# Evolving Bots: The New Generation of Comment Bots and their Underlying Scam Campaigns in YouTube

Seung Ho Na
harry.na@kaist.ac.kr
KAIST
Daejeon, Republic of Korea

Sumin Cho
jeevas@kaist.ac.kr
KAIST
Daejeon, Republic of Korea

Seungwon Shin
claude@kaist.ac.kr
KAIST
Daejeon, Republic of Korea

## ABSTRACT

This paper presents a pioneering investigation into a novel form of scam advertising method on YouTube, termed "social scam bots" (SSBs). These bots have evolved to emulate benign user behavior by posting comments and engaging with other users, oftentimes appearing prominently among the top rated comments. We analyzed the YouTube video comments and proposed a method to identify SSBs and extract the underlying scam domains. Our study revealed 1,134 SSBs promoting 72 scam campaigns responsible for infecting 31.73% of crawled videos. Further investigation revealed that SSBs exhibit advances that surpass traditional bots. Notably, they targeted specific audience by aligning scam campaigns with related video content, effectively leveraging the YouTube recommendation algorithm. We monitored these SSBs over a period of six months, enabling us to evaluate the effectiveness of YouTube's mitigation efforts. We also uncovered various strategies they use to evade mitigation attempts, including a novel strategy called "self-engagement," aimed at boosting their comment ranking. By shedding light on the phenomenon of SSBs and their evolving tactics, our study aims to raise awareness and contribute to the prevention of these malicious actors, ultimately fostering a safer online platform.

## CCS CONCEPTS

• **Information systems** → **World Wide Web**.

## KEYWORDS

Web, YouTube, Bot, Scam, Measurement

**ACM Reference Format:**
Seung Ho Na, Sumin Cho, and Seungwon Shin. 2023. Evolving Bots: The New Generation of Comment Bots and their Underlying Scam Campaigns in YouTube. In *Proceedings of the 2023 ACM Internet Measurement Conference (IMC '23), October 24–26, 2023, Montreal, QC, Canada.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3618257.3624822

## 1 INTRODUCTION

The proliferation of bots has posed significant challenges to online social networks (OSNs). As OSNs have become an integral part of daily life, malicious bots in social media have become elevated threats to the general public. They have demonstrated their capacity to disseminate disinformation while under the radar of bot detection efforts: evidence shows of their attempt in spamming propaganda during political events such as the 2016 US election and the "Brexit" referendum [18, 20]. Consequently, there has been a pressing demand for bot detection methods to counteract these malicious activities, only to be faced with *social bots*.

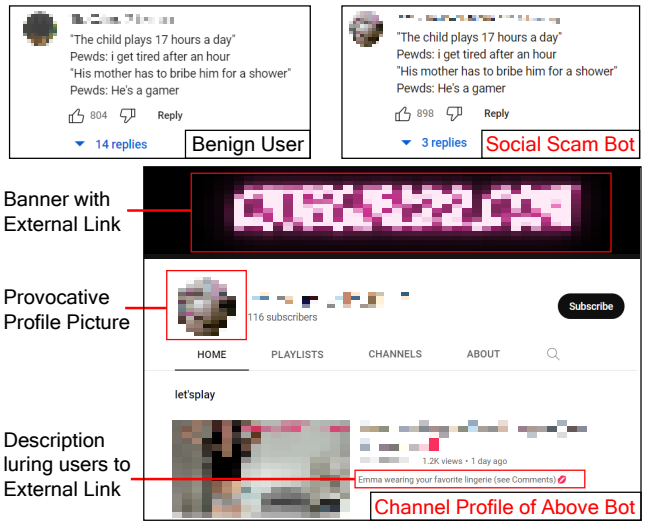



**Figure 1: Example of a social scam bot (SSB) in YouTube. SSBs imitate benign users (top of figure) but lure users to scam campaigns in their personal channel (bottom of figure).**

Social bots are bots with the ability to emulate normal user behavior to a certain extent, resulting in their assimilation into the communities of benign users [17]. By infiltrating into the feed of benign users based on their legitimate appearance, social bots directly expose benign users to malicious posts (e.g. waver opinions with disinformation). Given the severity of social bots, many studies were conducted to understand social bot behavior [26, 32, 38] and propose detection methods of these automated entities [9, 12, 19]. In general, previous studies have predominantly focused on text-based platforms such as Facebook, Twitter, and Reddit, where textual content serves as the primary source of information.

YouTube is an OSN for video content. As the OSN that accounts for 25% of global mobile traffic, it is the world's largest video sharing platform [1]. YouTube is also harmed by its own variety of exploitative behavior from social bots. An example of a social bot is shown in Figure 1. These social bots are different from basic spam bots in that they operate by *socializing* and exposing themselves to


*IMC '23, October 24–26, 2023, Montreal, QC, Canada*

ACM ISBN 979-8-4007-0382-9/23/10…$15.00
https://doi.org/10.1145/3618257.3624822

benign users through comment interactions. Their comments are well-structured and relatable; these comments are on-topic with the video and garner likes and responses from benign users. But when examining the personal channel, users are enticed with sentences to external links that redirect them away from YouTube. These links have been examined to be scam sites that deceive users into entering their personal information as well as personal financial information (Section 4).

In this work, we discover and analyze the social scam bots (SSBs) of YouTube. We propose a scanning method for SSBs utilizing YouTuBERT, our language model pretrained on YouTube comment data, and DBSCAN [15], and from these SSBs expose the scam campaigns that are behind these malicious entities. Our method of scanning for SSBs is efficient and minimizes channel profile visits by a crawler. From the collected SSBs and their scam campaigns, we conducted a comprehensive measurement study to grasp the impact of SSBs and characterize their features. We discover the advanced characteristics of SSBs in comparison to traditional spam bots, and evaluate YouTube's mitigation attempts of these SSBs. Further case studies expose strategies of these scam campaigns to boost exposure and evade termination.

Our contributions are as follows:

- We are the first work to systematically measure the scale of social scam bots in YouTube and expose its potential threat to YouTube users.
- We propose an effective process of distinguishing SSBs and discovering their scam campaign affiliations from comment data. Our work adheres to ethical considerations by reducing the number of visited YouTube channel profiles to 2.46% of total commenting users.
- We propose YouTuBERT, a large language model (LLM) pretrained on YouTube comments. When identifying bot candidates, utilizing YouTuBERT embeddings show the most robust performance with $\epsilon$ values.
- From our measurement study, we identify 72 existing scam campaigns, consisting of 1,134 SSBs and study their impact on YouTube. Furthermore, we report SSB behavior deduced from statistical evidence to help characterize their preference and assess the YouTube's own mitigation efficacy.
- We identified the strategies employed by SSBs, including self-engagement, a tactic that leverages the YouTube algorithm and volume of SSB for more visibility and exposure of the SSB comments.

## 2 RESEARCH QUESTION AND MOTIVATION

Social bots, as described by previous literature, are bots that emulate normal user behavior that allows their assimilation into benign user communities [17]. YouTube is a platform where the only user interaction is through commenting and replying on videos. Furthermore, as a viewer, profile characterization (e.g., decorating one's own profile, interaction history, sharing content) is not a main purpose of this OSN, which mean profiles lack information. The only direct social cues are comments with other users in videos. We define the bots described in Figure 1 as *social scam bots*.

*Definition 2.1 (Social Scam Bot (SSB)).* An SSB is a malicious bot that advertises and promotes a scam campaign in its profile page.

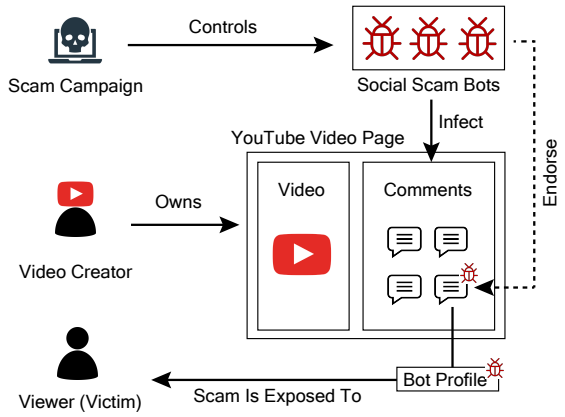



**Figure 2: Concept flow of related entities of the SSB phenomena.**

They mimic humans to operate socially in the comments section of YouTube videos, oftentimes gathering many likes and replies from other users.

SSBs can be classified as a type of social bot. They are able to gather attention from benign users by engaging in comments that semantically match the video context, much more advanced than primitive spamming bots [6]. While traditional spamming bots rely on indiscriminate and repetitive posting of irrelevant or promotional content [6], SSBs are adept at generating comments that semantically align with the context of the video. This is done by basing their comments on other benign user comments. Figure 2 shows a concept diagram of how scam campaigns expose themselves to viewers. Scam campaigns employ a group of SSBs that infect YouTube videos owned by creators. SSBs comment on the video section and these seemingly innocent comments are visible to viewers that happen to read comments. By clicking on the user profile or channel of the comment, the victims are exposed to external links to scam websites. On some occasions, SSBs endorse other comments by replying to comments made by other SSBs.

However subtle their process is, they are an immediate threat to the platform and its viewers. To understand these SSBs and their underlying scam campaigns, we focus on the following research questions:

*R1)* How can we discover and determine SSBs?
*R2)* How much impact do the SSBs have on YouTube and its viewers?
*R3)* How effective are the mitigations by YouTube?
*R4)* Does this phenomenon exhibit strategic characteristics?

These questions will be answered in the following areas of this work: *R1* is answered in Section 4 by explaining our method of finding and extracting SSBs from our collected dataset of YouTube comments, *R2* is answered through statistics and findings of the SSBs reported in Section 4 and Section 5.1, *R3* is answered by comparing the impact of active and terminated SSBs in Section 5.2, and *R4* is answered by our use case study of the strategies witnessed among scam campaigns in Section 6.

# 3 RELATED WORK

## 3.1 Social Bots in Social Network Platforms

Detection and blocking of social bots on OSN has been a significant challenge. Research has been conducted on various platforms, including Twitter [4, 5, 13, 21–23, 28–30, 33], Facebook [5, 39], WeChat [40], and Reddit [7].

Unlike other platforms, there is limited information available to be disclosed, making it difficult to find meaningful correlations on YouTube. Literature on Facebook and Wechat used account characterization information to detect malicious bots. Xu et al. [39] proposed a method for detecting malicious Facebook accounts by analyzing behavior patterns and information of accounts using entity classification. Similarly, Yuan et al. [40] proposed a method for detecting social bots on WeChat by analyzing the provided sign-up information. These studies were able to detect malicious accounts through account information and social graph analysis, and found correlations between malicious accounts.

Previous bot detection efforts shared valuable insight concerning characterization of social bots, but they were specific to their own domain and not applicable. Some have conducted detailed analyses of large-scale social botnets on Twitter [4] and have examined the difference between social bots and humans [28], utilizing data sets limited to Twitter and its unique 'Retweet' function, making it difficult to apply the results to other platforms. Ahmed et al. [5] analyzed Twitter and Facebook features, using statistical results to identify spam campaigns. Through the use of classical machine learning techniques, they identified seven characteristics that distinguished spam accounts from normal accounts. However, this method cannot be used for YouTube, as it takes advantage of the unique characteristics of Twitter and Facebook.

Some existing studies have found bots by analyzing the message semantics. Heidari et al. [22, 23] collected profile information of Twitter accounts and used a word embedding model to analyze the context of tweets posted by accounts with similar profile information to classify bot and human accounts. Similarly, Benjamin et al. [7] proposed a method for detecting social bots through crowd reactions to messages. These methods have the advantage of being applicable to various platforms, but the SSBs encountered on YouTube copy or base their comments on other benign comments and acquire semantic similarity, making it difficult to use the mentioned methodologies.

Despite the abundance in previous literature on social bots, none have studied the SSBs of YouTube. The structural differences of the platform and functional and behavioral differences of the YouTube SSBs are hurdles in the analysis of this phenomena. This study is the first to systematically analyze SSBs on YouTube.

## 3.2 Abusive Behavior on YouTube

Although not social bots, abusive behavior has also been studied in YouTube. Alberto et al. [6] proposed Tubespam, a method for spam filtering on YouTube by extracting the characteristics of comments used on the platform. Tubespam filters spam comments by checking for the presence of links or specific keywords in the comments. The spam filters are ineffective to the social bots of YouTube because they do not rely on specific keywords or links; they exhibit social skills of a normal user.

Previous studies examining video spam on YouTube [10, 11, 37, 41] have discussed the abundance of clickbait spam, where misleading or sensationalized titles and thumbnails are used to entice users to click on a video. Bouma-Sims et al. [8] conducted a study on scam videos on YouTube. They collected videos using specific search queries associated with potentially fraudulent content, identified fraudulent videos among the collected videos, and analyzed them. They found that many scams on YouTube lure users off the site in order to earn money, and classified the types of scams based on the content of the scam videos.

Although sharing the same platform as the mentioned literature, the SSBs that we analyze are a different variant of abusive behavior that has not yet been investigated. In this work, we conduct an extensive analysis on these social bots and uncover their underlying scam campaign operations.

# 4 COLLECTING SSB ACCOUNTS

In this section we describe our procedure of collecting and compiling our real world dataset of SSBs in YouTube. These SSBs from the wild are then analyzed to extract their scam campaign affiliations. The full workflow of extracting scam campaigns from the collected video comments is shown in Figure 3.

## 4.1 Data Crawling

The starting point of data collection is gathering data from the area where SSB operate: the comment section of YouTube videos. We adopted the top 1,000 United States YouTube creator list from HypeAuditor[1], an influencer marketing platform, to target an English speaking audience in the comment section on August, 2022. The platform offers channel-related information such as subscriber count, average views, and average comment count. From these channels, we selected the 50 most recent videos of each channel, and crawled up to 1,000 comments from each video (some videos have less than 1,000 comments). In addition, YouTube allows replying to comments, so these replies were crawled as well. For each set of replies to a single comment, 10 replies maximum were crawled for all comments with replies. We developed the crawler using Selenium[2] to allow browser interaction (e.g., scrolling down to load more comments, opening reply list).

YouTube makes available two comment sorting options: 'top comments' for sorting by YouTube's self-owned comment ranking algorithm, or 'newest first' for sorting in chronological order. We proceeded with the 'top comments' because it is the default comment presentation option. Note that this implies that the collected SSB were able to take advantage of the YouTube algorithm and were granted exposure by the algorithm.

The crawled data are summarized in Table 1. Of the 1,000 YouTube creators, 30 creators had their comments disabled due to YouTube's policy of improving child safety [36]. The videos of these creators, along with videos that had its comments removed purposely by the creator, contributed to a total of 4,678 comment-less videos. From the remaining videos, more than 22 million comments were collected from more than 12 million commenters.

[1] https://hypeauditor.com/
[2] https://www.selenium.dev/

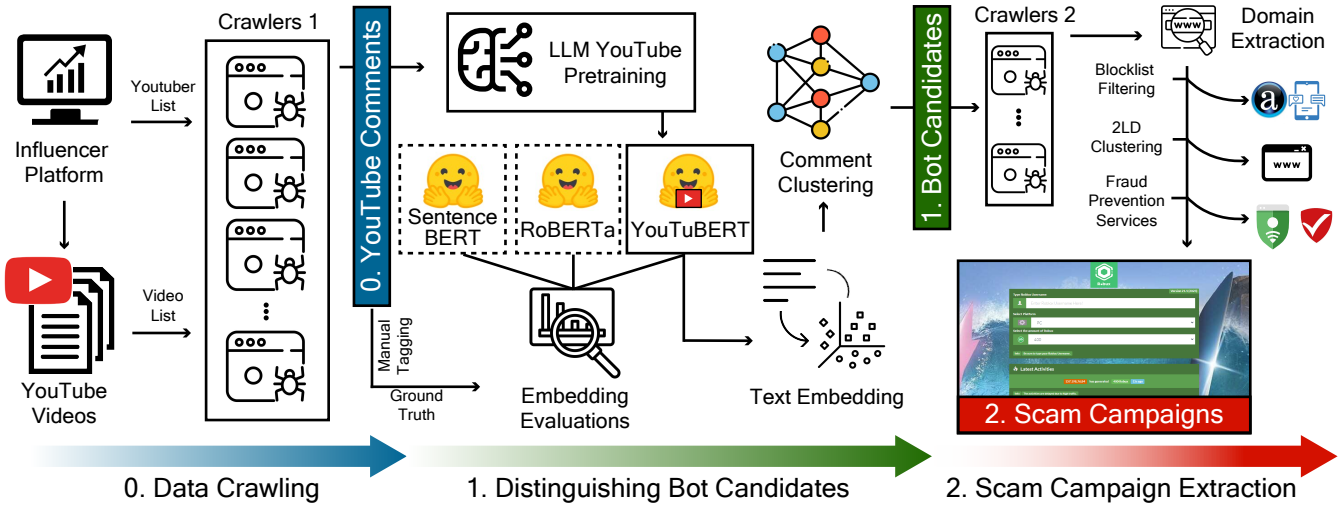



**Figure 3: Workflow to extract SSBs and their scam campaigns. The output is colored to match its respective phase.**

**Table 1: Dataset summaries.**

| | Full Dataset | Ground Truth |
|---|---|---|
| # of seed YouTube creators | 1,000 | - |
| # of crawled videos | 45,322 | - |
| # of total comments | 22,542,786 | 24,706 |
| # of total commenters | 12,517,762 | 22,232 |
| # of clusters (TF-IDF, $\epsilon = 1.0$) | 542,915 | 5,400 |
| # of clusters (YouTuBERT, $\epsilon = 0.5$) | 169,848 | - |
| # of verified SSBs | 1,134 | 333 |

## 4.2 Distinguishing Bot Candidates

As shown in Figure 1, SSBs can be distinguished specifically by inspection of the users' profile page. However, given the substantial size of our dataset (over 12 million unique commenters), conducting a comprehensive inspection of each individual user becomes impractical, necessitating the need for a reduction in the volume of inspections. To properly identify SSBs from the mass of YouTube comments, we take advantage of the observation that SSBs generally imitate other comments on the video (some would copy other comments while others modify the comment without changing its original context [34]). By employing a filtering mechanism to retain comments exhibiting similar semantics, we can effectively discern *bot candidates* with the potential of being SSBs. Note that we use the term bot candidate instead of SSB to recognize the presence of original comments by benign users during this filtering process. This crucial step encompasses the conversion of comment text into sentence embeddings, which encapsulate the semantic distance between sentences.

There are many open LLMs that can be used for text embedding. There is BERT [14], which leverages a transformer-based architecture to learn contextualized word representations by considering the surrounding context, and RoBERTa [25], an extension and improvement of BERT that opts out of the Next Sentence Prediction (NSP) task in pretraining. Sentence-BERT [31] is a language model specifically tailored to semantic textual similarity tasks in sentence pairs and is able to derive semantically meaningful sentence embeddings. However, the performance of these pretrained models cannot be guaranteed on our dataset of solely YouTube comments. In order to establish domain expertise within the context of YouTube, we undertake a domain-specific pretraining of the RoBERTa model using our dataset. This specialized training enables RoBERTa to acquire a deep understanding of the unique characteristics and linguistic nuances present in YouTube content.

In order to perform a comparative analysis and determine the optimal sentence embedding technique, a ground truth dataset is constructed. The comments associated with each video were initially transformed into TF-IDF vectors (to prevent bias on sentence embedding performance), with the entire collection of comments on the video serving as the corpus for this vectorization process. These vectors were clustered with DBSCAN using a generous value of $\epsilon = 1.0$ to allow easier formation of a cluster by bot candidates. From these 542,915 clusters 1% was randomly sampled to construct the ground truth clusters. Due to the relatively large $\epsilon$ value, clusters were bound to contain many benign comments along with the comments from SSBs. Consequently, the comments of these sampled clusters were tagged as either benign or bot candidate and was cross-validated by three security practitioners achieving an inter-annotator agreement Fleiss' Kappa value of 0.89[3]. The constructed ground truth dataset resulted in 3,464 comments tagged as bot candidates and 21,242 comments as benign.

[3]These practitioners adhered to a prearranged tagging standard that may involve visiting a users' profile page for confirmation. Details on the standard and process of tagging are described in Appendix B.

Three language models were compared on the ground truth dataset: Sentence-BERT, RoBERTa, and YouTuBERT. We chose RoBERTa over BERT because YouTube comments (in the order of top comments) lack chronological consistency; consecutive comments may not be related to each other. This characteristic aligns well with RoBERTa's omission of the NSP task. Sentence-BERT and RoBERTa are the pretrained language models that are made accessible through HuggingFace[4] and will be referred as original models to mitigate confusion with YouTuBERT, which is the RoBERTa based language model with 125M trainable parameters pretrained on our own comment dataset by masked language modeling for 32 hours on an NVIDIA TITAN RTX[5]. We compare performances of these models in regards to our filtering mechanism; if a comment is clustered, it will be considered as a bot candidate. Our main metric of comparison is the F1-score, which encompasses both precision and recall, allowing us to strike a balance between minimizing the quantity of accounts that necessitate further crawling (precision) and effectively identifying the SSBs (recall) from our bot candidates. Table 2 shows the results of each sentence embedding according to the cluster radius used in DBSCAN. Sentence-BERT has a slight advantage in F1-score among the three embedding choices, followed by YouTuBERT and RoBERTa. However, the performance of the original language models are sensitive to the value of $\epsilon$: there is a sharp loss in F1-score between the values $\epsilon = 0.2$ and $\epsilon = 0.5$ for Sentence-BERT and RoBERTa. This makes selecting $\epsilon$ a difficult task because of its unreliability due to the ground truth being only a small portion of the full dataset.

On the other hand, YouTuBERT demonstrates robustness and maintains consistent performance across various $\epsilon$ values. YouTuBERT retains robustness because of its adaptation and ability to learn contextual representations of the YouTube comments. Specifically, the utilization of an embedding pretrained on YouTube comments yields a finer-grained measure of semantic distance among YouTube comments compared to an open-domain embedding. This implies that the YouTuBERT enables a more detailed discernment of nuances and distinctions within the semantic space of YouTube comments than Sentence-BERT or RoBERTa. Therefore, we select YouTuBERT with the optimum F1-score as our sentence embedding using $\epsilon = 0.5$ for DBSCAN to filter our whole dataset for bot candidates. Note that our method of using YouTuBERT serves as a step in narrowing down bot candidates that are suspicious and needed for investigation; the 'precision' value does not pertain to the final performance of the whole workflow. Only the bot candidates that have been verified to promote a scam domain is labeled an SSB.

**Table 2: Performance of sentence embeddings on ground truth datatset of bot candidate. The values highlighted in bold represent the optimal performance for each metric across different embedding methods.**

| Method | $\epsilon$ | Prec. | Recall | Acc. | F1-Score |
|---|---|---|---|---|---|
| Sentence-Bert | 0.02 | **0.6378** | 0.8583 | **0.9118** | 0.7318 |
| | 0.05 | 0.6372 | 0.8606 | 0.9118 | **0.7323** |
| | 0.2 | 0.6126 | 0.9085 | 0.9066 | 0.7318 |
| | 0.5 | 0.2844 | 0.9778 | 0.6520 | 0.4407 |
| | 1.0 | 0.1402 | **1.0000** | 0.1402 | 0.2459 |
| RoBERTa | 0.02 | **0.6452** | 0.7870 | 0.9095 | 0.7091 |
| | 0.05 | 0.6449 | 0.7907 | **0.9096** | **0.7104** |
| | 0.2 | 0.6034 | 0.8265 | 0.8995 | 0.6975 |
| | 0.5 | 0.2189 | 0.9512 | 0.5173 | 0.3559 |
| | 1.0 | 0.1403 | **1.0000** | 0.1408 | 0.2461 |
| YouTuBERT | 0.02 | 0.6454 | 0.7702 | 0.9084 | 0.7023 |
| | 0.05 | **0.6455** | 0.7705 | 0.9085 | 0.7025 |
| | 0.2 | 0.6387 | 0.7771 | 0.9071 | 0.7011 |
| | 0.5 | 0.6369 | 0.8187 | **0.9091** | **0.7164** |
| | 1.0 | 0.5967 | **0.8782** | 0.8997 | 0.7106 |

## 4.3 Scam Campaign Extraction

As can be known by Figure 1, SSBs do not promote their scam addresses directly in their comments. Doing so would be a direct indicator of suspicious behavior considering that external links in the comment are discouraged by YouTube policies [2]. In our direct examination of SSB accounts, we identified five areas on the account channel page where SSB primarily places and advertises external links [6]. Therefore, we developed a second crawler to scrape the external link information from the respective areas. This information was saved only if the content was verified to contain a URL string through regular expression matching by the crawler.

Afterwards, the second-level domains (SLD) of the URL addresses are scanned through a blocklist filter to exclude known (and benign) domains that are commonly shared such as other OSN domains (e.g., Facebook, Twitter, Instagram). Additionally, the alternative domain names corresponding to each OSN domain were encompassed, i.e., Facebook uses both `fb.com` and `facebook.com`. The blocklist was further enriched by filtering the top 1,000 websites from the Alexa Top Sites service. Subsequently, the SLDs that satisfied the filter criteria are subjected to clustering analysis. Clusters exhibiting a size of less than 2 are excluded from the analysis, associating singular presence with personal websites rather than malicious domains propagated by SSBs. The remaining SLDs underwent scrutiny via a select pool of diverse online fraud prevention resources including ScamAdviser, an online scam database with 154K daily users, ScamWatcher, a community website with a database of more than 74K scams, and Google Safe Browsing. From 74 SLDs, a total of 72 SLDs were confirmed to be scam domains[7].

The 72 scam campaigns were categorized into 6 categories, described in Table 3. The scam categories that covered the majority of scam campaigns were romance scams and game voucher scams—87.5% of the scam campaigns pertained to these categories. However, despite the similarity in campaign size, these scam categories have a major difference in video infection performance. Romance scams were able to infect 28.80% of all videos in our dataset, while game voucher scams infected 4.88% of all videos. The rest of the scam categories were not as invasive and each showed up on less than 1% of the crawled videos.

[4]https://huggingface.co/
[5]Details on training YouTuBERT are included in Appendix C.
[6]Description of these identified areas can be found in Appendix D
[7]The full scam domain list can be found in Appendix E

**Table 3: Scam domain categories. Asterisks imply double counts, where some SSBs contained multiple scam domains.**

| Scam Category | Description | # of Campaigns | SSB Count | Infected Video Count (%) |
|---|---|---|---|---|
| Romance | Scams disguised as escort/dating services to lure victims of their personal/financial information. | 34 | 566 | 13,051 (28.80%) |
| Game Voucher | Scams targeting gamers for their personal information and game credentials in exchange for game currency coupons. | 29 | 444 | 2,212 (4.88%) |
| E-commerce | Scams baiting victims with market products at a highly discounted price for personal/financial information. | 3 | 15 | 97 (0.21%) |
| Malvertising | Fake ads used for phishing victims into downloading malicious software. | 1 | 6 | 61 (0.13%) |
| Miscellaneous | - | 4 | 15 | 234 (0.52%) |
| Deleted | Domains that were suspended by URL shortening services after user reports of malicious behavior. | 1 | 93 | 447 (0.99%) |
| Total | - | 72 | 1,139* | 16,102* (35.53%) |

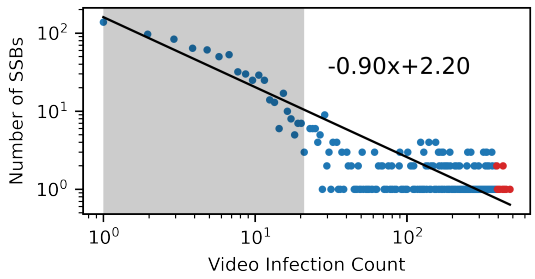



**Figure 4: Histogram scatter plot of number of SSBs and videos infected in log-log scale. The plot follows the power law distribution, where the shaded region marks the area of 80% of the SSBS. The instances in red show higher infection counts than SSBs in shaded region.**

The accounts that contained a scam URL address in their channel account page are SSBs in our dataset. From our dataset of 45,322 videos 14,380 (31.73%) were infected by one or more SSBs, where we were able to obtain 1,134 active SSB accounts. Figure 4 shows the histogram scatter plot of SSBs and their video infection count. By following the power law, the vast majority of SSBs were conservative in their video infection: 50% of SSBs infected less than 7 videos. On the other hand, bots from the long tail of the graph were highly active. The top 18 most active SSBs (1.57% of total SSBs) were responsible for more video infections than the lower 75% of SSBs (top two most active SSBs infected more than lower 50% of SSBs). The SSB causing the most video infections was observed to have commented on 479 videos, which is 1.1% of the total crawled videos.

It is important to note that our approach does not constitute a detection framework, as its methodology is accompanied by a risk of false negatives, meaning that not all SSBs and scam domains from our dataset will be captured. This limitation is evident in the recall results displayed in Table 2. Our methodology primarily aims to identify SSBs while concurrently minimizing the number of accounts requiring scrutiny (details in Appendix A Ethics). Therefore, it is crucial to emphasize that our work serves as a conservative estimate, representing a lower bound of the overall impact of SSBs and their underlying scam domains in the YouTube environment.

## 5 FINDINGS ON NATURE OF SSBS

### 5.1 Target Standards of SSB

One aspect towards understanding this SSB phenomena is studying their targets. The targets of SSBs are at three scales: the creator, the video, and the comment.

**Creators with more subscribers and comments are targeted.** Table 4 shows the linear regression results using the ordinary least squares model (Equation 1) on the number of bot infections by creator features.

$$Count_{SSB} = a * C_{Sub} + b * V_{avg} + c * L_{avg} + d * C_{avg} + Const \quad (1)$$

Of the listed features, 'number of subscribers' and 'average number of comments' were the features that showed rejection of the null hypothesis and therefore statistically significant ($p$-value of less than 0.001). Note that this is a much stricter bound than the common $p = 0.05$, which would have implied all variables of Table 4 to be statistically significant. We use stricter conditions to increase the confidence level, and hence reproducibility, of our findings. Coefficients for both of these features denoted a positive correlation with number of SSBs. Specifically, with each 1 million subscribers, the number of bot infections of a YouTube creator increased by 0.8, and with each 1 thousand comments, the number of bot infections increased by 3. However, the R-squared value was 0.081 due to the data points being noisy with high-variability. By having more subscribers and average comments, a creator is statistically supported to have more SSB attacks on their channel.

**Table 4: Regression results. Features *subscribers* and *avg. comments* have a p-value of less than 0.001, denoting positive correlation to the number of SSB attacks in regards to statistical significance.**

| | Coef. | std. err | p |
|---|---|---|---|
| Constant | 28.75 | 2.949 | < 0.001 |
| # of Subscribers ($C_{Sub}$) | 8.56e-07 | 2.43e-07 | < 0.001 |
| Avg. Views ($V_{avg}$) | 5.32e-06 | 1.84e-06 | 0.004 |
| Avg. Likes ($L_{avg}$) | -0.0001 | 4.1e-05 | 0.001 |
| Avg. Com. ($C_{avg}$) | 0.0030 | 0.001 | < 0.001 |

**Videos within the gaming category exhibit a higher propensity for being targeted.** Videos were labeled into 23 different categories[8] using the information from HypeAuditor. Video categories were multilabels, in which we applied multiple categorical variable regressions of video category to infected number of SSBs. From the 23 categories, only one category 'video games' was statistically significant. In other words, our findings indicate that there is no significant correlation between the video category and the number of SSBs, except in the case of the 'video games' category. Note that the game voucher scams are the second largest scam category that we identified in Table 3. Because these scams are specifically targeting the gaming community, these scams are more likely to appear in videos of the gaming category. Table 5 lists the distribution of video categories that game voucher scams have commented on. In total approximately 93.76% of the videos were of the category 'video games,' 'animation,' and 'humor.' Notably, it is intriguing that these particular video categories predominantly attract an audience that overlaps with the demographic targeted by game voucher scams. Similarly, categories such as 'news & politics' and 'education' are typically perceived as having lesser appeal to the aforementioned audience. Consequently, these categories exhibit relatively lower levels of engagement in terms of comments from game voucher scams.

Even when examining the overall distribution of scam domain categories that have infected 'video game' videos, we observe a consistent pattern. Among the total scam domains infecting 'video game' videos, 88.84% are romance scams, while 10.19% are game voucher scams[9]. Although the percentage of game voucher scams may appear relatively small, this value is approximately 5.76 times higher than the average percentage of game voucher scams across all categories. This disparity suggests that game voucher scams exhibit a higher degree of focus and specialization when targeting 'video game' videos. Consequently, these scams pose a particular risk to a younger audience engaged in the gaming community.

**SSB show preference to the more recent and highly liked top comments.** From the 169K clusters formed from DBSCAN, a total of 44,207 clusters were valid groups with an original comment and one or more SSBs. 1,300 (2.9%) clusters were invalid, being composed of only SSBs and missing an original comment. This means that a top 1,000 comment posted by a benign user was used in 97.1% of these clusters found in our dataset. From these clusters, the original comments had an average of 707 likes, whereas comments by SSBs had an average of only 27 likes, despite their semantic similarity. Also, SSBs would select comments that are relatively 18.4 times more liked than the average comment in the comments section. In the temporal aspect, SSBs selected comments that have been on average 1.82 days since posted. SSBs select comments that have already been chosen by the YouTube algorithm to appear in the 'top comments', which are generally comments that have been allowed the time for sufficient engagement. We also studied the order (indexed from 1 to 1,000) in which the comment appeared to compare the outcome of the original comment and SSB comments. We found that 44.6% (20,220 cases) of original comments copied by SSBs had an index of 20 or less, which is the number of comments that can be viewed on a PC without scrolling (first default batch of comments). This indicates that SSBs tend to base their comments on popular comments with higher visibility in order to increase their own exposure. As a result, in 21.2% (9,650 cases) of videos, comments written by SSBs were found to be displayed higher than the original comments they copied. In these cases, the SSB comments scored higher on the recommendation algorithm of YouTube than the original comments, suggesting SSBs to have taken advantage of the algorithm. Additionally, comments written by SSBs had an index of 20 or less in 8.2%(3,720 cases) of videos, meaning that SSBs were easily found in the first default batch of comments that the video loads for the viewer in 8.2% of the videos we checked.

**Table 5: Video categories that game voucher scams have commented on. The three primary categories encompassing a significant portion of the infected videos predominantly focus on engaging a younger audience.**

| Category | # of Videos | Percentage (%) |
|---|---|---|
| Video games | 3,522 | 59.44 |
| Animation | 1,480 | 24.98 |
| Humor | 553 | 9.33 |
| Others | 370 | 6.24 |
| *News & Politics* | 2 | 0.03 |
| *Fashion* | 1 | 0.02 |
| *Education* | 0 | 0.00 |
| *Remaining Others* | 363 | 6.19 |
| Total | 5,925 | 100.0 |

**The majority of SSBs have placed comments in the first default batch.** The SSB comment counts that were placed within the top 100 are shown in Figure 5. Top 100 imply the comments that are visible on the comment section within 4 reloads of comment batches. This can be represented as the performance of SSBs: how well they take advantage of the YouTube algorithm. Comment counts show a positive skewness of 1.531, and the number of responsible SSBs show a positive skewness of 1.152. The green bars in Figure 5 denote newly discovered SSBs that have not been observed to have higher ranking comments. Very few new-to-prior SSBs authored comments with an index of greater than 20. This indicates that SSBs have high probability of landing a comment

[8]Full list of categories shown in Appendix F

[9]For a complete overview of scam domain distribution across all categories, refer to Appendix Table 9

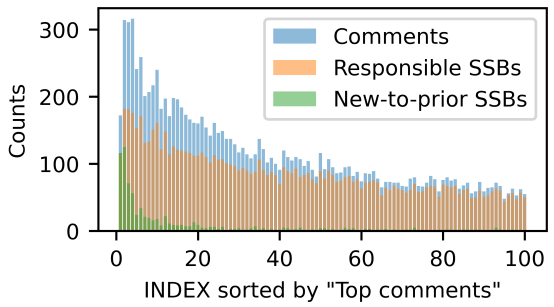



**Figure 5: Number of SSBs and their comments found for each index ranked by "Top comments." Responsible SSBs are the number of SSBs that wrote the comments (blue), and new-to-prior SSBs are the SSBs that have not been observed in previous indices (higher ranks).**

within the default batch. Specifically, 603 SSBs (53.17%) have been in the default batch (top 20 comments), 778 SSBs (68.61%) are in the top 100, and 1,039 SSBs (91.62%) are within the top 200 comments. The majority of SSBs successfully take advantage of the YouTube algorithm.

## 5.2 Prevalency of SSBs

YouTube's policies mention that violations of the community guideline will result in account terminations [2]. Through YouTube's own detection method or user reports, accounts that are determined as abusive, such as SSBs, can be banned. We monitored the termination process of the collected SSBs through periodic visits to the SSB channel page.

**SSBs have a half-life of approximately 6 months.** We observed our SSB profiles over the course of 6 months (October 30th, 2022 - April 30th, 2023) and recorded their termination status. The termination status of accumulated bots by their domains are shown in Figure 6. Only the top 10 domains are shown, with the rest being summed to a single set. On our dataset that was collected on August 1, 2022, out of 1,134 bots, 590 bots were observed to be active on April 30th, 2023. Over the course of 7 monthly examinations (a period of 6 months), 47.9% of SSBs were able to be detected and terminated. Therefore, the half-life (time required for one-half of the SSB count to decay) after identification is approximately 6 months. Of the domains, the three domains with the most terminations were robyoc.online (-68), royal-babes.com (-58), and 1vbucks.com (-46), the first being a romance scam and the followings to be game voucher scams. Furthermore, the game voucher scams were the scam categories with most terminations (-63.3%), where the average termination count of the rest of the scam categories were three times less than game voucher scams (-21.84%). The game voucher scams are mostly based on "robux", the virtual currency for Roblox and "vbucks", the virtual currency for Fortnite.

**SSBs with more expected exposure have succeeded in evading YouTube detection.** We propose to measure the quality of an SSB by its expected exposure: *the potential number of people that have been exposed to the specific scam URL address.* The expected exposure is defined by the following:

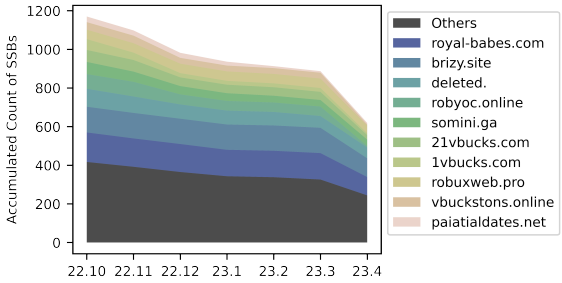



**Figure 6: Termination of SSBs by YouTube. Over 6 months, 47.97% of SSBs were banned.**

**Table 6: Active and banned SSBs. Active SSBs have a higher average expected exposure than banned SSBs.**

| | Active | Banned |
|---|---|---|
| # of Bots | 590 | 544 |
| Infected # of YouTube Creators | 558 | 552 |
| Avg. subscribers | 49.8M | 42.8M |
| Infected # of Videos | 9,575 | 9,110 |
| Avg. Expected Exposure | 15.4K | 12.0K |

$$E[Exposure(SSB)] = \sum_{v} W(v) R_{eng}^2(Y(v)) \quad (2)$$

The expected exposure of an SSB is the weighted sum of the YouTube video views ($W(v)$) by the engagement rate ($R_{eng}$) of the video creator ($Y(v)$), where $v$ is the infected videos by the SSB. Engagement rate is defined as 'the quantity or ratio of responses and interactions that content on social media generates from users' [24, 35]. Content that brings forth more interactions by the viewers will have higher engagement rate. The engagement rates of each YouTube creator were crawled from GRIN[10], an influencer marketing software that provides an online engagement rate calculator. In our scenario, for a benign viewer to be exposed to the scam domain, two engagements are involved: first the viewer must click on the SSB profile, and secondly click on the scam URL address to be redirected to the scam domain. Hence, the engagement rate is squared to properly address the sequence of actions in Equation 2.

Table 6 shows the metrics of active SSBs compared to the banned SSBs. Our analysis reveals that for active SSBs, a single SSB is responsible for an average of 16.2 video infections, while banned SSBs exhibit a slightly higher infection rate of 16.7 video infections per SSB. This finding suggests that YouTube has identified and taken action against SSBs that pose a greater risk, represented by a slightly elevated infection rate. Conversely, when considering the average expected exposure calculated from each individual SSB, we observe that the active SSBs have an accumulated value 1.28 times greater than that of the banned SSBs. This indicates that the active SSBs, despite their lower infection rate per SSB, have a broader reach and potential impact in terms of the number of individuals exposed to their scam URLs. The results indicate that

[10]https://grin.co/

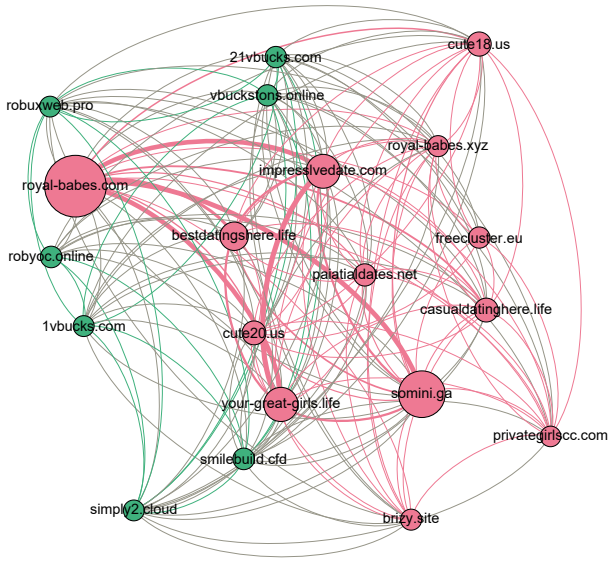



**Figure 7: Top 20 scam campaign graph based on video infections. Campaigns are connected by the number of overlapping videos (edge width) in which they infect. Node size represents the number of SSBs in control.**

YouTube may have failed in identifying SSBs that were particularly aggressive in promoting their scam campaigns. It is important to note that YouTube has prioritized the safety of content consumed by minors, which could explain the discrepancy. The utilization of the expected exposure metric, which incorporates the number of views as a measure of potency, enables the ranking of SSBs based on their potential disruption to the broader audience and not only minors. This suggests that this concept of expected exposure could aid in termination of SSBs that pose a disturbance to the general audience.

## 5.3 Scam Campaigns of SSBs

**There is tight competition between scam campaigns on video infection.** Figure 7 shows the graph of top 20 scam domains ranked by overlapping number of videos, where edges imply a shared video infection. Shared videos between romance scams are colored in pink while shared videos between game voucher scams are colored in green. The connections between these two scam categories are depicted in grey. Graph density is the ratio between edges present and the maximum number of edges that a graph can contain. The graph density of the whole graph is 0.92, where 1 implies a complete graph. Furthermore, the partitions of the graph in Figure 7 by domain category also have high graph densities: 0.93 for romance scams and 0.90 for game voucher scams. Viewing the graph as a bipartite graph between romance and game voucher scams, the graph density is 0.91.

The high graph density observed across various graph divisions indicates that a significant proportion of scam campaigns share a common video, irrespective of their specific scam category. This finding suggests intense competition among scam campaigns to secure a prominent position in the comment section of the targeted videos. The infected videos, on average, had 1,490K views and 67.4K likes, surpassing the corresponding metrics of all videos in our dataset, which averaged at 834K views and 38.4K likes. This notable difference in engagement features of infected videos likely fuels the competition between scam campaigns, as videos with higher engagement are more attractive in terms of gaining exposure through the YouTube algorithm. Although these findings do not represent the entirety of scam campaigns, they uncover the influence of SSBs targeting videos that offer greater potential for exposure of their domain addresses.

# 6 CASE STUDY: SCAM STRATEGIES

The SSBs exhibited a significant spread across multiple videos, aiming to redirect YouTube viewers to their respective URL links, ultimately leading them to engage with the underlying scam campaigns. We were able to observe two preventative measures that SSBs strategized to bypass abusive behavior detection by YouTube: URL shortening and self-engagment. Table 7 shows the top 10 scam campaigns ranked by expected exposure. 9 out of these campaigns used preventative measures, with the scam campaign 'somini.ga' being the unique domain to use both measures.

## 6.1 URL Shorteners are Still Effective

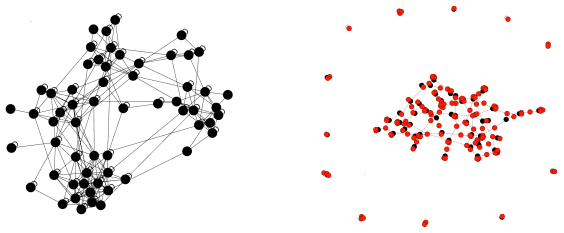
URL shorteners are online tools that convert a URL address into a shorter and more manageable version. Once a URL address is registered, the shortened address offers redirection using a 301 redirect. Fortunately, these shortening services offer preview functions that allow people to check the URL address that the shortened link redirects to. Using these features, we were able to collect and reveal the scam domains of the shortened addresses. From our dataset, 24 out of 72 scam campaigns controlling 644 SSBs (56.8%) used 9 different URL shortening services. The most used shortening services were bitly [11] with 434 SSBs and tinyurl [12] with 143 SSBs.

By using URL shortening services, scam campaigns masked their SLD information from the victim. This strategy is a widely recognized and researched approach that has been observed and analyzed over an extended period of time [27]. Most of the scam domains of Table 7 contain suspicious phrases and words (e.g., 'royal-babes', 'your-great-girls', 'bestdatingshere') that alert and cause the victim to distrust the website. Other cases of SSBs inserted the URL address in the form of hyperlinks. Interestingly enough, all scam campaigns that used URL shortening services were promoting their external links in visible text and not hyperlinks. Also, using URL shorteners serves as a disposable form of defense against being labelled as an indicator of abuse. In the case that YouTube accumulates a blocklist of abusive URL links, using URL shorteners will allow the scam domains be safe from being blocked because these domains never appeared on the platform. Furthermore, these shortened URLs can be easily renewed into a different shortened URL link, disabling any blocklist efforts. Despite the prominent and longstanding history of employing URL shorteners for spam detection evasion, its continued effectiveness persists without being outdated.

[11] https://bitly.com/
[12] https://tinyurl.com/app/

**Table 7: Top 10 scam campaigns ranked by expected exposure. 9 out of 10 of these scam campaigns strategize their campaigns by using URL shorteners and self-engaging SSBs. 'somini.ga' SSBs used self-engagement and was the scam domain with the highest percentage of SSB comments in the default 20 comments shown by YouTube.**

| Scam Campaign | Category | # of SSBs | # of Video Infections | Expected Exposure | URL Shortener | # of Self-Engaging SSBs | Within Default Comment Batch |
|---|---|---|---|---|---|---|---|
| royal-babes.com | Romance | 153 | 7956 | 5.438M | ✓ | - | 1,259 |
| somini.ga | Romance | 63 | 4780 | 3.667M | ✓ | 60 | 1,210 |
| brizy.site | Romance | 133 | 617 | 1.383M | ✓ | - | 0 |
| your-great-girls.life | Romance | 25 | 2604 | 821.4K | ✓ | - | 21 |
| impresslvedate.com | Romance | 20 | 2550 | 759.9K | ✓ | - | 199 |
| bestdatingshere.life | Romance | 22 | 1599 | 689.2K | ✓ | - | 150 |
| cute18.us | Romance | 8 | 829 | 556.8K | - | 2 | 147 |
| cute20.us | Romance | 5 | 756 | 436.5K | - | - | 137 |
| paiatialdates.net | Romance | 29 | 740 | 417.2K | ✓ | - | 58 |
| privategirlscc.com | Romance | 5 | 169 | 148.7K | ✓ | - | 13 |



(a) 'somini.ga' SSB reply graph. (b) Other domain SSB reply graph.

**Figure 8: SSB reply graphs of domain 'somini.ga' and others. Because 'somini.ga' has many self-engagements, the graph structure differs.**

## 6.2 Self-Engaging SSBs

The second strategy that a few scam campaigns employed was the use of self-engaging SSBs. Self-engaging SSBs are SSBs that reply on another SSB's comment. This form of endorsing the comment may help the comment get more exposure and appear higher on the list of comments. Among the observed scam campaigns, it was found that the scam campaign identified as 'somini.ga' exhibited the highest level of self-engaging behavior. As listed in Table 7, 60 out of the 63 SSBs demonstrate self-engagement. SSBs controlled by 'somini.ga' were observed to have replied on each others' comments to boost engagement of the comment for better exposure. 'cute18.us' is also a scam domain that has been using 2 self-engaging SSBs to promote their comment.

The SSB reply graphs are shown in Figure 8, where edges are formed when SSBs reply to another SSB's comment. The first graph is the visualized graph of the commenting SSBs of 'somini.ga', and the second graph shows the commenting SSBs that are not 'somini.ga'. For both graphs, black nodes denote SSBs that have been replied to by another SSB, while red nodes denote the remaining SSBs (that showed replying activity). All SSBs promoting 'somini.ga' have acquired replies made by another SSB. On the other hand, the majority of SSBs from other domains have been shown to engage in replying to other SSBs, while only a few of these SSBs were shown to have been replied to. Unlike 'somini.ga,' other scam domains were not replying to SSB comments that ended up within the top 1,000 comments.

Analyzing the graphs quantitatively, 'somini.ga' SSB reply graph (Figure 8a) is dense with many links compared to the SSB reply graph of other domains (Figure 8b). The graph density of Figure 8a is 0.138, more than 10 times denser than the graph density of Figure 8b, 0.010. Another key trait is that the reply graph of 'somini.ga' is a single connected component, denoting tight enforcement relationships between controlled SSBs, while on other domains there are 13 weakly-connected components. This is because 'somini.ga' SSBs reply to each other frequently while other scam campaigns refrain from interactions between SSBs. As a result of this strategy of self-engagement, SSBs advocating the domain 'somini.ga' successfully posted 1,210 comments within the default batch of 20 comments that YouTube videos automatically load. The scam campaign of 'royal-babes.com' has more comments on the default comment batch (1,259 cases), but considering the number of total infected videos, 'somini.ga' was the campaign with the highest rate of entering the default batch.

Self-engagement occurs in a scheduled manner. Upon examining the replies of SSBs to comments made by other SSBs, it was observed that an overwhelming majority, specifically 99.56% of self-engagement instances, featured an SSB reply as the first reply to the comment. It is noteworthy that none of the self-engagements occurred between distinct scam domains. This observation provides evidence to support the notion that scam domains employ bots for private purposes rather than utilizing them as a public resource between scam domains. This observation aligns with the finding presented in Section 5.3. Given that scam domains are compelled to vie for enhanced exposure, it is plausible that strategies like self-engagement would be sourced internally, or 'intra-sourced', confined solely to the scam domain.

Not only does self-engagement using SSBs help comments get visibility, it also helps the profiling of an SSB to imitate a benign user. Self-engagement within SSBs controlled by scam campaigns characterizes these SSBs as a group of YouTube users that comment regularly on videos, and interacts with other users by replying to

their comments. Semantically, because they are comments based on other users, there is no noticable difference and action-wise, because of its ordinary set of comment and replying actions, this strategy of self engagement complicates previous social bot detection methods [7, 21–23] that rely on semantic or structural information. Furthermore, SSB replies are difficult to differentiate from benign user replies because they are as semantically similar to the starting SSB comment as benign replies. Using YouTuBERT, we observed the cosine similarity between the SSB comment and the SSB reply to be 0.944, whereas the similarity between the SSB comment and benign replies were 0.924. In context, SSB replies would be related to the SSB comment, even more so than the comments of actual users. The occurrence of self-engagement serves as proof of the evolving tactics employed by SSBs to circumvent detection. By actively engaging with their own comments, bots are adapting their behavior to appear more natural and avoid raising suspicion. This adaptive behavior highlights the need for continuous improvement in bot detection mechanisms to effectively identify and mitigate such evolving strategies.

# 7 DISCUSSION

## 7.1 Sophistication of SSBs

SSBs on the YouTube platform exhibit distinct behavior compared to other types of bots commonly studied. Traditional bots found on YouTube are typically simplistic in nature, programmed to carry out specific tasks such as artificially increasing video views [10, 11, 37, 41] or spamming predefined comments in the comment section [6]. In contrast, SSBs demonstrate a more sophisticated and nuanced behavior that goes beyond simple task execution. SSBs understand viewer demographics and have been adapting throughout the years.

**SSBs exhibited a possibility to target the viewing audience.** Our analysis from Section 5.1 revealed that SSBs demonstrated a deliberate targeting strategy, where videos in the 'video game' category had more infections from 'game voucher' scams. This strategic choice can be attributed to the inherent ineffectiveness of game voucher scams for individuals who do not engage with the specific game being promoted. Thus, SSBs recognized this limitation and concentrated their efforts on videos within the gaming category where they could potentially attract victims who were more likely to be interested in such game-related offers. Conversely, 'romance' SSBs exhibited a broader distribution across various video categories within our dataset. Given the prevalence of romance scams as the most dominant category, these scams were deemed effective in targeting a wide range of YouTube users, as romance-related content is likely to appeal to a larger audience base.

**SSBs have effectively exploited the YouTube recommendation algorithm.** The strategic tactics employed by SSBs, particularly their self-engagement behavior, have proven highly successful in maximizing the exposure of their comments (Table 7). The self-engagement technique utilized by SSBs of 'somini.ga' resulted in its comment appearing in the default comment batch in one out of every four instances. The significant level of exposure achieved by self-engaging SSBs raises concerns, particularly considering that the YouTube recommendation algorithm remains undisclosed and inaccessible to the public. These findings suggest that self-engaging SSBs have effectively executed a black-box attack on the recommendation algorithm. SSBs are aggressors targeting the integrity and functionality of the YouTube platform itself.

## 7.2 Mitigations

YouTube recognizes the presence of SSBs and issues warnings to publishers of guideline violating content and links, and takes measures such as account termination and channel closure [2, 3]. Deduced from our analysis, we propose three insights that mitigation methods should regard for improved SSB detection.

**Utilizing shortened URLs as indicators can help identify SSB based suspicious activities.** As outlined in Section 6.1, the presence of shortened URLs can serve as a valuable signal for detecting potential abusive activities. Our analysis revealed that out of the 742 shortened URLs examined, 644 of them were identified as scam URLs, representing 56.8% of the SSBs detected through our workflow. Incorporating the straightforward feature of 'having a shortened URL' into the profile information of YouTube accounts could have successfully flagged more than half of the SSBs identified in our study. Moreover, URL shortening services themselves often maintain guidelines and measures to prevent the proliferation of abusive links. Recognizing that the ultimate harm lies in the external address (destination link) itself, if URL shortening services were to terminate and refuse redirection of shortened links associated with SSBs, the impact of these scams could be mitigated. By effectively communicating reports of abusive behavior on YouTube to URL shortening services, even if SSBs remain active, their ability to cause further harm would be curtailed. This underscores the importance of monitoring and scrutinizing shortened URLs as a means of identifying and preventing exposure of potential scam campaigns.

**Leveraging the top 20 comments enables efficient detection of SSBs.** The top 20 comments of a YouTube video represent the default batch of comments that are initially loaded for a video. Our findings in Section 5.1 revealed that 53.17% of SSBs were able to place their comment in the first default batch of comments. In other words, just by confirming the entities that arise in the first default batch, practitioners can detect and terminate more than half of the SSBs that were extracted from 1,000 comments. Note that more than 50% of SSBs originated from only 2% of the dataset. Therefore, instead of monitoring all user accounts for external links, focusing on accounts that have posted comments in the top 20 for any video can significantly reduce the number of SSBs that need to be monitored. By implementing a detection mechanism that specifically targets the top 20 comments, practitioners can achieve efficient identification and mitigation of SSBs. This approach capitalizes on the fact that a substantial portion of SSBs tend to emerge within this limited subset of comments, allowing for targeted monitoring and intervention.

**Preparations need to be made for SSBs employing large language models.** Currently, SSBs generate comment text by using a skeleton comment as a base, which allows us to employ semantic similarity techniques for identifying potential bot candidates. However, with the rapid advancement of LLMs such as ChatGPT[13],

[13] https://chat.openai.com/

text generation has become increasingly sophisticated and efficient. It can be expected that SSBs will leverage LLMs to generate their comments, utilizing not only existing comments but also video/audio/caption data as a source of inspiration. This development poses a significant challenge as traditional semantic-based detection methods (including our filtering method using YouTuBERT) may become less effective in identifying such SSBs. To address this evolving threat, additional meta-information can be incorporated into the detection and termination process. For instance, factors such as subscriber lists and commenting activity could be considered alongside text-based analysis, allowing methods utilizing graph information. Unfortunately, this information is not disclosed to the public and can only be utilized by YouTube. By leveraging a wider range of information and adapting detection techniques accordingly, it will be possible to detect and mitigate the activities of future SSBs that utilize advanced LLM-based comment generation techniques.

## 8 CONCLUSION

In this study, we conducted an extensive investigation into the activities of social scam bots (SSBs) and their associated scam campaigns on the YouTube platform. We developed a workflow capable of identifying SSBs and extracting their scam campaigns for analysis using YouTuBERT, our pretrained language model on YouTube comments. Our findings revealed that a total of 72 scam campaigns were responsible for infecting approximately 31.73% of the videos included in our dataset. We observed that SSBs exhibited higher levels of intelligence compared to traditional bots. They demonstrated an understanding of the target video's audience, strategically matching scam domains to increase the likelihood of success. Furthermore, the majority of SSBs demonstrated expertise in generating comments that appeared within the top 20 default comments for a given video. By use case study of the domain 'somini.ga', we were able to identify two strategies of scam campaigns: URL shorteners and self-engagement. Through self-engagement, SSBs effectively manipulated the YouTube recommendation algorithm, leading to increased visibility and potential victim engagement. We believe that this work can serve as a foundation for addressing and mitigating the phenomenon of SSBs and hope it contributes to the development of effective countermeasures and ultimately enhance the safety and security of the YouTube platform.

## REFERENCES


[1] [n. d.]. YOUTUBE USER STATISTICS 2022. https://www.globalmediainsight.com/blog/youtube-users-statistics/#stat. Accessed: 2022-11-01.

[2] 2023. YouTube policies - External links policy. https://support.google.com/youtube/answer/9054257. Accessed: 2023-01-17.

[3] 2023. YouTube policies - Spam, deceptive practices, & scams policies. https://support.google.com/youtube/answer/2801973. Accessed: 2023-01-17.

[4] Norah Abokhodair, Daisy Yoo, and David W McDonald. 2015. Dissecting a social botnet: Growth, content and influence in Twitter. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 839–851.

[5] Faraz Ahmed and Muhammad Abulaish. 2013. A generic statistical approach for spam detection in online social networks. *Computer Communications* 36, 10-11 (2013), 1120–1129.

[6] Túlio C Alberto, Johannes V Lochter, and Tiago A Almeida. 2015. Tubespam: Comment spam filtering on youtube. In *2015 IEEE 14th international conference on machine learning and applications (ICMLA)*. IEEE, 138–143.

[7] Victor Benjamin and TS Raghu. 2022. Augmenting Social Bot Detection with Crowd-Generated Labels. *Information Systems Research* (2022).

[8] Elijah Bouma-Sims and Brad Reaves. 2021. A First Look at Scams on YouTube. *arXiv preprint arXiv:2104.06515* (2021).

[9] Chiyu Cai, Linjing Li, and Daniel Zeng. 2017. Detecting social bots by jointly modeling deep behavior and content information. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1995–1998.

[10] Vidushi Chaudhary and Ashish Sureka. 2013. Contextual feature based one-class classifier approach for detecting video response spam on youtube. In *2013 Eleventh Annual Conference on Privacy, Security and Trust*. IEEE, 195–204.

[11] Rashid Chowdury, Md Nuruddin Monsur Adnan, GAN Mahmud, and Rashedur M Rahman. 2013. A data mining based spam detection system for youtube. In *Eighth international conference on digital information management (ICDIM 2013)*. IEEE, 373–378.

[12] Stefano Cresci. 2020. A decade of social bot detection. *Commun. ACM* 63, 10 (2020), 72–83.

[13] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*. 963–972.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[15] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *kdd*, Vol. 96. 226–231.

[16] European Commission. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). https://eur-lex.europa.eu/eli/reg/2016/679/oj

[17] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[18] Yuriy Gorodnichenko, Tho Pham, and Oleksandr Talavera. 2021. Social media, sentiment and public opinions: Evidence from# Brexit and# USElection. *European Economic Review* 136 (2021), 103772.

[19] Christian Grimme, Dennis Assenmacher, and Lena Adam. 2018. Changing perspectives: Is it sufficient to detect social bots?. In *International conference on social computing and social media*. Springer, 445–461.

[20] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 US presidential election. *Science* 363, 6425 (2019), 374–378.

[21] Maryam Heidari and James H Jones. 2020. Using bert to extract topic-independent sentiment features for social media bot detection. In *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE, 0542–0547.

[22] Maryam Heidari, James H Jones, and Ozlem Uzuner. 2020. Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter. In *2020 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 480–487.

[23] Maryam Heidari, James H Jones Jr, and Ozlem Uzuner. 2022. Online User Profiling to Detect Social Bots on Twitter. *arXiv preprint arXiv:2203.05966* (2022).

[24] Roope Jaakonmäki, Oliver Müller, and Jan Vom Brocke. 2017. The impact of content, context, and creator on user engagement in social media marketing. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, Vol. 50. IEEE Computer Society Press, 1152–1160.

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[26] Luca Luceri, Ashok Deb, Adam Badawy, and Emilio Ferrara. 2019. Red bots do it better: Comparative analysis of social bot partisan behavior. In *Companion proceedings of the 2019 World Wide Web conference*. 1007–1012.

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

[28] Michele Mazza, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. 2022. Investigating the difference between trolls, social bots, and humans on Twitter. *Computer Communications* 196 (2022), 23–36.

[29] Guanyi Mou and Kyumin Lee. 2020. Malicious bot detection in online social networks: arming handcrafted features with deep learning. In *International Conference on Social Informatics*. Springer, 220–236.

[30] Yew Chuan Ong. 2020. *Characterising and Detecting Social Bots*. Ph. D. Dissertation. University of Sheffield.

[31] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[32] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications* 9, 1 (2018), 1–9.

[33] Nisha P Shetty, Balachandra Muniyal, Arshia Anand, and Sushant Kumar. 2022. An enhanced sybil guard to detect bots in online social networks. *Journal of Cyber Security and Mobility* (2022), 105–126.

[34] Aashish Singh. [n. d.]. YouTube users increasingly concerned with adult comments (sex bots) spamming replies on videos. https://piunikaweb.com/2022/12/19/youtube-users-concerned-with-adult-comments-spamming-replies/. Accessed: 2022-11-01.

[35] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE second international conference on social computing*. IEEE, 177–184.

[36] The YouTube Team. [n. d.]. More updates on our actions related to the safety of minors on YouTube. https://blog.youtube/news-and-events/more-updates-on-our-actions-related-to/. Accessed: 2022-11-01.

[37] Ashutosh Tripathi, Kusum Kumari Bharti, and Mohona Ghosh. 2019. A study on characterizing the ecosystem of monetizing video spams on youtube platform. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*. 222–231.

[38] Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the international AAAI conference on web and social media*, Vol. 11. 280–289.

[39] Teng Xu, Gerard Goossen, Huseyin Kerem Cevahir, Sara Khodeir, Yingyezhe Jin, Frank Li, Shawn Shan, Sagar Patel, David Freeman, and Paul Pearce. 2021. Deep entity classification: Abusive account detection for online social networks. In *30th {USENIX} Security Symposium ({USENIX} Security 21)*.

[40] Dong Yuan, Yuanli Miao, Neil Zhenqiang Gong, Zheng Yang, Qi Li, Dawn Song, Qian Wang, and Xiao Liang. 2019. Detecting fake accounts in online social networks at the time of registrations. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*. 1423–1438.

[41] Savvas Zannettou, Sotirios Chatzis, Kostantinos Papadamou, and Michael Sirivianos. 2018. The good, the bad and the bait: Detecting and characterizing clickbait on youtube. In *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 63–69.

# A ETHICS

Our workflow of YouTube comment processing (Figure 3) involves visiting and foraging information from active YouTube accounts. We took awareness into the ethical considerations in data collection.

Automated crawling of OSN accounts is a sensitive matter, regulated by the General Data Protection Regulation (GDPR) [16]. In this, identifying bot candidates using YouTuBERT serves to prevent unreasonable crawling in mass. In order to streamline the account inspection process, the second crawler was designed to only scrape user accounts exhibiting suspicious behavior. By focusing solely on bot candidates, the number of accounts subjected to scrutiny was significantly reduced, optimizing the efficiency of the inspection procedure. This filtering step to acquire bot candidates reduced the number of channel page visits by the second crawler to only 2.46% of the total number of commenters. Visiting of account channel pages only occurred with reason.

During channel visit, note that any user account statistics (e.g., number of views, location, join date) that may lead to personally identifier information (PII) were not compiled. When the crawler distinguishes a URL address in the previously mentioned regions of the page, the address string is the *only* information that is compiled—the crawler does not visit the actual HTML of the external link. We acknowledge that the URL address itself might be PII because it is common for users to post links to personal websites (e.g., other OSN accounts), which is why the blocklist is arranged to delete and exclude these OSN domain addresses from analysis. We assembled our dataset under the conservative approach of omitting any possible PII external links.

# B MANUAL TAGGING STANDARDS

Three annotators were given guidelines in determining bot candidates for construction of the ground truth dataset. Annotators were asked to label each comment when supplied with the comment text, commenter username, commenter channel address, and cluster information to which the comment belonged to. All annotators were educated of the guidelines before performing the task. The tagging guidelines stated to label bot candidates for the comments that show the following characteristics:

- Identical comments within the same cluster
- Nearly identical comments that seem modified (addition or deletion of words, sentences, or punctuation marks)
- Scam-related words or phrases explicitly shown in their username
- The same text has already been verified as a bot candidate
- Commenter channel address containing prompts to scam domains (e.g., Figure 1)

The final label was selected through majority voting of the tagging results provided by each annotator, which had an inter-annotator agreement Fleiss' Kappa value of 0.89. This value denotes near-perfect agreement between annotators. After examination of the dataset with these guidelines, the annotators tagged 3,464 comments out of the total 24,706 as bot candidates.

# C PRETRAINING YouTuBERT ON YouTube

When comparing the sentence embeddings, we utilized the pretrained model `all-MiniLM-L6-v2` for Sentence-BERT and `roberta-base` for RoBERTa. To adjust YouTuBERT to the YouTube comment domain, we leverage the original RoBERTa model instead of training from scratch. This offers the benefit of utilizing the general English representation learned from the original model. Initializing the model with `roberta-base`, our YouTube comment text corpus is fed into the model for pretraining. As with the original RoBERTa training method, we use the byte-pair encoding tokenization vocabulary. YouTuBERT is fine-tuned on 3 epochs consisting of 313,500 steps with the learning rate of $5e-5$. Figure 10 illustrates the convergence of the training loss, indicating the successful fine-tuning of the model.

# D FINDING LINKS ON CHANNEL PAGES

During our examination of SSB accounts, we identified external links in two areas on the HOME tab of the channel page and three

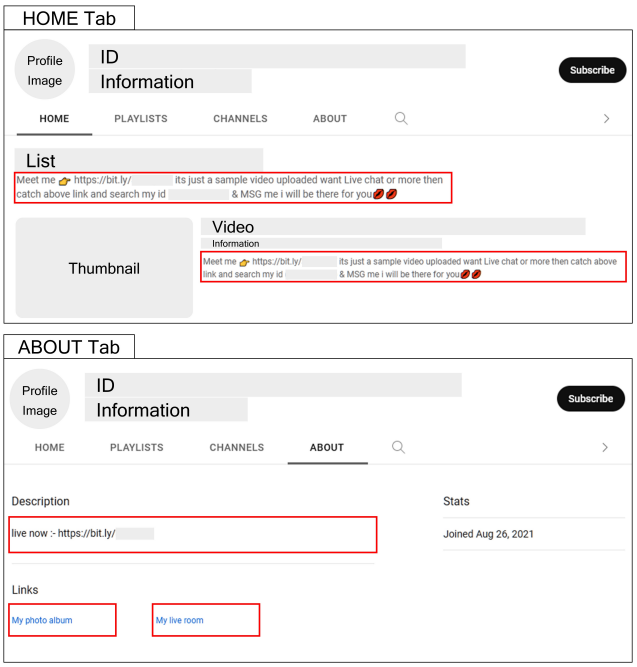



**Figure 9: Five areas from which links were collected on the account channel page**

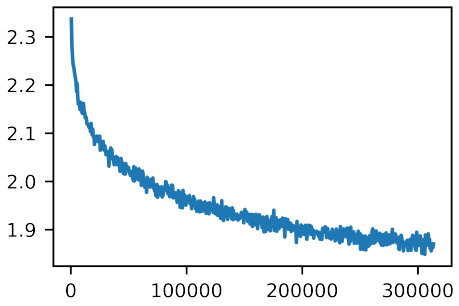



**Figure 10: Loss graph of pretraining YouTuBERT.**

**Table 8: Scam domains and their verifying services. Highlighted in bold are domains detected in three or more scam verification services.**

| Scam Verification Service | Scam Campaigns | # of Verified Scams |
|---|---|---|
| ScamAdvisor | **verifyus.net**, tamsu69.com, cogratulations.com, **chuaks.fun**, **cute20.us**, robuxgo.xyz, **topunlocker.net**, babe19.com, **cute18.us**, **v-buxy.club**, robuxcode.org, **cute20.us**, vikinq.bond, cardgen.online, skinnet.bond, meetbabes.xyz, 42web.io, simply2.cloud, vbuckstons.online, shrinke.me, rbxton.online, rbxai.com, **golead.pl**, **appfile.cc**, crycrox.xyz, **bestdatingshere.life**, **paiatialdates.net**, **privategirlscc.com**, **casualdatinghere.life**, robyoc.online, **21vbucks.com**, somini.ga, spnsrd.me, **gmai.com**, **tiltok4you.com**, **royal-babes.xyz**, rbxworld.cf | 37 |
| ScamWatcher | usheethe.com, date30.com, game-z.tech, smilebuild.cfd, royal-babes.com, sweet18.us, your-great-girls.life, teenisyours.com, timbantinh69.com, livegirls19.com, **bitly.com.vn**, **guserverification.xyz**, fuckme99.com, rovloxes1.blogspot.com, robuxweb.pro, cute25.xyz, **22robux.com**, playzone.top, mioin.space, modgang.com, impresslvedate.com, brizy.site, chonbantinh.xyz | 51 |
| Google Safe Browsing | e-reward.gb.net | 6 |
| URLVoid | thesmartwallet.com, havebucks.com, monglitch.monster, 1vbucks.com, lovegirl4you.life, freecluster.eu, lonely-chat.xyz, dirtyflirt0.com | 37 |
| IPQualityScore | agift.info, shewantyou.net, fuck27.com | 15 |

areas on the ABOUT tab. These areas are marked with red boxes in the accompanying Figure 9. Given the unrestricted nature of user profiles, which allows them to include external links within the description section, we gathered pertinent information pertaining to these external links from these areas.

## E SCAM DOMAINS AND THEIR VERIFICATIONS

Initially, the candidates were subjected to a blocklist filtering, after which their domains were cross-referenced with the following online fraud prevention resources. ScamAdvisor[14] is an online platform equipped with a robust database specifically designed to detect and classify scams. The platform employs sophisticated algorithms and methodologies to assess the trustworthiness of websites, contributing to the identification and mitigation of potential fraudulent activities. In ScamAdvisor, each url or domain is rated with a metric called a 'Trustscore' between 0 and 100. Ratings of less than or equal to 50 were classified as scam domains. ScamWatcher[15] is an online community that operates as a collaborative platform for reporting and sharing information about scams. As an active community the platform relies on community participation to actively monitor and expose emerging scams. ScamDoc[16] is their web tool that evaluates reliability of an email or website url. Similarly, ScamDoc assesses each website with a 'trust index' between 0% and 100%, with ratings of less than or equal to 50% being classified as scam domains. Google Safe Browsing[17] is a widely utilized web security service provided by Google. It flags potentially malicious websites or URLs by constantly crawling and analyzing the web. Their 'Check site status' service informs the user if a website URL is unsafe, which was used as verification of being a scam domain. URLVoid[18] is an online service that examines a multitude of factors such as historical data, blacklists, and other security indicators. URLVoid offers a summary of searching their database of 40 separate scanning engines. If there was a hit ($\geq 0$), it was labeled as a scam domain. IPQualityScore[19] is a platform specializing in IP intelligence and fraud prevention. They offer domain reputation reports, and any domain that was concluded with 'High Risk' was labeled a scam domain. Table 8 presents a compilation of SLDs identified as scams, along with the corresponding sources that have attributed the scam

[14]https://www.scamadviser.com/
[15]https://www.scamwatcher.com/
[16]https://www.scamdoc.com/
[17]https://safebrowsing.google.com/
[18]https://www.urlvoid.com/
[19]https://www.ipqualityscore.com/

**Table 9: Distribution ratios of scam domain categories over video categories. Highlighted in bold denote values that surpass one standard deviation.**

| Category | Romance | Game Voucher | E-Commerce | Malvertising | Miscellaneous | Deleted |
|---|---|---|---|---|---|---|
| Video games | 0.8884 | **0.1019** | 0.0014 | 0.0007 | 0.0023 | 0.0053 |
| Beauty | 0.9724 | 0.0022 | 0.0007 | 0.0000 | 0.0000 | 0.0246 |
| Design/art | 0.9466 | 0.0126 | 0.0007 | 0.0007 | 0.0021 | **0.0372** |
| Health & Self Help | 0.9863 | 0.0082 | 0.0000 | 0.0027 | 0.0027 | 0.0000 |
| News & Politics | 0.9909 | 0.0014 | 0.0007 | 0.0000 | 0.0007 | 0.0063 |
| Education | 0.9859 | 0.0000 | 0.0020 | 0.0000 | 0.0040 | 0.0081 |
| Humor | 0.9313 | **0.0430** | 0.0013 | 0.0005 | 0.0039 | 0.0200 |
| Fashion | 0.9721 | 0.0007 | 0.0014 | 0.0000 | 0.0007 | **0.0252** |
| Sports | 0.9905 | 0.0032 | 0.0032 | 0.0011 | 0.0021 | 0.0000 |
| DIY & Life Hacks | 0.9603 | 0.0060 | 0.0000 | 0.0009 | 0.0026 | **0.0302** |
| Food & Drinks | 0.9731 | 0.0072 | 0.0006 | 0.0036 | 0.0030 | 0.0126 |
| Animals & Pets | 0.9331 | 0.0213 | **0.0091** | **0.0182** | 0.0000 | 0.0182 |
| Travel | **0.9938** | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0062 |
| Animation | 0.9198 | **0.0715** | 0.0018 | 0.0008 | 0.0025 | 0.0036 |
| Science & Technology | 0.9786 | 0.0105 | **0.0040** | 0.0025 | 0.0036 | 0.0007 |
| Toys | 0.8670 | **0.0472** | **0.0043** | 0.0000 | **0.0665** | 0.0150 |
| Fitness | 0.9708 | 0.0065 | 0.0000 | **0.0097** | 0.0065 | 0.0065 |
| Mystery | 0.9588 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | **0.0412** |
| ASMR | 0.9665 | 0.0152 | 0.0000 | 0.0000 | 0.0030 | 0.0152 |
| Music & Dance | 0.9730 | 0.0081 | 0.0012 | 0.0005 | 0.0073 | 0.0098 |
| Daily vlogs | 0.9623 | 0.0132 | 0.0004 | 0.0011 | 0.0053 | 0.0178 |
| Autos & Vehicles | 0.9828 | 0.0043 | 0.0000 | 0.0043 | 0.0000 | 0.0086 |
| Movies | 0.9608 | 0.0233 | 0.0019 | 0.0013 | 0.0025 | 0.0102 |
| Mean | 0.9593 | 0.0177 | 0.0015 | 0.0021 | 0.0052 | 0.014 |
| Standard Deviation | 0.0317 | 0.0249 | 0.0020 | 0.0040 | 0.0132 | 0.0111 |

label to each domain. Evidently, instances arose when multiple verification services independently confirmed the identification of scam SLDs. To maintain conciseness, only the initial occurrence of each duplication was listed.

## F VIDEO CATEGORIES

The full list of video categories mentioned in Section 5.1 are shown in Table 9. Along with the full list, this table shows the distribution ratios of scam domain categories of each video category. In all the cases, romance scams were the major portion of scam domains that were infected on all videos. Upon examining the mean and standard deviation of each scam domain, it is apparent that a significant prevalence of game voucher scams is observed within the video games and animation categories. This observation raises concerns due to the notably elevated occurrence of such fraudulent activities in these specific domains that generally target a younger audience.