instead of full (and often long) threads.

In the end, we find that 95 out of the 3,510 candidate English tweets are bombing tweets. From Figure 3 we see that zoombombing on Twitter is less pervasive than on 4chan. In particular, of the 3,039 Zoom related candidate tweets, 75 are labeled as bombing, and 20 of the 157 Google Meet tweets are bombing. We found no bombing tweets for the other eight meeting tools.

Phase II: Characterizing zoombombing

While labeling threads and tweets as bombing or not is vital to understanding the problem, it does little to characterize the actual bombing activity itself. In this phase we aim to understand the *process* of a zoombombing event by analyzing the behavior that goes on in bombing threads.

We began by having four annotators go through the labeled bombing threads/tweets as determined by the Phase I labeling. This was a relatively loose process where the goal was to get a general sense of what is going on. Next, the annotators met and discussed their observations. In general there was agreement between the annotators of a clear trend of insider complicity in bombing of online classes in particular. After several rounds of discussion, we derived four, high level properties relevant to zoombombing threads and tweets: 1) thread structure (only applicable to 4chan threads), 2) link information, 3) invitation information, and 4) interaction (only applicable to 4chan threads).

Thread structure: New threads on 4chan are created when a so called "Original Poster" creates an "Original Post" and the thread constitutes replies to this post (**NB:** unlike other platforms, 4chan threads are *flat*) [23]. Thus, the first post in a thread usually represents the topic of the thread. We coded the following characteristics of a thread:

1) Whether the content of the first post is a zoombombing invitation. This indicates whether or not the thread was created primarily to act as a bombing thread as opposed to organically evolving into one.
2) The length of the thread (i.e., the number of posts), which indicates the thread's popularity.
3) The number of bombing invitation links, which is indicative of how the thread evolved with respect to bombing.

Link information: According to our definition of a bombing thread/tweet, both 4chan and Twitter posts need to include a video conference invitation link or meeting ID to be considered a bombing thread. For certain meeting platforms (e.g., Zoom) we can derive two additional pieces of information from meeting links directly: 1) *institutional information* (i.e., who is hosting the meeting) and 2) *password protection*.

For some platforms, we can automatically identify password-protected links by looking at a password parameter in the URL (e.g., https://zoom.us/j/123456789?pwd=12345aAbBcC678). When coding messages manually, we also look at the presence of passwords in the text of posts. Institutional information provides us additional information on the victims of attacks. To gather this information, we need to manually look at the URL (e.g., http://UNIVERSITY.zoom.us/j/XXXXXX), and search for its associated institution. We record each institution, its type (e.g., University), and country.

Invitation information: As noted previously, there are plenty of legitimate reasons to post a link to a video conference, and thus a posted link itself is not sufficient to say that an attack has occurred; this is why we require additional text calling for an attack. During our initial examination, we noticed that there was often additional information embedded in the bombing invitation itself, e.g., temporal details as well as hints at the existence of insiders.

> "[ZOOMURL] this class is up the tuesdays at 11:00 am UTC-5 crash this class plz."

For temporal information, we manually read the bombing invitation and labeled the meeting time according to three codes 1) *future event*, where the poster indicates the attached link will be active at some point in the future, 2) *live event*, where the poster indicates the meeting link is active and that bombers should join "now," and 3) *not sure*, where there was no clear indication of when the link would be active.

This temporal information indicates whether or not a bombing attack has been planned, or if it is an opportunistic attack.

Our preliminary analysis indicated that many zoombombing invitations are created by insiders, for example students in the case of college classes. To better understand insider complicity, we label each bombing post or thread as either 1) *insider* or 2) *non-insider*. To be labeled as *insider*, the bombing invitation should include text like "my teacher" or "our class," provide a password for the video conference (either explicitly in post text or implicitly in the link to the meeting), or give suggestions on what names bombers should select when joining the call (a tactic used to make it harder for legitimate meeting attendees/hosts to determine that joining bombers are not supposed to be there). Annotators recorded the details of what led to any *insider* label applied. Again, we conservatively label threads as *non-insider* if there is any doubt.

Interaction: For 4chan, we are able to collect entire threads discussing zoombombing. For these threads, we read the whole thread and record the following characteristics of the thread discussion:

- Time interval: the interval between the bombing invitation post and the first interaction post by other users (this characteristic is programmatically calculated);
- Problem feedback: participants reporting problems about their zoombombing attempts, for example being unable to join the meeting room, or being kicked out by the host;
- Toxic speech: participants insulting the host of the meeting with profanity, hate speech, etc.;
- Crime scene feedback: reports on successful attacks with details on what happened during the disrupted meeting;

For phase II, four raters independently rated 20 randomly chosen threads from 123 bombing 4chan threads and 20 random tweets from 95 bombing tweets from Twitter. Inter-rater reliability showed a perfect agreement in both sampled datasets (Fleiss' Kappa 1.0). We then split the rest of the dataset into four groups, with each rater coding one group.

V. Quantitative Analysis

To better understand the zoombombing phenomenon, we first start by quantitatively analyzing the 123 4chan threads and 95 tweets that we identified as part of the coding process, comparing them with posts and threads containing non-bombing meeting links. We focus our analysis on three aspects: 1) understanding which services are targeted the most by zoombombing 2) examining how zoombombing unfolds temporally and 3) using natural language processing techniques to quantify the content of zoombombing threads.

A. Targeted services

We observe that the platforms with a larger user base (see Table I) seem to attract more zoombombing attacks. In particular, we find 129 bombing links on Zoom, 66 on Google Meet, 10 on Webex, 7 on Jitsi, 3 on Skype, 2 on GoToMeeting, and 1 on Teams, while there are none for Hangouts, Bluejeans, and Starleaf.

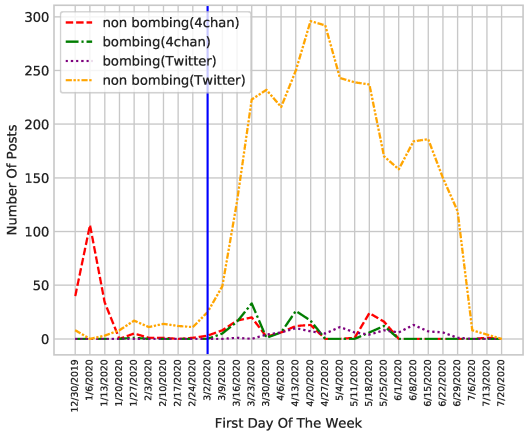

Figure 4: Number of posts per week for bombing & non-bombing threads and tweets. The vertical line indicates the beginning of the COVID-19 lockdown in the United States (on the week of 3/2/2020, when several West Coast US universities started going online.)

B. Temporal Analysis

Figure 4 plots the weekly occurrences of bombing and non-bombing posts on Twitter and 4chan. From the figure, we see that posts with meeting links became more prevalent (especially on Twitter) as the COVID-19 shutdown began in March 2020 (shown in the figure with blue line[1]). On 4chan, we observe a spike in benign posts containing meeting links around New Years Eve 2020, attributable to users organizing social gatherings as well as increased activity of a far-right group on the following week. An example of a non-bombing thread that appears repetitively, including a Google Hangout link on New Years Eve is the following:

> "JOIN OR YOUR MOTHER DIES :3 [HANGOUTURL]"

Generally speaking, zoombombing as a phenomenon barely existed before the quarantine. We observe a decline of the phenomenon in June 2020, potentially linked to school holidays; this is in line with the fact that we observe that most calls for zoombombing target school lectures and college classes, as discussed later in Section VI-A.

Next, we plot the number of posts per hour of the day for 4chan posts and tweets with bombing links in Figure 5. On Twitter, we find that zoombombing activity does not exhibit clear diurnal patterns. On 4chan, bombing posts are mostly shared from 08:00 to 23:00 UTC. We did not encounter any zoombombing tweet that specified a location and only 13 zoombombing posts had country information on 4chan (8 USA, 1 Indonesia, 1 Bulgaria, 1 Turkey, 1 Chile and 1 Italy). Considering the lack of diurnal patterns in Figure 5, we infer that zoombombing calls are not a localized problem.

Temporal analysis of 4chan threads. To better understand zoombombing behavior, we analyze threads on 4chan with

[1] https://www.insidehighered.com/news/2020/03/09/colleges-move-classes-online-coronavirus-infects-more

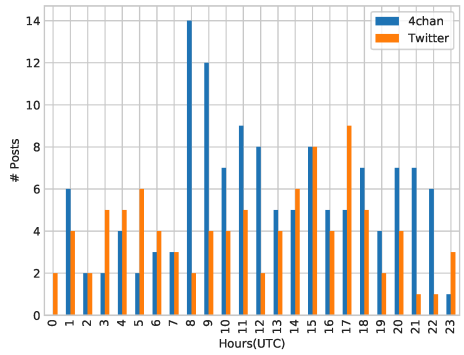

Figure 5: Hour Distribution of zoombombing posts. Note that we did not discard multiple posts that contain the same zoombombing link.

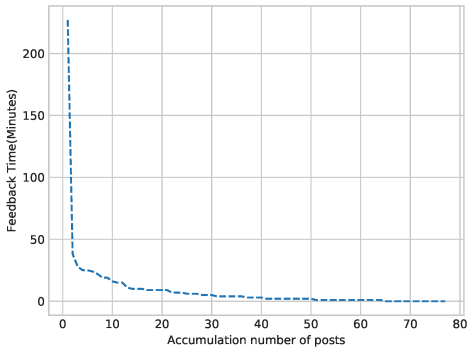

Figure 7: Feedback time between the posting of a zoombombing invitation on 4chan and the first reply to the thread.

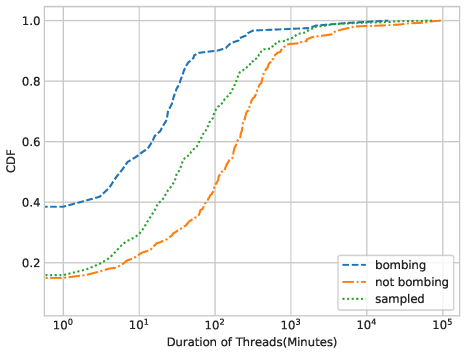

Figure 6: Duration of threads on 4chan.

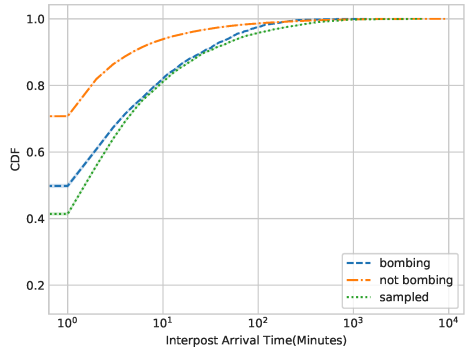

Figure 8: CDF of Interpost Arrival Times for bombing & non-bombing threads

post containing zoombombing links. This allows us to get a quantitative understanding of how discussion of zoombombing activity unfolds on the platform. Our 123 zoombombing threads have 2,693 total posts in them. We compare these 123 threads to the 311 threads (44,528 posts) that included a meeting link but were *not* bombing threads. Finally, we also compare to a baseline of 4chan posts chosen by sampling threads at random (without replacement) on a per-day basis such that we have the same number of baseline threads per day as we have threads where a meeting link was posted.

Figure 6 plots the cumulative distribution function (CDF) of the duration of threads in our dataset (defined as the difference in the timestamp of the last post and the timestamp of the original post). Recall that threads on 4chan are ephemeral, and once a thread is not active for a while it gets pruned and no further posts can be made [23]. From the figure, we observe that bombing threads have a shorter lifetime than other threads: 50% of bombing threads are active for less than 5 minutes, compared to 30 minutes for randomly sampled threads, and two hours for non-bombing threads. That said, we do have a long tail with about 10% of bombing threads lasting over 2 hours, compared to 7 hours for sampled threads and 12 hours for non-bombing threads.

In our threat model, threads become an aggregation point for attackers, and so understanding the feedback Charlie receives from the bombers he is trying to recruit is important. Thus, Figure 7 plots the delay between the bombing link being posted on 4chan and the first reply. From the figure, we see that 79% of zoombombing threads receive their first reply within 10 minutes. One explanation for this is that calls for zoombombing might be time sensitive; indeed in Section VI-B we show that many attackers are inviting bombers to join live meetings/classes. We then look at the interpost arrival time between each post in a thread. Similarly, Figure 8 plots the CDF of interpost arrival times, which is the time between consecutive posts in threads, for bombing and non-bombing threads. For most threads the elapsed time between consecutive posts in bombing threads is similar to sampled threads while being higher compared non-bombing threads. One explanation for this is that non-bombing meeting links tend to be posted to organize social gatherings, and thus tend to show up in more popular, faster moving threads. An alternative explanation is that while the zoombombing attack is happening 4chan users are slower in replying in the thread because they are busy performing malicious activities in the meeting itself.

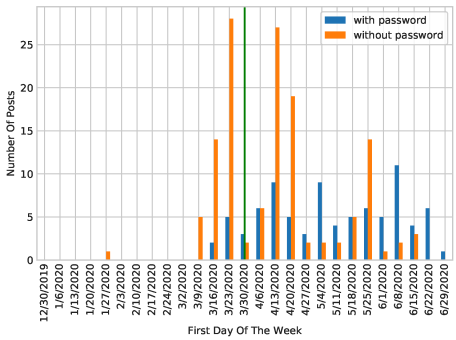

Figure 9: Occurrences of zoombombing links with and without passwords. The green line indicates the week Zoom required passwords if a user tries to enter a meeting using just the meeting ID instead of the meeting invite link (on the week of 3/30/2020)

C. Characteristics of zoombombing links

In this section we focus on what we can learn by analyzing the zoombombing links, in particular whether they contain information about the victim organizations and if they include a password as a URL parameter.

Targeted organizations. We want to understand what organizations are victims of zoombombing. Two of the services (Zoom and Webex) that we study allow organizations to set up a subdomain that identifies them (for example https://virginia.zoom.us/j/123456789 to identify the University of Virginia on Zoom and https://pacificbuddhistacademy.my.webex.com for the Pacific Buddhist Academy). We find that most zoombombing links posted on 4chan and Twitter are generic and do not contain subdomains that are specific to any organization: only 12 links contain specific subdomains to 10 institutions, and two links contain specific subdomains to one institution on Twitter. In particular, we find that 8 zoombombing links on 4chan belong to education institutions while there are none on Twitter. One of these is a high school located in the US (Evergreen PS in Washington), four are universities in the US (e.g., Arizona State University), and three are universities outside the US (e.g., Concordia in Canada). In Section VI-A we will show that the text of zoombombing posts often further identifies the institution or organization that the zoombombing link belongs to.

Password protection. As we discussed in Section II-B, two of the ten online meeting services (Zoom and Webex) allow hosts to protect their meetings using passwords. In the case of Zoom, the password can be embedded in links as a URL parameter (for example https://zoom.us/j/123456789?pwd=12345aAbBcC678). We find that 20 of the 123 bombing invitations on 4chan and 64 of the 95 on Twitter include a password. This is interesting because having a password by default was added by Zoom after the quarantine started, with the explicit goal of curbing zoombombing. In fact, we find that zoombombing posts containing passwords are concentrated toward the latter part of our timeline (see Figure 9). The week Zoom started to enable passwords by default is shown in the figure with a green line[2]). This is a worrying trend, since, as we confirm in Section VI-A, it indicates that many attacks are called for by insiders who have legitimate access to the meetings. This calls into question existing security measures and provides the impetus for rethinking these mitigation strategies.

Bombing				Non Bombing			
4chan		Twitter		4chan		Twitter	
Word	**Sim.**	**Word**	**Sim.**	**Word**	**Sim.**	**Word**	**Sim.**
virtual	0.834	zoomcodes	0.860	nihilist	0.628	live	0.264
lecture	0.820	boys	0.819	cia	0.561	virtual	0.249
lesson	0.777	zoin	0.814	join	0.552	pm	0.247
class	0.774	zoomclasse	0.812	neo	0.549	zoom	0.239
crash	0.755	girls	0.802	program	0.505	link	0.239
join	0.697	pm	0.792	nazi	0.502	join	0.229
webex	0.685	raiding	0.785	goat	0.482	please	0.208
meeting	0.682	random	0.771	glownigger	0.478	detail	0.195
conference	0.681	shit	0.771	fbi	0.455	march	0.192
password	0.675	join	0.769	autistic	0.374	reminder	0.178

Table III: Top 10 most similar words (by cosine similarity) related to online meeting links in Bombing & Non Bombing Threads and Tweets.

D. Content Analysis

After looking at timing information and at the characteristics of URLs, we focus on analyzing the language of social media posts/threads containing zoombombing invitations on Twitter and 4chan. To this end, we leverage *word embedding* models (i.e., word2vec [35]) to quantitatively learn about the context in which zoombombing links are discussed. Intuitively, this allows us to identify common themes used in discussions where the links appear. To build our models, we first replace all meeting links with the keyword "meetinglink."

For both 4chan and Twitter, we train two word2vec models, one for posts (and threads in the case of 4chan) containing zoombombing links, and one for posts and threads with benign meeting links. On 4chan, we use a window size of 7 and limit our vocabulary to words that appear at least 5 or 84 times for bombing and non-bombing threads, respectively, maintaining the ratio of total posts left after preprocessing. To avoid the effect of common/unnecessary words in our model, we remove stop words, punctuation, other URLs, mentions, posts with only one word, and exact quotes of previous posts in the case of threads. We also lemmatize the posts and convert all text to lowercase. On Twitter we apply the same pre-processing techniques as 4chan, as well as removing emojis, numbers, non-alphanumeric characters from words, and some Twitter-related keywords like RT and FAV. Since tweets are usually shorter than 4chan posts, to build our word2vec models we use a window size of 5. We keep words that appear at least 7 times for non-bombing tweets and words that appear at least once for bombing tweets to maintain ratios as we do for 4chan.

Since online meeting links do not have a fixed position in posts, but attackers place them arbitrarily as a word inside of a sentence, we use the Continuous Bag-Of-Words Model (CBOW) [35] for training our models.

Most representative words. After building our models, we want to identify the words that are "closer" to zoombombing

[2]https://www.businessinsider.com/zoom-security-passwords-waiting-rooms-stop-zoombombing-2020-4

and non-bombing links on both Twitter and 4chan. To do this, we look for the most similar words to "meetinglink" with similarity defined as the cosine similarity of the embedding vectors of words in our trained models.

As seen in Table III, the most representative words for zoombombing and non-bombing content are very different. On 4chan, we notice that most zoombombing words are related to education (e.g., "lecture," "class") or business meetings (e.g., "meeting," "conference"). On Twitter, we observe references to education as well ("zoomclass") as well as keywords related to attacks (e.g., "raiding"). For non-bombing content, on Twitter we observe that most keywords are related to conference meetings, reflecting the fact that public meeting URLs are often posted on the platform. On 4chan, we observe that non-bombing meeting URLs are often related to trolling and political discussion.

Visualizing discussion themes. We next aim to identify recurring "themes" in zoombombing content. To this end, we visualize the relationship between the words related to online meeting links following the methodology of Zannettou et al. [12]. From our trained word2vec models, we create a two-hop ego network centered around "meetinglink" where words are nodes, and edges are weighted with the cosine similarity between the those two words; we keep any edge whose weight is greater than or equal to a pre-defined threshold, and visualize this as a graph. For each graph, we elect the threshold as the value that results in a graph with 100 nodes (for ease of representation). We then detect communities of words using the Louvain algorithm [4], and display them using Gephi's ForceAtlas2 algorithm [24].

Figures 10 and 11 show the results of this analysis for zoombombing invitations in 4chan threads and Twitter posts, respectively. Intuitively, each colored community can be interpreted as a "theme" that features prominently in these posts. Looking at the 4chan graph (Figure 10) we see that many of the themes feature educational topics (e.g., the red community with "spanish," "course," and "skype" and the purple community with "university," "college," and "class"). We also note a community (orange) where users talk about security issues/conspiracies as we can infer from words like "ccp," "tiktok," "spyware," and "ban." This indicates that conspiratorial content is not only commonplace in regular discussion on 4chan, but is also featured in zoombombing content in particular. See the following post for example:

> "If you do the research you'll see our MSM is in bed with the CCP. This is being utilized for propaganda purposes just like tiktok. I work with a bunch of regressed and they all love posting on tiktok. The users of these applications have close to zero foresight when it comes to Intel collection in any fashion from any party. Kind of we are fucked because Jews take chinese money as investments in their companies."

On Twitter (Figure 11) we again see themes that cover online classes (e.g., the green community with "class," "history," "math"). We also see a number of keywords used as hashtagsto ensure calls for zoombombing obtain more visibility (e.g., "zoomcodeclass," "zoombomb," "zoomraids").

For completeness, we report the graphs for non-bombing threads on 4chan and non-bombing tweets on Twitter in Figures 12 and 13 respectively. From the figures, we see the themes in these cases are more varied. For example, on 4chan we see a community of keywords related to World War II and Nazi Germany, while on Twitter there is a community related to research and webinars.

VI. Qualitative Analysis: Forum Content

Our quantitative analysis highlighted several interesting aspects of zoombombing invitations and their discussion. In particular, we found evidence that online classes in particular are targeted by attacks, and we found several meeting passwords included in invitations, which could be an indicator that attacks are called for by insiders who have legitimate access to the meeting rooms. When dealing with online activity carried out by humans, however, quantitative analysis can only identify general trends, and lacks the nuance required to provide a better understanding of the problem. In this section, we answer deeper questions via a more thorough qualitative analysis informed by our quantitative results. As explained in Section IV we conduct this analysis by having the four authors of the paper manually annotate the dataset. Where appropriate, our analysis covers zoombombing posts on Twitter and 4chan, while for some of the analysis (e.g., analyzing back and forth communication between attackers) we rely only on 4chan threads. Based on our threat model (see Section II-A), we analyze attacks across four phases: i) *Call for attack*, ii) *Coordination*, iii) *Delivery*, and iv) *Harm*.

A. Phase I: Call for attack

In this phase, an attacker posts a call for an attack on an online platform.

Targeting the class room. In Section V-C we showed that we could quantitatively identify 8 academic institutions targeted by zoombombing attacks on 4chan. In addition to information that can be directly extracted from the URL of the bombing link, many bombing posts include additional text indicating that online classes are the target. For example, "lecture," "teacher," "class," etc. show up regularly in these threads. We find that 91 of our 123 zoombombing threads on 4chan target online classes. Of the 32 remaining threads, three target business meetings, and the target of the remainder could not be conclusively determined. On Twitter, we find that 56 of our 95 bombing calls target schools.

Evidence of insiders' complicity. In Section V-C we showed that 11 zoombombing links on 4chan include passwords, indicating that the call for attack was from legitimate participant in the meeting (e.g., a student in the class). When annotating the threads, we find 9 additional zoombombing threads including a password in the body of messages. In total, this accounts for 20 of our 123 threads on 4chan. For Twitter, we showed that 64 of 95 tweets include a password in the zoombombing link.

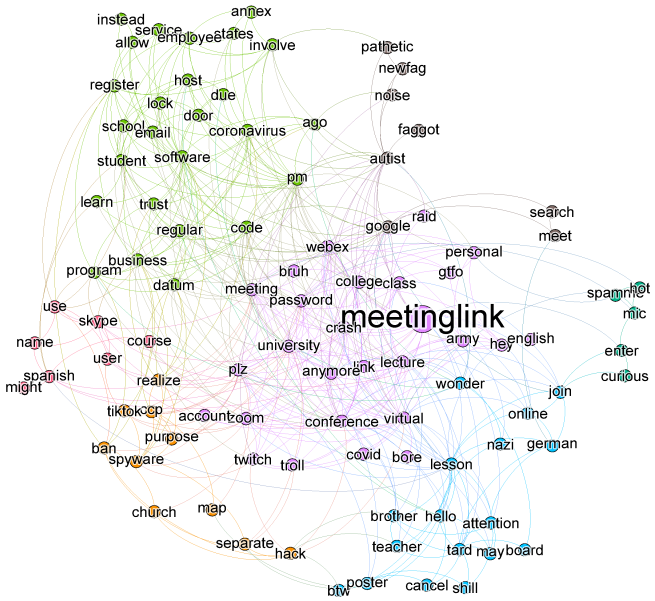

Figure 10: Words and themes associated with zoombombing links on 4chan.

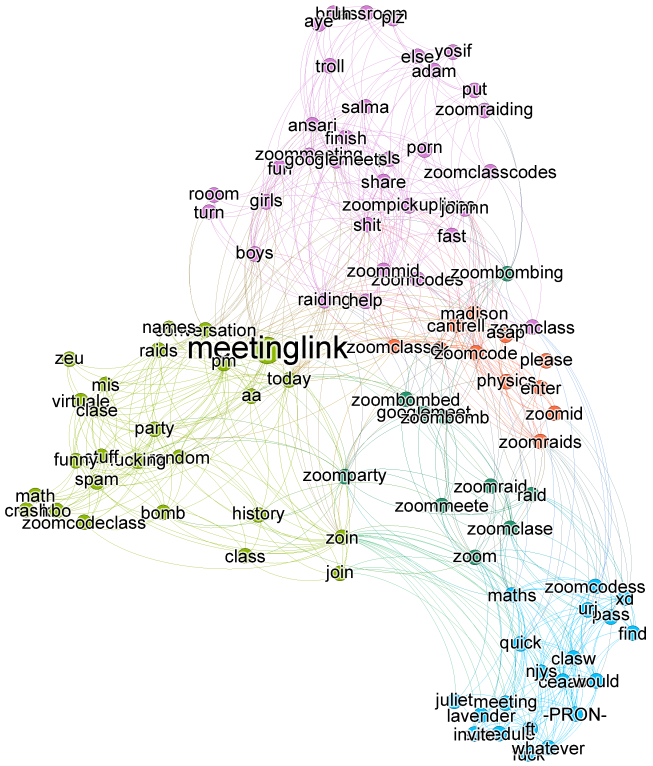

Figure 11: Words and themes associated with zoombombing links on Twitter.

There are additional indicators that can be used to qualitatively determine if an attack is called by an insider. We next look for two indicators: 1) whether the language of the call for the attack suggests it is called for by an insider and 2) whether whoever calls for an attack shares knowledge about the meeting only an insider would have.

For the first aspect, we look for language like "my lecture," "my colleague's presentation," "my company's meeting," etc. 58 zoombombing threads on 4chan and 19 zoombombing tweets include language indicating the attack is called for by an insider. In many cases, the users calling for the attack provide additional information that only an insider would know. In 8 zoombombing threads and 8 zoombombing tweets, the attacker asks others to use a certain name when joining the meeting to avoid being identified as an intruder and removed.

> "[GOOGLEMEETURL] name yourself [PARTICIPANTSNAME] all caps or she wont let you in." "Also please use real-sounding names."

In 11 threads we learn that the attacker is an insider from their interaction with other users.

> "Same school as you, different major. Someone wrote "NIGGERS" in my zoom class with the annotate function and started a zoom fight."

Together, with all information from both meeting links and post text, we identify 86 out of 123 zoombombing threads on 4chan that appear to have been posted by insiders (38/54 for Zoom, 35/46 for Google Meet, 8/10 for Cisco Webex, 3/3 for Skype, 0/2 for GoToMeeting, 2/7 for Jitsi, and 0/1 Teams). For Twitter, we find that 78 out of the 95 zoombombing tweets were posted by insiders.

Failed calls to attack. While 100 (out of 123) of our threads did start with an invitation to bomb, 46 of these 100 threads received no further replies. I.e., the call for an attack seems to have been stillborn. For the threads with replies, 54 (out of 77) were started with an invitation to bomb and 23 (out of 77) were created with more general topics of interest (e.g., politics, COVID-19, etc.) which were later converted into bombing threads. Threads with general topics tend to attract more posts than bombing threads.

B. Phase II: Coordination

After posting an invite to a zoombombing, attackers coordinate to carry it out. To better understand this, we look for temporal information on when the attack should be carried out in both 4chan threads and tweets.

Crimes of opportunity. Considering that most of the zoombombing links target online classes, and that these occur at regularly scheduled times, there is a question as to how much premeditation goes into a bombing attack. On the surface, it seems plausible that attacks could be planned days, and even weeks in advance. To dig deeper, we looked at the text posted along with a link and determined whether or not the invite was for a live meeting, or one that was scheduled to take place in the future. I.e., are attackers asking people to bomb *right now*

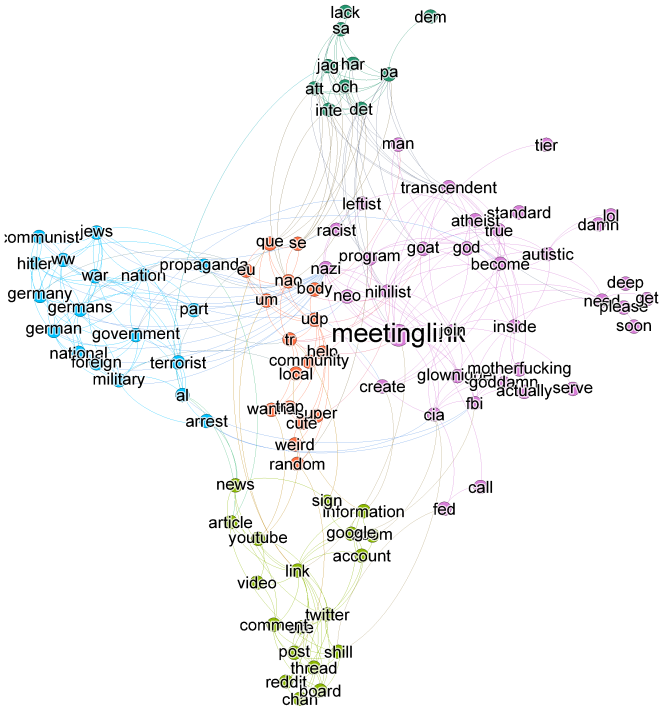

Figure 12: Words associated with online meeting links on non-bombing threads on 4chan.

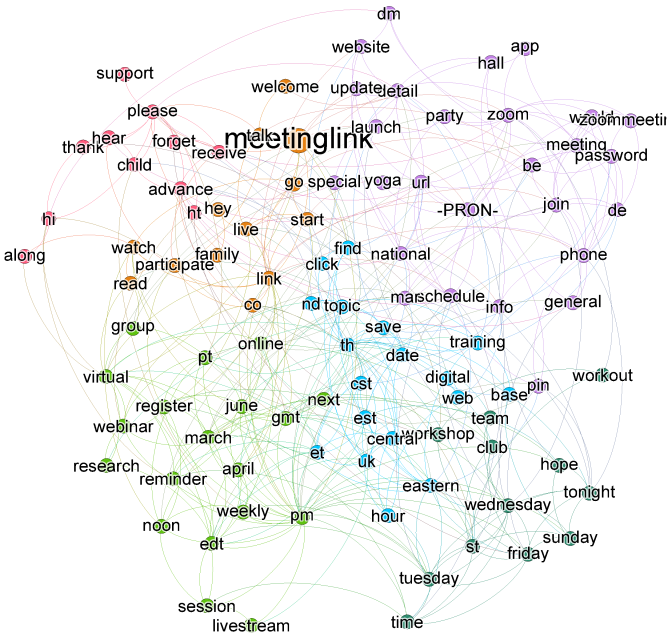

Figure 13: Words associated with online meeting links on non-bombing tweets on Twitter.

or planning a bombing that is going to happen later? We found that 115 of 123 bombing links on 4chan and 93 of 95 links on Twitter came along with a clear implication that the meeting was live at the time of posting. We find 8 future links among 123 links on 4chan and 2 out of 95 links on Twitter. A future link example from 4chan is:

> "RAID THIS BOOMER Wednesdays 10:00-10:45 [INSTITUTIONAL ZOOMURL]"

Refusing to participate. We find 20 threads on 4chan where users openly refuse to join into the attack, calling it unethical or referring to the fact that 4chan users are not the insider's personal army (NYPA – Not Your Personal Army). This indicates that not all users on 4chan are willing to participate in these attacks, and is particularly interesting because it is a possible explanation for at least some failed attacks: users do not reply because they reject the idea of being a troll in the service of another user.

> "[ZOOMURL]please spam this online class"
> "I'm not downloading shit"
> "Nypa faggot"

C. Phase III: Delivery

In this phase, the attackers join the online meeting and begin their harassing and disruptive actions. As part of our analysis, we find discussion of how the attacks went down in replies within the bombing threads on 4chan.

Quick action. We compare the time interval between when the link is posted and the first feedback on the attack. Of 123 bombing threads on 4chan, we find 37 with clear feedback related to the bombing. According to this analysis, a zoom-bombing attack often finishes within 20 minutes. An example of attack feedback on 4chan is as follows:

> 19:51:59 "Join a teachers zoom [ZOOMURL]"
> 20:05:18 "What the fuck is this? Who are these people?"
> 20:07:43 "quickly screencap it. They kicked me out instantly."

Problem feedback. For 24 threads we find participants reporting problems with the zoombombing invitation.

> "Raid our school live call class, i believe in you faggots. [GOOGLEMEETLINK]"
> "It says someone has to allow me to join, some shit like that"
> "this meeting has been locked by the host. Sad!"

D. Phase IV: Harm

Finally, we want to understand the toxic speech that happens during attacks, together with what actions attackers carry out.

Toxic speech. We find 14 4chan zoombombing threads containing toxic content including racism, sexism, or hateful words.

> "[SKYPEURL] Anyone wanna join our online lesson? Our teacher is black. Its gonna be in 20 mins."
> "NIGGER." "That is absolutely a 'he', no matter how the swine identifies."
> "What the fuck, I swear I spotted a beard on that chin."

On Twitter, we did not find any toxic tweets among the 95 zoombombing tweets. However, recall that on Twitter we only retrieve the call for attack and do not have any feedback (e.g., the replies to those tweets).

Crime scene feedback. On 4chan, we find 15 threads containing feedback from the zoombombing attack, providing us with a better view of what happens during these attacks. Here are some examples:

> "Hard working he's probably the kind of teacher who sits reverse on a chair and is up to date with the cool kids."
> "HAHAHAHA that was great."
> "Party's over my dudes, IT is here shutting down the stream, we had a good laugh."
> "Did you hear me saying nigger?"
> "Ayone heard me farting."
> "Yeah everyone heard and saw the chat and vc lmao."
> "I didn't hear that, maybe not loud enough but there was a bunch of rambling about the numbers on screen and then someone started farting and the class was just dying of laughter."
> "Nice bro."
> 'Totally lmfao. Best class disruption ever."

VII. Discussion

In this paper we presented a data-driven analysis of the emerging phenomenon of zoombombing. Our findings improve the understanding of who the people calling for zoombombing attacks are and how they operate. In the following, we first discuss the implications of our findings to existing mitigations against zoombombing, and propose some best practices to protect online meeting rooms. We then discuss the limitations of our study and some future work directions.

Implications for zoombombing mitigation. After the rise in popularity of online meeting tools, researchers have been looking at the privacy risks linked to online meeting [26]. At the same time, researchers, law enforcement, and the online meeting providers themselves have been publishing best practices to avoid zoombombing [6, 11, 55]. These include not posting meeting links publicly, protecting meeting rooms to control who can get in, and reducing the capabilities of participants, like muting them upon joining as well as disabling screen sharing and screen annotations.

The main assumption behind existing guidelines to prevent zoombombing is that attackers will actively seek out meeting links online, or that they will bruteforce their ID. Given this threat model, protecting meetings with passwords makes sense. However, our findings show that most of the calls for attacks that we observe come from insiders. This makes password protection ineffective, because the insider will share the password with the other attackers. Having participants join a waiting room and vet them before letting them in can be a more effective mitigation, although it inevitably increases the workload of meeting hosts, requiring moderators specifically checking the meeting room in the case of large meetings. Our analysis however shows that insiders often share additional information with potential attackers, for example instructing them to select names that correspond to legitimate participants in the meeting. This reduces the effectiveness of a waiting room, because it makes it more difficult for hosts and moderators to identify intruders.

Providing a unique link for each participant reduces the chances of success of zoombombing attacks. If the meeting service still allows multiple people joining with the same link, at least this gives some accountability, since the meeting host can identify who the insider was based on the unique link used by attackers to join. An even better mitigation is to allow each link to be used by a single participant at a time. This way, as long as the insider joins the meeting unauthorized people will not be able to join using the same link. While this mitigation makes zoombombing unfeasible, not all meeting services have adopted it. At the time of writing, only Zoom and Webex make available per-participant links that allow a single user to join at a time. To do this, Zoom requires participants to log in, and checks if the unique link is the same that was sent to that email address as a calendar invite. We encourage other meeting platforms to adopt similar access control measures to protect their meetings from insider threats. We also note that other similar mitigations are possible, like having meeting links expire after they are used once.

Additionally, we find that zoombombing attacks usually happen in an opportunistic fashion, with insiders asking others to join meetings happening in real time. This reduces the effectiveness of proactive measures like monitoring social media for calls for future attacks.

Limitations and future work. As with any data-driven study, our study is not exempt from limitations. We only have a 1% sample of Twitter available, therefore our zoombombing

results related to Twitter are a lower bound of the actual extent of the problem. Additionally, API limitations prevent us from collecting replies to zoombombing tweets, allowing us to only get a partial picture of how attacks unfold on the platform. On 4chan, users are anonymous. We therefore cannot trace per-user behavior, and this prevents us from observing serial offenders calling for multiple attacks over time. As an additional limitation, it is possible that zoombombing attacks are organized by other platforms other than Twitter and 4chan. While we believe that these two services provide a representative overview of behaviors and motives, attackers on other platforms might operate differently than what we observed in this paper. Finally, our analysis is limited to calls for attacks and responses to such calls on social media, but we are unable to observe what happens in the actual meeting rooms. Future work could develop alternative study designs that allow analyzing the attack on the online meeting platform itself, for example by collecting and analyzing recorded online meetings that were bombed, or by interviewing victims of zoombombing. This would also allow a better understanding of the mental and emotional toll on zoombombing victims.

VIII. Related Work

Coordinated malicious activity on social media. The security community has extensively studied automated malicious behavior on social media, mostly focusing on bots sending spam [17, 21, 51] and on malicious accounts colluding to inflate each other's reputation [10, 46, 48]. The mitigation systems proposed to detect and block this type of activity rely on the fact that these operations are large scale, rely on automated methods, and are carried out by single entities. Therefore, synchronization features can be used to distinguish between benign and malicious activity [7, 45, 54]. Alternatively, systems have been proposed that identify common traits in massively created fake accounts, for example an anomalous fraction of followers to friends or a large set of accounts created around the same time [3, 9, 44, 50, 51].

More recently, the community's focus expanded to looking at coordinated malicious campaigns that are not carried out by automated means, but rather by humans controlling a small number of inauthentic accounts. This includes conspiracy theories being pushed on social media [41, 42] and influence campaigns by foreign state actors [1, 52]. While not as automated as large-scale bot activity, these campaigns still show coordination, which can be leveraged for detection [31].

Coordinated online harassment and aggression. A closer line of work to the problem studied in this paper looks at coordinated behavior geared toward harassing victims online. Kumar et al. [28] measure the problem of *brigading* on Reddit, where the members of one sub-community (*subreddit*) organize to disrupt another community by posting offensive messages and prevent it from continuing its normal operation.

Hine et al. [23] study the activity of 4chan's Politically Incorrect Board (/pol/), showing that members of that community often call for attacks against people who post videos on YouTube and end up harassing the poster in the comments section of the video. Mariconti et al. [33] develop a multimodal machine learning system able to predict which videos are likely to receive this kind of hate attack in the hope of aiding moderation efforts.

Zannettou et al. [53] investigate a similar phenomenon, studying the effect of posting a URL to a news article on 4chan and Reddit. They show that posting URLs to certain types of news outlets results in a sudden increase in the hate speech on the comments to that article.

Snyder et al. [40] study the problem of *doxing*, in which attackers post information about a victim, calling for people to attack that person through multiple media (e.g., on multiple social networks or through email), sometimes even transcending to the physical world.

Tseng et al. [47] analyze five forums in which miscreants share and discuss tools and techniques that can be used to spy on their partners and further harass them.

Our work builds on previous research on coordinated harassment by studying the emerging problem of zoombombing. Unlike previously studied threats, we show that zoombombing attacks are often called by insiders; this has important implications when designing security mitigations against the problem.

IX. Conclusion

In this paper, we perform the first data-driven study of calls for zoombombing attacks on social media. Our findings indicate that these attacks mostly target online lectures, and are mostly called for by insiders who have legitimate access to the meetings. We find that insiders often share confidential information like meeting passwords and the identity of real participants in the meeting, making common protections against zoombombing ineffective. We also find that calls for zoombombing usually target meetings happening in real time, making the proactive identification of such attacks challenging. To protect against the threat, we encourage online meeting services to allow hosts to create unique meeting links for each participant, although we acknowledge that this has usability implications and might not always be feasible.

X. Acknowledgments

We would like to thank the anonymous reviewers for their useful comments. This work was funded by the NSF under grant 1942610.

References

[1] A. Badawy, E. Ferrara, and K. Lerman. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2018.

[2] M. Bailey, D. Dittrich, E. Kenneally, and D. Maughan. The Menlo Report. *IEEE Security & Privacy*, 2012.

[3] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on Twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, 2010.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[5] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 2006.

[6] B. Brown. Notes on running an online academic conference or how we got zoombombed and lived to tell the tale. *Interactions*, 2020.

[7] Q. Cao, X. Yang, J. Yu, and C. Palow. Uncovering large groups of active malicious accounts in online social networks. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2014.

[8] CNBC. Zoom has added more videoconferencing users this year. "https://www.cnbc.com/2020/02/26/zoom-has-added-more-users-so-far-this-year-than-in-2019-bernstein.html".

[9] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. Botornot: A system to evaluate social bots. In *The Web Conference (WWW)*, 2016.

[10] E. De Cristofaro, A. Friedman, G. Jourjon, M. A. Kaafar, and M. Z. Shafiq. Paying for likes? understanding facebook like fraud using honeypots. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2014.

[11] FBI. Fbi warns of teleconferencing and online classroom hijacking during covid-19 pandemic. "https://www.fbi.gov/contact-us/field-offices/boston/news/press-releases/fbi-warns-of-teleconferencing-and-online-classroom-hijacking-during-covid-19-pandemic".

[12] J. Finkelstein, S. Zannettou, B. Bradlyn, and J. Blackburn. A quantitative approach to understanding online antisemitism. In *AAAI International Conference on Web and Social Media (ICWSM)*, 2019.

[13] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 1971.

[14] J. L. Fleiss, B. Levin, and M. C. Paik. *Statistical methods for rates and proportions*. john wiley & sons, 2013.

[15] C. I. Flores-Saviaga, B. C. Keegan, and S. Savage. Mobilizing the trump train: Understanding collective action in a political trolling community. In *AAAI International Conference on Web and Social Media (ICWSM)*, 2018.

[16] J. Fox and W. Y. Tang. Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies. *New Media & Society*, 2017.

[17] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. Detecting and characterizing social spam campaigns. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2010.

[18] Google. Google Meet add more users. https://www.theverge.com/2020/4/28/21240434/google-meet-three-million-users-per-day-pichai-earnings.

[19] Google. Mute or remove video meeting participants. "https://support.google.com/meet/answer/7501121?co=GENIE.Platform%3DDesktop&hl=en".

[20] Google. There is no current feature for 'mute all'. https://support.google.com/meet/thread/35068017?hl=en.

[21] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: the underground on 140 characters or less. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2010.

[22] S. Hinduja and J. W. Patchin. Bullying, cyberbullying, and suicide. *Archives of suicide research*, 2010.

[23] G. E. Hine, J. Onaolapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *AAAI International Conference on Web and Social Media (ICWSM)*, 2017.

[24] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLOS ONE*, 2014.

[25] Jitisi. There is no current feature for 'mute all'. "https://community.jitsi.org/t/option-to-mute-unmute-participants-by-moderator/15062".

[26] D. Kagan, G. F. Alpert, and M. Fire. Zooming into video conferencing privacy and security threats, 2020.

[27] B. Kaleli, M. Egele, and G. Stringhini. On the perils of leaking referrers in online collaboration services. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA)*, 2019.

[28] S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky. Community interaction and conflict on the web. In *The Web Conference (WWW)*, 2018.

[29] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *The Web Conference (WWW)*, 2010.

[30] T. Lauinger, K. Onarlioglu, A. Chaabane, E. Kirda, W. Robertson, and M. A. Kaafar. Holiday pictures or blockbuster movies? Insights into copyright infringement in user uploads to one-click file hosters. In *International Symposium on Recent Advances in Intrusion Detection (RAID)*, 2013.

[31] L. Luceri, S. Giordano, and E. Ferrara. Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us election. In *AAAI International Conference on Web and Social Media (ICWSM)*, 2020.

[32] J. M. MacAllister. The doxing dilemma: seeking a remedy for the malicious publication of personal information. *Fordham L. Rev.*, 2016.

[33] E. Mariconti, G. Suarez-Tangil, J. Blackburn, E. De Cristofaro, N. Kourtellis, I. Leontiadis, J. L. Serrano, and G. Stringhini. "You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. *ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2019.

[34] L. McLean and M. D. Griffiths. Female gamers' experience of online harassment and social support in online

gaming: a qualitative study. *International Journal of Mental Health and Addiction*, 2019.

[35] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. 2013.

[36] A. Nagle. *Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right*. 2017.

[37] A. Papasavva, S. Zannettou, E. De Cristofaro, G. Stringhini, and J. Blackburn. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In *AAAI International Conference on Web and Social Media (ICWSM)*, 2020.

[38] Reuters. Cisco webex draws record users. "https://www.reuters.com/article/us-cisco-systems-webex/ciscos-webex-draws-record-324-million-users-in-march-idUSKBN21L2SY".

[39] C. M. Rivers and B. L. Lewis. Ethical research standards in a world of big data. *F1000Research*, 2014.

[40] P. Snyder, P. Doerfler, C. Kanich, and D. McCoy. Fifteen minutes of unwanted fame: Detecting and characterizing doxing. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2017.

[41] K. Starbird. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *AAAI International Conference on Web and Social Media (ICWSM)*, 2017.

[42] K. Starbird, A. Arif, and T. Wilson. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, 2019.

[43] Statista. Numbers of skype. "https://www.statista.com/statistics/820384/estimated-number-skype-users-worldwide/".

[44] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In *Annual computer security applications conference (ACSAC)*, 2010.

[45] G. Stringhini, P. Mourlanne, G. Jacob, M. Egele, C. Kruegel, and G. Vigna. Evilcohort: Detecting communities of malicious accounts on online services. In *USENIX Security Symposium*, 2015.

[49] Wikipedia. Zoombombing. "https://en.wikipedia.org/wiki/Zoombombing".

[46] G. Stringhini, G. Wang, M. Egele, C. Kruegel, G. Vigna, H. Zheng, and B. Y. Zhao. Follow the green: growth and dynamics in Twitter follower markets. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2013.

[47] E. Tseng, R. Bellini, N. McDonald, M. Danos, R. Greenstadt, D. McCoy, N. Dell, and T. Ristenpart. The tools and tactics used in intimate partner surveillance: An analysis of online infidelity forums. In *USENIX Security Symposium*, 2020.

[48] J. Weerasinghe, B. Flanigan, A. Stein, D. McCoy, and R. Greenstadt. The pod people: Understanding manipulation of social media popularity via reciprocity abuse. In *The Web Conference*, 2020.

[50] C. Yang, R. C. Harkreader, and G. Gu. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In *International Symposium on Recent Advances in Intrusion Detection (RAID)*, 2011.

[51] D. Yuan, Y. Miao, N. Z. Gong, Z. Yang, Q. Li, D. Song, Q. Wang, and X. Liang. Detecting fake accounts in online social networks at the time of registrations. In *ACM Conference on Computer and Communications Security (CCS)*, pages 1423–1438, 2019.

[52] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, and J. Blackburn. Who let the trolls out? towards understanding state-sponsored trolls. In *ACM conference on web science*, 2019.

[53] S. Zannettou, M. ElSherief, E. Belding, S. Nilizadeh, and G. Stringhini. Measuring and characterizing hate speech on news websites. In *ACM conference on web science*, 2020.

[54] Y. Zhao, Y. Xie, F. Yu, Q. Ke, Y. Yu, Y. Chen, and E. Gillum. Botgraph: Large scale spamming botnet detection. In *USENIX Symposium on Networked Systems and Design (NSDI)*, 2009.

[55] Zoom. How to keep uninvited guests out of your zoom event. "https://blog.zoom.us/keep-uninvited-guests-out-of-your-zoom-event/".