# Developing Socio-Computational Approaches to Mitigate Socio-Cognitive Security Threats in a Multi-Platform Multimedia-Rich Information Environment

**Nitin Agarwal, Ph.D.**
Maulden-Entergy Chair and Distinguished Professor of Information Science
Founding Director, COSMOS Research Center
University of Arkansas – Little Rock
2801 S University Ave., Little Rock, Arkansas, 72204
UNITED STATES

nxagarwal@ualr.edu

## *ABSTRACT*

*Growing weaponization of social media is influencing peacekeeping, tactical, operational, and strategic operations. At the strategic and operational levels, social media platforms that are manipulated by adversarial campaigns can shift international and regional opinions about the use of military force or validity of military operations in a region. At the tactical level, social media propaganda could potentially be used to persuade susceptible targets to disrupt or delay military operations through protests or other "non-lethal" resistance. Narratives can be easily manipulated and influenced by bots, trolls, and other influence operation TTPs. Moreover, most users of social media cannot or will not differentiate between legitimate and non-legitimate accounts. Since such influence operation TTPs can be employed clandestinely in a low-cost, low-risk context, military leaders can expect to encounter an increased amount of adversary-generated AI-amplified social media-driven propaganda. Furthermore, socio-cognitive threats are increasingly becoming – 1) a collective phenomenon and 2) multimedia online information environment centric. This study aims to advance social, behavioural, and cultural science and enhance situation awareness and sensemaking (HFM ET-356 House Model) by detecting, examining, evaluating, measuring, and predicting the cognitive threat level/impact of the adversarial information campaigns to strengthen community resiliency further. The models and approaches presented in this study are validated and demonstrated in real-world use cases (e.g., COVID-19, the Indo-Pacific region). A multi-model multi-theoretic approach is developed that blends computational modelling, big 'social' data, and social science theoretical principles to characterize adversarial information campaigns conducted in an integrated online information environment (OIE). Specifically, the research identifies key actors, groups/mobs, narratives, TTPs (and their impact), in mainstream platforms as well as emerging and multimedia-rich platforms to characterize information actors (producers and consumers) and campaign dynamics for proactive decision-making to mitigate cognitive threats.*

## 1.0 INTRODUCTION

Social media is characterized as a powerful online interaction and information exchange medium. The US National Security Strategy identifies how social media's power has been harnessed by potential adversarial state actors, paid trolls, and extremist organizations to conduct disinformation campaigns, provoke hysteria, and coordinate nefarious acts. Due to afforded anonymity and perceived less personal risk of connecting and acting online, such deviant acts are becoming increasingly common. Online deviant groups have grown in parallel with online social networks (OSN), whether it is black hat hackers using Twitter to recruit and arm attackers, announce operational details, coordinate cyberattacks [1], and post instructional or recruitment videos on YouTube targeting certain demographics; or state/non-state actors' and extremist groups' (such as ISIS') savvy use of social communication platforms to conduct phishing operations, such as viral

retweeting a message containing an image which if clicked unleashes malware [2]. Social media's vulnerabilities revolve around the lack of effective global policy enforcement and, most importantly, its affordability to develop a false or misleading narrative before facts are released. As Multi-Domain Operations (MDO) continue to evolve, the digital realm fosters a large opportunity for expansion, evidenced by the accelerated growth of USCYBERCOM as well as the newly created USSPACECOM: both have rapidly evolving operations in the physical, informational, and cognitive domains.

As evidenced by several recent events, narratives on social media could be easily weaponized and propagated online at frighteningly fast speeds to sow discord. Such insidious threats that attempt to influence beliefs and behaviors need to be considered as modern weapons of cyber warfare. Several multi-year studies conducted by the Collaboratorium for Social media and Online Behavioral Studies (COSMOS), published by NATO, US Army, Royal Military College of Canada, Baltic Defense Journal, and others, have identified the use of cyber influence tactics in –

1) Terrorists/violent extremists' information campaigns;
2) Anti-NATO information campaigns during various NATO exercises;
3) COVID-19-related conspiracy theories, disinformation, and scams; and
4) Pro-Chinese/anti-West influence campaigns in the Indo-Pacific region.

These studies supported the NATO Research and Technology Group (RTG)-248 and Human Factors and Medicine (HFM)-293 that demonstrated the highly sophisticated and ever-evolving nature of the complex information operations to propagate disinformation, which changed themes, narratives, and messages in favor of the adversary. In today's interconnected world, adversaries employ advanced communication tactics via a sophisticated orchestration of a variety of existing and emerging social media platforms. Narratives are framed on blogs, video blogs (or vlogs), Reddit, 4/8chan (i.e., payload) and distributed widely through platforms like Facebook, Twitter, VKontakte, MeWe, Weibo, etc. Emerging platforms like Gab, Discord, Parler, Rumble, BitChute and the likes help the payload to reach niche communities [3]. Information actors are relying increasingly on multimedia centric social platforms because of – (1) our society's shifting information consumption habits from purely text-based to multimedia-based platforms, and (2) compounded impact of the information through multiple modalities (e.g., image, audio, video, text, algorithmically served content) on our cognition [4]. Social media is moving faster than our lawyers, scientists, and policymakers are able to review, recommend, and decide on the way forward.

Further, in recent years, cognitive attacks (e.g., crowd polarization, adversarial influence campaigns) have been increasingly conducted as flash events, where adversaries self-organize, get together in cyberspace, perform an unanticipated act (such as coordinating a misinformation campaign), and quickly disperse, almost akin to a cyber manifestation of a flash mob. Hence, we call them deviant forms of cyber flash mobs, or deviant cyber flash mobs [5], [6]. Deviant flash mobs have taken a darker twist as criminals exploit the anonymity of crowds, using social networking to coordinate everything from robberies to fights to general chaos [7], [8]. More recently, the term "mob" has been increasingly used to remark an electronically orchestrated violence such as the recent attack on the State Capital in Washington, D.C. by protesters that lead to property damages, disruption, injuries, and deaths [9], [10], [11]. In another recent incident, an army of small investors from all over the world used Reddit to coordinate "flashmob investing" [12] to create stock market frenzy causing GameStop's stock value to rise from $20 to $483 in less than a month [13]. Such events raise the importance of systematically studying such behaviors. Modern Information and Communication Technologies (ICTs) provide affordable and easy to use means of communications (such as social network platforms, viral emails, and SMS) that facilitates and ease the process of recruiting, training, and looking for a specific sector of the society, e.g., specific gender, age, political affiliation, interest, and cultural background. This, in turn, has led to an increase in the occurrences of emerging socio-technical behaviors [14], including flash mobs. These deviant flash mobs

have become emerging social cyber threats. With the increased number of organized violent "protests" in recent months, there is a need to systematically study such events and be able to build predictive models to distinguish between benign and deviant/violent/vicious flash mobs.

Misinformation, conspiracy theories, and scams pertaining to COVID-19 serve as a profound example for the deviant flash mob phenomenon mentioned above. There are similarities between misinformation about COVID-19 and other misinformation cases that we have studied for NATO, US, EU, Singapore, Canada, etc. Like in other cases, the motivation for spreading COVID-19 misinformation is monetization or to provoke hysteria. Bad actors or scammers are spreading misinformation to further their political agenda or trying to profit off this adversity. For instance, Russian state media pushed the conspiracy theory that the virus is manufactured by the western rich and elites to suppress the poor; Chinese state media pushed the conspiracy theory that US Army created the virus to bring Chinese economy to a grinding halt. And then there are scammers who are selling fake masks, fake cures, using fake websites to ask for private/sensitive information from people by posing as government websites. However, there is a significant difference between COVID-19 and other misinformation campaigns that we have studied before. Being a global and rapidly evolving crisis, the nature of misinformation is also extremely diverse and super-fast. Other misinformation campaigns were specific to an entity, event, region, elections, military exercises. However, misinformation about COVID-19 has both global as well as regional narratives. While fake masks, fake cures, etc. affect a global audience, the regional narratives include promoting medicines for bovine coronavirus as cure for human coronavirus affecting rural/agriculturalist regions. Moreover, the misinformation about COVID-19 ranges from health to policy to religion to geopolitical affairs, i.e., high topical diversity [15], [16], [17].

Given the increasing volume, velocity, and variety of flash event style cognitive threats, research is warranted to study such campaigns and their organization. In the case of COVID-19, the problem of misinformation is worse than the pandemic itself. That's why it is called infodemic or more specifically, misinfodemic. Like the pandemic, misinformation cases are also rising exponentially. These cases are more difficult to track than the epidemic, as they can originate in the dark corners of the internet. To make matters worse, we cannot enforce lockdown on the Internet to stop the spread of this infodemic. Because during crises Internet is usually the only mode of communication. There are some quarantine efforts. For instance, social media companies like Facebook, YouTube and retail companies like Amazon are doing their best to block such content, suspend bad actors or scammers who are spreading misinformation or trying to profit off this adversity. Social media platforms are banning misinformation videos and users posting these videos about COVID-19 and the COVID-19 vaccine. These videos and content are sometimes taken down within minutes to several days. However, users are still posting these videos and getting around the algorithms until someone reports them. Users that have been banned have moved to freedom of speech platforms to post their content. When searching for COVID-19 and vaccine information on YouTube, misinformation results were not easy to find. The solution to finding these videos on other platforms was to explore BitChute, which is considered a freedom of speech platform. The search is not as robust as YouTube to filter down results, but we could get what was needed. On BitChute, the keywords to search were COVID, COVID-19, vaccine, COVID vaccine. We filtered by relevance and newest first. After gathering the data from BitChute, we were able to find other platforms users were using, from Rumble, Parler, Minds, MeWe, and more. The collected information was then used to find the same content on other social media platforms. But such cases are simply too many and growing too fast. What makes this problem worse is the fact that the information spreads like a wildfire on the Internet, especially the false or misinformation. As Winston Churchill said, a lie gets halfway around the world before the truth has a chance to get its pants on. It is very much true in the age of social media. Many studies have concluded that misinformation travels faster than its corrective information, and the more questionable the misinformation is the faster it travels. This is simply because on social media people usually have a lot more virtual friends than they do in their real life. So, if they share or retweet some misinformation, wittingly or unwittingly, they expose all their virtual friends to the misinformation.

Using a variety of tactics, techniques, and procedures (TTPs), information actors deploy information maneuvers and seek to influence others [18]. Understanding these mechanisms of socio-digital influence of the rapidly evolving Online Information Environment (OIE), particularly the multimedia-rich OIE, is very important for understanding modern information conflicts, especially in environments of social instability with limited observatory capabilities, such as war zones and revolutions. COSMOS is pioneering methodologies that cut across various disciplines to diagnose novel pathologies of online social media. More specifically, the study focuses on characterizing the OIE, identifying & documenting information maneuvers and TTPs, and developing models of assessing influence operations in the cyber space.

Government and other regulatory bodies need systematic, accurate, explainable, and science-driven approaches to measure the impact/level of a cognitive threat for proactive planning. Toward this direction, this study aims to use mathematical and computational social science concepts to measure the power of adversarial information campaigns, i.e., is a campaign gaining or losing prominence? Such a power measure could help prioritize the investigation and countermeasure development. Can we characterize the campaign dynamics (e.g., accelerating vs. decelerating phase) to assist effective and timely policymaking? Can we characterize information producers (initiators, amplifiers, sustainers) and consumers (susceptible, exposed, infected, skeptic, etc.) with a theoretical grounding in social sciences to help develop community resilience? The focus of the study is on analyzing how decentralized online individual actions transform into collective actions. Further, what necessary conditions are required that lead to the emergence of the deviant flash mob phenomenon and, subsequently, its sustenance? Answers to these questions will enable us to explore the predictive capabilities of the theoretical model we have developed [5], [6], [19], [20], [21].

The rest of the paper is organized as follows. Section 2 discusses the characterization of information maneuvering tactics observed on multimedia-rich OIE, along with methodologies to assess their impact. These include a mix of content-driven, network-driven, and algorithmic/AI affordance-driven tactics. Section 3 describes our collective action (CA) based cyber campaign characterization framework to afford socio-computational modeling of coordinated cognitive attacks. The CA framework is combined with the Deviant Cyber Flash Mob (DCFM) model (which measures a mob/crowd's utility, interest, control, and power) and an operations research-based model called focal-structure analysis (FSA) to identify key units within the social networks that are the most powerful in mobilizing and coordinating cognitive attacks. Section 4 discusses our approaches for population characterization as susceptible (S), exposed (E), infected (I), or skeptic (Z) that allows decision makers to inoculate the right group of individuals for effective mitigation of cognitive attack. These approaches leverage the Diffusion of Innovation theory and epidemiological theories to characterize information actors and consumers. These models further show how toxicity spreads like a contagion, polarizes communities, and could lead to a breakdown of discourse, thus providing indicators of cognitive attacks and ways to measure the impact. Section 5 concludes our study with directions for future work.

All our research-driven models are transitioned to usable software in an ongoing basis. The software tools, viz., BlogTracker (https://btracker.host.ualr.edu) [22], VideoTracker (https://vtracker.host.ualr.edu) [23], and COVID-19 Misinformation Tracker (https://cosmos.ualr.edu/covid-19) [16] were recognized as top 10 solutions in the 2021 NATO's Innovation Hub challenge on "The Invisible Threat: Countering Cognitive Warfare". The COVID-19 Misinformation Tracker was deployed in real-time in partnership with the Arkansas Office of the Attorney General to educate the public about the false claims in the misinformation campaigns. It was recognized by the World Health Organization (WHO) as one of the key technological innovations developed across the world to address COVID-19 pandemic, and it was integrated into a NATO Science and Technology publication (HFM-293) [17]. These tools are also part of the US Department of State's Global Engagement Center's innovation hub on fighting foreign-based propaganda.

# 2.0 CHARACTERIZATION OF INFORMATION MANEUVER TACTICS IN MULTIMEDIA ONLINE INFORMATION ENVIRONMENT

In recent years, several frameworks have been proposed to communicate and characterize information campaigns and incidents of information manipulation and interference online, while some have been incorporated into the existing institutional frameworks by government agencies and industry. Some of the proposed frameworks aimed at systematic documentation and communication of reported campaigns across institutions (viz., ABCDE framework, SCOTCH framework, Disinformation ABC, the US Department of Justice's Framework to Counter Malign Foreign Influence Operations), while others focused on the characterization of behaviors, tactics, techniques, and procedures (TTPs) employed by threat actors (viz., US Department of Homeland Security's Disinformation Kill Chain framework, DISARM/AMITT framework, RICHDATA framework). Also, a few frameworks characterize the narrative and rhetorical characteristics of influence in general (viz., 4Ds of disinformation, BEND framework). The existing frameworks offer great value in overall understanding and timely detection of influence operations. However, almost all the proposed frameworks cannot address the emerging developments in the modern information environment and pose a number of limitations to the characterization of high-risk hostile influence operations. In this study, we emphasize the use of multimedia environments, including video, audio, real-time broadcasting, and real-time audio chat environments for online influence operations as a high-priority area of focus for an additional computational and analytical framework. All in all, we posit that the expansion of the current set of frameworks with a multi-modal approach toward such social media platforms would improve the efforts to monitor, track, detect, and mitigate the emerging threats in the information environment. The state-of-the-art characterization frameworks have several limitations as mentioned below.

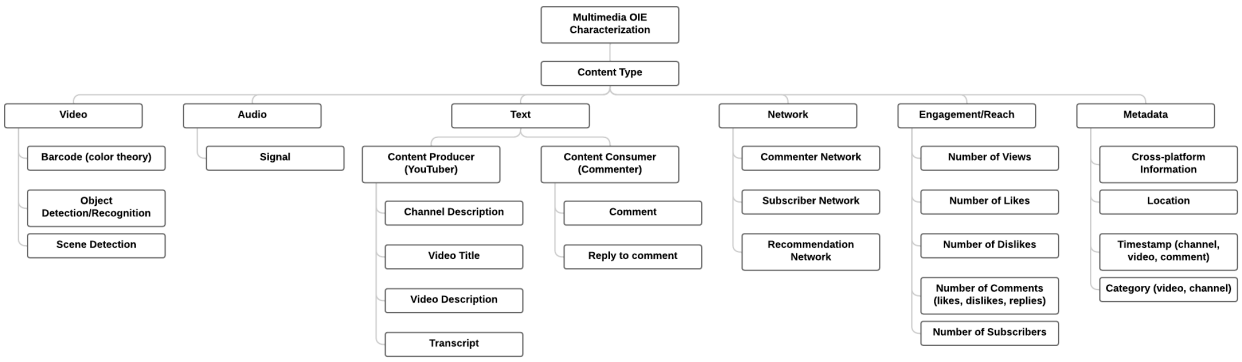



**Figure 1: Taxonomy of Content Types.**

Most critical limitation of existing characterization frameworks stems from the fact that they do not consider the dynamic characterization of information actors (producers and consumers) and the information campaign process. To develop a rich understanding of the online information environment (OIE), all three elements are critical, i.e., tactics, information actors, and the campaign's dynamics. Our characterization framework considers the information actor and the campaign dynamics to develop effective countermeasures. Each analytical and conceptual framework mentioned above offers value to the assessment of online influence operations. However, the relevant literature and policy-relevant toolkit still lack a framework that would, in combination with some of the existing frameworks, address the emerging characteristics of multimedia platforms and cross-platform dynamics in the contemporary information environment. Most of the existing analytical workflows rely on text-based analysis of threat actors and information manipulation incidents, with limited, time-consuming, and mostly qualitative assessments of the artifacts, content, or actors present in multimedia environments. In addition, most of the current knowledge regarding the online influence operations relies on the given analytical workflows and text-based computational tools, which in turn pose significant limitations to the overarching understanding of how current hostile operations are conducted, or

which artifacts they use. Furthermore, existing frameworks focus on platforms that are primarily user-centric (e.g., Twitter, Facebook, WhatsApp). Emerging multimedia-based platforms (e.g., YouTube, TikTok, Parler, Clubhouse) are primarily content-centric. TTPs of influence operations are shaped by the platform architectures, and what is available as manipulative tools. By mostly relying on the text-based tools that are mostly applied to data originating from a certain list of social media platforms such as Twitter, the overwhelming majority of the current frameworks often fall short of offering a comprehensive view of how multiple modalities of signals from video, audio, text, network, user/content-engagement, and platform affordances are combined. Further, time series data analysis needs to be considered for a near real-time evaluation. Our characterization framework considers all the above-mentioned data elements as well as time series information. There is not much focus on impact assessment in existing frameworks. In our approach, we address these gaps. Further, our approaches have rigorous theoretical and mathematical grounding.

We developed a characterization framework for multimedia-rich social platforms. The framework addresses the gaps in the current frameworks (discussed above) and presents novel ways of characterization of multimedia-rich social platform-based behaviors. The framework also provides analytical tools to not only identify such behaviors but also measure their impact. Our framework emphasizes multimedia environments to complement the existing frameworks of hostile influence operations. The analytical framework of influence operations across multimedia environments should address the multi-dimensionality and richness of data, behavior, content, and signal types that emanate from such platforms. Therefore, the computational and analytical components of our framework include, but are not limited to, the content types shown in Figure 1 and analysis types shown in Figure 2.

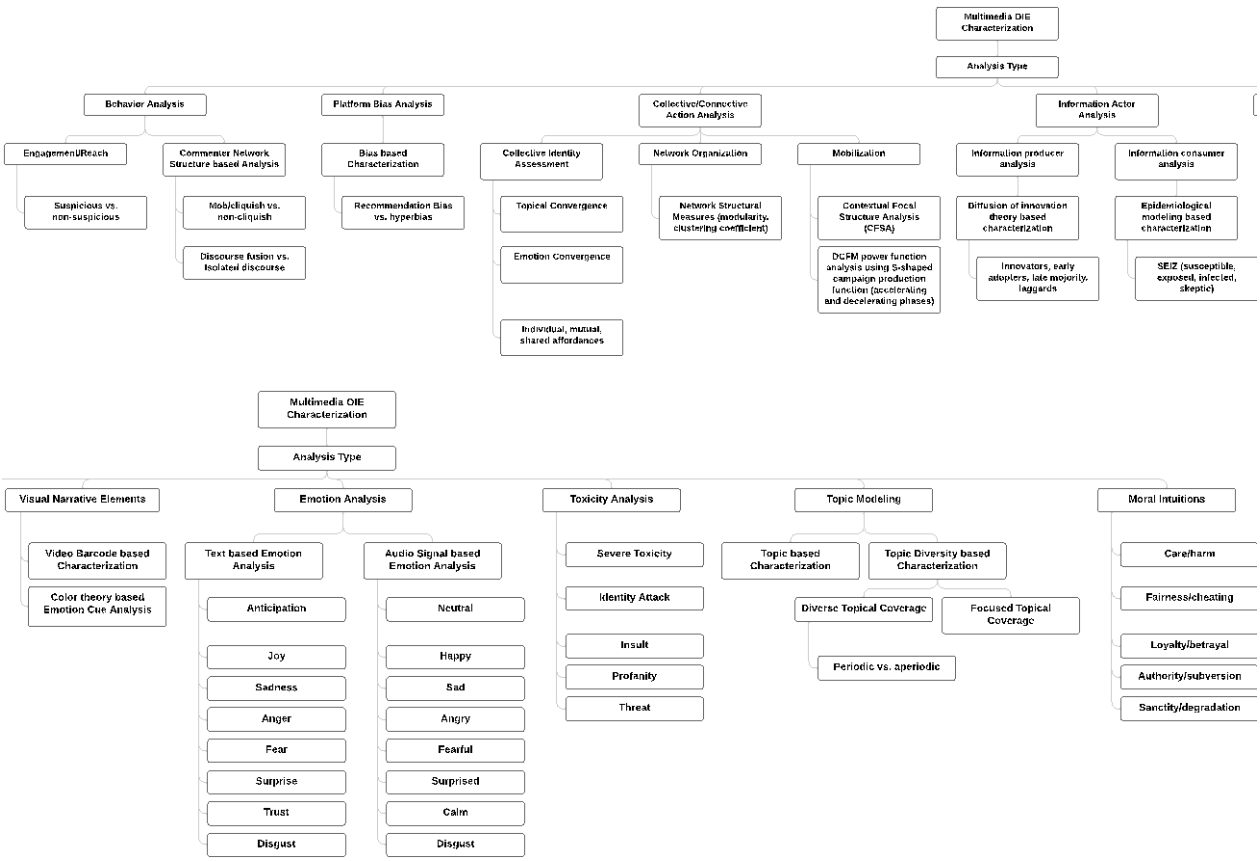



**Figure 2: Taxonomy of Analysis Types. The taxonomy is split into two parts. The top part displays behavior analysis, platform bias, collective/ connective action, and information actor analysis. The bottom part displays visual narrative element analysis, emotion, toxicity, topic modeling and moral intuition analysis.**

Various multimedia OIE-specific tactics have been identified in our studies. These tactics are characterized by their detection and impact assessment, i.e., network (N), content (C), and/or platform affordance (P). The list of these tactics is presented in Figure 3, along with their respective characterization (N, C, and/or P). While some tactics leverage content, network, or platform affordances, others depend on a combination of two or all three facets. For instance, content cloning/repurposing leverages the content facet; however, coordinated content push leverages both the content and network facets. On the other hand, coordinated deep fake tactic leverages all three facets, i.e., network, content, and platform affordance. Figure 3 illustrates this multi-faceted characterization of multimedia-rich OIE-specific tactics by positioning the tactic closer to the facets leveraged by the said tactic.

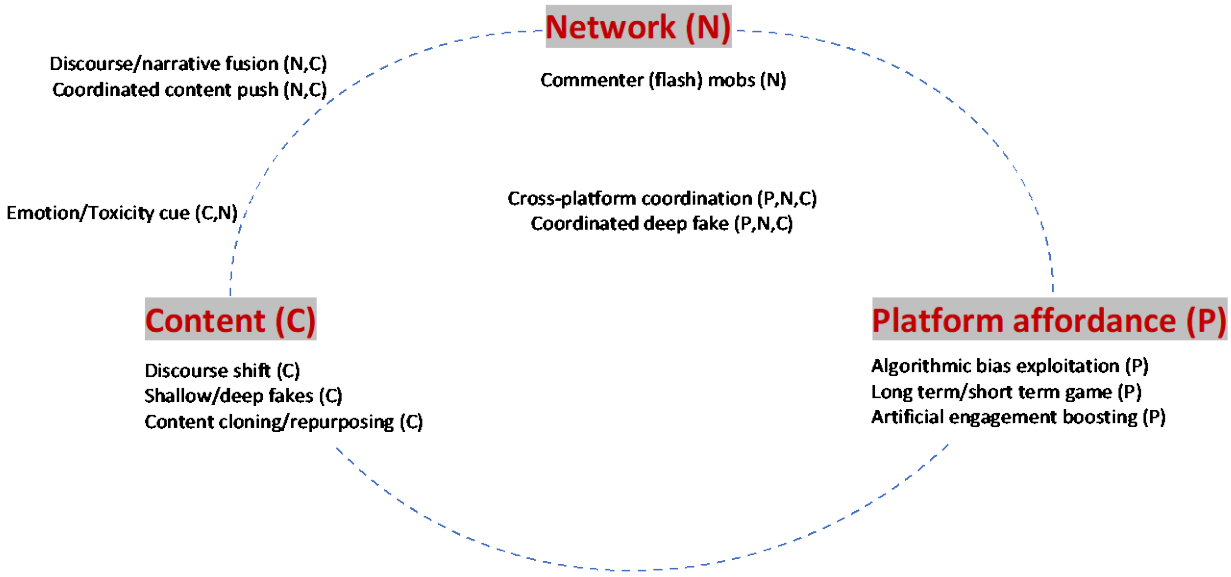



**Figure 3: Multimedia OIE-specific Tactics and their Characterizations for Network, Content, and/or Platform Affordance-based Impact Assessment.**

For each tactic, we develop analytics to detect and measure its impact. These analytics are presented as a playbook to show what data elements are needed, how to process them, and the analytical methodology to detect and measure the impact. Figure 1 and Figure 2 present characterization of the rich data environment of multimedia OIE and characterization of multimodal multimethod analytical methodologies for multimedia OIE, respectively. Next, we will discuss our analytical approaches to detect and measure the impact of a few of these TTPs.

## 2.1 Commenter Flash Mob Tactic

A type of coordinated inauthentic behavior, commenter flash mob is a tactic where a seemingly arbitrary group of commenters collectively comment on a (or a set of) video(s). Such an act seems uncoordinated to an outsider but there is a sophisticated coordination in their activity, like a flash mob. Their comments may or may not be relevant to the video content. Some examples of coordinated commenter flash mobs are shown in Figure 4 (left). As a contrasting example, a regular commenter behavior is shown in Figure 4 (right). Some of the commenter flash mobs shown in Figure 4 (left) comprises of groups of commenters who co-commented on over 100 videos. Such commenter flash mobs can create a perception of a content being popular thereby helping increase its virality [24], [25].

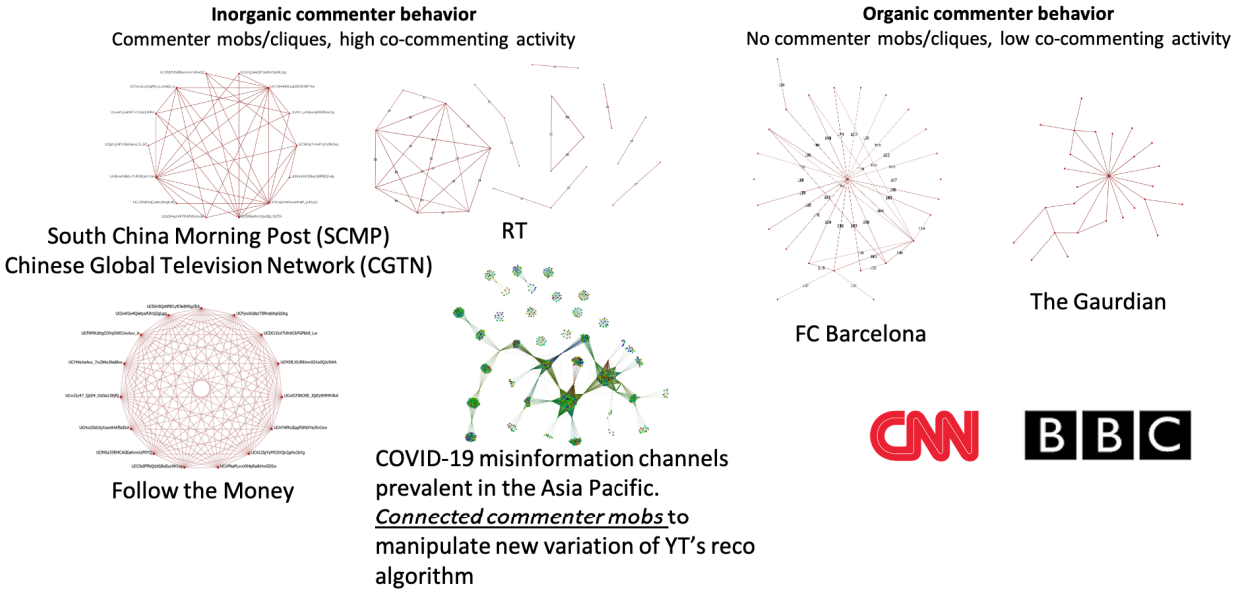



**Figure 4: Commenter Flash Mob Behavior (left) and Regular Commenter Behavior (right).**

Table 1 shows the data elements and analytical methodologies needed to detect and assess commenter (flash) mob tactic in multimedia OIE.

**Table 1: Data elements and analytical methodologies needed to detect and assess commenter flash mob tactic.**

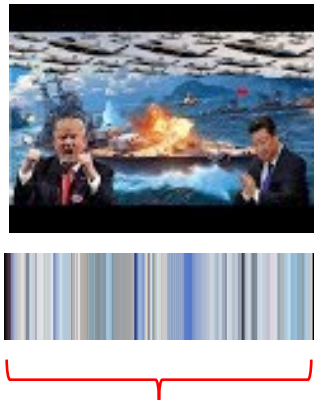
| Data Elements | Analytical Methodology |
|---|---|
| Commenter Network | Commenter Network Structure Analysis |
| Subscriber Network | Subscriber Network Analysis |

## 2.2 Coordinated Content Push Tactic

A type of coordinated behavior, coordinated content push is a tactic where a group of information actors (e.g., channels on YouTube) post identical or similar content. The group of information actors could post identical videos, similar videos, portions of a video embedded in another, may use similar content production styles, similar background/foreground images in their video, same/similar audio, speech, text, etc. [26] Figure 5 shows two different channels that pushed same videos with identical titles. Further, the group members may also follow each other (as subscribers) and/or deploy other tactics such as commenter flash mobs. Often content from the group of channels is recommended by the platform and shows up in the 'related videos' stream of each other.

Figure 6 shows a group of information actors on YouTube that deployed coordinated content push tactic. These channels posted highly similar to identical videos, with same/similar titles, same/similar audio transcript, and used robotic voice narration. Further, these channels also deployed commenter flash mob tactic. Due to these tactics, content from these channels show up in each other's related video stream [25].

**Video ID:** OM5vaF2kzPA
**Title:** China vs US The War in the South China Sea already Start
**Channel:** Breaking News TV



**Video ID:** GsCmudyXY2o
**Title:** China vs US The War in the South China Sea already Start
**Channel:** DOT COM US

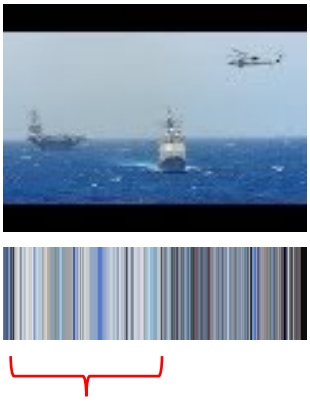


**Figure 5: Different YouTube Channels Pushing Same Content. The Channel on the Left Posted a Video Embedded in a Video Posted by the Channel on the Right. Video Barcode Approach developed by COSMOS was used to Detect Content Similarity [26].**

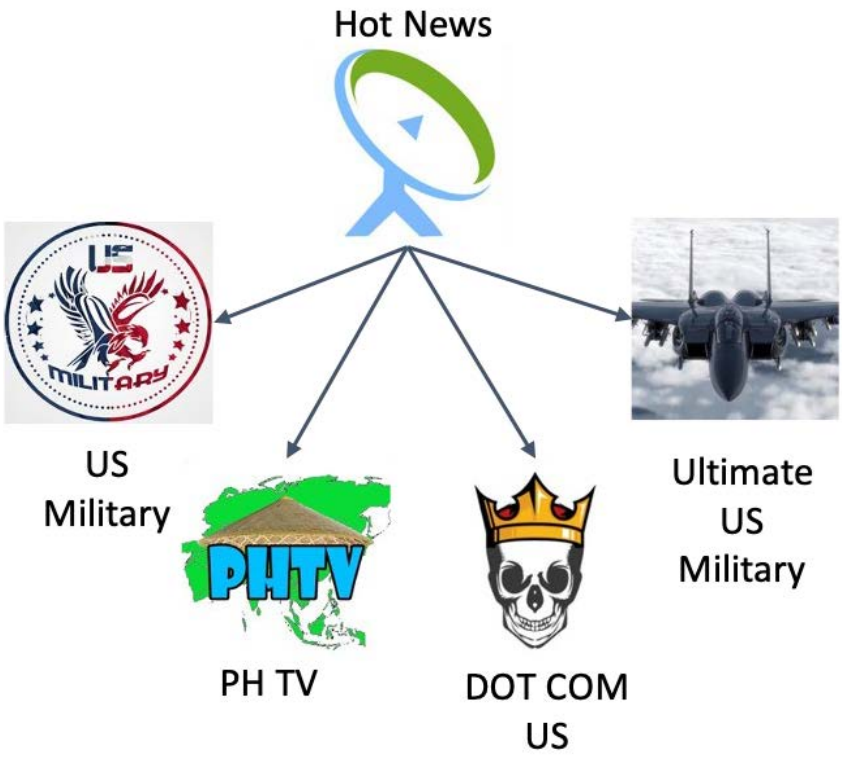



**Figure 6: Group of Information Actors (Channels) on YouTube Deploying Coordinated Content Push Tactic.**

Table 2 shows the data elements and analytical methodology taxonomy needed to detect and assess coordinated content push from multimedia OIE.

**Table 2: Data Elements and Analytical Methodologies needed to Detect and Assess Coordinated Content Push Tactic.**

| Data Elements | Analytical Methodology |
|---|---|
| Video Barcode | Video Barcode based Characterization |
| Transcript | Text Similarity based Characterization |
| Audio | Audio Signal Similarity based Characterization |
| Text | Topic based Characterization; Topic Diversity based Characterization |
| Network-Commenter<br>Network-Subscribers<br>Network-Recommendations | Commenter Network Structure based Analysis;<br>Subscriber Network Structure based Analysis;<br>Bias based Characterization |

## 2.3 Discourse/Narrative Fusion Tactic

Discourse/narrative fusion is a tactic where different topics (that may or may not be related) are injected in a discourse in an attempt to conflate issues, raise popularity/visibility of one topic through the other, bridge related echo chambers, and fabricate a perception of relevance of the topic or support for the topic in society [27]. This tactic could also result in exploitation of algorithmic bias. Narrative/discourse mixing could also help in detection of code-switching tactics.

Commenter networks stretch beyond a channel. Cross-channel commenter networks allow identification of discourse fusion tactics. Figure 7 (left) shows an example of isolated discourse on YouTube. The discourse tends to stay localized to a video or a channel. Commenters prefer to stick to a channel or a video. They comment and leave. Isolated discourse is largely observed on mainstream news and politics category channels. People rarely watch the same news on different YouTube channels much less commenting on the same news story on different channels, hence a very sparsely brokered discourse network. On the other hand, Figure 7 (right) shows an example of discourse fusion tactic, where people have bridged various discourse communities by co-commenting. In this example, people have co-commented on a variety of COVID-19 misinformation promoting videos, anti-vaccination ideology promoting videos, and conspiracy theory promoting content on YouTube. Thereby, creating a highly brokered discourse network that pushes YouTube's algorithms to consider such content related and show them as such in related video streams.

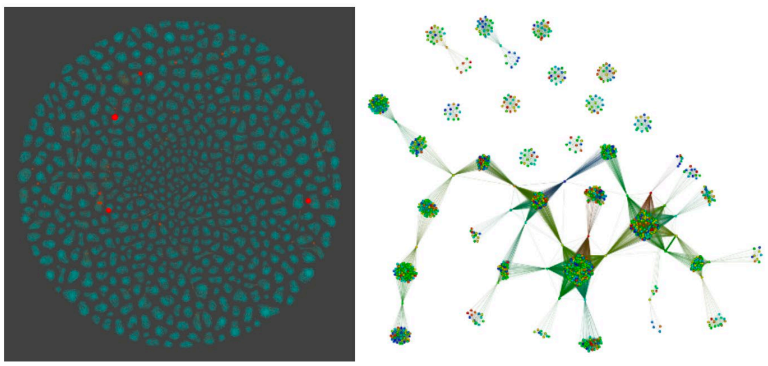


**Figure 7: Types of discourse on YouTube. On the left, is an isolated discourse, i.e., no discourse fusion. On the right, is an example of discourse fusion.**

Table 3 shows the data elements and analytical methodology taxonomy needed to detect and assess discourse/narrative fusion from multimedia OIE.

**Table 3: Data Elements and Analytical Methodologies needed to Detect and Assess Discourse/Narrative Fusion Tactic.**

| Data Elements | Analytical Methodology |
|---|---|
| Network-Commenter<br>Network-Subscribers<br>Network-Recommendations | Commenter Network Structure based Analysis; Subscriber Network Structure based Analysis; Bias based Characterization |
| Text | Topic based Characterization; Topic Diversity based Characterization (novelty, resonance, transience) |
| Video | Visual Narrative Element analysis |
| Audio | Audio Signal Similarity based Characterization |

## 2.4 Toxicity Tactic and Community Dynamics

Toxicity in discourse is an important signal for an I&W system as it could forecast brewing unrest, impending attacks, and any hateful rhetoric that may lead to a violent outcome. Toxicity has been shown to be contagious in our prior work (Figure 8 and Figure 9) [28]. It is important to study what happens to a community when toxicity spreads, as it helps in assessing **connective action properties of an emerging cyber-social threat**. In our studies, increase in toxicity has led to splintering of communities, as seen in Figure 10, Figure 11, and Figure 12. Granger Causality test is used to determine causality relation between toxicity and community dynamics (number of communities over time). In our preliminary study for COVID pro-vaccine and anti-vaccine dataset, we found that toxicity's effect on community dynamics was most pronounced on second and third day. This suggests a lag of 2-3 days before an increase in toxicity may lead to fracturing of a community [29]. In some ways, rise in toxicity could serve as an early indicator of polarization and collective identity emergence.

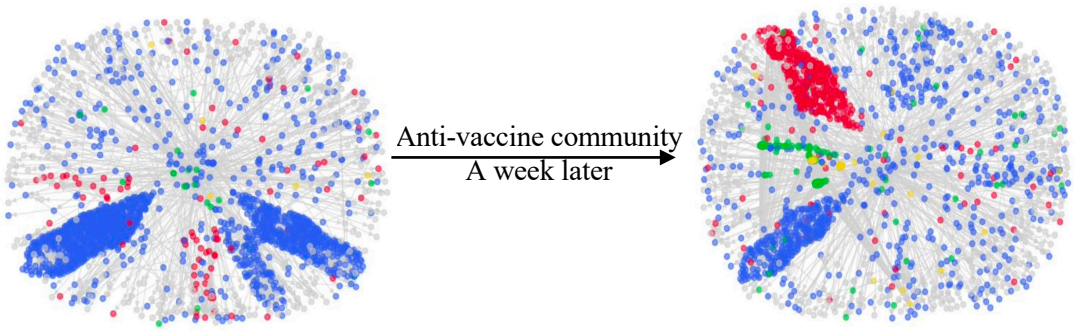



**Figure 8: Segregation of toxic users (tweeters) in an anti-vaccine community.**

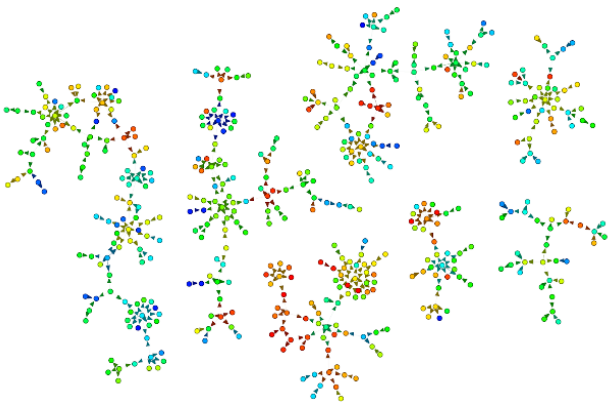


**Figure 9: Segregation of toxic commenters (YouTube) in a COVID-19 conspiracy theory espousing channel.**

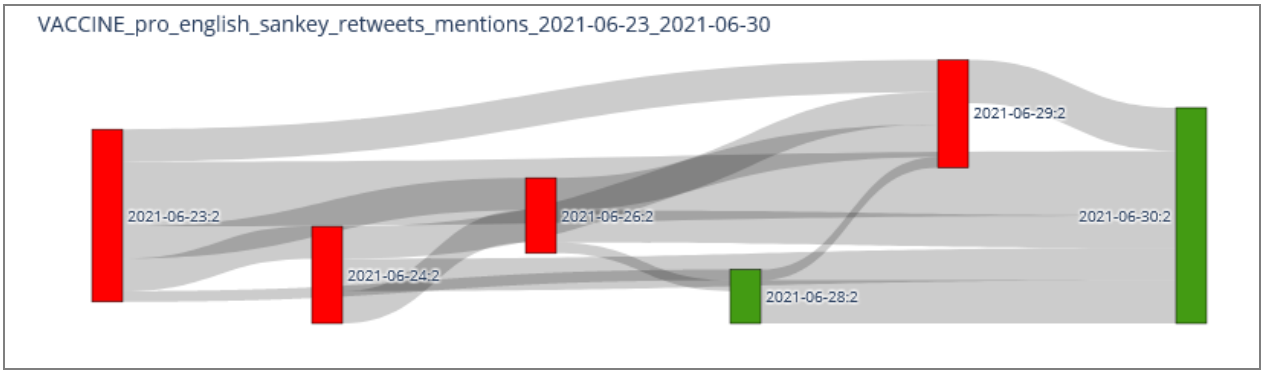



**Figure 10: Pro-vaccine community Sankey flow from 6/23/21 to 6/30/21.**

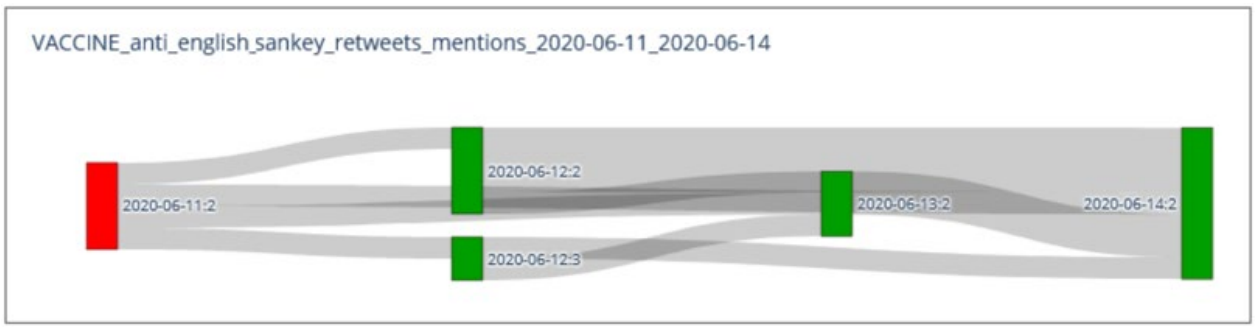



**Figure 11: Anti-vaccine community Sankey flow from 6/23/21 to 6/30/21.**

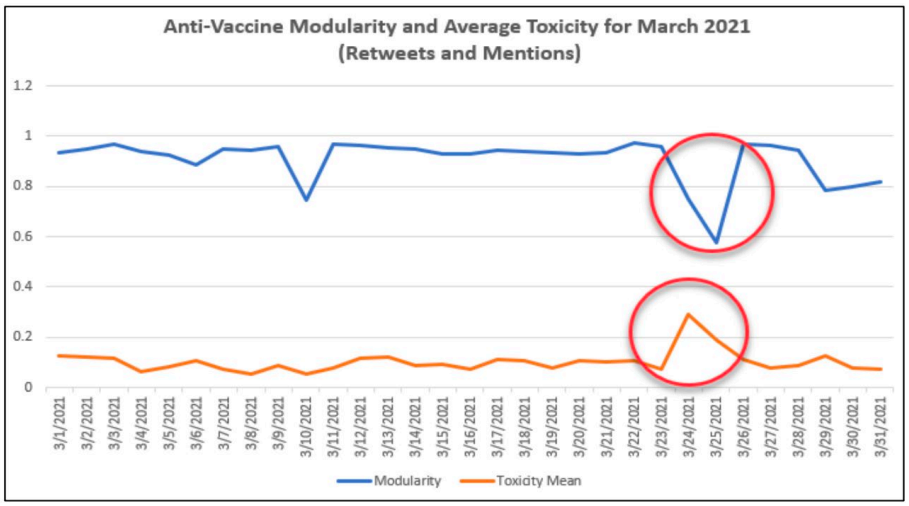



**Figure 12: Collating toxicity indicator and network structural measures such as modularity.**

Further, social bots can be synchronized to cause viral contagion such as toxicity! This is an important signal, as it not only examines the presence of bots but their impact on the discourse and thereby community dynamics. Bot activity is observed to be positively correlated with toxicity. Increase (or decrease) in number of bots is strongly correlated with increase (or decrease) in toxicity (Figure 13) [30]. We can collate the three signals i.e., toxicity, network structure measures, and bot frequencies to detect deliberate amplification of the process of enhancing collective identity of a group or manufacturing polarization in an online community. This shows one way of putting together multiple signals derived from social media to offer rigorous and more timely view in the behaviors of crowds on social media. This example also offers a view into the complexity of scenarios – not all scenarios may provide same set of meaningful signals.

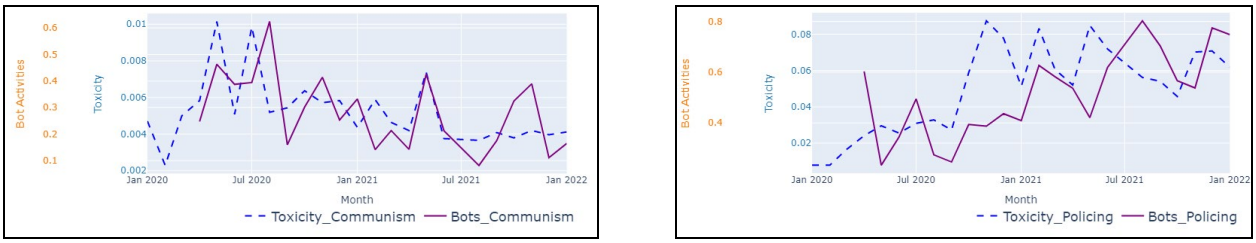



**Figure 13: Presence of bots and toxicity is strongly correlated. (Trimmingham & Agarwal, 2022).**

Table 4 shows the data elements and analytical methodology taxonomy needed to detect and assess toxicity and its impact on community dynamics from multimedia OIE.

**Table 4: Data Elements and Analytical Methodologies needed to Detect and Assess Toxicity and its impact on Community Dynamics.**

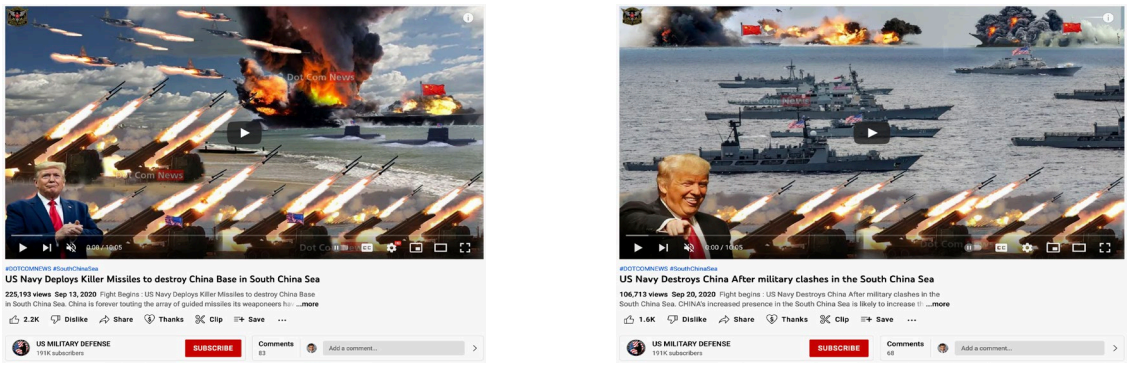
| Data Elements | Analytical Methodology |
|---|---|
| Network-Commenter | Commenter Network Structure based Analysis; |
| Audio-Signal | Audio Signal Based Emotion Analysis |
| Barcode (Color Theory) | Color Theory based Emotion Analysis |
| Object Detection/Recognition (faces) | Facial Emotion Analysis |

| Data Elements | Analytical Methodology |
|---|---|
| Text | Text based Emotion Analysis;<br>Toxicity Analysis |
| Text | Topic based Characterization; Topic Diversity based Characterization |

## 2.5 Content Cloning/Repurposing Tactic

A type of tactic, where an information actor (e.g., channel on YouTube) would appropriate content (e.g., a video) in part or in entirety from other channels. Content similarity could range from being highly similar to completely identical. Similar content could include video, audio, text, or all of these. Such content cloning/repurposing has a two-fold advantage. First, it helps in content amplification, thereby increasing the chances of the content being recommended by the platform's content recommendation/related feed curation algorithms. Second, if offensive content is removed/suspended by the platform citing platform's policy violation, then the copy of the content ensures continued access and distribution.

An example of content repurposing is shown in Figure 14. A pair of highly similar videos are shown. Our video barcode approach was used to detect content similarity [26]. Figure 15 shows the barcode of a pair of videos with portions of high similarity identified.




(a) (b)

**Figure 14: Content repurposing by different YouTube channels. Videos shown in (a) and (b) are highly similar.**

**Figure 15: Barcodes of a pair of videos (Figure 14 (a) and (b)) with portions of high similarity identified.**

Barcode based video similarity assessment could be used to identify clusters of videos cloning/ repurposing content from other videos/channels. Figure 16 shows clusters of videos with highly similar to completely identical content. Different colors correspond to various clusters. Highly similar videos are placed very close to each other, while identical videos are superimposed on one another. These clusters are interpreted using topic models on the text available through the videos' titles, description, and transcripts. Cluster interpretations are shown in Table 5.

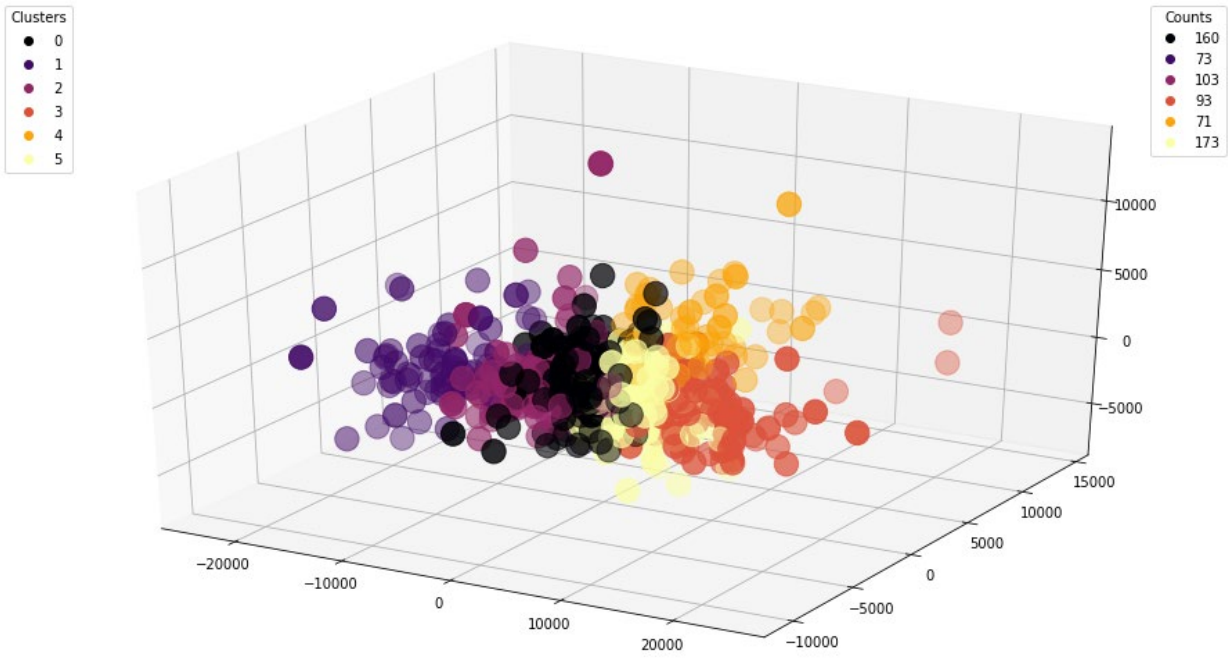



**Figure 16: Clustering of videos based on barcode-based similarity assessment.**

**Table 5: Interpretations of the 6 clusters shown in Figure 16 using topic models on the text available through the videos' titles, description, and transcripts.**

| Cluster Number | Characteristics of the Cluster | Channels/Video Title |
|---|---|---|
| Cluster 0 | Daylight videos with fewer individuals | CGTN, BBC News, News China TV (News anchors); China reopens its embassy in Nicaragua |
| Cluster 1 | Videos with fireworks, bright light, and events | Taikonauts capture city lights from space on Chinese New Year's Eve; Oil depot near Kyiv ablaze after explosions light up predawn sky |
| Cluster 2 | Videos focused on war and crime scenes | South China Morning Post; Global Research TV; CGTN; Kosovo Launches Border Offensive Against Serbia |
| Cluster 3 | Videos focused on reports on the military, fight scene-based content | ABC News In-depth; GlobalResearchTv Propaganda and fake footage from the Ukraine invasion; War Propaganda and Media Lies within the Syria Conflict |
| Cluster 4 | Videos focused on reporting events about celebration | CGTN; Narasi; University students embrace Beijing 2022 with dance: Selamat Tahun Baru Imlek 2022 |
| Cluster 5 | Videos focused on sky or cloud-like scene content view, whose dominant color is blue | CGTN; Russia says using precision weapons to disable military targets, not targeting civilians |

## 2.6 Artificial Engagement Boosting Tactic

A tactic that is often deployed by an information actor to gain views, likes, comments, and subscribers through inorganic means, that are often driven by bots provided by services like YouTube Booster (https://youtube-booster.space/). Artificial engagement boosting tactic could generate several thousands of views, likes, comments, and/or subscribers for a channel, that would start placing the video higher in the search results and increase the chances of being recommended to viewers.

Artificial engagement boosting services are highly sophisticated, as they boost videos via placing then in playlists, increase views, likes, comments, subscribers, emulate different traffic sources (suggestions, YouTube search, home YouTube page, channel, direct visits and from other sites), generate profiles, farming (warm up) profiles and accounts, watching / skipping ad, monetization clicking, rewind the first seconds of the video, smart rewinds, ability to set watch time, multithreading, proxy support (http / socks: regular, mobile change IP by interval or API), anonymity to ensure maximum security to exclude view drops and bans, highest emulation of human behavior (e.g., moving the mouse over the progress bar, reading description, comments, sometimes opens suggested videos, copies a link, enable pause, highlights text on a page, and much more. In general, everything is like an ordinary user does.), ability to pass PVA when logging into a Google account, solving captcha in Google, support of accounts with 2fa, scheduler, ability to set different settings for individual videos, and ability to increase views without Google accounts.

A few examples of artificial engagement boosting tactics deployed by YouTube channels is shown in Figure 17. To detect such sophisticated engagement boosting tactics, an advanced machine learning based algorithm is developed that scales well with the amount of data available at YouTube and needed for processing/detecting such behaviors. Using a combination of techniques involving rolling window correlation analysis, anomaly detection, peak detection, rule-based supervised classification and unsupervised clustering, we analyze user engagement statistics of these channels to detect suspicious engagement trends within the channels and assign a score to these periods of suspicion. Figure 17 shows engagement profiles of the YouTube channels that have the highest mean suspicion scores according to our model [24]. These figures depict instances of anomalous engagement behaviors in these channels.

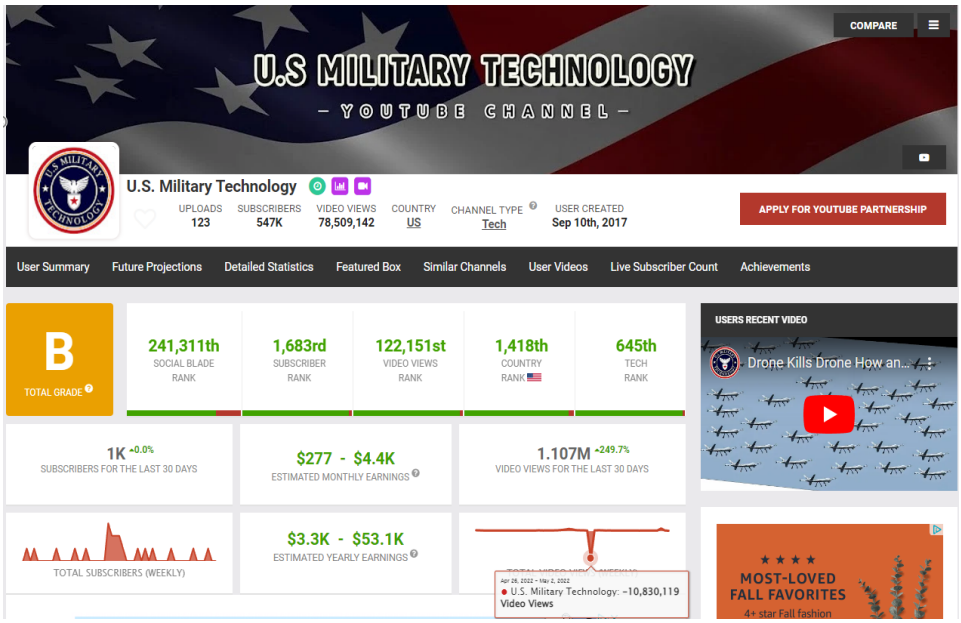



**Figure 17: U.S Military Technology. This channel has spikes of subscribers which seem suspicious when compared with the view count. Moreover, a negative number of views were recorded for the channel that suggests some videos were removed from the channel.**

Our model uses six indicators to rank suspicion. These indicators represent the correlation values between pairs of engagement statistics. An example of a suspicious activity is a negative correlation between views and subscribers of a channel over time, which suggests increase in views while there is a decrease in subscribers – a behavior that is very uncommon on YouTube. These indicators are assigned weights indicating their relative importance in detection of suspicious behaviors. The pairs of engagement scores with their relative weights are: Views vs Subscribers: 0.35; Videos vs Comments: 0.25; Subscriber vs Comments: 0.2; Views vs Videos: 0.1; Views vs Comments: 0.06; and Subscriber vs Videos: 0.04.

While it is helpful to analyze individual channels using the six indicators based on engagement scores, it is however, an unscalable approach for a large collection of channels. We developed a PCA + clustering-based approach to automatically characterize / group channels similar in terms of engagement behaviors (Figure 18(a) and Figure 18(b). This analysis helps in not only identifying suspicious channels (based on color grades) but also detecting groups of channels that deploy similar tactics. All 3 outliers show off the chart inorganic behaviors (5 out of 6 indicators). One America News Network (alt-right news network) being the most inorganic (Figure 18(a)). Figure 18(b) shows a method for characterization of YouTube channels based on engagement scores. This analysis helps in not only identifying suspicious channels but also detecting groups of channels that deploy similar tactics. Cluster 1 is mostly news networks (similar engagement profiles). Cluster 2, 3, and 5 show high anomaly scores. These clusters also consist of several misinformation riddled anti-US military news channels, prominent in the Indo-Pacific region. Coordinated inauthentic behavior was identified in a channel syndicate working up an anti-US military discourse in the Indo-pacific region. This was identified using DBSCAN clustering analysis, particularly within members of clusters 2, 3 and 5.

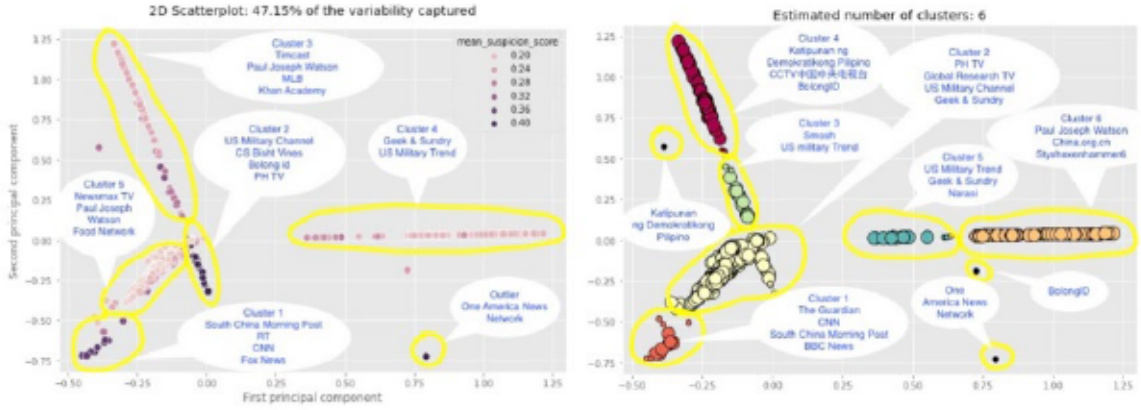



**Figure 18: Detecting suspicious engagement behaviors on a set of YouTube channels and characterization based on engagement scores.**

To evaluate scalability of the approach, we tested it on a collection of 3,517 YouTube channels posting content relevant to the Indo-Pacific region. Many of these channels are known to post content with false/misleading narratives. These channels have posted tens of thousands of videos and millions of comments. Figure 19(a) shows identification of channels that deployed artificial engagement boosting tactics through a color gradient (darker color depicts more inorganic engagement behaviors). The channel with highest suspicion score, viz., 'Breaking News TV' (suspicion sore 0.72 highlighted in Figure 19(a)) has been suspended by YouTube citing platform's terms and conditions (Figure 20). Figure 19(b) shows a clustering of channels with similar engagement anomaly profiles. A total of 10 clusters were obtained. These clusters range from highly suspicious (inorganic) to non-suspicious (organic) engagement profiles and the members of these clusters, i.e., channels fit those profiles.

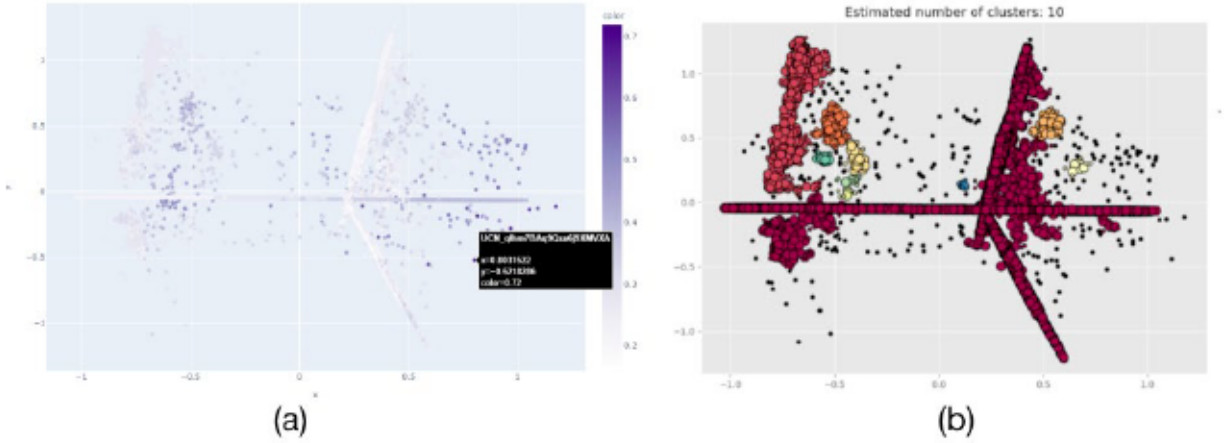



**Figure 19: Engagement analysis of 3,517 Indo-Pacific Channels. Detecting suspicious engagement behaviors on these YouTube channels and characterization based on engagement scores.**

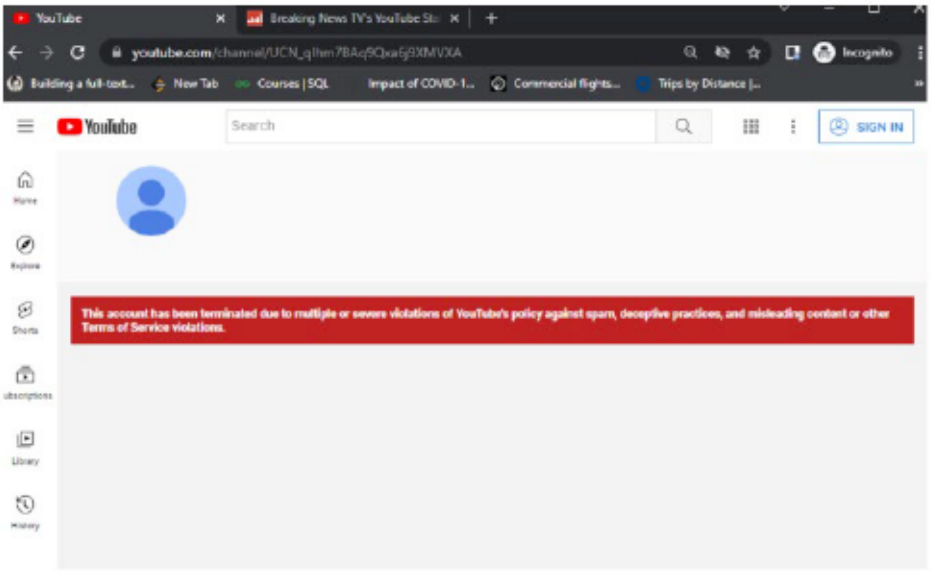



**Figure 20: YouTube channel with the highest suspicion score ('Breaking News TV' (suspicion sore 0.72 highlighted in Figure 19(a)) has been suspended by YouTube.**

## 2.7 Algorithmic Bias Exploitation Tactic

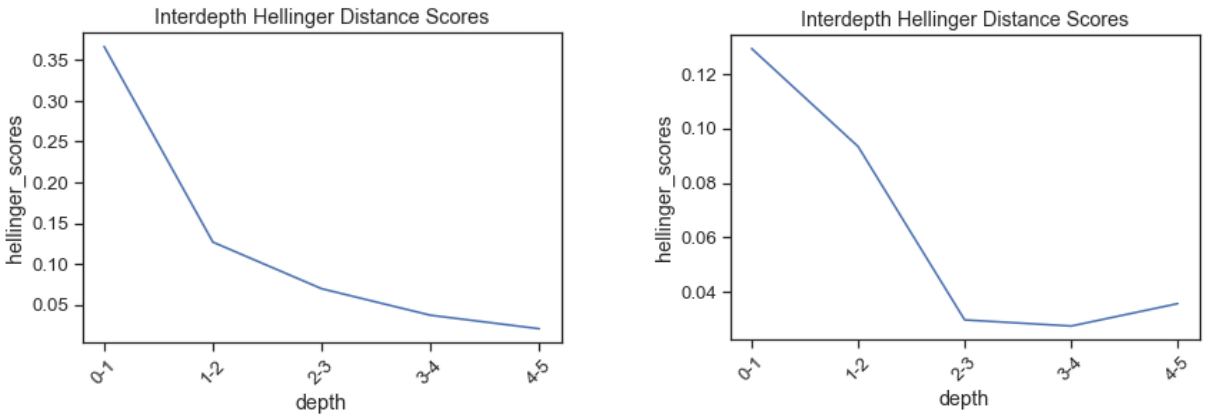
Algorithmic content curation and biases serve more of the same content, increasing motivation of potential participants to act. Influence campaigns use affordances through online social networks to influence rapid adoption, while reducing costs to join and providing anonymity. Since the costs to join a campaign are low, the decision to adopt the idea, or join the movement requires little exposure. Affordances correspond to "action possibilities and opportunities that emerge from actors engaging with a focal technology". Leonardi distinguished between these types of actions by labeling them as individualized, shared, and collective affordances [31]. However, tackling platform vulnerabilities and algorithmic bias vis-a-vis adversarial information operations is a challenging domain that requires adaptive, dynamic, and agile analytical capabilities.

Our study focuses on developing a conceptual and computational framework to study platform vulnerabilities of online social networks vis-à-vis the inherent, structural, and emergent algorithmic biases that are exploited by a variety of threat actors. Adversarial actors, ranging from state actors to terror groups and private companies that serve as "disinformation as a service (DaaS)" [32] entities, exploit algorithmic biases and weaknesses in platform architectures to carry out effective hostile influence operations. Documented sets of adversarial tactics, techniques, and procedures target the inherent cognitive biases that affect human behavior as well as platform designs and algorithmic choices that define how much and how far a piece of information spreads across communities and how online social networks take different forms. Our methodologies assess, understand, and predict platform vulnerabilities against threat actors that exploit algorithmic bias. Applying network science, statistical disparity metrics, natural language processing, and machine learning methods to diversified datasets, we aim to increase the existing knowledge regarding algorithmic bias issues and their effects on vulnerabilities against the TTPs that aim to multiply the effectiveness of adversarial campaigns.

Existing studies have shown that the recommendation algorithm has an inherent tendency of bias toward a small fraction of videos, and it pushes users into mild ideological echo chambers. This study aims to analyze the relevance of content and emotion at different recommendation depths and, to determine the extent to which YouTube's recommendation bias could be exploited, and how to evaluate the impact of this tactic. We develop analytics to measure drifts emanating from recommendation bias and evaluate on the Uyghur narrative and a counter-narrative called 'Cheng Ho' pushed by pro-Chinese media to project China as an advocate of religious freedom [33]. A list of 50 videos on Cheng Ho was collected with the help of SMEs in the Indo-Pacific region. These 50 videos served as the seed for the recommendations, with 58,825 videos collected through 5 depths of recommendations. We computed the topic drift on the recommendation depths and discovered that the recommendations led us further away from the original topic. Furthermore, observing the eigenvector centrality values of videos within the recommendation network of different depths, we saw the evolution of influential videos as their relevance to Cheng Ho diminished. The results showed how YouTube's recommendation system discards the topics of the seed videos by subtly introducing a new but still pro-China topic in the network through influential videos. This new topic is about economic growth and religious freedom in China targeting Indonesia's younger demographic by focusing on current events and pop culture.

Looking at the results of the inter-depth Hellinger distance score of English videos in Figure 21(a), we observed a significant distance in the topic similarity between the seed and depth one. It indicates that the content between them is dissimilar, but as we traverse through the recommendation depths, the distance shrinks showing increased content similarity between adjacent depths. It tells us that the depths are more similar to each other than to the seed. For the inter-depth Hellinger distance score of Indonesian videos in Figure 21(b), we observed a similar trend as seen in Figure 21(a), but with less distance between depths. Showing that the recommendation depths share more similarity between themselves and smaller similarity with the seed.

Next, we analyzed each depth's network to identify the influential nodes pulling recommendations. Network visualization of the recommendation graph at each hop/depth is depicted in Figure 22. Colors depict the various content communities extracted using network modularity. The videos in each depth were ranked with their eigenvector centrality value (depicted by their size), which determines a node's transitive influence in the network. By examining the network's influential nodes, we could determine the kinds of videos that affect recommendations and how a video's influence evolves as we go through the depths. We observed how the influential nodes in the network evolve from relevant videos to irrelevant videos about the Cheng Ho topic. The topics of the influential nodes drift from the original narrative and a new topic ("Cha Guan") is subtly introduced at depths 2 and 3, with a complete shift at depth 4. "Cha Guan" is a segment on the "Asumi" channel, a media-tech institution aimed at Indonesia's younger demographic, with a focus on current events and pop culture. Notably, the videos on this channel discussed economic growth and religious freedoms in China. Still observing the evolution of influential videos in the network, we noticed a shift in language at depth 5, from Indonesian to English. In addition to the language change amongst influential videos at depth 5, the distribution of eigenvector scores was significantly lower.




**(a)** **(b)**

**Figure 21: Inter-depth Hellinger Distance of English and Indonesian videos.**

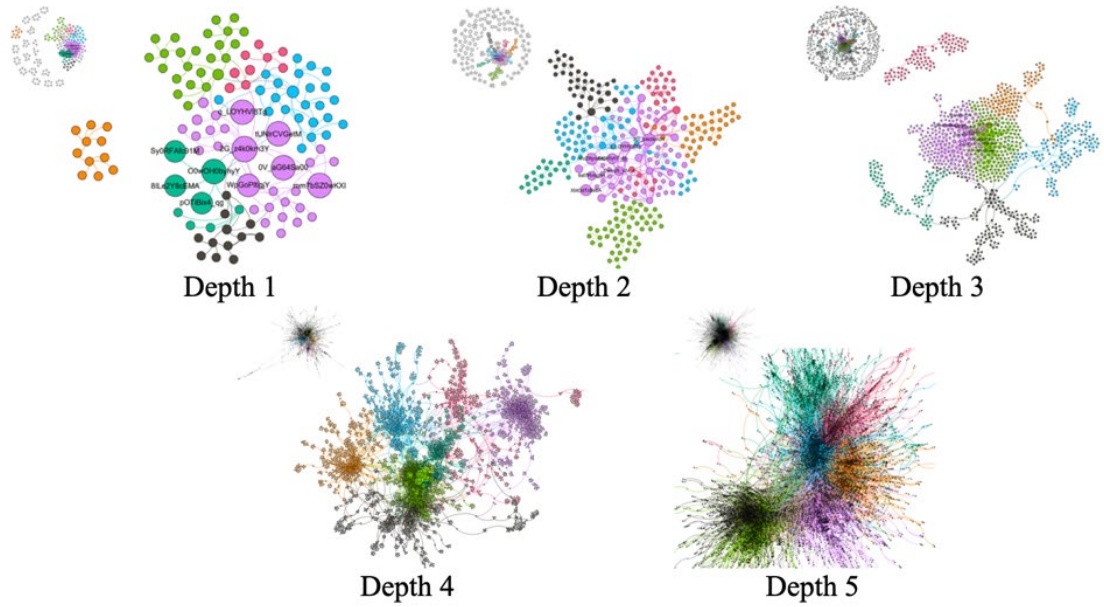



**Figure 22: YouTube video recommendation graphs for each depth.**

The results from the topic drift analysis show high content similarity in the recommendations between depths, but low content similarity with the seeds. Accordingly, the increased similarity between recommendation depths and their divergence from the seed videos is indicative of bias toward a coordinated pool of videos with no relevance to the original topic. Analyzing the network with the eigenvector centrality measure revealed influential videos at each depth. The results showed how YouTube's recommendation system discards the topics of the seed videos by subtly introducing a new but still pro-China topic in the network through influential videos. This new topic is about economic growth and religious freedom in China, targeting Indonesia's younger demographic by focusing on current events and pop culture. The emotion analysis performed on the data showed joy as the most prominent emotion, which aligns with emotions attached to the Cheng Ho narrative and the new narrative that is being pushed through the recommendation system. Overall, the methodology developed in this effort helps identify, track, and measure the impact of the Algorithmic Bias Exploitation tactic.

# 3.0 COLLECTIVE ACTION BASED FRAMEWORK TO CHARACTERIZE COORDINATED COGNITIVE ATTACKS

As disinformation campaigns are increasingly becoming a collective phenomenon, we have developed a collective action (CA) based cyber campaign characterization framework that operationalizes and evaluates processes like collective identity formation, network organization, and mobilization leveraging various theories (viz., social identity, deindividuation, information manipulation theory, motivated reasoning theory, resource mobilization theory, collective action, social movement spillover theory, among others) [34], [35]. The CA framework is combined with the Deviant Cyber Flash Mob (DCFM) model (which measures a mob/crowd's utility, interest, control, and power) [6], [21] and an integrated model called focal-structure analysis (FSA) (which measures the groups' influence in the entire network by solving a dual-optimization problem involving the individual-level measure of the users' betweenness centrality value, and the group-level measure utilizing the spectral modularity method) [36], [37] to identify deviant actors pushing disinformation narratives (Figure 1) [38]. This allows analysts to know which adversarial campaigns to target and develop counter-measures.

The study aims to examine the processes of collective identity formation, mobilization, and network organization and identify various factors that influence the processes. These factors can be classified under cognitive, relational, and environmental categories. We identify concepts/ theories that underpin the factor-to-process relation, thereby providing a richer understanding of the CA dynamics. This multi-theoretic conceptualization of the relation between CA, its processes, and the underlying factors enables systematic and rigorous operationalization and CA modeling in multimedia OIE with explainable and interpretable outcomes. Next, we describe the proposed CA framework. To characterize the role of multimedia OIE in CA-based campaigns, it is essential to understand the underlying factors that influence their outcome. Our CA framework identifies three processes, viz., collective identity formation, mobilization, and network organization. We examine these processes and identify the factors that influence the outcomes of each subprocess. These factors are classified under cognitive, relational, and environmental categories. The relationship between the CA-based campaigns, processes, and underlying factors is illustrated in Figure 23.

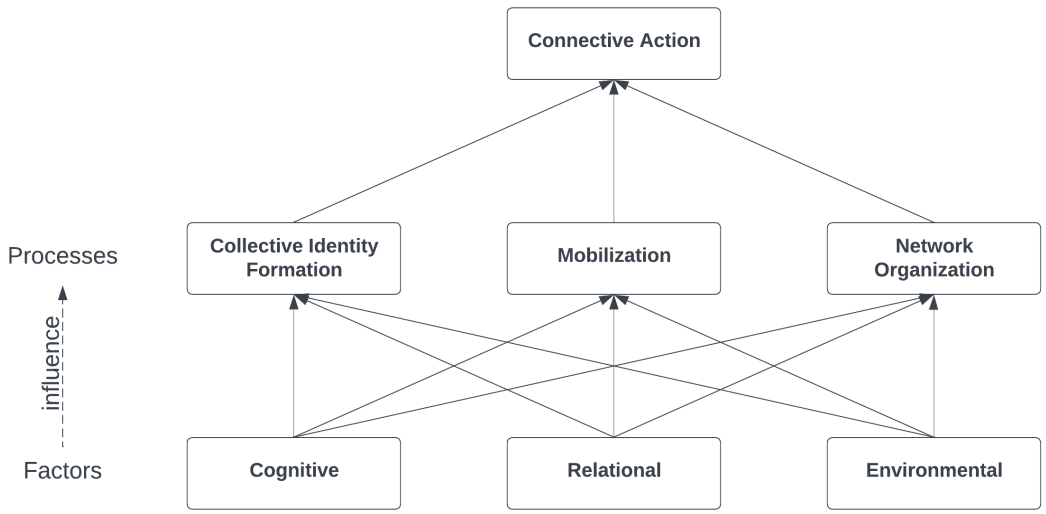



**Figure 23: Collective action (CA) framework showing relations between the processes and influencing factors.**

According to social network analysis theory, the connection and ties between participants are critical for the diffusion of information, individual recruitment, and coalition building. Social media platforms provide various affordances (posting, resharing/retweeting, replying, mentioning, etc.) that lead to different interdependencies among individuals, such as pooled, sequential, and reciprocal interdependence. In pooled interdependence, each user contributes to the collective action without requiring participation from other users. This is represented through the posting of original tweets, independent of others posting contributions to the event. Sequential interdependence considers the output of a particular user as a requirement for the input to another user. This content is usually uni-directional, and as such, the retweet functionality is representative of sequential interdependence. Reciprocal interdependence looks at the outputs of users as they become inputs for others, in a bi-directional flow among actors over time. This can be represented by actors using the "@" mention feature, which allows for a user to directly reply or respond to the actor [31], [39]. It is, therefore, important to identify the key sets of individuals that have the power to mobilize crowds and lead a collective action.

Focal structures are key sets of individuals who may be responsible for coordinating events, protests, or leading citizen engagement efforts on social media networks. Discovering focal structures that are able to promote online social campaigns is important but complex. Unlike influential individuals, focal structures can affect large-scale, complex social processes. In our prior work, we applied a greedy algorithm and bi-level decomposition optimization solution to identify focal structures in social media networks. However, the outcomes lacked a contextual representation of the focal structures that affected interpretability. To address this limitation, we have developed the contextual focal structure analysis (CFSA) model that enhances the discovery and interpretability of the key coordinating units of the network [36]. The model provided context in terms of the content shared by the focal structures through their communication network. The model utilizes multiplex networks, where one layer is the user network based on communications and social network relations (such as mentions, replies, friends, followers, etc.), and the second layer is the content co-occurrence network (such as hashtags, topics, comments, etc.). The two layers have interconnections based on the user content relations. To the best of our knowledge, this work is the first effort to identify influential sets of individuals instead of a set of influential individuals and reveal their interest.

Our model combined two well-known social network analysis methods viz., centrality methods and modularity method to bridge the shortcoming in traditional community detection methods used in graph theory. The resultant combination is a bi-level linear optimization problem to realize/observe the interactions between the user-level and network-level as presented in Figure 24. We explored hidden in tensive groups and ranked them for further investigations. We found key information spreaders in a complex social network by using a bi-level decomposition optimization method in a YouTube channel spreading fake news about the South China Sea conflict where the authors monitored the impacts of suspending these key sets of spreaders from the network [40], [41]. In this research, we identified intensive sets of aggressors, measured their power utilizing the deviant cyber flash mob detection method, and then analyzed the network's changes when a focal structure was suspended from the network. We implemented a comprehensive decomposition optimization model for locating key sets of commenters spreading conspiracy theory in online social networks [42], as depicted in Figure 25. In addition, we combined advanced computational social science and graph theoretic techniques to reveal adversarial information operations in online social networks [43]. We studied computational social science techniques to identify coordinated cyber threats to smart city static infrastructure networks [44]. The model revealed –

1) Significant patterns and trends of popular hashtags such as "#China," "#SouthChinaSea," "#NavyPartnerships," and "#United_States", and part of these hashtags were linked to accounts disseminating information concerning oil and gas exploration and drilling operations, mostly undertaken by the NATO alliances such as the US effort to gain the strongest influence in the region vs. China.
2) Sets of commenters who frequently posted about wars, the connection to the United States, how oil/natural resources are in danger. We saw commenters trying to provoke hysteria about impending US navy's attacks. These focal sets were spreading fake information about weapons and other aggressive topics.

3) Focal sets of commenters posting about western countries' political leaders and leaders of organizations such as UN, ASEAN, EEZ for not taking any serious actions to end the Russia-Ukraine conflict. These focal commenters sets were trying to push for radical behaviors against other countries and other organizations.

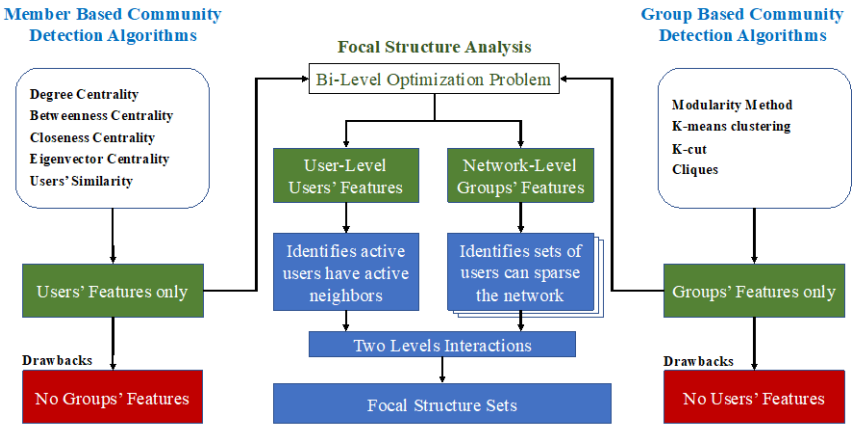



**Figure 24: Focal Structure Analysis, a bi-level maximization network model, i.e., identifying authoritative individuals and identifying communities.**

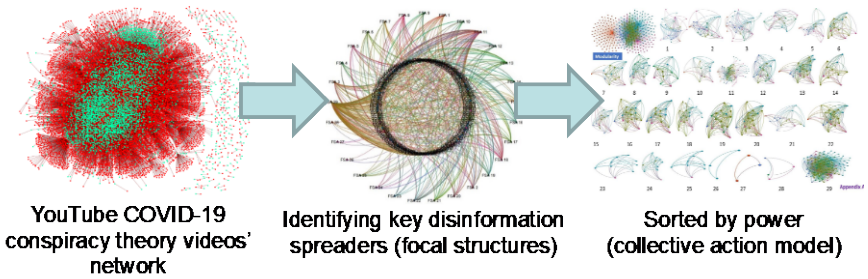



**Figure 25: Multi-criteria optimization formulation to identify focal structures (e.g., key disinformation spreaders on YouTube) and measure their power, utilizing collective action framework, DCFM, and FSA approach. Research showed powerful coordination among conspiracy groups.**

## 4.0 CHARACTERIZING INFORMATION ACTORS ENGAGED IN COGNITIVE ATTACKS

Characterization of population is important as it allows decision makers to inoculate the right group of individuals for effective mitigation of cognitive attack. We developed information actor and consumer characterization approaches leveraging diffusion of innovation theory [45] and epidemiological models (SEIZ – pronounced as 'size') [46]. The model is resilient to presence of bots in the data [47].

We leverage the diffusion of innovations (DOI) theory to show how users adopt into social media campaigns. The psychological and social behaviors that are understood from the DOI model can be applied to social networks to categorize social media users into adoption categories and to use diffusion curves to determine if a campaign has reached its tipping point (see Figure 26). We demonstrate this application of DOI theory to social networks in [48], where we show how the spread of misinformation follows a trajectory

pattern similar to the adoption of new technologies. The framework considers these two approaches to understanding the diffusion of information in a collective action campaign. This view can be used to help shape collective identity and understand whether a collective response is a function of aggregation effects or specific interaction conditions within a group. This research also considers how online social networks are used for the purpose of social media campaigns. The sociotechnical theories of affordance and interdependence amongst users are applied to understand how individual user participation impacts the rate of adoption into an information campaign. The interactions between users in a collective action campaign can provide insights to help determine how future actors will react as information propagates through the network. Leonardi distinguished between these types of actions by labeling them as individualized, shared, and collective affordances [31]. The interdependencies created between users creates a production function that allows us to measure the diffusion of information in a cyber campaign. The initial observations from our analysis support a mathematical characterization for connective action campaigns and show how affordances can be used to identify the inter-user message strategy [49].

We leverage the diffusion of innovations (DOI) theory to examine the trajectory of the overall campaign as it evolves. The second derivative test is then applied to identify changes in the slope of the s-curve. The second derivative test is a method in multivariable calculus used to determine if a critical point of a function is a local minimum, maximum or saddle point. This empirical approach provides an initial quantitative method to find the points at which the slope of the curves shifts from the accelerating stage to the point at which the campaign reaches maximum diffusion or critical mass, and when the slope flattens out or enters the decelerating stage. Since DOI theory has been applied to various disciplines, a key benefit to this approach is that the s- curves are platform independent. Understanding these processes provides a foundation for predicting when an information campaign reaches critical mass.

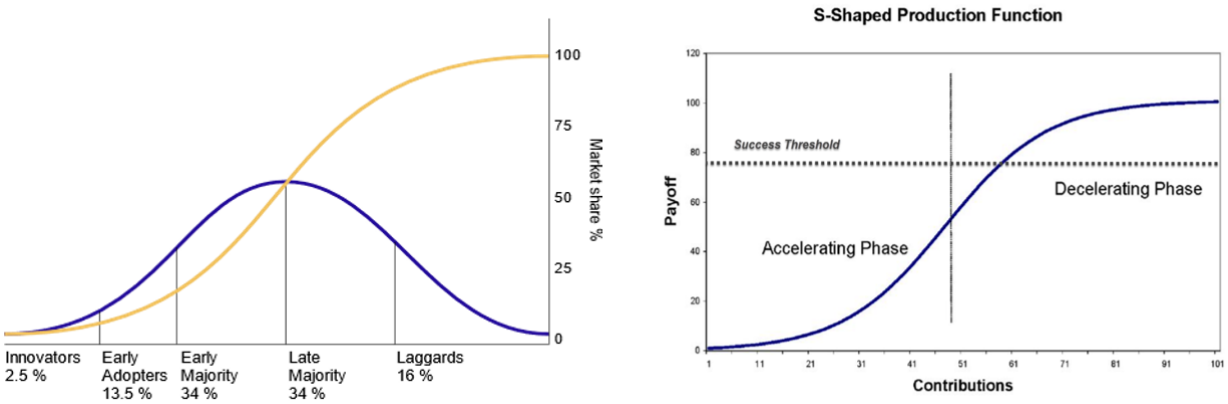



**Figure 26: Diffusion of Innovation (DOI) model and characterizing stages for user adoption (left). The model of s-curve campaign trajectory (right) helps in characterizing its accelerating and decelerating phases.**

In fully adopted information campaigns such as Figure 27, we observe an s-shaped phenomenon. Applying the concept of adoption stages from DOI, we see there is an initial growth stage, accelerating stage, and then a decelerating stage. The COVID-19 misinformation campaign #BillGatesVirus (Figure 27, left) provides a classic example of the s-curve feature showing the full life cycle of how the campaign started slowly, experienced rapid growth, and then slows toward the end of the campaign. Another COVID-19 misinformation campaign, viz., #NoVaccineForMe (Figure 27, right) depicts multiple/cascaded s-shaped adoption, suggesting continuous monitoring of misinformation campaigns for future action. We use DOI theory as a guiding principle and apply the concept to suggest new categories for information actors that align with the cascaded adoption phenomenon, viz., initiators, amplifiers, and sustainers.

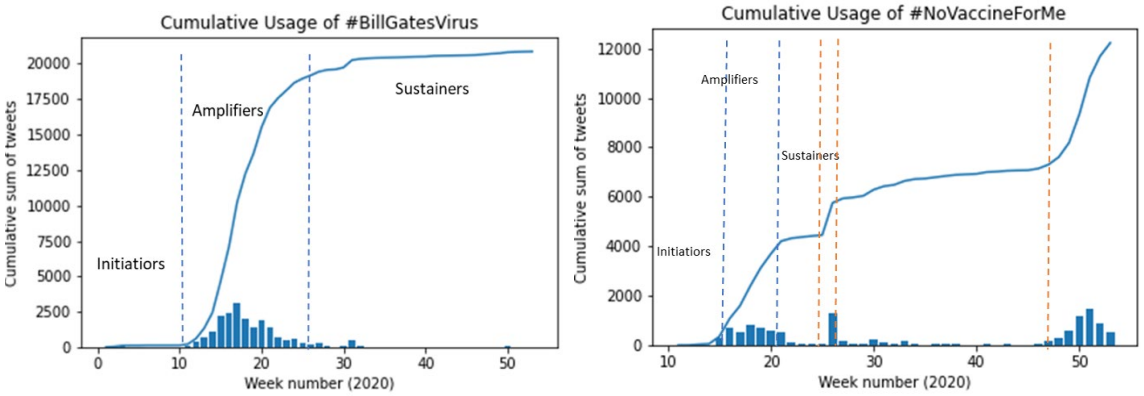



**Figure 27: Information actor characterization based on s-shape connective action process dynamics. Full Adoption Cycle for #BillGatesVirus (left) and Cascaded Adoption Cycle for #NoVaccineForMe (right).**

Applying a mathematical model to understand the trends and dynamics of the spread of influence and opposing narratives on online social networks can help develop techniques to quickly identify and mitigate the spread. The approach presented here attempts to understand the influence that the propagation of information on social media has on the beliefs and behaviors of information consumers and characterize them based on their actions. two most common epidemiological models, viz., SIR and SEIZ are evaluated.

Basic epidemiological models such as SIR are limited by accounting for only one possible path from the Susceptible (S) compartment, which is to enter the Infected (I) compartment. In the case of information as the infection, however, users in the Susceptible compartment have additional paths possible for transition. They can transition to the Infected compartment by deciding to post the information using the specific hashtag or the video. They can also decide not to post but continue to follow the infected user. They therefore move into an Exposed (E) compartment, meaning that they are still at risk. In addition, some users may need time to decide if they believe the information and should spread the hashtags or videos related to them. Users in the Susceptible compartment can indicate that they are decidedly skeptical (Z) of the information. Finally, some users in the Susceptible compartment show no indication of having been exposed to the information. None of these additional possibilities are considered via the basic SIR epidemiological model but are accounted for in the more robust SEIZ model. We applied both the SIR and SEIZ model to our data to illustrate the comparative ability to model the propagation of information on OSN. When applying the SEIZ epidemiological model to Twitter data with the objective of analyzing the propagation of information, the composition of the compartments can be considered as follows: Infected consists of the users who have posted a video. Susceptible consists of users who follow the Infected users. Exposed consists of users exposed to the video (or a hashtag-specific tweet) and, after some time, also commented on the video. Skeptic consists of users who have been exposed to the video but have decided not to engage.

We applied SIR and SEIZ epidemiological models to various datasets with COVID-19 context (e.g., face masks, lockdowns, vaccine, 5G, Bill Gates). Figure 28 shows comparative analysis of SIR (left) and SEIZ (right) models for diffusion of the anti-lockdown narrative. The error reported for SIR and SEIZ models is 26.4% and 6.9%, respectively. In this analysis, we applied epidemiological models to a COVID-19 dataset [50]. Our findings showed that the SEIZ model can efficiently fit the spreaders of misinformation and legitimate narratives more than the other epidemiological models. For example, in cases where the campaign cycle adheres to the s-shape adoption characteristics more strictly (e.g., #Nofacemask and #Wearafacemask), the error for the 'I (Infected)' compartment is relatively low compared to the datasets that do not conform strictly to the s-shaped adoption characteristics (e.g., #lockdownkills) [50], [51].

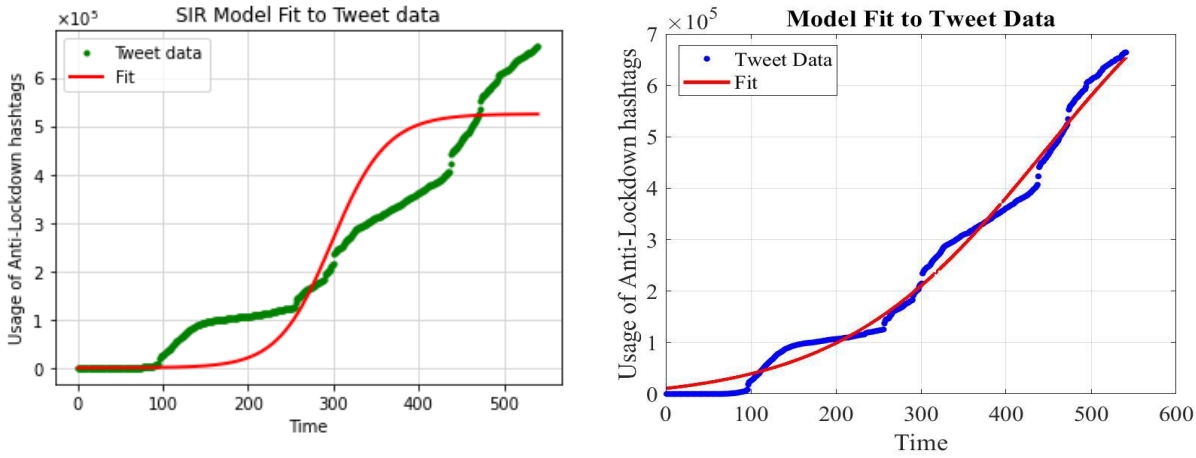



**Figure 28: SIR model fit (left) for anti-lockdown narrative with error = 26.4% (left) and SEIZ model fit (right) with error = 6.9%.**

## 5.0 LOOKING AHEAD

Currently, the United States and our Allies are in the infancy of where participatory media, technology, and policy meet — a lawless Wild West of social media exists—calling for rigorous studies on socio-technical behavioral modeling, content moderation and algorithmic transparency in social media, cyber-threat assessment, cyber-diplomacy, and social computing technologies. Defense planners require a comprehensive understanding of social networks to create effective strategies to control and contain adversarial cyber influence operations. However, reliable and sufficient information about these networks is difficult to come by, often containing missing or erroneous data. Models and analysis tools are required for understanding the flows of expertise, funds (including cryptocurrencies), information, goods, and materials. Tools are also required to understand the command-and-control architectures and communication tactics of adversaries and to develop effective strategies for the containment and disruption of cognitive attacks. This study highlights the need for rigorous and theoretically grounded methods to examine socio-cognitive security threats that are increasingly becoming – 1) a collective phenomenon and 2) multimedia online information environment centric. Toward this direction, we develop multi-model, multi-theoretic approaches that blend computational modeling, big 'social' data, and social science principles to characterize adversarial information campaign dynamics, information actors (producers and consumers), groups/mobs, narratives, TTPs (and measure their impact) in mainstream platforms as well as emerging and multimedia-rich platforms in an integrated fashion, and validate the developed models in real-world use cases (e.g., COVID-19, the Indo-Pacific region). These contributions have also underscored the need for more work as the information environment, behaviors, and cognitive threats co-evolve.

## 6.0 REFERENCES

[1] Al-Khateeb, S., Conlan, K., Agarwal, N., Baggili, I., and Breitinger, F. (2016). Exploring Deviant Hacker Networks (DHN) On Social Media Platforms. Journal of Digital Forensics, Security and Law, 11(2), pp. 7-20.

[2] Calabresi, M. (2017). Inside Russia's Social Media War on America. Time. URL: http://time.com/4783932/inside-russia-social-media-war-america/ Last accessed: 08/16/2023.

[3] Agarwal, Nitin, & Bandeli, K. K. (2018). Examining strategic integration of social media platforms in disinformation campaign coordination. Defence Strategic Communications, 4(1), 173.

[4] Fenstermacher, L., Uzcha, D., Larson, K., Vitiello, C., & Shellman, S. (2023, June). New perspectives on cognitive warfare. In Signal Processing, Sensor/Information Fusion, and Target Recognition XXXII (Vol. 12547, pp. 162-177). SPIE.

[5] Al-khateeb, S. and Agarwal, N. (2019) Deviance in Social Media and Social Cyber Forensics: Uncovering Hidden Relations Using Open Source Information (OSINF). SpringerBriefs in Cybersecurity. Springer, 2019. ISBN: 978-3-030-13689-5.

[6] Al-khateeb, S. and Agarwal, N. (2021). Flash Mobs: A Multidisciplinary Review. Journal of Social Network Analysis and Mining (SNAM). Springer. 2021 DOI: 10.1007/s13278-021-00810-7

[7] Steinblatt, H. (2011). E-Incitement: A Framework for Regulating the Incitement of Criminal Flash Mobs. Fordham Intell. Prop. Media & Ent. LJ, 22, 753.

[8] Tucker, E. and Watkins, T. (2011) More Flash Mobs Gather with Criminal Intent. NBC News. August 9, 2011. https://www.nbcnews.com/id/wbna44077826

[9] Washington Post. (2021) Woman Dies after Shooting in U.S. Capitol; D.C. National Guard Activated after Mob Breaches Building." Washington Post, January 6, 2021. https://www.washingtonpost.com/dc-md-va/2021/01/06/dc-protests-trump-rally-live-updates/ Last accessed: 08/16/2023.

[10] Barry, D., McIntire, M., and Rosenberg, M. (2021) 'Our President Wants Us Here': The Mob That Stormed the Capitol. The New York Times, January 9, 2021, https://www.nytimes.com/2021/01/09/us/capitol-rioters.html Last accessed: 08/16/2023

[11] McNerney, H., Spann, B., Mead, E., Kready, J., Marcoux, T., and Agarwal, N. 2022. Assessing the influence and reach of digital activity amongst far-right actors: A comparative evaluation of mainstream and 'free speech' social media platforms. For(e)Dialogue, Vol 4, Issue 1, 2022. DOI: 10.21428/e3990ae6.60c47409.

[12] Pratley, N. (2021) The Reddit Flash Mob Won't Be Able to Work the GameStop Magic on Silver. The Guardian. February 1, 2021. http://www.theguardian.com/business/nils-pratley-on-finance/2021/feb/01/reddits-flash-mob-wont-be-able-to-work-the-gamestop-magic-on-silver Last accessed: 08/16/2023.

[13] Brignall, M. (2021) How GameStop Traders Fired the First Shots in Millennials' War on Wall Street. The Guardian. January 30, 2021. http://www.theguardian.com/business/2021/jan/30/how-gamestop-traders-fired-the-first-shots-in-millenials-war-on-wall-street Last accessed: 08/16/2023.

[14] Mohilever, Y. S. (2012) Taking over the City: Developing a Cybernetic Geographical Imagination-Flash Mobs & Parkour. In Theatre Space after 20th Century, Thematic Proceedings of the 4th International Scientific Conference in the Cycle Spectacle–City–Identity, R. Dinulović, M. Krklješ and O. Gračanin Eds. Novi Sad: Department of Architecture and Urbanism, Faculty of Technical Sciences, 196.

[15] Marcoux, T., and Agarwal, N. 2021. Narrative Trends of COVID-19 Misinformation. 4th International Workshop on Narrative Extraction from Texts (Text2Story) co-located with 43rd European Conference on Information Retrieval. March 28 – April 1, 2021. Lucca, Italy.

[16] Marcoux, T., Galeano, K., Galeano, R., DiCicco, K., Al Rubaye, H., Mead, E., Agarwal, N., and Galeano, A. (2021). A Public Online Resource to Track COVID-19 Misinfodemic. Social Network Analysis and Mining – Special Issue on Tackling COVID-19 Infodemic. 2021. Springer. DOI: 10.1007/s13278-021-00748-w

[17] Agarwal, N., Mead, E., Spann, B., and Donoven, K. (2022). Developing Approaches to Detect and Mitigate COVID-19 Misinfodemic in Social Networks for Proactive Policymaking. In Coronavirus and Disinformation. Editors: Dr. Ritu Gill (DRDC Canada) and Dr. Rebecca Goolsby (US DoD/ONR). NATO Research and Technology Group (RTG HFM-293). Springer. 2022. https://doi.org/10.1007/978-3-030-94825-2_3

[18] Agarwal, N., Liu, H., Tang, L., & Yu, P. S. (2008, February). Identifying the influential bloggers in a community. In Proceedings of the 2008 international conference on web search and data mining, pp. 207-218.

[19] Al-khateeb, S. and Agarwal, N. (2014). Developing a Conceptual Framework for Modeling Deviant Cyber Flash Mob: A Socio-Computational Approach Leveraging Hypergraph Constructs. Journal of Digital Forensics, Security and Law, Vol. 9. No. 2, pp. 113-127.

[20] Al-Khateeb, S., and Agarwal, N. (2014a). Modeling flash mobs in cybernetic space: Evaluating threats of emerging socio-technical behaviors to human security. Proceedings - 2014 IEEE Joint Intelligence and Security Informatics Conference, JISIC 2014, 7(1), 328. https://doi.org/10.1109/JISIC.2014.73

[21] Al-khateeb, S. and Agarwal, N. (2015). Analyzing Deviant Cyber Flash Mobs (DCFMs) of ISIL on Twitter. In the Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP15), March 31-April 3, Washington D.C.

[22] Akinnubi, A., Agarwal, N., Shaik, M., Okeke, V., & Sunmola, A. (2023). Powering Blogosphere Analytics with BlogTracker: COVID-19 Case Study. In Cyber Security and Social Media Applications (pp. 1-27). Cham: Springer Nature Switzerland.

[23] Marcoux, T., Adeliyi, O., Banjo, D. S., Gurung, M. I., & Agarwal, N. (2023). Exploring Online Video Narratives and Networks Using VTracker. In Cyber Security and Social Media Applications (pp. 115-126). Cham: Springer Nature Switzerland.

[24] Baris Kirdemir, Oluwaseyi Adeliyi, and Nitin Agarwal. Towards Characterizing Coordinated Inauthentic Behaviors on YouTube. The 2nd Workshop on Reducing Online Misinformation through Credible Information Retrieval (ROMCIR 2022) held with the 44th European Conference on Information Retrieval (ECIR 2022). April 10-14, 2022, Stavanger, Norway.

[25] Muhammad Nihal Hussain, Samer Al-Khateeb, Serpil Tokdemir and Nitin Agarwal. Analyzing Disinformation and Crowd Manipulation Tactics on YouTube. In the Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), International Workshop on Social Network Analysis Surveillance Technologies (SNAST), August 28-31, 2018, Barcelona, Spain. Springer.

[26] Recep Erol, Rick Rejeleene, Richard Young, Thomas Marcoux, Muhammad Nihal Hussain, and Nitin Agarwal. YouTube Video Categorization Using Moviebarcode. The Sixth International Conference on Human and Social Analytics (HUSO 2020), October 18-22, 2020, Porto, Portugal.

[27] Baris Kirdemir and Nitin Agarwal. Exploring Bias and Information Bubbles in YouTube's Video Recommendation Networks. The 10th International Conference on Complex Networks and their Applications (COMPLEX NETWORKS 2021), November 30 – December 2, 2021. Madrid, Spain.

[28] Adewale Obadimu, Tuja Khaund, Esther Mead, Thomas Marcoux, and Nitin Agarwal. Developing a Socio-Computational Approach to Examine Toxicity Propagation and Regulation in COVID-19 Discourse on YouTube. Information Processing and Management Special issue on Dis/Misinformation Mining from Social Media. Vol. 58, Issue 5, 2021. Elsevier. DOI: 10.1016/j.ipm.2021.102660

[29] Karen Watts DiCicco. 2022. Toxicity and the Effect on Digital Communities. Ph.D. Dissertation. University of Arkansas at Little Rock. Advisor(s) Agarwal, Nitin. Order Number: AAI29164168.

[30] Connice Trimmingham and Nitin Agarwal. Leveraging Topic Modeling and Toxicity Analysis to Understand China-Uyghur Conflicts. 15th International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS 2022), Springer. September 20-23, 2022, Pittsburgh, USA.

[31] Leonardi, P. M. 2013. "When Does Technology Use Enable Network Change in Organizations? A Comparative Study of Feature Use and Shared Affordances," MIS Quarterly (37:3), pp.749-775.

[32] Fisher M., (2021) Disinformation for hire, a shadow industry, is quietly booming. The New York Times. (2021), July 25. Retrieved 2/15/2023 from https://www.nytimes.com/2021/07/25/world/europe/disinformation-social-media.html Last checked: August 21, 2023.

[33] Mert Can Cakmak, Obianuju Okeke, Ugochukwu Onyepunuka, Billy Spann, and Nitin Agarwal. Analyzing Bias in Recommender Systems: A Comprehensive Evaluation of YouTube's Recommendation Algorithm. In Proceedings of the International Workshop on Mining and Analyzing Social Networks for Decision Support (MSNDS 2023) co-located with the ACM/IEEE International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2023), November 6-9, 2023, Kusadasi, Turkey.

[34] Al-khateeb S., Wigand R.T., Agarwal N., Misinformation Campaigns. Applying Motivated Reasoning and Information Manipulation Theory to Understand the Role and Impact of Social Media in the Digital Transformation. In: Alm, N., Murschetz, P.C., Weder, F., Friedrichsen, M. (eds) Die digitale Transformation der Medien. Springer Gabler, Wiesbaden. (2022). DOI: 10.1007/978-3-658-36276-8_18

[35] Spann B., Agarwal N., A Computational Framework for Analyzing Social Behavior in Online Connective Action: A COVID-19 Lockdown Protest Case Study. Americas Conference on Information Systems (AMCIS). August 10-14, (2022), Minneapolis, Minnesota, USA.

[36] Alassad M., Agarwal N., Contextualizing Focal Structure Analysis in Social Networks. Journal of Social Network Analysis and Mining. Springer. (2022). DOI: 10.1007/s13278-022-00938-0

[37] Alassad M., Spann B., Agarwal N., Combining Advanced Computational Social Science and Graph Theoretic Techniques to Reveal Adversarial Information Operations. Journal of Information Processing and Management, (2020). Elsevier. DOI: 10.1016/j.ipm.2020.102385

[38] Alassad M., Hussain M.N., Agarwal N., Decomposition Optimization Method for Locating Key Sets of Commenters Spreading Conspiracy Theory in Complex Social Networks. Central European Journal of Operations Research. (2021). Springer. DOI: 10.1007/s10100-021-00738-5

[39] Vaast, E., Safadi, H., Lapointe, L., Negoita, B. 2017. Social Media Affordances for Connective Action: An Examination of Microblogging Use During the Gulf of Mexico Oil Spill, MISQ 41, 1179–1205. https://doi.org/10.25300/MISQ/2017/41.4.08

[40] Alassad, M., Hussain, M. N. and Agarwal, N. (2019) 'Finding Fake News Key Spreaders in Complex Social Networks by Using Bi-Level Decomposition Optimization Method', in International Conference on Modelling and Simulation of Social-Behavioural Phenomena in Creative Societies. Springer International Publishing, pp. 41–54. doi: 10.1007/978-3-030-29862-3_4

[41] Alassad, M., Agarwal, N. and Hussain, M. N. (2019) 'Examining Intensive Groups in YouTube Commenter Networks', in in proceedings of 12th International Conference, SBP-BRiMS 2019, pp.

224–233. doi: 10.1007/978-3-030-21741-9_23

[42] Alassad, M., Hussain, M. N. and Agarwal, N. (2021) 'Comprehensive decomposition optimization method for locating key sets of commenters spreading conspiracy theory in complex social networks', Central European Journal of Operations Research. Springer, pp. 1–28. doi: 10.1007/s10100-021-00738-5

[43] Alassad, M., Spann, B. and Agarwal, N. (2021) 'Combining advanced computational social science and graph theoretic techniques to reveal adversarial information operations', Information Processing & Management. Elsevier Ltd, 58(1), p. 102385. doi: 10.1016/j.ipm.2020.102385

[44] Alassad, M. et al. (2021) 'Using Computational Social Science Techniques to Identify Coordinated Cyber Threats to Smart City Networks', in. Springer, Cham, pp. 316–326. doi: 10.1007/978-3-030-64217-4_35.

[45] Spann B., Mead E., Maleki M., Williams T., Agarwal N., Applying Diffusion of Innovations Theory to Social Networks to Understand the Stages of Adoption in Connective Action Campaigns. Online Social Networks and Media, Elsevier. Vol. 28, March (2022). DOI: 10.1016/j.osnem.2022.100201

[46] Maleki M., Arani M., Buchholz E., Mead E., Agarwal N., Applying an Epidemiological Model to Evaluate the Propagation of Misinformation and Legitimate COVID-19-related Information on Twitter. International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction, and Behavior Representation in Modeling and Simulation (SBP-BRiMS 2021), July 6-9, (2021), Washington D.C., USA.

[47] Maleki M., Mead E., Arani M., Agarwal N., Measuring the Interference Effect of Bots in Disseminating Opposing Viewpoints Related to COVID-19 on Twitter Using Epidemiological Modeling. The 56th Hawai'i International Conference on System Sciences (HICSS), January 3-6, (2023), Maui, Hawaii, USA.

[48] Spann, B., Mead, E., Maleki, M., Agarwal, N., and Williams, T. (2022). Applying Diffusion of Innovations Theory to Social Networks to Understand the Stages of Adoption in Connective Action Campaigns. Online Social Networks and Media, Elsevier. Vol. 28, March. DOI: 10.1016/j.osnem.2022.100201

[49] Spann, B., Agarwal, N., Johnson, S., and Mead, E., (2020). Modeling Protester Orchestration through Connective Action: A COVID-19 Lockdown Protest Case Study. 2020 International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction, and Behavior Representation in Modeling and Simulation (SBP-BRiMS 2020), October 18-21, 2020, Washington D.C.

[50] Maleki, M., Arani, M., Buchholz, E., Mead, E., & Agarwal, N. (2021). Applying an Epidemiological Model to Evaluate the Propagation of Misinformation and Legitimate COVID-19-Related Information on Twitter. International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, 23–34. https://doi.org/10.1007/978-3-030-80387-2_3

[51] Maleki, M., Mead, E., Arani, M., & Agarwal, N. (2021). Using an Epidemiological Model to Study the Spread of Misinformation during the Black Lives Matter Movement. https://doi.org/10.48550/arxiv.2103.12191