Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM 2021)

ABOME: A Multi-platform Data Repository of Artificially Boosted Online Media Entities

Hridoy Sankar Dutta, Udit Arora, Tanmoy Chakraborty
IIIT-Delhi, India
hridoyd@iiitd.ac.in, udita@iiitd.ac.in, tanmoy@iiitd.ac.in

Abstract

The rise of online media has incentivized users to adopt various unethical and artificial ways of gaining social growth to boost their credibility within a short time period. In this paper, we introduce ABOME, a novel multi-platform data repository consisting of artificially boosted (also known as blackmarket-driven *collusive entities*) online media entities such as Twitter tweets/users and YouTube videos/channels, which are prevalent but often unnoticed in online media. ABOME allows quick querying of collusive entities across platforms. These include details of collusive entities involved in blackmarket services to gain artificially boosted appraisals in the form of *likes*, *retweets*, *views*, *comments*, *follows* and *subscriptions*. ABOME contains data related to tweets and users on Twitter, YouTube videos and YouTube channels. We believe that ABOME is a unique data repository that can be used as a benchmark to identify and analyze blackmarket-driven fraudulent activities in online media. We also develop SearchBM, an API and a web portal to identify blackmarket entities.

Introduction

The past decade has seen a momentous rise in Online Social Networks (OSNs) such as Twitter, YouTube, and Facebook, which help people connect for personal and business interactions. These platforms now boast billions of active users, thereby making them an important component of today's human social fabric. People share and form thoughts and opinions about events, products, and other people on these platforms. This makes online media an attractive platform for people who wish to propagate their opinions to spread their agenda, such as promoting a product or political ideology. Therefore, gaining a stronger influence on online media platforms carries a high level of economic benefit.

However, in order to spread an opinion, users require a large reach across the network. This reach can either be acquired *organically* – by posting quality content over time and gaining popularity, or *inorganically* – by certain online media-driven blackmarket services that allow users to boost the reach of their content artificially. *Collusion in online media* involves users artificially gaining social reputation, which violates the Terms of Service (ToS) of the online media platform. These users approach blackmarket services to artificially inflate their social status. This results in entities such as Twitter tweets/users or YouTube videos/channels to appear more attractive to the end-users, thus leading to activities such as fake promotions, campaigns, and misinformation. The blackmarket services support various online media services ranging from online social networks to other platforms such as rating/review platforms, video-sharing platforms and even recruitment platforms (Dutta and Chakraborty 2020).

A substantial number of studies have investigated the phenomena of **fake** (Alsaleh et al. 2014; Gupta et al. 2013; Gupta, Kumaraguru, and Chakraborty 2019; Cresci et al. 2015; Stringhini et al. 2013), **fraudulent** (Giatsoglou et al. 2015; Liu, Hooi, and Faloutsos 2017; Li et al. 2016; Shah et al. 2014; Hooi et al. 2016; Chavoshi, Hamooni, and Mueen 2016a) and **spam** (Benevenuto et al. 2010; Yardi et al. 2010; Thomas et al. 2011) activities. We encourage the readers to go through (Kumar and Shah 2018; Pierri and Ceri 2019) for detailed surveys on false and fake information on the web.

However, there are relatively fewer studies on the detection and analysis of collusive activities that result in an artificial boosting of social growth. Our recent investigations (Chetan et al. 2019; Dutta et al. 2018; Dutta and Chakraborty 2019; Arora, Paka, and Chakraborty 2019; Dhawan et al. 2019; Dutta et al. 2020; Arora et al. 2020; Sankar Dutta et al. 2020) revealed that existing fraud detection strategies are not suitable for blackmarket-driven collusive entity detection. These studies reported that collusive users are not bots or fake users; rather, they are normal users showing a mix of organic and inorganic activities with no synchronicity across their behaviors. ABOME would help researchers analyze blackmarket-driven collusive activities and build systems to detect them.

ABOME consists of multi-platform datasets for collusion in online media collected from two major blackmarket services - YouLikeHits and Like4Like. ABOME comprises datasets of two types – *historical data* and *time-series data*. The former type of dataset is divided into two parts – the first part consists of 23,522 collusive retweets and 18,368 collusive follower requests for Twitter; the second part consists of 58,091 collusive likes, 25,106 comments, and 7,847 subscriptions requests for YouTube. The latter type of dataset consists of time-series data of 2,350 Twitter users and 4,989

tweets collected from blackmarket services.

ABOME is unique for the following five reasons:

- To the best of our knowledge, ABOME is the first public dataset of collusive entities in online media such as Twitter tweets/users and YouTube videos/channels. We believe these datasets have tremendous research potential in the field of analysis and detection of collusive behavior in online media.
- ABOME comprises of two types of datasets: *historical* and *time-series* data.
- ABOME provides abundant textual and temporal information of collusive entities for Twitter and YouTube.
- ABOME has an API and a web portal, SearchBM to discover collusive entities using text search queries.
- ABOME protects user privacy and can be used in a wide range of research areas, such as fraudulent entity detection, diffusion modeling, social-growth prediction, etc.

The entire dataset, along with a smaller sample, is available at the following link: (Dataset URL: https://zenodo.org/record/4437987)
(Dataset DOI: http://doi.org/10.5281/zenodo.4437987).
We also provide a datasheet for our dataset according to Datasheets for Datasets recommendations (Gebru et al. 2018) as supplementary material.

Blackmarket Services

Websites such as YouLikeHits (https://www.youlikehits.com/), Like4Like (https://www.like4like.org/), TraffUp (https://traffup.net/), JustRetweet (https://www.justretweet.com/) allow social media users to gain appraisals *inorganically* in different forms. They provide services for various **online social networks**, e.g., Facebook (followers, likes, shares, comments), Twitter (followers, retweets, likes), Instagram (followers, likes, comments). Other than OSNs, the blackmarket syndicates also provide service to **video subscription-sharing platforms**, e.g., YouTube (views, subscribers, likes, comments), Vimeo (plays, followers), **music-sharing platforms**, e.g., SoundCloud (plays, followers, likes, reposts, comments), ReverbNation (fans), **business and employment-oriented platforms**, e.g., LinkedIn (followers, connections, endorsements). Organically gaining higher reach on online media is difficult and time-consuming, which boosts the allure of these blackmarket services. With higher reach comes stronger influence, and with stronger influence comes higher economic value. This influence can be used for product promotions or opinion propagation. Fig. 1 shows an example of one such blackmarket service which provides collusive appraisals.

(Shah et al. 2017) divided the blackmarket services into two types based on the model of service - *Premium* and *Freemium*. Customers in the premium services need to pay a cash lump sum to receive blackmarket services. Freemium services may not demand customers to pay; rather, these services create a community of customers where each member gains appraisals by appraising the content of other customers registered on these blackmarket services. When a customer appraises some content through the blackmarket portal, they

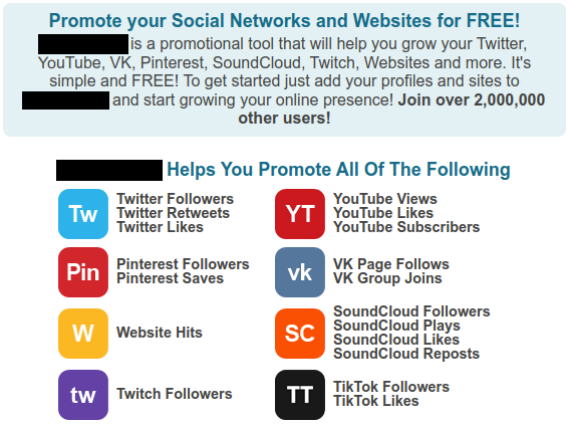

Figure 1: Example blackmarket service providing collusive appraisals to online media platforms such as Twitter, Pinterest, YouTube, VK, SoundCloud, Twitch (name of the blackmarket redacted).

earn *credits*, which they can use later to gain appraisal for their own content. Such freemium services are also called *credit-based freemium services*. These services are a menace and pose a threat to the credibility of social media platforms. Unlike bots, detecting users who engage in these activities is difficult because they may display a mix of organic and inorganic behavior - whereby they appraise some content related to their interest as a genuine user and also appraise some content to gain credits on a freemium service.

We made the first attempt to investigate blackmarket customers on Twitter (Dutta et al. 2018). We collected data related to users engaged in producing fake retweets from four blackmarket services and annotated each user into one of the four categories – *bots, promotional customers, normal customers, and genuine users*. Finally, we ran several classifiers on a set of 64 features to perform multi-class and binary classification to detect collusive users. We further extended this work to show how users involved in premium blackmarket services exhibit unusual properties as compared to those involved in freemium services (Dutta and Chakraborty 2019). (Chetan et al. 2019) proposed CoReRank, an unsupervised approach to detect collusive users and suspicious tweets submitted to blackmarket services based on two intrinsic traits – the credibility of users and the merit of tweets. CoReRank considers a directed bipartite graph of (re)tweeters and (re)tweets in order to incorporate interdependency of user-level and content-level traits such as network properties, behavioral properties, and topical similarity. (Arora, Paka, and Chakraborty 2019) detected tweets submitted to blackmarket services using a multitask learning approach. (Arora et al. 2020) proposed a multi-view learning-based approach to detect collusive retweeters by utilizing various attribute and network views based on the user's posts and interactions on the social graph. We encourage the reader to go through (Dutta and Chakraborty 2020) for a comprehensive survey on analyzing and detecting collusive activities in online media platforms.

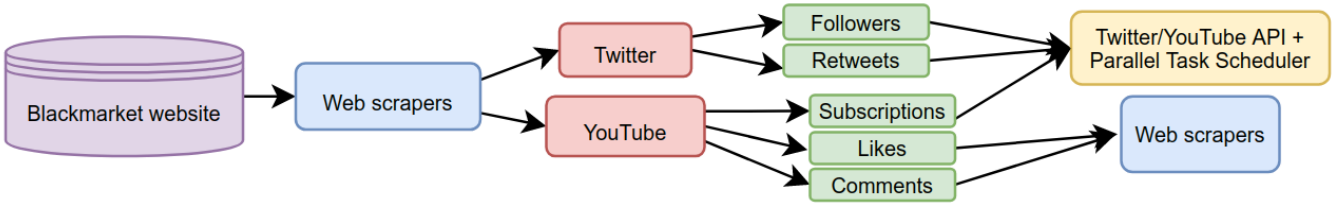

Figure 2: The process of collecting ABOME dataset from the blackmarket service.

Data Collection

In this section, we first comment on our ethics and data privacy statement. We then describe the process of creating our datasets in detail.

Ethics and Data Privacy Statement

The entire data collection process has been carried out through Twitter API[1], YouTube API[2] and web scrapers. We did not seek explicit permission from YouLikeHits and Like4Like to scrape the content because these blackmarket sites do not themselves act in line with the Terms & Conditions (T&C) of the services they connect with[3]. They allow users to boost their influence on these platforms artificially, thereby going against the T&C of Twitter[4] and YouTube[5]. Further, since the data we posted is anonymized, the privacy and identity of these users will not be compromised. We abode by the terms, conditions, and privacy policies of our Institute Institutional Review Board (IRB) approval[6].

Collecting Data from Blackmarket Services

After careful IRB approval, we developed web scrapers for parsing two of the most popular blackmarket websites - YouLikeHits and Like4Like. The parser for YouLikeHits used Python's BeautifulSoup library to parse the HTML DOM of the website and got the details of the users and content that are posted for appraisals. The parser for Like4Like used Selenium (https://www.seleniumhq.org/) to run a headless web browser within which the website is loaded and BeautifulSoup was then used to parse the details of users and content posted for appraisal.

Data Anonymization

The data is anonymized by removing all Personally Identifiable Information (PII) and generating pseudo-IDs corresponding to the original IDs. A consistent mapping between the original and pseudo-IDs is used to maintain the integrity and usefulness of the data.

[1]https://developer.twitter.com/en/docs

[2]https://developers.google.com/youtube/v3

[3]We don't reveal the identity of users/tweets (Twitter id, Tweet ids, YouTube channel ids, etc.)

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

[5]https://support.google.com/youtube/answer/3399767?hl=en

[6]Permission to collect data from blackmarket services is mentioned in our IRB approval.

Collecting Data at Scale from Twitter and YouTube

We focused on collecting data from credit-based freemium services. We divide the datasets into two parts:

- **Historical data**: This consists of all the data for Twitter and YouTube from YouLikeHits gathered via sequential querying of the website's URLs (the historical data for Like4Like was not available, which is why it is missing in our collected dataset) between the period March-June, 2019. The details of the sequential querying technique are explained in the last part of this section.
- **Time-series data:** This consists of time-series data (collected every 8 hours) of Twitter users and tweets collected from two blackmarket services – YouLikeHits and Like4Like between the period of March-June, 2019.

Historical data: Since the entities we collected were a few years old in many cases, we focused on collecting the relatively static properties (such as the profile description and tweet content) and information related to those entities. We collected the following historical data from YouLikeHits:

1. **Twitter Retweets:** Tweet ids of tweets that have been posted on YouLikeHits in order to gain retweets, and their tweet objects using the Twitter API as well as the last 100 retweets of these tweets (we were only able to collect the last 100 retweets due to Twitter API limitations).
2. **Twitter Followers:** User ids of accounts that have been posted on YouLikeHits in order to gain followers, and their user objects using the Twitter API.
3. **YouTube Likes and Comments:** Video ids of videos that have been posted on YouLikeHits in order to gain likes, and their corresponding metadata, which is detailed in the next section.
4. **YouTube Subscriptions:** Channel ids of YouTube channels that have been posted on YouLikeHits in order to gain subscribers, and their corresponding metadata (which is detailed in the next section).

Time-series data: Since the entities we collected here were recent and fetched periodically, we collected dynamic information (such as the follower/followee network of Twitter users) related to these entities along with the static information. To collect time-series data, we developed a parallel task scheduler that can use multiple Twitter API keys to fetch a large volume of data at high speed. The task scheduler used the Tweepy library in Python for the API requests and ran in parallel the requests being made using the Python multiprocessing module. Multiple processes were created,

and each process was assigned a given API key to work with. The code for the Parallel Task Scheduler is available at: https://tinyurl.com/y2ztmoqg. We used the Parallel Task Scheduler to collect time-series data after every 8 hours. We collected the following data between the period of March – June 2019:

1. **Retweets:** We parsed the tweet ids of the tweets that have been posted by blackmarket customers in order to gain retweets of their own tweets, and collected the tweet objects, retweets of these tweets, and the timelines of the authors of these tweets.
2. **Twitter Followers:** We parsed the user ids of the user accounts which have been posted by blackmarket customers in order to gain followers on their accounts and collected their timelines, follower and followee networks.

The scheduler was designed in such a way that it was able to collect data for all Twitter entities after every k hours. In our case, we used $k = 8$ to collect data every 8 hours and ignored users who had more than 50,000 followers or followees due to the Twitter API rate limits. Figure 2 summarizes the steps followed to produce ABOME dataset from the blackmarket services.

Data Description

We release two different types of datasets as a part of ABOME - historical data and time-series data, as explained in the previous section.

Historical Data

We collected the metadata of each entity present in the historical data.

Twitter: We collected the following fields for retweets and followers on Twitter:

• user_details: A JSON object[7] representing a Twitter user.

• tweet_details: A JSON object[8] representing a tweet.

• tweet_retweets: A JSON list of tweet objects representing the most recent 100 retweets of a given tweet.

The details of the fields obtained from Twitter API can be found in the Twitter API Documentation (https://developer.twitter.com/en/docs.html).

YouTube: We collected the following fields for YouTube likes and comments:

• is_family_friendly: Whether the video is marked as family friendly or not.

• genre: Genre of the video.

• duration: Duration of the video in ISO 8601 format (duration type). This format is generally used when the duration denotes the amount of intervening time in a time interval.

• description: Description of the video.

• upload_date: Date that the video was uploaded.

• is_paid: Whether the video is paid or not.

• is_unlisted: The privacy status of the video, i.e., whether the video is unlisted or not. Here, the flag *unlisted* indicates that the video can only be accessed by people who have a direct link to it.

• statistics: A JSON object containing the number of dislikes, views and likes for the video.

• comments: A list of comments for the video. Each element in the list is a JSON object of the text (*the comment text*) and time (*the time when the comment was posted*).

We collected the following fields for YouTube channels:

• channel_description: Description of the channel.

• hidden_subscriber_count: Total number of hidden subscribers of the channel.

• published_at: Time when the channel was created. The time is specified in ISO 8601 format (YYYY-MM-DDThh:mm:ss.sZ).

• video_count: Total number of videos uploaded to the channel.

• subscriber_count: Total number of subscribers of the channel.

• view_count: The number of times the channel has been viewed.

• kind: The API resource type (e.g., *youtube#channel* for YouTube channels).

• country: The country the channel is associated with.

• comment_count: Total number of comments the channel has received.

• etag: The ETag of the channel which is an HTTP header used for web browser cache validation.

The historical data is stored in five directories named according to the type of data inside it. Each directory contains JSON files corresponding to the data described above.

User Type	Total	Suspended	Verified
Retweet requests	36,029	12,507	-
Follower requests	23,152	4,784	114

Table 1: Summary of Twitter users for which historical information was collected from Freemium blackmarket services.

Type	Total	Suspended
Like requests	69200	11109
Comment requests	30131	5025
Subscription requests	11282	3435

Table 2: Summary of YouTube videos and channels for which information was collected from Freemium blackmarket services.

Time-series Data

We also collect the following time-series data for retweets and followers on Twitter:

[7]https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object

[8]https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object

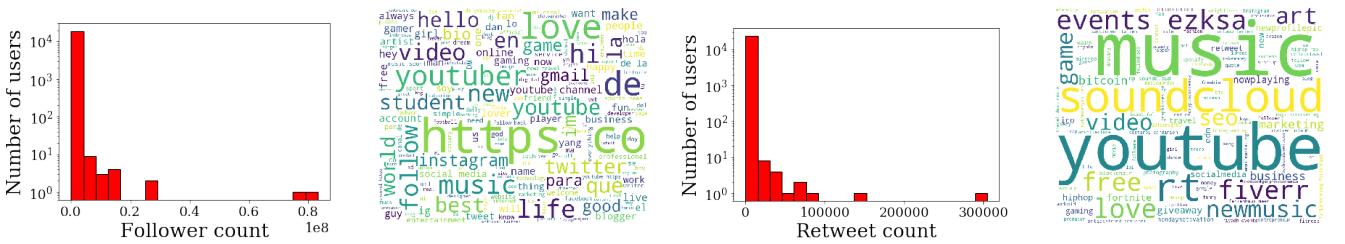

Figure 3: (a) Distribution of follower count, and (b) wordcloud aggregated over description text for users registered in blackmarket for collusive follower appraisals. (c) Distribution of retweet count, and (d) wordcloud aggregated over tweet text for tweets submitted in blackmarket for collusive retweet appraisals. Note that for clarity we remove common stopwords and single-letter words.

• `user_timeline`: This is a JSON list of tweet objects in the user's timeline, which consists of the tweets posted, retweeted and quoted by the user. The file created at each time interval contains the new tweets posted by the user during each time interval.

• `user_followers`: This is a JSON file containing the user ids of all the followers of a user that were added or removed from the follower list during each time interval.

• `user_followees`: This is a JSON file consisting of the user ids of all the users followed by a user, i.e., the followees of a user, that were added or removed from the followee list during each time interval.

• `tweet_details`: This is a JSON object representing a given tweet, collected after every time interval.

• `tweet_retweets`: This is a JSON list of tweet objects representing the most recent 100 retweets of a given tweet, collected after every time interval.

The time-series data is stored in directories named according to the timestamp of the collection time. Each directory contains sub-directories corresponding to the data described above.

Tables 1 and 3 detail the observations of the historical data and the time-series data for Twitter. It can be seen that only a very small fraction of the user accounts and tweets are no longer available on Twitter. We also observed some blackmarket customers who are marked as 'Verified' by Twitter. Table 2 details the observations of the historical data for YouTube collected from the blackmarket services. It can be seen that only a very small fraction of the channels and videos have already been removed from YouTube. This further motivates the need to develop techniques to analyze and detect collusive entities on online media platforms.

User Type	Total	Suspended	Verified
Retweet requests	4,989	492	-
Follower requests	2,350	297	28

Table 3: Summary of Twitter users for which time-series information was collected from Freemium blackmarket services.

Analysis of ABOME Data

In this section, we provide an analysis of the ABOME dataset to gain useful insights that will assist in demonstrating the opportunities opened by this new dataset.

Twitter Data. Figure 3(a) and Figure 3(b) shows the follower count distribution and wordcloud generated from the description text of users registered in blackmarket for collusive follower appraisals. We found that the maximum and average follower count was 83.28 million and 26432.28 respectively. In Fig. 3(b), we clearly see the presence of social media keywords such as "follow", "youtube", "gmail" etc. Figure 3(c) and Figure 3(d) shows the retweet count distribution and wordcloud generated from the text of tweets submitted in blackmarket for collusive retweet appraisals. We found that the maximum and average retweet count was 304442 and 131.08 respectively. Also in Fig. 3(d), along with the presence of social media keywords, we also observe some advertising keywords such as "free", "seo" etc. For the collusive tweets submitted for collusive retweet appraisals, we analyze the machine-detected language of the tweet text using langdetect library[9]. We observe that 28.55% of these tweets are written in non-english languages. The presence of multi-lingual tweets in the ABOME dataset further adds to its contributions that enable researchers to explore cross-lingual learning and also develop new tools for languages other than English for various NLP-based tasks in the area of anomaly detection research.

YouTube Data. Figure 4(a) show the distribution of (a) like count, (b) comment count, and (c) subscriber count for videos/channels submitted in blackmarket for collusive like, comment and subscription appraisals. The inset in (a) and (b) shows the corresponding distribution of the duration of the videos. Figure 5 shows the number of channels submitted to blackmarket services for collusive subscription requests from different regions. We first extract the `country` parameter for each YouTube channel in our dataset and convert into a new parameter `continent` using the PyCountry library[10]. Asia is the top region accounting for 52.6% of the total channels, followed by Europe, with 23.8% of the total channels. For the videos

[9]https://pypi.org/project/langdetect/
[10]https://pypi.org/project/pycountry/

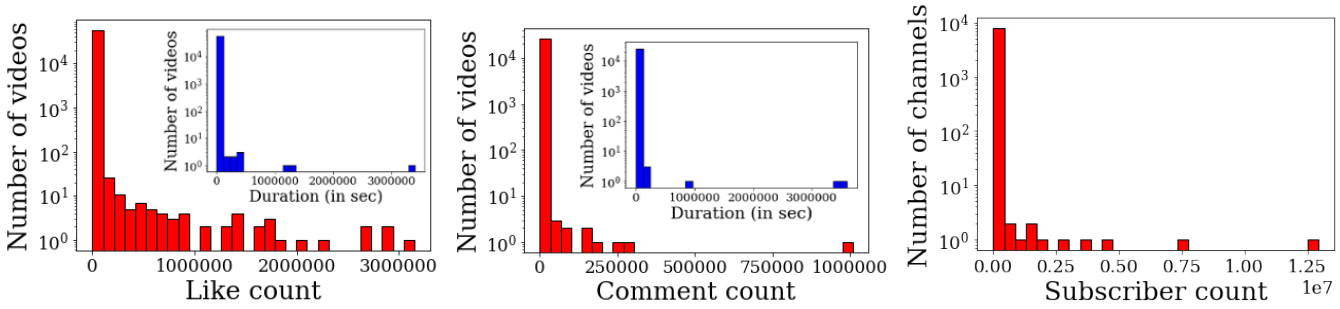

Figure 4: Distribution of (a) like count, (b) comment count, and (c) subscriber count for videos/channels submitted in blackmarket for collusive like, comment and subscription appraisals. The inset in (a) and (b) shows the distribution of the duration of the videos.

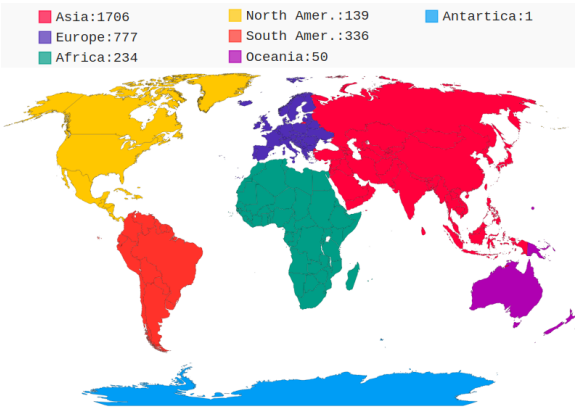

Figure 5: Region-wise count of YouTube channels submitted to blackmarket services for collusive subscription requests. Asia is the top region accounting for 52.6% of the total channels, followed by Europe, with 23.8% of the total channels.

submitted for collusive comment requests, we measured the sentiment of the comments received by the videos. The sentiment was measured using Python TextBlob library[11]. Unsurprisingly, we found that more than 95% of the comments received by these videos have a positive sentiment. Figure 6 shows a genre-wise representation of the count of views, likes, and dislikes for YouTube videos. Most of the videos for collusive requests are from the genre 'Non-profits & Activism.' The possible reason behind such a trend is that this genre allows organizations to upload videos with free premium services such as donate buttons, call-to-action overlays, live-streaming, and goal tracking, which are preferred ways to reach new as well as old audiences.

SearchBM: A Search Engine for Collusive Entity Discovery

To better aid the collusive entity discovery process and provide a better understanding of how and where the entities have been used, we developed `SearchBM`, an API and a

[11]https://textblob.readthedocs.io/en/dev/quickstart.html

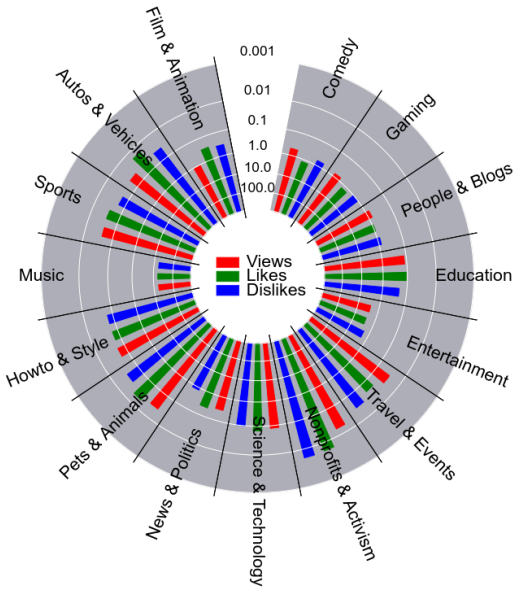

Figure 6: Genre-wise representation of views, likes and dislikes for YouTube videos registered in blackmarket services. 'Non-profits & Activism' is the top genre for collusive appraisals because of its three unique perks: (i) call to action overlays, (ii) a donation button, and (iii) Google Ad grants for paid advertising campaigns.

web portal for our end-users to effectively query within the `ABOME` dataset. Users can directly search for a query text using the interface of `SearchBM`. The query is then sent to our backend server. The backend of the server is developed using Python-Flask (http://flask.pocoo.org/). Currently, the API can accept a given text at the following request paths:

- (`<text>`,`/collusive_twitter_retweets`): This checks for presence of the query text in the tweets submitted in blackmarket services for gaining collusive retweets.
- (`<text>`,`/collusive_yt_channels`): This checks for presence of the query text in the description of YouTube channels submitted in blackmarket services for gaining collusive subscriptions.

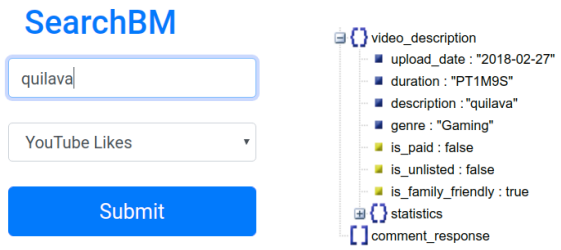

Figure 7: `SearchBM` entity query form. End-users have to enter the query text in the textbox and select one of the types from the drop-down menu.

• (`<text>,/collusive_yt_likes`): This checks for presence of the query text in video description of YouTube videos submitted in blackmarket services for gaining collusive likes.
• (`<text>,/collusive_yt_comments`): This checks for presence of the query text in the video description and user comments of YouTube videos submitted in blackmarket services for gaining collusive comments.

The API returns a JSON object indicating the presence of the text in our datasets. If the entity is found in our database, the API returns the details of the entity from our dataset. Note that to maintain anonymity, we only show specific attributes of the entity and not the identifiers (tweet/user identifier for Twitter data and video/channel identifier for YouTube). We also provide a web interface where end-users can enter the query text and select one of the services - *Twitter retweets, YouTube likes, YouTube comments and YouTube subscriptions* and get the corresponding details via the API. Fig. 7(a) shows the interface of `SearchBM`. End-users have to enter the query text in the textbox and select one of the types from the drop-down menu. On clicking the submit button, the query parameters are sent as a GET request to our API, which returns a JSON object indicating the presence of text in the collusive entity. The front-end uses a JSON viewer, as shown in Fig. 7(b) to display the entity details.

Research Opportunities Using the ABOME Dataset

We believe that `ABOME` can can benefit in many threads of anomaly detection research. We discuss a few examples below:

1. **Fraudulent user/entity detection:** A great deal of work has been devoted to fraudulent user/entity detection in online media platforms. The task of detecting fraudulent users includes identifying fake users (Gupta et al. 2013; Atodiresei, Tănăselea, and Iftene 2018; Fire et al. 2014), spammers (Miller et al. 2014; Benevenuto et al. 2010), bots (Chavoshi, Hamooni, and Mueen 2016a,b; Dickerson, Kagan, and Subrahmanian 2014), collusive users (Dutta et al. 2018; Dutta and Chakraborty 2019; Chetan et al. 2019; Arora, Paka, and Chakraborty 2019; Dutta et al. 2020), sockpuppets (Kumar et al. 2017) etc. Most of the above algorithms only detect individual users. However, in reality, it is seen that the anomalous phenomena also occur in groups. The group detection task (finding a group of users that jointly exhibit fraudulent behavior) is more difficult as compared to the individual detection task due to the variation present in the inter-group dynamics. In our case, the users of freemium blackmarket services perform actions (retweet/like/comment) on collusive entities in order to gain credits. Therefore, it is almost certain that users in `ABOME` must have interacted with each other in order to gain credits. We believe that the availability of the `ABOME` dataset can foster fraudulent user/entity detection approaches (both individual and group) with the advantage of adding the topical as well as the temporal dimension.
2. **Mining connectivity patterns:** Understanding connectivity patterns of an underlying network is a well-studied problem in the literature. It include tasks such as inferring lockstep behavior (Beutel et al. 2013), dense block detection (Shin, Hooi, and Faloutsos 2016), detecting core users (Shin, Eliassi-Rad, and Faloutsos 2016), identifying the most relevant actors in a network (Borgatti 2006), sudden appearance/disappearance of links (Eswaran et al. 2018) etc. Using the `ABOME` dataset, researchers can create various networks among the users/entities present in the dataset to investigate various structural patterns of the network.
3. **Modeling temporal evolution:** As `ABOME` dataset contains time-series data, it can be used for various temporal modeling tasks. Some example tasks include detecting time periods containing unusual activity (Giatsoglou et al. 2015), identifying repetitive patterns in time-evolving graphs (Zeidanloo and Manaf 2010) etc.
4. **Diffusion modeling:** The `ABOME` dataset can be represented as networks based on the actions performed by the collusive users on the content of other users of the services. It could be then use to study multiple diffusion modeling tasks such as influence maximization (selecting a seed set to maximize the influence spread) (Jendoubi et al. 2017; Mei, Zhao, and Yang 2017), predicting information cascade (Rattanaritnont, Toyoda, and Kitsuregawa 2011), measuring message propagation and social influence (Ye and Wu 2010; Brown and Feng 2011) etc.
5. **Event-specific studies:** As `ABOME` contains data obtained from multiple sources and spans over a long period of time, it may consist of information from many major events (Atefeh and Khreich 2015), which can be easily extracted for event-centric studies. Researchers can also check how these users/entities were involved in manipulating the popularity of events by artificially inflating the social growth of users/entities in online media (Zhang et al. 2016).
6. **Multi-lingual studies:** In the previous section, we mentioned the presence of multilingual texts in the `ABOME` dataset. We anticipate that the multilingual data will be useful for a broad range of Natural Language Processing

(NLP) tasks in the anomaly detection domain.

How ABOME Is a FAIR-Compliant Dataset?

In this section, we explain how we have made the ABOME dataset compliant to the four FAIR data principles: *Findable, Accessible, Interoperable and Re-usable*.

To make the ABOME dataset *Findable* and *Accessible*, our dataset is publicly available on Zenodo which allows downloading the entire dataset with the following citation: Hridoy Sankar Dutta, Udit Arora & Tanmoy Chakraborty. (2021). ABOME: A Multi-platform Data Repository of Artificially Boosted Online Media Entities [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4437987.

To make the ABOME dataset *Interoperable* and *Re-usable*, the dataset files are provided in standard JSON (JavaScript Object Notation) format that can be easily parsed using any standard JSON parser and can be exported to other data formats like CSV (Comma Separated Values), XML (Extensible Markup Language) etc. We also provide a readme file to optimize the re-use of the dataset.

Conclusion

Collusive entity detection is an important problem that has largely been overlooked. To the best of our knowledge, ABOME is the first dataset in the literature that consists of multi-platform data related to blackmarket-driven collusive entities collected from two credit-based freemium services - YouLikeHits and Like4Like. In addition, we also designed an API and a web portal, SearchBM to discover collusive entities using text search queries. We believe that the datasets released in this paper will provide more opportunities for researchers to advance the development of technologies in detecting collusive entities in online media, thereby creating an adequate social space. We also encourage researchers working in the domain of privacy and security in OSNs to propose interesting tasks and use ABOME as a benchmark.

Acknowledgements

The work was partially supported by ECR/2017/00169 (SERB). T. Chakraborty would like to thank the Ramanujan Fellowship (SERB) and the CAI, IIIT-Delhi.

References

Alsaleh, M.; Alarifi, A.; Al-Salman, A. M.; Alfayez, M.; and Almuhaysin, A. 2014. Tsd: Detecting sybil accounts in twitter. In *2014 13th International Conference on Machine Learning and Applications*, 463–469. IEEE.

Arora, U.; Dutta, H. S.; Joshi, B.; Chetan, A.; and Chakraborty, T. 2020. Analyzing and Detecting Collusive Users Involved in Blackmarket Retweeting Activities. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11(3): 1–24.

Arora, U.; Paka, W. S.; and Chakraborty, T. 2019. Multitask learning for blackmarket tweet detection. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 127–130.

Atefeh, F.; and Khreich, W. 2015. A survey of techniques for event detection in twitter. *Computational Intelligence* 31(1): 132–164.

Atodiresei, C.-S.; Tănăselea, A.; and Iftene, A. 2018. Identifying fake news and fake users on Twitter. *Procedia Computer Science* 126: 451–461.

Benevenuto, F.; Magno, G.; Rodrigues, T.; and Almeida, V. 2010. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, 12.

Beutel, A.; Xu, W.; Guruswami, V.; Palow, C.; and Faloutsos, C. 2013. Copycatch: stopping group attacks by spotting lockstep behavior in social networks. In *Proceedings of the 22nd international conference on World Wide Web*, 119–130.

Borgatti, S. P. 2006. Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory* 12(1): 21–34.

Brown, P. E.; and Feng, J. 2011. Measuring user influence on twitter using modified k-shell decomposition. In *Fifth international AAAI conference on weblogs and social media*.

Chavoshi, N.; Hamooni, H.; and Mueen, A. 2016a. DeBot: Twitter Bot Detection via Warped Correlation. In *ICDM*, 817–822.

Chavoshi, N.; Hamooni, H.; and Mueen, A. 2016b. Identifying correlated bots in twitter. In *International Conference on Social Informatics*, 14–21. Springer.

Chetan, A.; Joshi, B.; Dutta, H. S.; and Chakraborty, T. 2019. CoReRank: Ranking to Detect Users Involved in Blackmarket-Based Collusive Retweeting Activities. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 330–338. ACM.

Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; and Tesconi, M. 2015. Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems* 80: 56–71.

Dhawan, S.; Gangireddy, S. C. R.; Kumar, S.; and Chakraborty, T. 2019. Spotting Collusive Behaviour of Online Fraud Groups in Customer Reviews. *arXiv preprint arXiv:1905.13649* .

Dickerson, J. P.; Kagan, V.; and Subrahmanian, V. 2014. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, 620–627. IEEE.

Dutta, H. S.; and Chakraborty, T. 2019. Blackmarket-Driven Collusion Among Retweeters–Analysis, Detection, and Characterization. *IEEE Transactions on Information Forensics and Security* 15: 1935–1944.

Dutta, H. S.; and Chakraborty, T. 2020. Blackmarket-driven Collusion on Online Media: A Survey. *arXiv preprint arXiv:2008.13102* .

Dutta, H. S.; Chetan, A.; Joshi, B.; and Chakraborty, T. 2018. Retweet us, we will retweet you: Spotting collusive retweeters involved in blackmarket services. In *2018 IEEE/ACM*

International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 242–249. IEEE.

Dutta, H. S.; Dutta, V. R.; Adhikary, A.; and Chakraborty, T. 2020. HawkesEye: Detecting fake retweeters using Hawkes process and topic modeling. *IEEE Transactions on Information Forensics and Security* 15: 2667–2678.

Eswaran, D.; Faloutsos, C.; Guha, S.; and Mishra, N. 2018. Spotlight: Detecting anomalies in streaming graphs. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1378–1386.

Fire, M.; Kagan, D.; Elyashar, A.; and Elovici, Y. 2014. Friend or foe? Fake profile identification in online social networks. *Social Network Analysis and Mining* 4(1): 194.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumé III, H.; and Crawford, K. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* .

Giatsoglou, M.; Chatzakou, D.; Shah, N.; Faloutsos, C.; and Vakali, A. 2015. Retweeting activity on twitter: Signs of deception. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 122–134. Springer.

Gupta, A.; Lamba, H.; Kumaraguru, P.; and Joshi, A. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, 729–736. ACM.

Gupta, S.; Kumaraguru, P.; and Chakraborty, T. 2019. MalReG: Detecting and Analyzing Malicious Retweeter Groups. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, 61–69. ACM.

Hooi, B.; Shah, N.; Beutel, A.; Günnemann, S.; Akoglu, L.; Kumar, M.; Makhija, D.; and Faloutsos, C. 2016. Birdnest: Bayesian inference for ratings-fraud detection. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, 495–503. SIAM.

Jendoubi, S.; Martin, A.; Liétard, L.; Hadji, H. B.; and Yaghlane, B. B. 2017. Two evidential data based models for influence maximization in twitter. *Knowledge-Based Systems* 121: 58–70.

Kumar, S.; Cheng, J.; Leskovec, J.; and Subrahmanian, V. 2017. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web*, 857–866.

Kumar, S.; and Shah, N. 2018. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559* .

Li, Y.; Martinez, O.; Chen, X.; Li, Y.; and Hopcroft, J. E. 2016. In a world that counts: Clustering and detecting fake social engagement at scale. In *Proceedings of the 25th International Conference on World Wide Web*, 111–120. International World Wide Web Conferences Steering Committee.

Liu, S.; Hooi, B.; and Faloutsos, C. 2017. Holoscope: Topology-and-spike aware fraud detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1539–1548. ACM.

Mei, Y.; Zhao, W.; and Yang, J. 2017. Influence maximization on twitter: A mechanism for effective marketing campaign. In *2017 IEEE International Conference on Communications (ICC)*, 1–6. IEEE.

Miller, Z.; Dickinson, B.; Deitrick, W.; Hu, W.; and Wang, A. H. 2014. Twitter spammer detection using data stream clustering. *Information Sciences* 260: 64–73.

Pierri, F.; and Ceri, S. 2019. False News On Social Media: A Data-Driven Survey. *arXiv preprint arXiv:1902.07539* .

Rattanaritnont, G.; Toyoda, M.; and Kitsuregawa, M. 2011. A study on characteristics of topicspecific information cascade in Twitter. In *Forum on Data Engineering (DE2011)*, 65–70.

Sankar Dutta, H.; Jobanputra, M.; Negi, H.; and Chakraborty, T. 2020. Detecting and analyzing collusive entities on YouTube. *arXiv* arXiv–2005.

Shah, N.; Beutel, A.; Gallagher, B.; and Faloutsos, C. 2014. Spotting suspicious link behavior with fbox: An adversarial perspective. In *2014 IEEE International Conference on Data Mining*, 959–964. IEEE.

Shah, N.; Lamba, H.; Beutel, A.; and Faloutsos, C. 2017. The many faces of link fraud. In *2017 IEEE International Conference on Data Mining (ICDM)*, 1069–1074. IEEE.

Shin, K.; Eliassi-Rad, T.; and Faloutsos, C. 2016. Corescope: Graph mining using k-core analysispatterns, anomalies and algorithms. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 469–478. IEEE.

Shin, K.; Hooi, B.; and Faloutsos, C. 2016. M-zoom: Fast dense-block detection in tensors with quality guarantees. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 264–280. Springer.

Stringhini, G.; Wang, G.; Egele, M.; Kruegel, C.; Vigna, G.; Zheng, H.; and Zhao, B. Y. 2013. Follow the green: growth and dynamics in twitter follower markets. In *Proceedings of the 2013 conference on Internet measurement conference*, 163–176. ACM.

Thomas, K.; Grier, C.; Song, D.; and Paxson, V. 2011. Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, 243–258. ACM.

Yardi, S.; Romero, D.; Schoenebeck, G.; et al. 2010. Detecting spam in a twitter network. *First Monday* 15(1).

Ye, S.; and Wu, S. F. 2010. Measuring message propagation and social influence on Twitter. com. In *International conference on social informatics*, 216–231. Springer.

Zeidanloo, H. R.; and Manaf, A. B. A. 2010. Botnet detection by monitoring similar communication patterns. *arXiv preprint arXiv:1004.1232* .

Zhang, Y.; Ruan, X.; Wang, H.; Wang, H.; and He, S. 2016. Twitter trends manipulation: a first look inside the security of twitter trending. *IEEE Transactions on Information Forensics and Security* 12(1): 144–156.