Towards Characterizing Coordinated Inauthentic Behaviors on YouTube

Conference Paper · April 2023

CITATIONS
25

READS
113

3 authors, including:

Oluwaseyi Adeliyi
University of Arkansas at Little Rock
8 PUBLICATIONS **35** CITATIONS
SEE PROFILE

Nitin Agarwal
University of Arkansas at Little Rock
305 PUBLICATIONS **3,375** CITATIONS
SEE PROFILE

Towards Characterizing Coordinated Inauthentic Behaviors on YouTube

Baris Kirdemir , Oluwaseyi Adeliyi, and Nitin Agarwal

COSMOS Research Center, UA Little Rock, Little Rock, AR, USA

Abstract

Online social networks and information consumed by their users have been increasingly targeted and altered by manipulative campaigns in recent years. YouTube is one of the most popular websites in the world. However, most of the existing systems and studies to detect and characterize manipulative influence campaigns focus on other platforms, while very little is known about the potential ways to detect and mitigate such attacks on YouTube. Furthermore, although recent literature is significantly developed in terms of assessing and detecting individual suspicious accounts across online social networks, more research is needed to detect, predict, and characterize large-scale coordinated campaigns in different contexts. This makes the analysis and detection of coordinated suspicious and inorganic activities on the given platform vital for researchers, policymakers, journalists, and more. In this paper, we report our study in progress to assess and characterize such suspicious activity on the level of YouTube channels, combining multiple layers of rolling window correlation analysis, anomaly detection, peak detection, rule-based supervised classification, network feature engineering, and unsupervised clustering approaches. Overall, the experimental dataset amounted to 39 channels, 936,247 videos, 99,415,476 comments, 115,825,225 subscribers and over 51 billion views. The results show that channels exhibiting inauthentic activities are characterized by a relatively lesser number of peaks in their anomaly patterns. However, the magnitude of these peaks is usually higher compared to that of less suspicious channels. Also, coordination assessment based on network structures and features produces promising results for identifying clusters of suspicious behaviors across channels.

Keywords [1]

Coordination, inauthentic behavior, YouTube, anomaly detection, social network analysis, network features

1. Introduction

Coordinated inauthentic campaigns use a wide variety of tactics, techniques, and procedures (TTPs) to manipulate information as well as networks of communication across social media platforms. Given their potential as well as proven effects on human behavior, beliefs, emotions, and attitude, such campaigns lead to social, political, economic, financial, and psychological harm, covering both online and offline realms. Within the last decade, many studies have used a variety of computational methods to describe, explain, and predict inauthentic activities and manipulative campaigns online, with a special focus on the detection and characterization of social media accounts [5, 9, 18], dynamics of disinformation diffusion [6, 30, 34], characterization of TTPs and narratives [1, 4, 24, 32], and assessment of their broader implications. In particular, a significant portion of such data-driven systems and scientific research aimed to detect automated or semi-automated social media accounts. Nevertheless, the literature and existing counter-disinformation toolkits still lack comprehensive approaches that would enable a better understanding of how coordinated inauthentic campaigns occur and how they can be assessed, characterized, and detected on less-studied yet popular and influential social media platforms.

[1] ROMCIR 2022: The 2nd Workshop on Reducing Online Misinformation through Credible Information Retrieval, held as part of ECIR 2022: the 44th European Conference on Information Retrieval, April 10-14, 2022, Stavanger, Norway

YouTube, as one of the most popular social media platforms across the world, has been a particularly influential medium of choice for video sharing, news consumption, content monetization, political activism, and cross-platform information dissemination. Thus, for many inauthentic, hostile, and coordinated manipulation campaigns, the platform constitutes an important channel [15]. Coordinated manipulative activities often involve inorganic boosting of explicit engagement metrics such as views, likes, and comments, as such services are also provided openly or in the online black markets by many commercial entities [28]. However, due to a variety of reasons (i.e., data availability), most of the existing scientific literature and systems for countering online information manipulation focus or rely on data collected from other platforms, such as Twitter, that have a significantly different platform architecture in comparison to YouTube.

In addition, the timely detection and characterization of coordination remain a significant research problem overall. Existing literature has documented significant progress in the detection of individual automated accounts (social bots, mostly on Twitter) and characterization of disinformation events by tracing their artifacts and historical data. On the other hand, cross-campaign variety of TTPs and evolving technologies of manipulation decrease the performance, effectiveness, and accuracy of approaches focusing on individual actors and automated accounts. Behavioral and content-based characteristics of individual accounts change over time and across campaigns, leading to a drop of performance in machine learning systems. Also, coordinated information manipulation includes both human and automated accounts with varying distributions. Therefore, as also argued by Cresci [5] and Khaund et al. [18], future studies should move beyond the current approaches and improve the understanding and timely detection of anomalous coordination.

In this paper, we present our study in progress for the characterization of coordinated inauthentic or suspicious behaviors on YouTube. Using a sample of YouTube channels and their historical data, we explore the performance and utility of two distinct but interrelated methodological approaches in the prediction of suspicious coordinated activity on the channel level. First, we demonstrate a multi-step time-series analysis of engagement trends and a combination of unsupervised and supervised machine learning experiments. Second, we use co-commenter networks as an implicit artifact of coordination and build a set of features that would signal high-level coordination and suspicious behavior. Furthermore, we report and discuss the initial findings of unsupervised clustering of YouTube channels based on a time series analysis of engagement trends and network features. We posit that by improving the current understanding of coordinated suspicious activities and their artifacts, our study may lead to the development of an accurate and effective methodology for the timely detection of harmful manipulative campaigns on YouTube. Next, we introduce a survey of relevant literature.

2. Related Work

YouTube is one of the most popular online social media platforms for disseminating information, but it has also been exploited by bad actors for spreading false or misleading narratives [10]. Researchers have studied disinformation campaigns on various social media platforms, associated narratives [31], and their influence on human behavior [3]. Recent studies have extended this research into YouTube, analyzing crowd manipulation strategies on the platform [15]. In a bid to uncover inorganic behaviors within YouTube channels, the researchers analyzed the video posting behavior of a YouTube channel with conspiracy theory videos as well as its user engagement statistics. YouTube's policy states that the platform does not permit any activities that increase the number of views, likes, comments, or other metrics artificially, either by serving videos to unsuspecting viewers or through other automated systems [36]. However, these activities have grown in popularity on YouTube in recent years, increasing the engagement statistics of some channels and fueling disinformation campaigns. The problem is compounded when such behaviors tap into the biases of a platform's search and recommendation algorithms [19, 20].

Dutta et al. analyzed collusive entities on YouTube, studying black market services and fraud/spam detection on online media platforms [7]. The researchers studied a major black-market service used to gain appraisals inorganically for YouTube channels, from video views to likes, subscribers, and comments. They also investigated spam comments from a collusion perspective. The popularity of YouTube videos is dependent on the amount of implicit and explicit engagement it receives from the

content consumers and with the advent of monetization on the platform, some content creators are motivated to use inorganic means to get appraisals for their content. Our research, however, does not study the motivations of the content creators, but the inauthentic behaviors exhibited by the YouTube channels and utilize tactics towards gaining inorganic engagement. These behaviors are uncovered through the analysis of user engagement statistics and network data from YouTube channels.

Recent literature indicates that the study of coordination would enable an improved understanding and more accurate characterization of modern manipulative information campaigns. Starbird et al. [29] argue that a significant portion of modern information manipulation campaigns is "participatory" and "collaborative" in nature, using target audiences also for further dissemination of amplified narratives. Moreover, modern information campaigns exhibit organizational and structural variance leading to changing characteristics of information dissemination. Kumar et al. [22], Hine et al. [14], and Cresci [5] demonstrate the significance of group-level coordination in the study of inauthentic and manipulative information campaigns. Recently documented coordination assessment methodologies also focused on inauthentic boosting of web links [12], mass astroturfing [20], and other deviant mob-like behaviors that interfere in political processes or dissemination of health-related information [13, 26].

Recent studies proposed several network-based approaches to examine coordination, with varying levels of maturity and readiness for training machine learning/classification models. Pacheco et al. [25] introduced an unsupervised and network structure-based methodology to detect coordinated communities on social media. The authors used a multi-step network analysis approach to uncover tightly knit clusters signaling coordination. Weber and Neumann [35] proposed a temporal window approach using user account interactions and metadata. They focus on group-level activity, "regardless of their degree of automation", that exhibits "anomalously high levels of coordinated behavior". Using a set of coordination behavior artifacts, they present a slightly different version of the previously reported FSA algorithm [27], FSA V, to distinguish Highly Coordinated Communities (HCCs) from the rest of the network. They validate the results of the given approach with three supervised classifiers to compare HCCs between campaigns and with ground truth data.

Moving beyond the detection of individual accounts to campaign-level activities, Vargas et al. [33] demonstrated one of the very few studies testing the feasibility of coordination network analysis and features to detect coordinated disinformation campaigns on Twitter. They trained a binary classifier to detect "Strategic Information Operations" (Twitter) against baseline activity of various "legitimate" coordinated networks. They established the baseline activity from "communities exhibiting varying levels of coordination" rather than random topic networks of Twitter accounts, aiming to have a better representation of real-world campaigns. Their results showed that supervised classifiers based on coordination network features perform well in predicting future instances of "same" coordinated campaigns. However, the performance scores drop significantly when predicting previously unseen campaigns, indicating the significance of the cross-campaign variety of tactics employed in information operations [33].

This study offers multiple original contributions to the existing literature. As stated in the previous sections, the existing literature still lacks an overarching methodological framework that would enable the assessment, characterization, and detection of coordinated suspicious behaviors on YouTube. Also, due to YouTube's unique platform architecture and its black-box characteristics, coordinated suspicious activity on the platform often includes multiple implicit and latent features. The following sections describe our multi-layered approach to tackling the given problem. We also present the initial results of the documented methodology and their potential implications for future work on the timely detection of such campaigns.

3. Data and Methods

To uncover channel-level suspicious activity on YouTube, we employed two primary methodologies consisting of a multi-step time-series analysis of engagement trends and the analysis of structural properties of co-commenter networks as potential artifacts of coordinated suspicious behavior. For the first, we processed metrics of video production and engagement (number of video postings, number of views, number of comments, and number of channel subscribers) through a multi-step analytical

pipeline including rolling window correlation analysis, anomaly detection, peak detection, rule-based classification, principal component analysis (PCA), and unsupervised clustering. Second, we constructed co-commenter networks using the comment data collected from YouTube channels and explored the utility of network structural features (group-level features, in particular) in the identification of suspicious clusters of YouTube channels.

3.1. Data Collection

We used the VTracker tool [23] and procedures described in Kready et al. [21] for collecting the number of daily video postings, the number of comments, and comment-specific data (comment content, commenter id, commented video id, timestamps), using the YouTube Data API [2]. These data were collected in accordance with YouTube's Terms of Service and data collection guidelines [37]. To collect daily subscriber counts and the number of daily video views we used Social Blade API [11] which also provides public YouTube data in accordance with YouTube's guidelines [38]. Overall, our experimental dataset consisted of 39 YouTube channels. The topical categories in the experimental dataset included news, defense and security, education, and entertainment, while the activity timeline ranged from November 2017 to July 2021. The categorical variety of the experimental channels ranged from highly popular news sources (Fox News, CNN) to football clubs (Barcelona FC) and suspicious channels previously discovered as actively engaging in geopolitical influence campaigns [15]. Data preprocessing included the elimination of missing values in engagement metrics, computation of the total number of views, the total number of subscribers, the total number of comments, and the total number of video postings for each channel. In addition, we used anonymized commenter ids and video ids for building the co-commenter networks, as discussed in the following sections.

Table 1
YouTube Dataset Statistics

Data Elements	Head 2
Channels	39
Videos	936,247
Comments	99,415,476
Commenters	21,067,211
Views	51,608,630,100
Subscribers	115,825,225

3.2. Characterization based on Engagement Trends

Rolling window Correlation Analysis. We grouped the data into rolling windows of 100 days and computed the pairwise correlation between total views, total subscribers, total comments, and total videos for each window. This was done to capture inauthentic behaviors such as a channel with decreasing subscribers but increasing views or, conversely, increasing subscribers but decreasing views. The resulting data comprised start and end dates for each window, with the values of the correlation pairs - views and subscribers, views and videos, views and comments, subscribers and videos, subscribers and comments, videos and comments.

Anomaly Detection. The output of the rolling window correlation analysis was used to train a long short-term memory (LSTM) model on the time series data. The LSTM model was run through each dataset with the Mean Squared Error loss function and Adam optimizer for loss function optimization. A batch size of 32 was used due to the size of each dataset (approximately 1000 data points), with a lookback size of 1. This lookback size was selected because a single data point represented a window of 100 days. Each model was then trained over 30 epochs, with an average training loss of 13%. We represented the losses from the computation as anomaly confidence scores and plotted the output on a 2-D plot. Figure 1 shows the sample output of the anomaly detection step for a channel; on the left we see a steady pattern in the correlation between the views and subscribers statistics for the channel and on the right, a peak can be seen in the correlation between the views and videos statistics within the

period of September 2019 and January 2020. This strong peak indicates an anomaly in the correlation between the views and videos. To capture these anomalous periods, we set an anomaly threshold (based on the anomaly confidence score) for each channel's correlation pairs to capture all data points that were placed above that threshold.

The given procedure resulted in an anomaly list for each channel capturing the channel id, start and end dates, duration (of the anomalous period), minimum correlation, and maximum anomaly score for all six correlation pairs (also referred to as indicators).

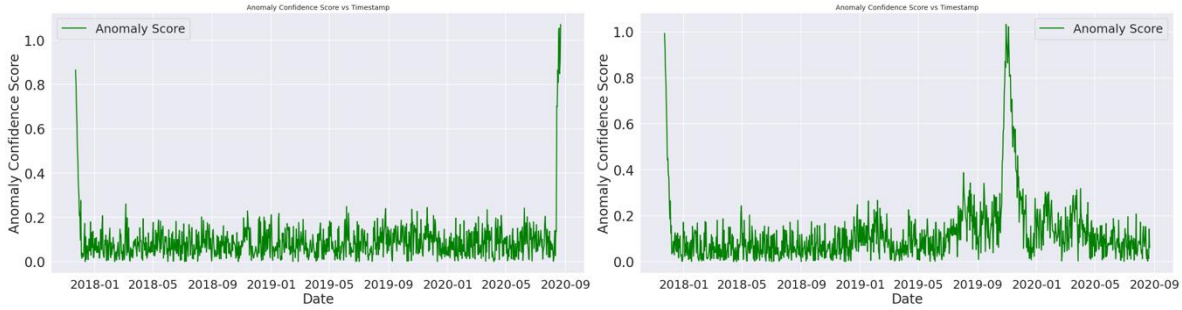

Figure 1: Channel anomalies for views vs. subscribers (left) and views vs. videos (right).

Table 2
A subset of channel statistics for views vs. subscribers and views vs. videos

Start date	End date	Duration	Average views	Average subscribers	Minimum correlation	Maximum anomaly score	Average Error Sum of Square (SSE)
2019-10-23	2020-03-10	139 days	208,501,054	393,125	0.877355	0.975601	0.017192
2020-08-14	2020-11-29	107 days	274,492,331	6,107	-0.3407	1.041448	0.39364

Peak Detection. We used the SciPy library [28] which takes a 1-D array and finds all local maxima by comparing neighboring values to detect the peaks in the data represented by the 2D plot from the anomaly detection. As a result of the noise in the chart, a lot of peaks were being detected. Hence, we applied smoothening to the data to reduce noise and capture the significant peaks. Figure 2 shows the peaks after smoothening. We then analyzed the number and intensity of the peaks across all the channels, characterizing peaks from very high intensity to very low intensity.

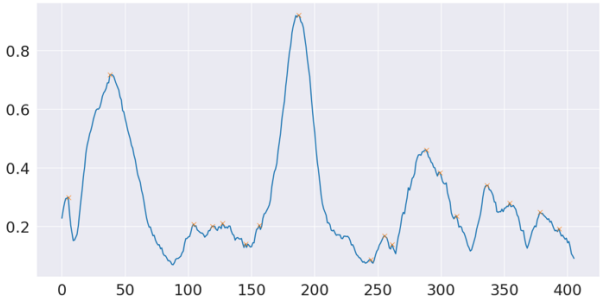

Figure 2: Peak detection after smoothening data.

Rule-based Classification. As a result of YouTube's actions on channels that exhibit inauthentic behaviors, obtaining a validation set comprising such channels could be a challenging task as these channels are sometimes taken down without a reason known to the public. Therefore, based on the features generated from the anomaly detection step, we annotated the dataset and created a rule-based classification algorithm to determine a suspicion score for each indicator, with a range of (0,1) where 0 represents the least suspicious and 1 represents the most suspicious. The suspicion scores across all six

indicators were then aggregated (by assigning weights to each indicator) to create a single suspicion score for each observation. Table 3 explains the weights assigned to each indicator based on their respective importance in detecting inauthentic behaviors. We determined these weights by examining the semantics behind each indicator and the degree to which a positive and negative correlation between the indicators points to suspicious behavior on YouTube. The weighting scheme used is explained as follows:

Views vs Subscribers: A positive correlation implies that views increased on the channel as subscribers also increased. This is expected of YouTube channels; however, subscribers may grow at a slower rate than views, but a positive correlation is still expected. A negative correlation implies that views decreased on the channel as subscribers increased or vice versa which is unusual behavior for a channel. While there are cases where unpopular channels upload a viral video, views could increase drastically while subscribers remain the same or grow at a slower rate compared to views, but rarely do we see these numbers go in the opposite direction. Therefore, this indicator was assigned a high ranking for judging inauthentic behavior.

Views vs Videos: A positive correlation implies that views increased on the channel as videos increased and vice versa. While a channel might upload multiple videos within a period, the views may not necessarily increase at the same upload rate and may take longer to catch up. Also, for videos that go viral, a channel might upload lesser videos for the specific video to gain more views while the video uploads remain the same. A negative correlation implies that videos decreased while views increased on a channel, which is plausible for channels that choose to upload lesser videos. It is also possible for the opposite to occur, where a channel racks up lesser views compared to video uploads. Therefore, this indicator was assigned a medium ranking for judging inauthentic behavior.

Views vs Comments: A positive correlation implies that views increased on the channel as comments increased or vice versa. This is a common scenario as more views are acquired, more user engagements are expected. However, for a negative correlation, we consider a channel where videos have been uploaded and views have been acquired, users could engage for longer periods of time even after the video has stopped gaining views. This is common on YouTube as commenters reply to comments even when views have stopped growing. Also, some videos could have comments disabled and still acquire a lot of views. Therefore, this indicator was assigned a low ranking for judging inauthentic behavior.

Subscribers vs Videos: A positive correlation implies that subscribers increased on a channel as videos increased or vice versa. For popular channels that are consistent with uploads, this is a common pattern as they increase upload rate and acquire more subscribers along the way. For a negative correlation, we consider a case where a channel could increase uploads but remain on the same number of subscribers or gain a lesser number of subscribers. There are also cases where viewers watch a video they like and subscribe to the channel while the channel owner has not uploaded a video in a while. Therefore, this indicator was assigned a low ranking for judging inauthentic behavior.

Subscribers vs Comments: A positive correlation implies that comments increased on a channel as subscribers increased or vice versa. This is common among channels as more user engagement is expected as the number of subscribers increases (because of more people watching the videos). For a negative correlation, we consider viral videos or videos recommended by YouTube's recommendation algorithm, where user engagement could continue to increase as more people watch the video, but these viewers may not necessarily subscribe to the channel. Subscribers can also increase with comments decreasing in a case where comments are disabled for some of these videos. However, these cases are less common. Therefore, this indicator was assigned a high ranking for judging inauthentic behavior.

Videos vs Comments: A positive correlation implies that comments increased on a channel as videos increased or vice versa. This is common among channels as user engagement increases as more videos are uploaded but in the case of a viral video, comments could increase quicker than more videos are uploaded to the channel. For a negative correlation, we consider a scenario where engagement could decrease as more videos are uploaded to a channel because of disabled comments for some videos. The alternate case of comments increasing while video uploads decrease is also possible but is not always common. Therefore, this indicator was assigned a high ranking for judging inauthentic behavior.

These rankings were used to assign weights to each indicator to combine the suspicion scores under each indicator and arrive at an overall suspicion score. To assign weights to the indicators, we used the indicator rankings to represent the importance of each indicator and then assigned each indicator a value

representing a fraction of the overall suspicion score. The suspicion score for each indicator was then multiplied by its weight and summed up to obtain a single suspicion score, with a range of (0,1).

Table 3
Indicators and their weight assignments.

Indicators	Weights
Views vs Subscribers	0.35
Videos vs Comments	0.25
Subscribers vs Comments	0.2
Views vs Videos	0.1
Views vs Comments	0.06
Subscribers vs Videos	0.04

Principal Component Analysis. The output of the anomaly detection returned anomalous periods with a large feature set and as a result, we used Principal Component Analysis (PCA) to reduce the dimensions of the dataset and created a scatterplot using the first two principal components.

Clustering. We visually identified clusters from the PCA scatter plot to group channels based on their engagement trends. We also utilized DBSCAN, a density-based clustering algorithm, to automatically identify the channel clusters [9]. This was done because of the distribution of the data points and also to verify that the clusters that were identified manually were computationally accurate.

3.3. Characterization based on Network Structures

To trace network artifacts of activity, we built a co-commenter network for each channel in our experimental list. We assume a connection (edge) between pairs of commenters if they commented on the same video. To declutter the network and focus on the suspicious network clusters, we filtered out commenter (node) pairs (low-weight edges) if they co-commented on less than 10 videos. This threshold number is adjustable in practice. Nevertheless, because of our initial experiments, we fixed the current threshold for co-commenter pairs to 10 to cover as many channels as possible and include edges that signal frequent interactivity, while eliminating the random chance of co-commenter pairings. Nodes are commenters in the final network of co-commenters, and the number of nodes (network size) varies between different channels. Edges represent the connections between nodes, i.e., co-commenting behavior of commenters.

Furthermore, we computed a set of network features in the co-commenter networks that would potentially enable the detection of suspicious clusters and behavior on the level of YouTube channels. We combined two major categories of features. The first set of metrics are extracted from the co-commenter networks by using well-established graph measures, including average degree, number of nodes, number of edges, average clustering coefficient, and modularity.

Second, we added an additional set of metrics by computing descriptive statistics and engineering features in relation to maximal cliques. A clique in a graph is a tightly knit sub-network in which all of its members are directly connected to each other. In network analysis terms, a maximal clique has the maximum number of members it can contain and can't be expanded further with additional members in the recursive computation process. Similar to the threshold we used for co-commenter networks, we counted maximal cliques only if they have at least five members. We observed a high level of variation between the channels in our experimental list in terms of the average clique sizes and the total number of cliques. Although the total number of cliques is correlated with the size of the full co-commenter graph, the variation of clique sizes implies increased and suspicious network activity in some channels. Finally, apart from a simple count of maximal cliques, we used additional clique-based features by computing the average degree of clique members, the average clustering coefficient of clique members, and the median clique size for each channel. In addition, we included comparative features showing the relationship between the given clique-based metrics with network measures extracted from the entire co-commenter network for each channel.

After building the co-commenter networks and computing the set of network and clique-based features as described, we ran a feature similarity analysis using the Pearson correlation values. Furthermore, using an unsupervised approach, we checked the feasibility of PCA, k-means clustering, and hierarchical clustering methods with normalized feature values to examine if the current set of network features and well-known clustering and dimensionality reduction algorithms lead to meaningful clusters of the channels in the experimental list, while also enabling unsupervised detection of channels with traces of suspicious and coordinated network activity. We note that our current analysis of co-commenter networks does not include any temporal limitations. In the following phases of this study, we may also experiment with features extracted from network activity in limited timeframes, adding a time series aspect to our current approach.

4. Analysis and Findings

We discuss our findings in this section. First, we discuss the findings from our engagement trend-based analysis and then the findings from network structural feature-based analysis will be discussed.

4.1. Characterization based on Engagement Trends

The scatter plot in Figure 3 shows how the observations are spread across the two principal components with a color map indicating the suspicion score of each observation. The suspicion score is computed using the method described in Section 3.2. From the scatter plot, we visually identified five clusters, as detailed in Table 4. This approach characterizes channels not only based on suspicion but also allows us to identify channels with similar engagement trends and channels that deploy similar tactics to grow their engagement statistics inorganically. The first cluster comprises mostly news networks such as CNN, Fox News, BBC News, and more. The third cluster, with the largest number of suspicious observations, consists of several misinformation-riddled defense channels, prominent in the Indo-Pacific region. We compared these results with the output from the DBSCAN clustering algorithm and we discovered the same number of clusters with highly similar cluster configurations as shown in Figure 4 below.

This experiment demonstrates that an unsupervised machine learning approach such as DBSCAN can help achieve the same channel clustering as a manual inspection can, thereby having a potential value in assisting human analysts. The most suspicious data point within the dataset is CIS News Network (a channel focused on India, Pakistan, and China relationships) with a suspicion score of 0.43. On further analysis of this channel, we saw that the channel had about 53,300 subscribers in August 2021, but the channel page shows videos with views running in millions as seen in Figure 5 below. Videos featured on this channel are pushing false narratives that are factually incorrect. This behavior was also spotted in a channel (BolongID) within Cluster 3, with a relatively high suspicion score of 0.4. A further look into this channel revealed a relatively small number of comments, compared to the number of views and subscribers of the channel. This channel includes videos with geopolitical content relating to Indonesia and China. As of August 2021, the channel had 670,000 subscribers, 246,465,788 views with a relatively small number of comments (4176), while in January 2022 the number of subscribers had dropped to 5320, and the view count had dropped to 448,577, showing an anomalous change in engagement metrics.

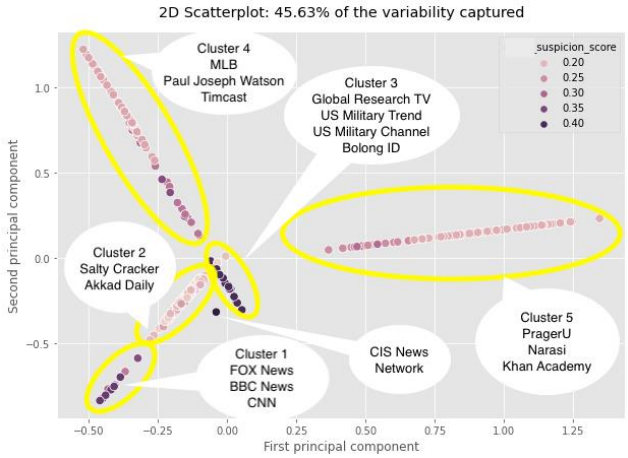

Figure 3: Scatterplot showing all clusters and their featured channels

Table 4

Clusters details.

Cluster	Number of observations	Highest suspicion score
1	12	0.37
2	287	0.27
3	36	0.4
4	108	0.34
5	121	0.29

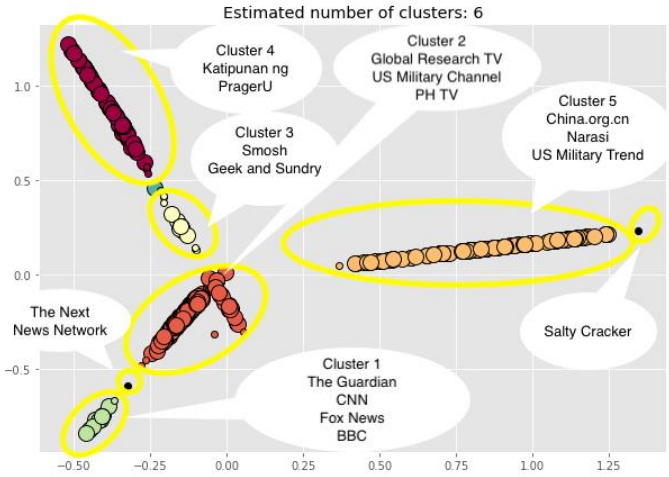

Figure 4: Scatterplot showing clusters identified using DBSCAN.

Figure 5: A recent set of highly suspicious view rates belonging to a video in our experimental dataset

To further reveal the behaviors within the channels, we analyzed the peaks extracted from the output of the LSTM model across all the channels. Before we applied our developed methodology, we took note of one channel - Defense Flash News - where we had detected an unusual peak in the daily subscriber trend in November 2020 (over +50,000 subscribers) that was later corrected by YouTube in December 2020 by removing suspicious subscribers (over -10,000 subscribers) as shown in Figure 6. Based on the behavior exhibited by this channel, we compared its anomaly plot generated from the LSTM model with that of a "well-known" channel - FC Barcelona. Using SciPy, we detected and represented the peaks within the plot using cross (x) marks as shown in Figures 7 and 8. This process was repeated across all the channels and we observed that the more suspicious channels had fewer peaks compared to the less suspicious channels, however, the magnitude of the peaks from the more suspicious channels were higher.

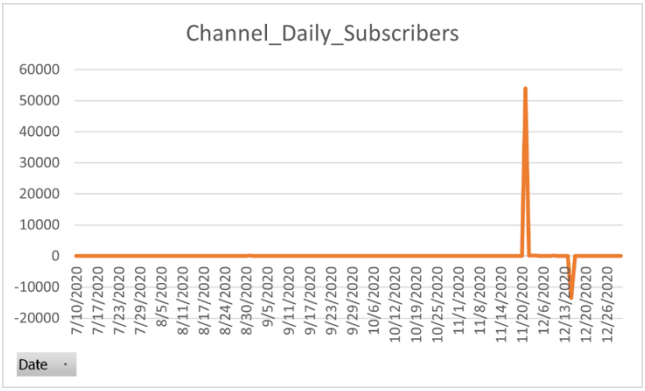

Figure 6: Defense Flash News new subscribers per day.

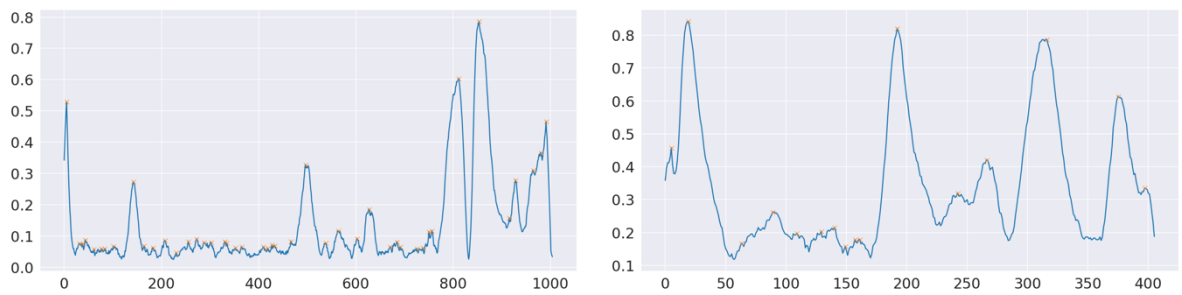

Figure 7: Anomaly plot representing peaks within FC Barcelona channel

Figure 8: Anomaly plot representing peaks within the Defense Flash News channel.

To validate this hypothesis, we analyzed the number and intensity of the peaks across each channel, characterizing peaks from Very high intensity to Very low intensity. The degree of peak intensity was determined as follows:

- Very high intensity - 2 or more standard deviations above the mean peak
- High intensity - 1 standard deviation above the mean peak
- Moderate intensity - The mean peak
- Low intensity - 1 standard deviation below the mean peak
- Very low intensity - 2 or more standard deviations below the mean peak

We sorted the channels based on the number of very high peaks and extracted the top 6 channels as shown in Table 5.

Table 5
Top 6 channels based on the number of very high peaks.

Channel name	Peak count	Very low peaks (percentage)	Low peaks (percentage)	Moderate peaks (percentage)	High peaks (percentage)	Very high peaks (percentage)
US Military Trend	221	0	37	42	7	14
US Military Channel	129	1	32	48	8	12
PragerU	295	0	34	49	6	11
Defense Flash News	123	0	35	43	11	11
Geek & Sundry	333	0	33	53	2	11
CIS News Network	359	0	29	56	4	11

We discovered some similarities between these top channels and the channels with a high suspcion score from the PCA plot, specifically the channels focused on defense, as well as CIS News Network. On the other end of the spectrum, Table 6 shows the channels with the least number of very high peaks. Channels on this spectrum tend to be “well-known” channels less likely to exhibit suspicious activities, validating our hypothesis.

Table 6
Bottom 6 channels based on the number of very high peaks.

Channel name	Peak count	Very low peaks (%)	Low peaks (%)	Moderate peaks (%)	High peaks (%)	Very high peaks (%)
Fox News	416	0	29	54	13	3
Styxhexenhammer666	397	0	25	61	10	4
FC Barcelona	386	0	27	57	11	4
BBC News	408	0	29	55	12	4
CGTN	329	0	27	60	9	4
The Guardian	412	0	29	57	10	4

4.2. Characterization based on Network Structures

Exploring the co-commenter networks and distributions of graph measures and clique-based feature values we described in Section 3, we observed variation between the channels we experimented with. Overall, co-commenter networks vary in size. This variation also corresponds with the total number of cliques, and clique size distributions in each channel. To compute the clique-size distributions, we simply counted the maximal cliques with each size (n>4) and plotted the final distribution. Accordingly, some channels have a long tail distribution of clique sizes with most of the cliques having less than 10 members, while several other channels tend to have bigger cliques and size distributions skewed right. Figure 9 shows the box plot for the median clique size in our experimental list of channels.

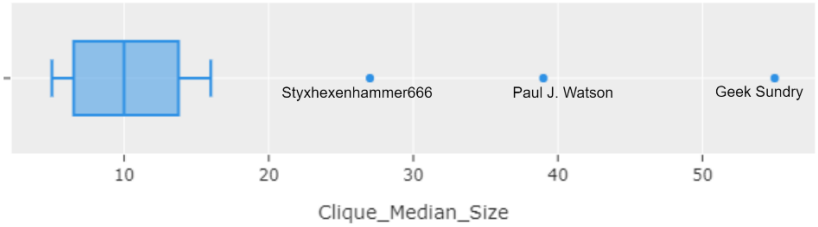

Figure 9: Clique median size in co-commenter networks for a sample of channels.

In addition, the feature similarity analysis and Pearson coefficient values show the pairwise correlation between a small number of features. Although the current sample size is small, the size of the co-commenter networks (threshold=10) seems to be strongly correlated with the total number of maximal cliques (n members > 4) in each corresponding co-commenter network. In the initial list of channels, the number of maximal cliques ranged from 16 to 4.64 million. Similarly, the number of nodes (co-commenters) ranged from 129 to 38,729. This variation of co-commenting and cliquish behavior supports our initial assumption that implicit network behaviors may correspond with the level of suspicious and coordinated behavior in relation to YouTube channels.

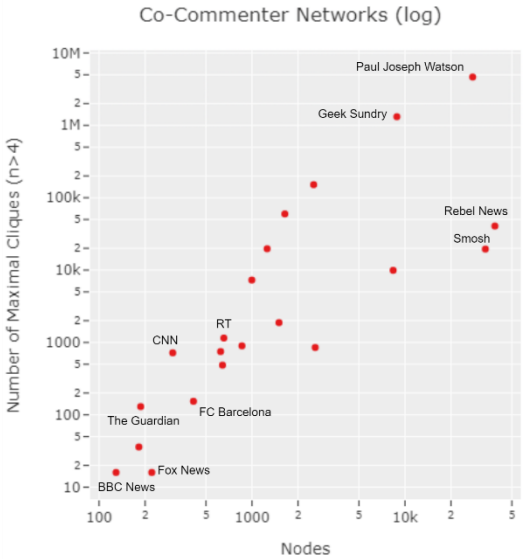

Figure 10: Number of nodes and number of maximal cliques for a sample of channels. The chart is in logarithmic scale.

One specific feature that may indicate the level of channel-level suspicious commenter behavior is the ratio of the number of unique commenters in cliques to the total number of nodes (co-commenters) in the network. As Figure 11 demonstrates, some co-commenter networks consisted of large numbers of co-commenters who were also members of maximal cliques. Thus, this feature indicates that on some channels, a higher number of nodes in the co-commenter network also form fully-knit cliques, signaling strong coordination.

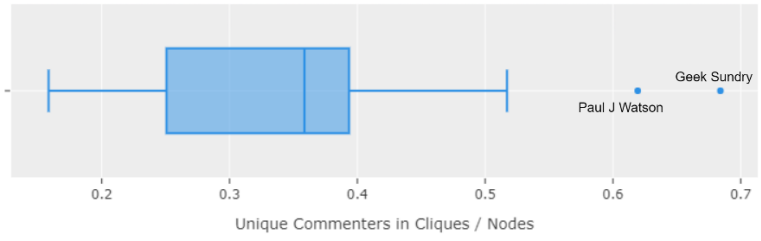

Figure 11: Number of unique commenters in cliques / total number of nodes in the co-commenter network for a sample of channels.

Similarly, the modularity of co-commenter graphs showed a wide range of variation between experimental channels. We observed that, in general, the networks with higher numbers of maximal cliques tend to have larger modularity. Nevertheless, the pairwise relationship between the modularity and the total number of maximal cliques is not necessarily linear. This contrasts the pairwise correlation between the average clustering coefficient (co-commenter network level) and median clique size, which seems to be more linear on a logarithmic scale.

Finally, to record the level of the degree difference between cliques and co-commenter networks overall, we simply divided the average degree of clique members by the average degree in the entire co-commenter network. Clique members tend to have higher degree centrality in comparison to the rest of the network. However, the given ratio does not seem to correlate with any other feature, as the outlier channels signaling suspicious behavior seem to have either very high levels of clique-based average degree or very high co-commenter participation in cliques. For example, cliques in the channel with the highest ratio of participation in cliquish behavior (0.68) have an average degree 7.75 times higher than the entire network. On the other hand, the channel with the highest difference in terms of the average degree (26 times higher), has only limited participation of co-commenters in cliquish behavior (0.22).

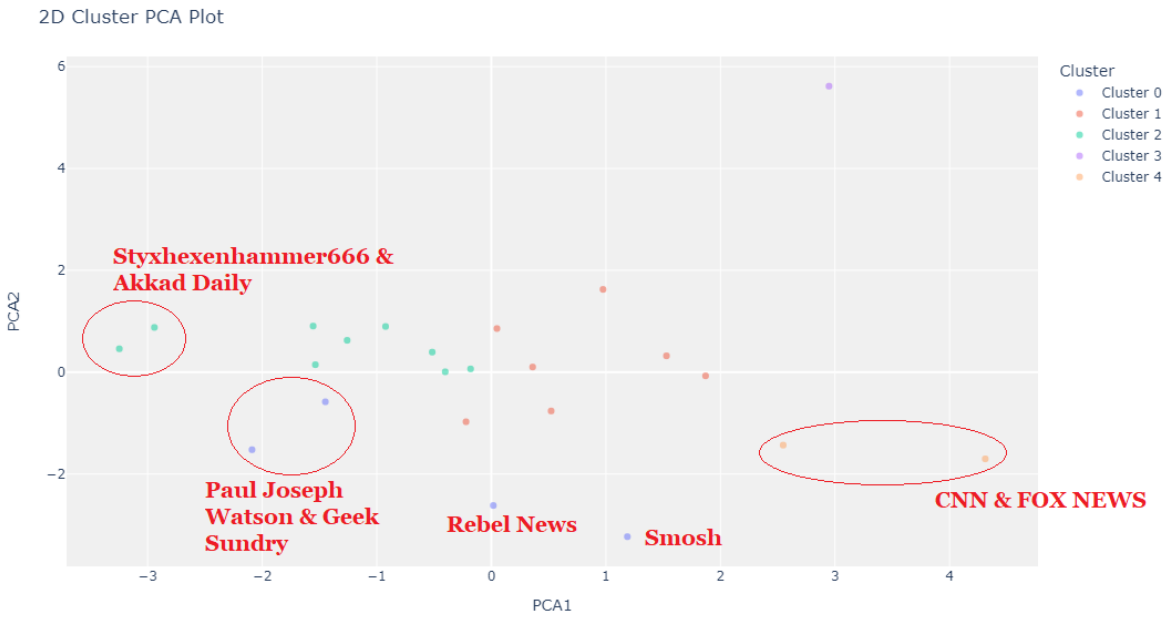

Figure 12: Post-PCA (network features) k-means clustering of a sample of channels.

Following the exploratory analysis of the co-commenter networks, maximal cliques, and additional network features we computed using the given pair of categories, we combined the feature similarity analysis with principal component analysis, k-means clustering, and hierarchical clustering to examine

of the current set of network features lead to meaningful clusters of channels in unsupervised settings. Given the Pearson correlation coefficients, we first reduced the number of features by removing multicollinearity above the threshold of 0.88. We then extracted five clusters of channels in a post-PCA k-means clustering setting. Finally, we also ran a simple hierarchical clustering algorithm in the final dataset. We observed that both unsupervised approaches extract clusters of similar channels. For example, post-PCA k-means clustering grouped news channels such as CNN and Fox News together, while clustering channels signaling high levels of coordinated activity grouped in other corresponding clusters.

5. Conclusion and Future Work

Characterization and detection of coordinated inauthentic campaigns on social media remain an open and significant problem. In this study, we aimed to explore new approaches to assess the latent and implicit characteristics of coordination that indicate the manipulation of information and communication networks on YouTube. We developed computational models to study suspicious and coordinated inauthentic behaviors exhibited by various channels. These models leverage a multi-step time-series analysis of engagement trends, network structural feature-based analysis, particularly the group behavior of co-commenter networks, and a combination of unsupervised and supervised machine learning experiments. Our models afford identification of suspicious behaviors overall as well as the precise time periods during which such behaviors are prominent in the channel's history. Furthermore, our models allow the identification of coordination among commenters and clusters of YouTube channels that exhibit similar behavioral profiles.

The results of our research show that channels exhibiting inauthentic activities are characterized by a relatively lesser number of peaks in their anomaly patterns. However, the magnitude of these peaks is usually higher compared to that of less suspicious channels. We also identified clusters of channels with similar engagement trends which is useful in detecting groups of channels that deploy similar tactics. In terms of user engagement statistics that characterize inauthentic behaviors in YouTube channels, the views and subscribers stand out as the most relevant indicator. This could suggest that channels that grow appraisals inorganically focus more on their views and subscriber count, however other indicators also characterize inauthentic behaviors. In addition, the models based on co-commenter networks and tightly knit sub-networks uncover clusters of channels that exhibit similar suspicious coordination of comments. In future phases, we aim to combine two methodological components and build a unified source of suspicious behavior signals. We posit that by improving the current understanding of coordinated suspicious activities and their artifacts, our study may lead to the development of an accurate and effective methodology for the timely detection of harmful manipulative campaigns on YouTube.

6. Acknowledgements

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

7. References

[1] Al-Khateeb, S., & Agarwal, N. (2015, March). Analyzing deviant cyber flash mobs of ISIL on Twitter. In International conference on social computing, behavioral-cultural modeling, and prediction (pp. 251-257). Springer, Cham.

[2] API Reference | YouTube Data API. (2021, July 2). Google Developers. Retrieved January 21, 2022, from https://developers.google.com/youtube/v3/docs

[3] Bovet, Alexandre and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. Nature communications 10, 1 (2019), 1–14.

[4] Carley, K. M. (2020). Social cybersecurity: an emerging science. Computational and mathematical organization theory, 26(4), 365-381.

[5] Cresci, S. (2020). A decade of social bot detection. Communications of the ACM, 63(10), 72-83.

[6] Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., ... & Quattrociocchi, W. (2016). The spreading of misinformation online. Proceedings of the National Academy of Sciences, 113(3), 554-559.

[7] Dutta, H. S., Jobanputra, M., Negi, H., & Chakraborty, T. (2021, August). Detecting and Analyzing Collusive Entities on YouTube. ACM Transactions on Intelligent Systems and Technology, 37(4), 111. https://doi.org/10.1145/1122445.1122456

[8] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD (Vol. 96, No. 34, pp. 226-231).

[9] Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. Communications of the ACM, 59(7), 96-104.

[10] Galeano, Katrin Kania, LTC Rick Galeano, Esther Mead, Billy Spann, Joseph Kready, and Nitin Agarwal. The Role of YouTube during the 2019 Canadian Federal Election: A Multi-Method Analysis of Online Discourse and Information Actors. Journal of Future Conflict, Issue 2, Fall 2020, pp. 1-22. Queen's University, Canada.

[11] Getting started with the Business API (Matrix) - Socialblade.com. (n.d.). Social Blade. Retrieved January 21, 2022, from https://socialblade.com/business-api

[12] Giglietto, F., Righetti, N., Rossi, L., & Marino, G. (2020). It takes a village to manipulate the media: coordinated link-sharing behavior during 2018 and 2019 Italian elections. Information, Communication and Society, 1–25.

[13] Graham, T., Bruns, A., Zhu, G., & Campbell, R. (2020). Like a virus: The coordinated spread of coronavirus disinformation. Report commissioned for the Centre for Responsible Technology.

[14] Hine GE, Onaolapo J, Cristofaro ED, Kourtellis N, Leontiadis I, Samaras R, Stringhini G, Blackburn J (2017) Kek, cucks, and God Emperor Trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In: ICWSM, AAAI Press, pp 92–101

[15] Hussain, M. N., Tokdemir, S., Agarwal, N., & Al-Khateeb, S. (2018, August). Analyzing disinformation and crowd manipulation tactics on YouTube. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1092-1095). IEEE.

[16] Information Operations. Twitter. Retrieved January 20, 2022. https://transparency.twitter.com/en/reports/information-operations.html

[17] Keller, F. B., Schoch, D., Stier, S., & Yang, J. (2020). Political Astroturfing on Twitter: How to coordinate a disinformation Campaign. Political Communication, 37(2), 256-280.

[18] Khaund, T., Kirdemir, B., Agarwal, N., Liu, H., & Morstatter, F. (2021). Social Bots and Their Coordination During Online Campaigns: A Survey. IEEE Transactions on Computational Social Systems.

[19] Kirdemir B., Agarwal N. (2022) Exploring Bias and Information Bubbles in YouTube's Video Recommendation Networks. In: Benito R.M., Cherifi C., Cherifi H., Moro E., Rocha L.M., Sales-Pardo M. (eds) Complex Networks & Their Applications X. COMPLEX NETWORKS 2021. Studies in Computational Intelligence, vol 1016. Springer, Cham. https://doi.org/10.1007/978-3-030-93413-2_15

[20] Kirdemir B., Kready J., Mead E., Hussain M.N., Agarwal N., Adjeroh D. (2021) Assessing Bias in YouTube's Video Recommendation Algorithm in a Cross-lingual and Cross-topical Context.

In: Thomson R., Hussain M.N., Dancy C., Pyke A. (eds) Social, Cultural, and Behavioral Modeling. SBP-BRiMS 2021. Lecture Notes in Computer Science, vol 12720. Springer, Cham. https://doi.org/10.1007/978-3-030-80387-2_7

[21] Kready, Joseph, Muhammad Nihal Hussain, and Nitin Agarwal. YouTube Data Collection Using Parallel Processing. IEEE Workshop on Parallel and Distributed Processing for Computational Social Systems (ParSocial 2020), May 22, 2020, New Orleans, Louisiana USA.

[22] Kumar S, Hamilton WL, Leskovec J, Jurafsky D (2018) Community Interaction and Conflict on the Web. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW'18, ACM Press, pp 933–943, DOI 10.1145/3178876.3186141

[23] Marcoux, Thomas, Nitin Agarwal, Recep Erol, Adewale Obadimu, and Muhammad Nihal Hussain. Analyzing Cyber Influence Campaigns on YouTube Using YouTubeTracker. In: Çakırtaş M., Ozdemir M.K. (eds) Big Data and Social Media Analytics. Lecture Notes in Social Networks. Springer, Cham. pp. 101-111. 2021. https://doi.org/10.1007/978-3-030-67044-3_5

[24] Nimmo B, François C, Eib CS, Ronzaud L, Ferreira R, Hernon C, Kostelancik T (2020) Exposing secondary infektion. Report, Graphika. https://secondaryinfektion.org/

[25] Pacheco, D., Hui, P. M., Torres-Lugo, C., Truong, B. T., Flammini, A., & Menczer, F. (2021, May). Uncovering Coordinated Networks on Social Media: Methods and Case Studies. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 15, pp. 455-466).

[26] Schafer, F., Evert, S., & Heinrich, P. (2017). Japan's 2014 General Election: Political Bots, Right-Wing Internet Activism, and Prime Minister Shinz Abe's Hidden Nationalist Agenda. Big Data, 5(4), 294–309

[27] Şen, F., Wigand, R., Agarwal, N., Tokdemir, S., & Kasprzyk, R. (2016). Focal structures analysis: identifying influential sets of individuals in a social network. Social Network Analysis and Mining, 6(1), 17.

[28] SciPy documentation — SciPy v1.9.0.dev0+1313.ecb800f Manual. (n.d.). SciPy. Retrieved January 22, 2022, from https://scipy.github.io/devdocs/index.html

[29] Singularex, 2019. The Black Market for Social Media Manipulation. Riga: NATO Strategic Communications Centre of Excellence.

[30] Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-26.

[31] Stewart, Leo G, Ahmer Arif, and Kate Starbird. 2018. Examining trolls and polarization with a retweet network.

[32] U.S. Department of State Global Engagement Center, 2022, Kremlin-Funded Media: RT and Sputnik's Role in Russia's Disinformation and Propaganda Ecosystem.

[33] Vargas, L., Emami, P., & Traynor, P. (2020, November). On the detection of disinformation campaign activity with network analysis. In Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop (pp. 133-146).

[34] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151

[35] Weber, D., & Neumann, F. (2021). Amplifying influence through coordinated behaviour in social networks. Social Network Analysis and Mining, 11(1), 1-42.

[36] YouTube. Fake engagement policy - YouTube Help. (n.d.). Google Support. Retrieved January 21, 2022, from https://support.google.com/youtube/answer/3399767?hl=en&ref_topic=9282365

[37] YouTube API Services Terms of Service. (2021, July 1). Google Developers. Retrieved March 4, 2022, from https://developers.google.com/youtube/terms/api-services-terms-of-service

[38] Terms of Service. (2021, November 9). Social Blade. Retrieved March 4, 2022, from https://socialblade.com/info/terms