A Dataset of Coordinated Cryptocurrency-Related Social Media Campaigns

Karolis Zilius[1], Tasos Spiliotopoulos[2], Aad van Moorsel[2]

[1]Newcastle University, UK,
[2]University of Birmingham, UK
kikarolis@gmail.com, a.spiliotopoulos@bham.ac.uk, a.vanmoorsel@bham.ac.uk

Abstract

The rise in adoption of cryptoassets has brought many new and inexperienced investors in the cryptocurrency space. These investors can be disproportionally influenced by information they receive online, and particularly from social media. This paper presents a dataset of crypto-related bounty events and the users that participate in them. These events coordinate social media campaigns to create artificial "hype" around a crypto project in order to influence the price of its token. The dataset consists of information about 15.8K cross-media bounty events, 185K participants, 10M forum comments and 82M social media URLs collected from the Bounties(Altcoins) subforum of the BitcoinTalk online forum from May 2014 to December 2022. We describe the data collection and the data processing methods employed and we present a basic characterization of the dataset. Furthermore, we discuss potential research opportunities afforded by the dataset across many disciplines and we highlight potential novel insights into how the cryptocurrency industry operates and how it interacts with its audience.

Introduction

The last years have seen the rise of Decentralized Finance (DeFi), a form of finance that does not rely on central financial intermediaries, such as banks and exchanges, to offer traditional financial services, and instead utilizes smart contracts on distributed ledgers (blockchains). The cryptographically secured digital tokens that are stored on these blockchains, called cryptoassets, include cryptocurrencies (like Bitcoin and Ethereum), utility tokens, security tokens and Non-Fungible Tokens (NFTs) (Karim et al. 2022). A simple application of this is that people can effectively invest in a company or project by acquiring and trading digital tokens of that company without intermediaries. Because these cryptoassets are freely traded, there is a lot of room for price speculation and manipulation (Nizzoli et al. 2020). Importantly, compared to traditional investing, this decentralized investing can be conducted by lay investors (Abramova et al. 2021) who are often influenced disproportionally by information they receive from social media (Glenski, Saldanha, and Volkova 2019; Jahani et al. 2018). This provides

a strong incentive for such crypto projects to create artificial "hype" in order to convince more social media users to buy their token, or in order to manipulate the price of a token in a certain way. In particular, they provide payments in the form of "bounties" to users willing to engage with the project on social media in a positive way, such as following, commenting, retweeting and creating content.

This paper presents a dataset of crypto-related bounty events that coordinate social media campaigns to create artificial hype around a crypto project, as well as the users that engage in these campaigns. Our dataset was crawled from the *Bounties(Altcoins)* subforum of the *Bitcointalk.org* online forum. While a lot of research has studied inauthentic, manipulative, deceptive and collusive content online, this type of bounty hunting phenomenon is unique due to the particularities of the cryptoassets context. In particular, the sentiment and the volume of information online have been found to strongly influence market prices and user investing decisions in traditional finance (Bollen, Mao, and Zeng 2011), and even more so in relation to cryptoassets (Domingo, Piñeiro-Chousa, and Ángeles López-Cabarcos 2020; Xie, Chen, and Hu 2020). In the context of cryptoassets, these coordinated campaigns aim to achieve this in two ways; directly, by influencing social media users to buy an asset based on artificial hype, and indirectly, by influencing trading algorithms that make investing decisions based on sentiment mined from social media (Garcia and Schweitzer 2015). What makes this practice particularly concerning is that bounty hunters are typically paid in the token that they are promoting. This gives them an incentive to "dump" (i.e., sell) the asset when the campaign ends, and since they have inside knowledge of exactly when the campaign will end, this practice can be considered equivalent to a "pump-and-dump" scheme (Hamrick et al. 2021) that will dramatically crash the price of the token causing the investors outside the campaign to lose the funds they invested.

While this type of bounty hunting phenomenon is unique, it builds on known practices that have been identified and studied earlier in the literature. These practices have been tailored specifically to the crypto space and are of varying ethical and legal positioning and justification. **Bounty programs**, in general, have been commonly used to manage rewards traditionally aimed at incentivizing positive and productive behavior, such as finding bugs in software code

(Ding, De Jesus, and Janssen 2019), providing worthy answers in question-answer platforms (Parnin et al. 2012) and verifying the bias and safety of AI systems (Brundage et al. 2020). The practices and ethics of **advertising**, in general (Schauster, Ferrucci, and Neill 2016), and social media promotion (Zeng, Kohno, and Roesner 2021), in particular, have been debated by researchers. Related work has highlighted particularly questionable practices, such as **astroturfing** (i.e., hidden coordinated information campaigns that mimic genuine user behavior by incentivizing agents to spread information online) for consumers (Kauppila and Soliman 2022; Lee, Tamilarasan, and Caverlee 2013) and citizens (Schoch et al. 2022). Dutta and Chakraborty (2021) have studied **blackmarket services** that are employed by "collusive" users to inflate the popularity of their online account and get appraisals for their content, while **spam** has been a topic of intensive research over the years (e.g., Benevenuto et al. 2010; Yardi et al. 2009).

From the perspective of **finance**, researchers have examined the inner workings, the nuances and the relationships between cryptocurrency pump-and-dump schemes (Hamrick et al. 2021), market price manipulation (Nizzoli et al. 2020), and Initial Coin Offerings (ICOs) (Ante, Sandner, and Fiedler 2018). From a **legal and regulatory** perspective, researchers are examining the ways and the extent to which social media content creators comply with the guidelines of the US Federal Trade Commission (FTC) which requires them to disclose their endorsements in order to prevent deception and harm to users (Mathur, Narayanan, and Chetty 2018). Specifically in the crypto space, scholars are examining the practices of crypto users and businesses in order to determine the status of cryptoassets as *securities* (i.e., investment contracts), something that would also place restrictions on discussions around cryptoassets and would require disclosure of any connections or endorsements (Henderson and Raskin 2019). Finally, such practices are in violation of the Terms of Service (ToS) of online platforms; for example Twitter's platform manipulation and spam policy[1] prohibits artificially amplifying content and makes specific mentions to coordinated activities for doing so. This mix of characteristics of other practices that is present in crypto-related social media bounty hunting muddles the way that campaign organisers, participants and even targets of the campaigns perceive the ethical and legal implications.

This paper presents a dataset collected from the *Bounties(Altcoins)*[2] subforum (message board) of the *Bitcointalk.org*[3] forum describing crypto-related bounty events, the participants, the actions taken and the rewards received from 13-May-2014 when the first thread was started to 31-Dec-2022. We collected and analyzed information about 15.8K unique cross-media bounty events, 185K bounty hunters (i.e., users that participated in such events), 10.01M comments, 18M spreadsheet lines, 56K images, 42.1M Tweets and 33.9M links to Facebook posts, among other in-

formation posted on the subforum. We further enhanced the dataset to make it accessible to a broader set of researchers and enable future research in this area by linking information collected across the forum threads and aggregating it to calculate useful metadata. This resulted in a rich set of descriptive metadata; for example, we provide 31 data fields describing each Bounty event and 17 data fields describing each participant.

While previous research has focused on detecting coordinated social media campaigns and attempting to infer the actors and the details of deceptive actions from their social media traces (e.g., Peng et al. 2017; Keller et al. 2020), our dataset provides comprehensive information on these campaigns and their organisation from their source, including the structure of incentives, exact instructions and timing of actions, and a full account of all participants. We expect that this unprecedented detail afforded by our dataset will be invaluable to researchers studying online coordinated deceptive practices.

We make the entire dataset, along with a smaller sample and supplementary material, publicly available at the following link: (Dataset URL: https://zenodo.org/record/7539178).

The Bitcointalk.org Forum and the Bounties(Altcoins) Subforum

The BitcoinTalk forum was created by the developer of Bitcoin in 2009, the same year that Bitcoin was developed[4]. While originally focusing on Bitcoin itself, the discussions on the forum gradually expanded to include alternative cryptocurrencies (altcoins). The forum currently has more than 3.5M registered users and 61.5M posts on 1.3M topics[5]. As the first and very popular community created around cryptocurrencies, the forum had an original intention of furthering adoption and supporting users. Areas of discussion include technical aspects of distributed ledgers, such as development and mining, economics, politics, as well as supporting users in using the technology, trading, and recovering from scams. As cryptocurrencies became more widespread and less technically-inclined users turned to the forum for information and advice, researchers utilized the BitcoinTalk forum to study crypto-related collective intelligence (Jahani et al. 2018), speculation (Xie, Chen, and Hu 2020) and scams (Vasek and Moore 2019).

The main BitcoinTalk forum comprises multiple subforums, each focusing on a separate aspect of cryptoassets. Each subforum contains many discussion threads initiated by different users. The earliest entries of the Bounties(Altcoins) subforum date back to 2014, although it didn't receive much attention until 2017. The subforum experienced a massive growth in popularity during 2017 – 2018, which aligns with the first steep increase in crypto adoption. More than 15,000 threads have been created since and it has evolved to the most popular subforum of BitcoinTalk[6].

Bounties(Altcoins) serves as a place to organize airdrops (cryptocurrency giveaways) and alert the public about up-

---

[1]https://help.twitter.com/en/rules-and-policies/platform-manipulation

[2]https://bitcointalk.org/index.php?board=238.0

[3]https://bitcointalk.org/

[4]https://bitcointalk.org/index.php?topic=5

[5]https://bitcointalk.org/index.php?action=stats

[6]see note 5

coming projects or ICOs, however it is dominated by *bounty events*. A bounty event involves the distribution of cryptocurrency or monetary rewards for the accomplishment of event-specific objectives. These rewards are typically distributed in the token being promoted, although it is not uncommon to use stablecoins (e.g., USDT) and even US dollars transferred via a service like Paypal. Each bounty event can have one or more *campaigns* which are allocated a percentage of the total reward pool. The vast majority of the campaigns require participants to use their social media accounts to spread awareness about the project, however event organizers may also be interested in producing quality content with campaigns such as translation (e.g., of a whitepaper or website content), copywriting, or visual content creation such as infographics or visual art.

Over time, certain norms have arisen to make it easier for users to participate in a bounty event. Although for some events the procedure differs slightly, most events follow a three-part structure:

1. A user shows their interest in participating by posting a "proof of registration" comment and including information about themselves (e.g., Telegram username, campaigns joined, crypto wallet address) which may differ slightly from event to event.

2. The participant completes the tasks defined in the campaign rules section. The event can last from a few weeks to a few months or even up to a year, therefore the tasks can be completed more than once (usually once a week).

3. The participant provides evidence of engagement with the project by posting "proof of participation" comments that include some required information (e.g., social media handle, social media interaction links) which is usually similar for all events. However, a small subset of events uses other means for providing evidence, such as Google forms.

Users that employ social media bot accounts are identified by the event organisers and are removed from the event. Participants are rewarded for their contribution at the end of an event. Each campaign has different amounts of stakes that are distributed to users which is proportional to the difficulty of the task, the quality of content, or the value provided by the user (e.g., a user with more Twitter followers will be allocated more stakes for the same task). The total reward can be calculated with the formula

$$User\ Rewards = \frac{Campaign\ Prize\ Pool}{Distributed\ Stakes} * User\ Stakes$$

Data Collection

In this section we describe our data-scraping methods. The process consisted of crawling *subforum* pages, *user* pages, *thread* pages, *images*, and *Google spreadsheets*. To achieve this, a Python crawler was developed which used HTML requests library[7] to fetch images and information from forum pages, BeautifulSoup[8] to parse HTML and gspread library[9] to access Google Sheets API endpoints.

---

[7]https://github.com/psf/requests
[8]https://github.com/waylan/beautifulsoup
[9]https://docs.gspread.org/en/v5.7.0/

Subforum and thread crawling The forum comprises multiple subforums, which contain a collection of threads divided into pages. Each subforum has a unique identifier; Bounties(Altcoins) ID is 238. Each page in the subforum contains up to 40 threads and has a unique identifier which starts at 0 for the first page and is increased by 40 for every subsequent page. Additionally, each subforum page contains a navigation strip, which can be utilized to retrieve the last page ID. This allowed us to generate a unique URL for each page by inserting the identifiers into *https://bitcointalk.org/index.php?board=238.page_id*. Crawling all pages resulted in retrieval of information from 15K unique events.

Comments crawling Each forum thread has a unique identifier retrieved during subforum crawling. Similarly to subforum page, each thread has a comment page ID which starts at 0 for the first page and is increased by 20 for each subsequent page and a navigation section which can be used to derive the last comment page ID. This allowed us to generate unique URLs for each comment page by inserting unique identifiers into *https://bitcointalk.org/index.php?topic=thread_id.page_id*. 10M comments were obtained after crawling each page.

User profile crawling Forum users are given a unique identification number which was retrieved during comment crawling. Inserting it into *https://bitcointalk.org/index.php?action=profile;u=user_id* allowed us to access each user's profile page. Information on 185K users was retrieved after crawling every page.

Image crawling Some events in the forum used images to convey information, which had to be retrieved, converted to text and inserted back into the main post for more accurate processing. The image URLs could be found in HTML `<img>` elements of the main forum post. In total, 64K unique image URLs were detected out of which 56K were downloaded successfully and 8K were no longer being hosted.

Spreadsheet crawling Some events used Google spreadsheets to track user progress and keep user information in one place. Spreadsheet URLs were collected from main posts using regular expressions to match spreadsheets domain[10] and cleaned to extract unique spreadsheet identifiers. Using this information, all rows were retrieved from Google sheets by querying Google Sheets API with gsread. This resulted in the retrieval of more than 20M distinct rows from Google spreadsheets.

Data Processing

In this section we provide an overview of how the collected data were processed and labeled and we explain the structure of the dataset.

Forum Threads

The primary focus of the dataset is bounty events. This section explains the methods used to extract information from

---

[10]https:///www.docs.google.com/spreadsheets

the main post as well as how such events were distinguished from other sorts of events in the subforum.

Categorization   Information from 15,870 unique events was fetched from the first thread creation on 13-May-2014 to 31-Dec-2022. Each thread was assigned one or more categories based on the title, main post comment and user replies. The category of *Bounty* was assigned to threads that:

1. Had seven or more out of 14 common Bounty event traits. Nine of the traits involved matching against a single word (Stake, Rewards, Campaign, General rules, Token, Bounty, Pool, Spreadsheet, Whitepaper). Five traits had multiple synonyms (Rules: 3, Reward allocation: 19, Proof: 2, Signature: 2 and Social media campaigns: 22); or

2. Had six out of the 14 common bounty event terms and more than 100 replies or the terms included *Bounty* and *Whitepaper*; or

3. More than 50% of comments were classified as registration / participation (see next section for details) and the event was closed (the main post, title or both removed) or had more than 200 replies.

Events containing the keyword *bounty* in the title but not meeting any of the requirements above were assigned a category of *Bounty(low quality)*. These categories had to be separated because titles were created by users and some threads that contained the *bounty* keyword in the title had nothing to do with bounty hunting.

Other categories included: *Moved* (event moved to another subforum), *ICO* (Initial Coin Offering), *Closed* (removed content), *Announcement*, and *Other*. 7,177 events were identified as Bounty events, 3,432 threads were identified as Bounty (Low Quality) and 5,261 events were assigned one of the other categories. While just fewer than half of the threads were categorized as Bounty events, 96.4% of the collected comments were posted in Bounty events and 1.2% to Bounty(Low Quality) events, demonstrating the importance of Bounty events in the subforum.

Content processing   Information about forum events was gathered from several places. Data which are included in every thread were collected from the subforum page which is well structured. This information includes *thread title*, *thread author*, *username* and *unique ID*, number of *replies* and *views*, and information about last comment: *date*, *author username* and *ID*. The thread page also contained structured information that was common between all events such as event creation date and main post in HTML format. Further analysis was only carried out on events assigned a category of Bounty.

All HTML `<a>` elements of the main post were inspected for links. By matching domain names with regular expressions, URLs were divided into the following categories: *spreadsheet*, *image* – embedded images that were converted to text, *forum* – bitcointalk.org domain, *social media* and *other* – the rest of the URLs.

The remaining event information was obtained from text. To overcome a BeautifulSoup parser limitation when converting HTML to text and to preserve post structure, a new line character was inserted after each element that was not `<br>` or inline. The text was split on new lines to aid the analysis process.

The *Reward pool* size section of the event was identified by matching keywords (e.g., "Bounty pool", "Bounty details"). Subsequent lines were checked for three common reward pool patterns (e.g., "$100 worth of ABC", "ABC 500", "200$ in ABC token").

Token names were detected in title, reward pool and text (near phrases such as "Token" or "Coin"). Each of these strings was tested against a set of conditions: two characters or longer, no words in a list of 27 phrases (e.g., "NFT", "USDT", "APY"), only contain characters A-Z and $, not a word in a USA or UK dictionary. Strings that passed all checks and were most frequent are included in the dataset.

Reward allocation was determined by matching phrases (e.g., "Reward allocation", "Budget"). Each subsequent line was inspected for % sign and campaign name. If the percentages of one or more campaigns added up to a 100%, then a list of campaign titles and reward allocation percentages was recorded in the database.

Bounty events would often have sections for event *general rules*, and *rules* and *rewards* for each campaign. Retrieving these sections comprised two steps; detecting in which line the section began and where it ended:

1. The general rules section was detected by matching the text for relevant phrases (e.g., "Bounty rules"). For each campaign this was achieved by matching a phrase (e.g., "Twitter campaign") followed by the keywords "Rules" and "Rewards".

2. Once the index of a line where the section began was found, the subsequent lines were checked for an ordered or unordered list. If found, this information outlined where the section began and finished.

Thread Comments

Categorization   The dataset contains five comment types: *mixed* (participation and registration in one), *participation*, *registration*, *author* (comments by the thread author) and *other*. Comment category was assigned with a 4-step process:

1. Check if userid of comment author and main post author match. If yes, assign author category, else next step.

2. Check how many out of six common proof of registration traits (headline "#proof of authentication", forum username, forum profile URL, telegram username, campaigns joined, wallet address) the comment has.

3. Check how many out of six common proof of participation traits (headline "week" or "day", social media URLs, social media campaign names, social media username, numbered list, keywords such as "Like" and "Retweet") the comment has.

4. If a comment passes a registration threshold of 4 traits (3 if there are fewer than 5 lines in the comment) or has # symbol with one of 75 synonyms for "proof of authentication" then the threshold for participation comment becomes 2 (it must contain social media links). If both
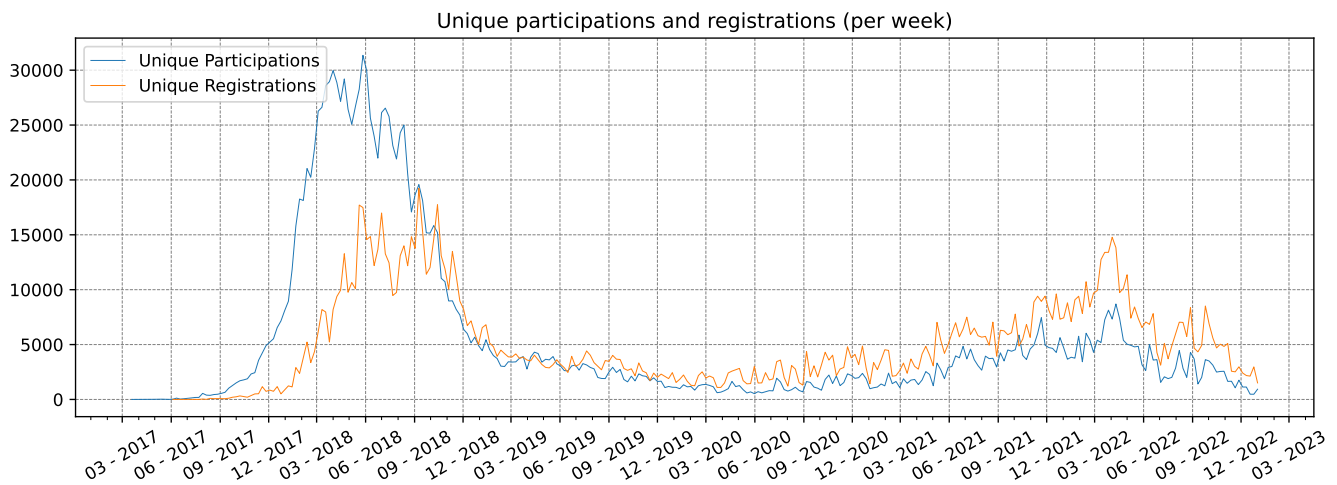
Figure 1: Number of unique participations and registrations per week (multiple comments of the same type by an individual in the same event are counted as one).

thresholds are passed, the comment is assigned a category of mixed else, it becomes registration. If registration threshold is not passed, the comment is considered as participation with threshold of 3 (must contain social media links or keywords: "week", "day"). If this threshold is passed the comment is assigned a category of participation, else it is assigned to other.

One limitation of this process is detecting registration comments which include little information, however it is very efficient at identifying and distinguishing between registration and participation comments. We report the results of a manual check of a random sample of 100 comments of each category for different threshold values in the Supplementary Material. We also provide the regular expressions used to match social media URLs.

Content processing   Each comment shares some mutual information retrieved from the forum: comment id, author username and id, post time. Additionally, the thread id where the comment was found is included in the dataset. Comments of type *other* and *author* were not processed and contain the comment text saved in HTML format. This choice was made to preserve the structure of the comment since it would be lost if the comment was a reply (i.e., contained one or more `<div class="quote">` elements).

On the other hand, *participation*, *registration* and *mixed* comments were processed extensively. Mixed comments include a Boolean check and get analyzed by both algorithms. Most of the information was extracted with conditional statements and regular expressions utilizing library re[11].

Registration comments include information such as *campaigns a user registered for*, *Telegram username*, *Twitter username*, *forum rank* and *post count* (specified by user), *line number* in spreadsheets and *crypto wallet address*. Ad-

ditionally, the dataset also contains *other information* which was rare and did not get assigned any label.

Participation comments include information about user *social media username* (with social media platform), *campaigns participated* in (based on participation links found) and all proof of interactions which include *TweetIDs* and social media links (*Twitter (without id)*, *Facebook*, *Instagram*, *Telegram*, *Reddit*, *YouTube*, *Medium*, *LinkedIn*, *Discord*, *Tik-Tok*, *Steemit*, *Image sharing*, *Other*)

Google spreadsheets   Each Google spreadsheet contains one or more sheets, which are processed independently. The process comprises two parts:

1. Many spreadsheets included banners of varying height, so we had to detect which row contains column names by iterating through the first 15 rows and checking which one has the most out of ten common column names (e.g., "timestamp", "twitter followers", "post count").

2. Label consequent rows based on column names.

Unidentified columns *labels* and *data* are also included in the dataset. Identified columns include: *timestamp*, *proof of registration post URL*, *forum rank*, *forum profile URL*, *forum username*, *wallet address*, *email address*, *stakes*, *twitter followers*, *twitter audit URL*, *post count*, *social media username*, *social media profile URL*.

Images to Text

Using Tesseract OCR[12], locally stored images were converted to text and inserted into respective HTML `<img>` elements. The process encountered a few limitations when dealing with intricate graphics pictures. Occasionally when dealing with such images, there were misinterpreted letters, random symbols, or parts of the text missing. The dataset contains the *text* generated from locally retrieved images and corresponding *image URLs*.

---

[11]https://docs.python.org/3/library/re.html

[12]https://github.com/tesseract-ocr/tesserac

Data category	Entries collected
Total threads in subforum	15,870
Bounty events	7,177
Total comments in subforum	10,024,001
Total comments in Bounty events	9,655,442
Bounty comments (participation)	6,926,577
Bounty comments (registration)	1,670,871
Bounty comments (mixed)	233,268
Bounty comments (author)	107,650
Bounty comments (other)	1,183,612
Social media links	82,854,734
Images (converted to text)	56,529
User profiles	185,709
Google spreadsheets	24,887
Lines in Google spreadsheets	20,182,746
Labeled Google spreadsheets entries	18,082,209

Table 1: Statistics of collected dataset.

Structure of the Processed Dataset

Our data collection and analysis process resulted in a PostgreSQL database with twelve tables. We exported the tables to twelve tsv files and added some additional annotations. The final dataset contains four tsv files of unprocessed information retrieved from the subforum, google spreadsheets, and image hosting services. To accommodate the largest audience possible and to aid further analysis by researchers, we also release eight processed tsv files resulting from the cleaning, merging and annotation conducted in the processing stage. In total, the files contain 159 fields (events – 31 fields, users – 17, comments_participation - 29, comments registration – 21, comments_author, comments_other and comments_raw – 7, rewards_and_rules – 4, spreadsheets – 18, spreadsheets_raw – 5, images_raw – 3, threads_raw – 10). More extensive description of each file and field can be found in the Supplementary Material that come with the dataset. Table 1 shows some statistics of the collected data.

Exploratory Analysis

Subforum activity over time Figure 1 shows forum activity in terms of *unique* user registrations and participations. If a user created more than one participation/registration comments in the same thread, they were only counted as one participation/registration. To preserve plot space, the visualization starts from 2017 as there are only ten threads that were created before that time. From the illustration we can observe that bounty events first started gaining traction in the second half of 2017 and peaked during the first half of 2018, which aligns with the timeline of the first massive bull market run[13] and corresponding increase in crypto adoption. During this period we can observe up to 30,000 unique participations every week. Since then, the forum has stayed active with some ups and downs in participation over time. After a substantial slump in 2019 and 2020 where activity fell to almost one tenth of the peak, activity started to slowly
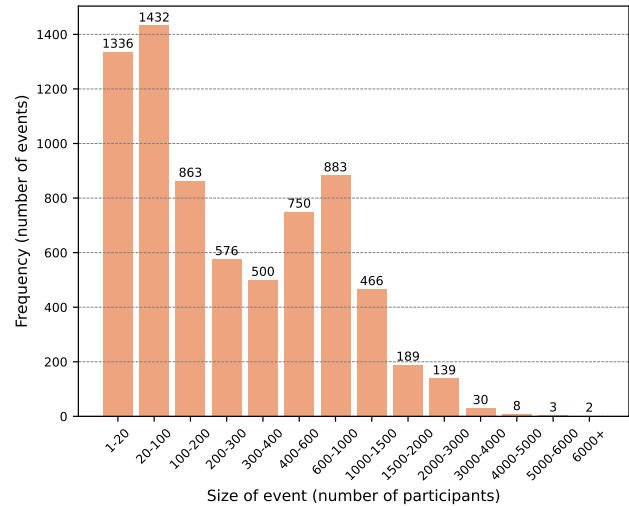
---

[13]https://coinmarketcap.com/charts/

Figure 2: Number of unique participants per Bounty event. Note that for visualization purposes, the bins in the horizontal axis are not equal.

recover in 2021 and falling again from 2022.

Size and activity of Bounty events Most Bounty events are substantially large to make a difference in promoting a project; 4,402 (61.4%) events had more than 100 unique participants and 837 (11.7%) had more than 1,000 unique participants. We found five outlier events where more than 5,000 forum users participated (Figure 2).

The number of activities that participants engaged in varied substantially across Bounty events. In the early days of the subforum the norms hadn't evolved yet and the comments did not have the structure that could help us crawl them successfully and categorize them as registration or participation comments. Still, a substantial number of events show relatively little activity with limited participation from users. On the other hand, we also identified a large number of events with several thousand participation comments; each one typically being a weekly comment including many activities across different campaigns (Figure 3).

Duration of events Figure 4 shows the duration of Bounty events. The duration is determined by calculating the time passed from the first post in the thread (which announces the beginning of the event) to the last (which typically ends the event). An event typically requires participants to post every week. However, sometimes, the event can get paused and the thread blocked for a few days. For this reason, we considered any period of more than three weeks between two consecutive comments as the end of an event.

We can observe that 350 (5%) events last less than a day and 30% end in the first month, which leads us to believe that the majority of Bounty events end successfully. A small number of events (92) kept operating for more than 300 days. These projects may have a long-term objective to remain in the cryptocurrency market and invest in a continuous bounty program.
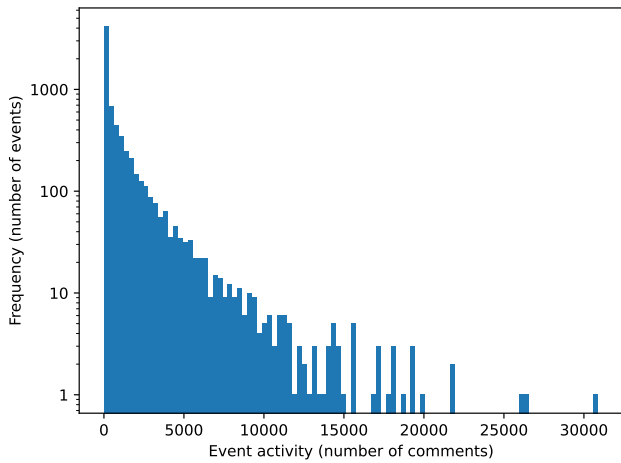
Figure 3: Number of participation comments per Bounty event. Note that the y axis is logarithmic.
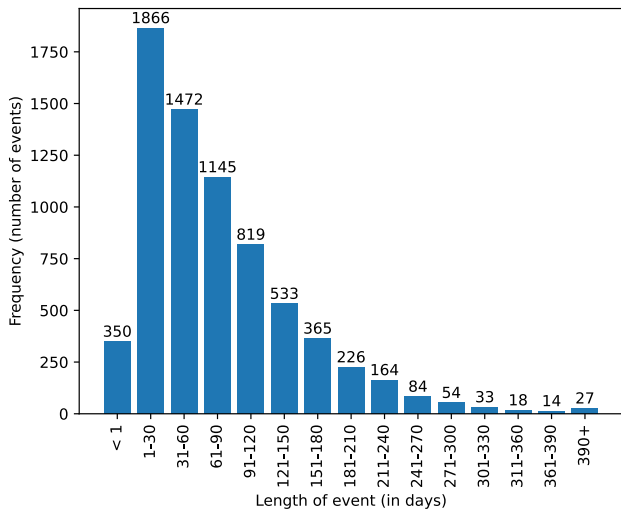

Figure 4: Number of consecutive days a Bounty event stayed active.

Bounty campaign popularity We discuss which campaigns are most likely to attract participating users and how many links grouped by campaign type have been collected from user comments.

Figure 5 illustrates how many unique participations for each campaign type were recorded in the subforum each week. Twitter and Facebook are the most utilized platforms across all time by a substantial margin. Telegram and Instagram started being utilized considerably in recent years.

In Figure 6 we can observe campaign popularity from another angle. Twitter and Facebook are still the most popular platforms with 42.1M and 33.9M URLs collected respectively, while all other platforms account for less than 9% of all links collected. No Telegram links were collected as these tasks did not require users to share a proof of par-

ticipation link. Campaigns that involved more complex or time-consuming tasks, such as creating YouTube videos or Medium blog posts are underrepresented compared to their relative importance.

Reward allocation Bounty events typically involve more than one campaign, necessitating the distribution of the reward pool across campaigns. Each campaign gets allocated a percentage of the total pool based on the project owners' preferences and the goal of the event. We were able to determine reward allocation that added up to 100% for only 2,705 of the 7,170 Bounty events, which we analyze below. Figure 7 shows the average reward allocation for each campaign type as a percentage of the full available reward pool. Campaigns that were encountered fewer than five times are grouped under the category "rare campaigns".

The campaign with the most funds allocated on average is *Signature*. Signature is an exclusive campaign to the forum that requires users to update their profile with a specific banner that is visible on all their posts, effectively acting as advertisements for the Bounty event. This suggests that event organizers may expect that forum users will consider campaigns with signatures to be more professional, trustworthy or popular, and thus will be more likely to join them.

The second most rewarded campaign type is *Content creation*. This campaign type includes activities such as .gif and meme creation, infographics, and copywriting. Together with the *Rare* category, which also attracts a substantial level of rewards, this suggests that Bounty events are particularly interested in unconventional, but also sophisticated and time-consuming activities of project promotion.

Twitter and Facebook are the third and fourth most rewarded campaigns, which is expected since they make up 91% of all links collected from proof of participation comments and are key channels for spreading information about Bounty events online.

Research Opportunities Using the Dataset

The BitcoinTalk forum has a long history and has had significant influence in the cryptocurrency space. As the largest subforum, and one that encourages users to engage with social media, create content and participate in discussions, the Bounties (Altcoins) dataset can provide useful insights to researchers of various interests. Here we discuss some unique research opportunities arising from our dataset.

Financial analysis Market price manipulation has long been of interest to researchers, both in the stock market and in the crypto space (e.g., Gandal et al. 2021). Further studies can use our dataset to study the effect of coordinated social media campaigns (Schou et al. 2022) by examining the price movement of the tokens included in our dataset. This can show whether the bounty campaigns have been successful and whether they were intended to con investors out of their investments by inflating the price and then "dumping" the token (Mirtaheri et al. 2021; Nizzoli et al. 2020). Further research can also attempt to correlate the activity in the forum or the prices of the promoted tokens with the movements of the wider economy, Bitcoin or other cryptoassets (Gandal et al. 2021) to examine how these affect online campaigns.
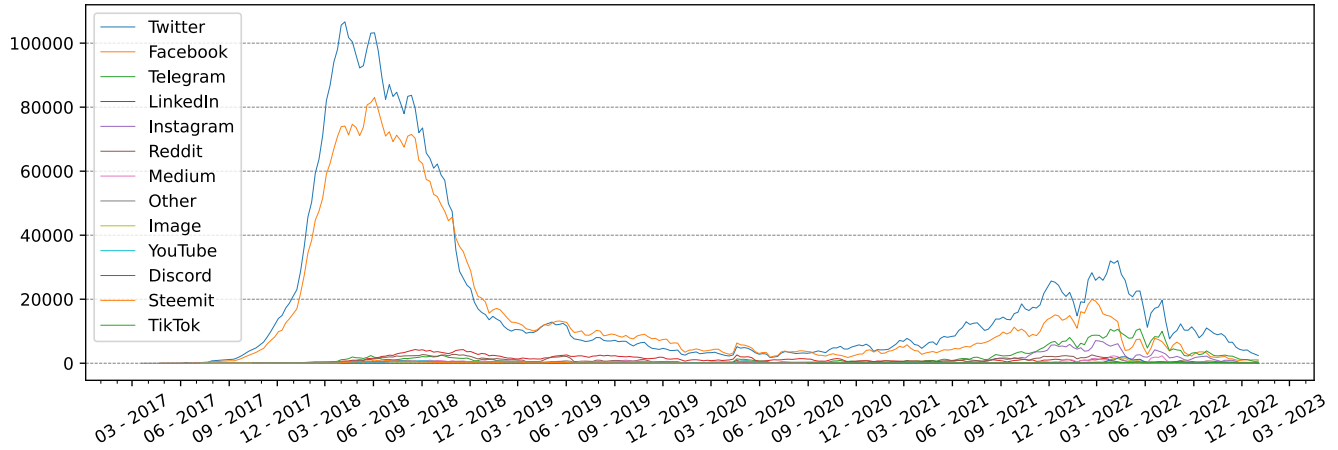
Figure 5: Number of unique participations (multiple participation comments from the same user in the same thread are counted as one) for different campaign types over time. The legend is sorted from most to least popular campaigns, top to bottom.
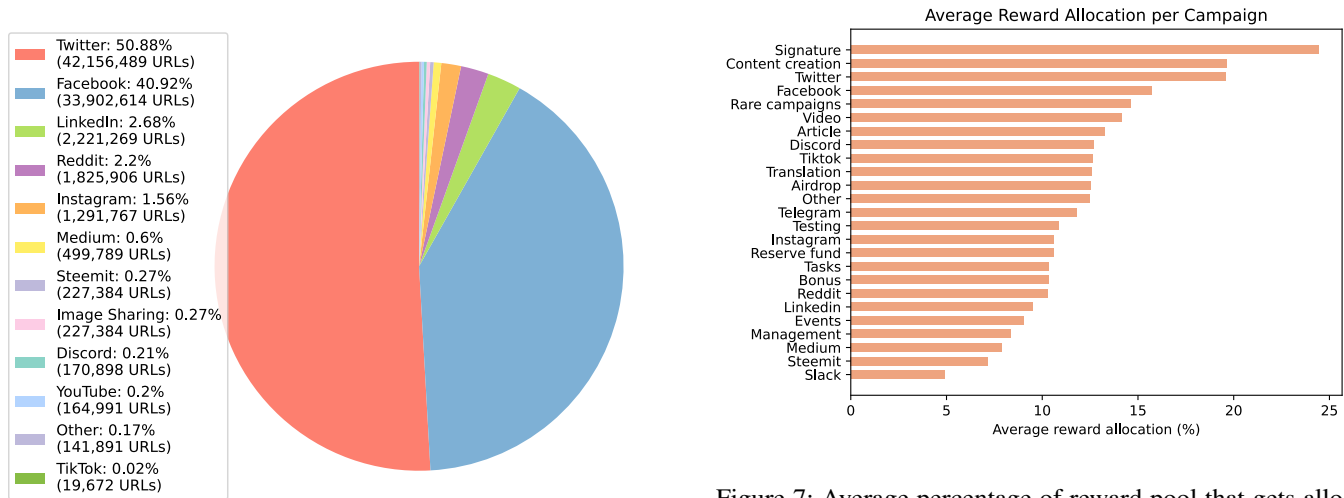

Figure 6: Number of links collected from proof of participation comments for different types of campaigns.

Figure 7: Average percentage of reward pool that gets allocated to different campaign types in Bounty events.

Social media and network analysis A lot of research has examined how and why people use multiple media (Spiliotopoulos and Oakley 2020; Alhabash and Ma 2017). Our dataset, which includes the details of multiple social media accounts for each user, can be valuable for social media researchers studying cross-media user posting practices. Researchers can also follow the social media links from the dataset to gain a better understanding of the effect of these campaigns. For example, how much of the discussion around a crypto project is organic and how much is driven by bounties? How many followers of a crypto project are genuine and how many are bounty hunters? Social network researchers can use the bounty hunter profiles provided in the dataset to determine details about the networks (Himelboim et al. 2017) of bounty hunters; How many are connected and follow each other within and across media? What cliques are they part of? Are their followers genuine or are they just following one another and effectively promoting the project to the converted?

Natural Language Processing Researchers can follow the social media links and download text content from the Tweets, Facebook posts, Reddit discussions and blog posts to study the specific language used by bounty hunters. Natural Language Processing (NLP) techniques (Hirschberg and Manning 2015) can identify the topics discussed and the linguistic characteristics of this content, in order to detect it and protect social media users. Sentiment analysis techniques (Valle-Cruz et al. 2022) can be used to quantify the effect of changes in online sentiment created by bounty campaigns on the crypto market (Xie, Chen, and Hu 2020).

Policy and regulation   Legal scholars and policy makers can use the dataset to reach a more nuanced understanding of the crypto space and the relationship between social media or online communities and cryptocurrency investing. Examination of the structure and organization of Bounty events and the practices of crypto bounty hunters will inform the evaluation of the current regulatory state and protect naive investors. This is particularly important as the calls for regulation of the crypto space become increasingly insistent (Henderson and Raskin 2019).

Broader Impact of the Work and Ethical Considerations

Overall, there are numerous benefits that can be derived from our dataset. We bring awareness to a large number of Bounty events and campaigns that can potentially lead inexperienced crypto investors to make poor financial decisions. Our dataset can be a useful resource for researchers that can study the details and structure of the campaigns, gain insights into how the cryptocurrency industry operates and how it interacts with its audience, and develop ways to protect users and inform policy makers.

Of course, these benefits need to be considered together with the potential risks of collecting the data and sharing the dataset, especially risks related to unanticipated secondary use (Salganik 2019). With regards to the data collection stage, while our dataset includes information and links from a range of social media and other online platforms, we only collected data from the BitcoinTalk forum. This data collection is in line with the Terms and Conditions of the forum, and, in fact, the forum specifically encourages data scraping from its boards[14]. The collected raw data were held in secure password-protected devices and cloud accounts.

A cross-media user dataset presents additional potential misuses compared to a dataset from a single social media platform, such as more extensive profiling and tracking, cyberstalking, and identity theft. In order to minimize the risks associated with sharing the dataset, we removed email addresses. However, we decided to keep other account information, such as social media account IDs and crypto wallet addresses. It was clear to us that these accounts were created for the purpose of participating in Bounty events and there is minimal overlap with forum users' personal social media accounts or other personal information. The removal of personal information and the use of non-personal (or "throwaway") social media accounts by the forum users keeps the risks associated with sharing the dataset relatively low. It is possible that a forum user in the future may decide to delete their account from the forum and their account information will remain in our dataset, but we consider that to be in line with the reasonable privacy expectations of public forum users, especially since a significant amount of content scraping takes place by other forum users in public[15]. We do not foresee that the release of our dataset can put any of the forum users in any additional danger or risk, as the activities taking place during these Bounty events are currently within

the law. Furthermore, while we did not employ any social media APIs to collect the data and, thus, we are not technically subject to any of their Terms and Conditions, we still decided to not share social media content (e.g., tweets) and to only share social media URLs and Tweet IDs. Finally, this work received ethical approval from our institution.

Compliance with FAIR principles   Our dataset adheres to the four FAIR data principles[16]. Our dataset is *Findable* as it is assigned a unique and persistent Digital Object Identifier (DOI): 10.5281/zenodo.7539178. The dataset is indexed in a searchable resource and contains rich metadata that describe it. The data and metadata are retrievable by their identifier in an open and free manner, making our dataset *Accessible*. The dataset is released in tsv format, making it *Interoperable*. It is extensively documented with metadata describing every data field in a text file that accompanies the data. We also provide further details of the regular expressions used for some of the data processing we conducted in a supplementary file and we also share files with the raw collected data in order to enable unanticipated future uses. This makes our dataset *Re-usable*.

Acknowledgments

References

Abramova, S.; Voskobojnikov, A.; Beznosov, K.; and Böhme, R. 2021. Bits Under the Mattress: Understanding Different Risk Perceptions and Security Behaviors of Crypto-Asset Users. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19.

Alhabash, S.; and Ma, M. 2017. A Tale of Four Platforms: Motivations and Uses of Facebook, Twitter, Instagram, and Snapchat Among College Students? *Social Media + Society*, 3.

Ante, L.; Sandner, P.; and Fiedler, I. 2018. Blockchain-Based ICOs: Pure Hype or the Dawn of a New Era of Startup Financing. *Journal of Risk and Financial Management*, 11(4): 80.

Benevenuto, F.; Magno, G.; Rodrigues, T.; and Almeida, V. 2010. Detecting spammers on twitter. In *CEAS 2010*, 12.

Bollen, J.; Mao, H.; and Zeng, X. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1): 1–8.

Brundage, M.; Avin, S.; Wang, J.; Belfield, H.; Krueger, G.; Hadfield, G.; Khlaaf, H.; Yang, J.; Toner, H.; and Fong, R. 2020. Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv:2004.07213*.

Ding, A. Y.; De Jesus, G. L.; and Janssen, M. 2019. Ethical hacking for boosting IoT vulnerability management. In *ICTRS '19*. New York, New York, USA: ACM Press.

---

[14]https://bitcointalk.org/index.php?topic=5208423.0

[15]see previous note

[16]https://force11.org/info/the-fair-data-principles/

Domingo, R.-S.; Piñeiro-Chousa, J.; and Ángeles López-Cabarcos, M. 2020. What factors drive returns on initial coin offerings. *Technological Forecasting and Social Change*, 153: 119915.

Dutta, H. S.; and Chakraborty, T. 2021. Blackmarket-Driven Collusion on Online Media: A Survey. *ACM/IMS Transactions on Data Science*, 2(4): 1–37.

Gandal, N.; Hamrick, J. T.; Moore, T.; and Vasek, M. 2021. The rise and fall of cryptocurrency coins and tokens. *Decisions in Economics and Finance*, 44(2): 981–1014.

Garcia, D.; and Schweitzer, F. 2015. Social signals and algorithmic trading of Bitcoin. *Royal Society Open Science*, 2(9): 150288.

Glenski, M.; Saldanha, E.; and Volkova, S. 2019. Characterizing speed and scale of cryptocurrency discussion spread on reddit. *The Web Conference, WWW 2019*, 560–570.

Hamrick, J.; Rouhi, F.; Mukherjee, A.; Feder, A.; Gandal, N.; Moore, T.; and Vasek, M. 2021. An examination of the cryptocurrency pump-and-dump ecosystem. *Information Processing & Management*, 58(4): 102506.

Henderson, M. T.; and Raskin, M. 2019. A regulatory classification of digital assets: toward an operational Howey test for cryptocurrencies, ICOs, and other digital assets. *Columbia Business Law Review*, 443.

Himelboim, I.; Smith, M. A.; Rainie, L.; Shneiderman, B.; and Espina, C. 2017. Classifying Twitter Topic-Networks Using Social Network Analysis. *Social Media + Society*, 3(1): 205630511769154.

Hirschberg, J.; and Manning, C. D. 2015. Advances in natural language processing. *Science*, 349(6245): 261–266.

Jahani, E.; Krafft, P. M.; Suhara, Y.; Moro, E.; and Pentland, A. S. 2018. ScamCoins, S*** Posters, and the Search for the Next Bitcoin. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 1–28.

Karim, S.; Lucey, B. M.; Naeem, M. A.; and Uddin, G. S. 2022. Examining the interrelatedness of NFTs, DeFi tokens and cryptocurrencies. *Finance Research Letters*, 47: 102696.

Kauppila, S.; and Soliman, W. 2022. "This Is the Future of Advertising!" Or Is It? New Insights into the Justifiability of Deceptive Crowdwork in Cyberspace. In *HICSS 2022*.

Keller, F. B.; Schoch, D.; Stier, S.; and Yang, J. 2020. Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign. *Political Communication*, 37(2): 256–280.

Lee, K.; Tamilarasan, P.; and Caverlee, J. 2013. Crowdturfers, campaigns, and social media: Tracking and revealing crowdsourced manipulation of social media. *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, 331–340.

Mathur, A.; Narayanan, A.; and Chetty, M. 2018. Endorsements on Social Media: An Empirical Study of Affiliate Marketing Disclosures on YouTube and Pinterest. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 1–26.

Mirtaheri, M.; Abu-El-Haija, S.; Morstatter, F.; Steeg, G. V.; and Galstyan, A. 2021. Identifying and Analyzing Cryptocurrency Manipulations in Social Media. *IEEE Transactions on Computational Social Systems*, 8: 607–617.

Nizzoli, L.; Tardelli, S.; Avvenuti, M.; Cresci, S.; Tesconi, M.; and Ferrara, E. 2020. Charting the Landscape of Online Cryptocurrency Manipulation. *IEEE Access*, 8: 113230–113245.

Parnin, C.; Treude, C.; Grammel, L.; and Storey, M.-A. 2012. Crowd documentation: Exploring the coverage and the dynamics of API discussions on Stack Overflow. In *Georgia Tech Technical Report GIT-CS-12-05*.

Peng, J.; Detchon, S.; Choo, K.-K. R.; and Ashman, H. 2017. Astroturfing detection in social media: a binary n-gram-based approach. *Concurrency and Computation: Practice and Experience*, 29(17): e4013.

Salganik, M. J. 2019. *Bit by bit: Social research in the digital age*. Princeton University Press.

Schauster, E. E.; Ferrucci, P.; and Neill, M. S. 2016. Native Advertising Is the New Journalism. *American Behavioral Scientist*, 60(12): 1408–1424.

Schoch, D.; Keller, F.; Stier, S.; and Yang, J. 2022. Coordination patterns reveal online political astroturfing across the world. *Scientific Reports*, 12(1): 4572.

Schou, P.; Bucher, E.; Waldkirch, M.; and Grünwald, E. 2022. We Did Start the Fire: r/wallstreetbets, 'Flash movements' and the Gamestop Short-Squeeze. In *HICSS 2022*.

Spiliotopoulos, T.; and Oakley, I. 2020. An exploration of motives and behavior across Facebook and Twitter. *Journal of Systems and Information Technology*, 22: 201–222.

Valle-Cruz, D.; Fernandez-Cortez, V.; López-Chau, A.; and Sandoval-Almazán, R. 2022. Does Twitter Affect Stock Market Decisions? Financial Sentiment Analysis During Pandemics: A Comparative Study of the H1N1 and the COVID-19 Periods. *Cognitive Computing*, 14(1): 372–387.

Vasek, M.; and Moore, T. 2019. Analyzing the Bitcoin Ponzi Scheme Ecosystem. In *Financial Cryptography and Data Security: Lecture Notes in Computer Science*, 101–112. Berlin, Heidelberg: Springer Berlin Heidelberg.

Xie, P.; Chen, H.; and Hu, Y. J. 2020. Signal or Noise in Social Media Discussions: The Role of Network Cohesion in Predicting the Bitcoin Market. *Journal of Management Information Systems*, 37(4): 933–956.

Yardi, S.; Romero, D.; Schoenebeck, G.; and Boyd, D. 2009. Detecting spam in a Twitter network. *First Monday*.

Zeng, E.; Kohno, T.; and Roesner, F. 2021. What Makes a "Bad" Ad? User Perceptions of Problematic Online Advertising. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM.