

DET: Detection Evasion Techniques of State-Sponsored Accounts

Charity S. Jacobs, Lynnette Hui Xian Ng, Kathleen M. Carley

Center for Informed Democracy & Social - Cybersecurity, Carnegie Mellon University
Pittsburgh, PA, United States
{csking, huixiann, carley}@andrew.cmu.edu

Abstract

This study analyzes two covert Chinese bot networks, employing tweet-based and account-based methods to find detection evasion tactics. We reveal the use of message artifacts that disguise spam, engagement strategies that mimic human interaction, and behavioral patterns suggesting algorithmic control. We uncover bot maintenance practices and algorithmic account naming conventions. These insights demonstrate the evolving strategies of inauthentic digital personas, enhance our understanding of online disinformation campaigns, and inform the development of digital manipulation countermeasures. Comparing campaigns in 2021 and 2023, we discover that the techniques used by state-sponsored actors shifted from text-based to image-based techniques, indicating the increased sophistication of these actors to evade the detection algorithms of the social media platform. This work provides insight into the tactics of covert bot networks and discusses possible advancements in detection techniques.

Introduction

In the digital age, social media has become a pivotal area for state actors to wield influence across borders, bypassing traditional channels of engagement. This domain is particularly crucial for countries like China, which uses these platforms to shape global narratives and advance its geopolitical agenda (DiResta et al. 2020; Ng and Carley 2023b). The rise of such digital influence efforts highlights the critical need to understand and counter state-sponsored disinformation campaigns.

The moderation policies of social media platforms regarding state-sponsored content have become contentious issues. The platform formerly known as Twitter has removed labels identifying government-affiliated or state-funded accounts (Reuters 2023), contrasting with continued, but inconsistent, labeling practices on Facebook, Instagram, and YouTube (Kofman 2024; Mac, Isaac, and Frenkel 2022). Weak labeling practices are related to uninformed interactions with state-sponsored content, particularly among users unfamiliar with the origins or biases of these sources (Moravec, Collis, and Wolczynski 2022). Furthermore, the reduction in platform moderation has coincided with a surge in bot activity,

which complicates the landscape of digital discourse (Perez 2024).

Analyzing covert state-sponsored information campaigns poses unique challenges due to the deliberate obfuscation of their origins and the evolution of their tactics to evade detection. The difficulty lies not only in attributing these campaigns to specific state actors, but also in identifying and mitigating evolving tactics designed to evade detection (Rocha et al. 2016). Current techniques to identify inauthentic state sponsored actors rely on bot detection techniques (Ng and Carley 2023a; Feng et al. 2022) or coordination identification techniques (Sharma et al. 2021). However, inauthentic campaigns continue to evolve and adapt as detection research progresses. Furthermore, conventional methods may not fully capture the nuanced strategies employed by state actors attempting to blend in with genuine social media users.

This paper investigates the shifting tactics of state-sponsored information campaigns, focusing on the People’s Republic of China’s (PRC) evolving approach to managing unattributed social media accounts. By comparing data from 2021 and 2023, we explore how these tactics have adapted in response to changes in social media moderation and detection technologies. We evaluate the presence of DETs (Detection Evasion Techniques) and compare the frequencies of deployment of DETs. Our analysis uses a combination of tweet and agent-based metrics to uncover different types of evasion deployed by these campaigns, offering insight into the complex dynamics of modern information warfare.

Methods

We compared two datasets to trace the evolution of PRC social media strategies, particularly focusing on the use of Detection Evasion Techniques (DETs).

2021 Dataset The first dataset, released by Twitter in 2021, encompasses a network attributed to China that was actively disseminating narratives in favor of the Chinese government’s stance regarding the Uyghur population in Xinjiang. This dataset has 16,009 tweets from 1,978 users and serves as our baseline to understand the tactics of PRC state-sponsored information campaigns (Blog 2021; Davidson 2021). Despite Twitter’s partial anonymization and data reduction, we parsed the dataset to a V1 Twitter JSON format to facilitate an in-depth analysis. This conversion al-

lowed us to recover and analyze mentioned and retweeted accounts to identify targeted or amplified accounts, illuminating the reach and influence strategies of the network.

2023 Dataset Our analysis for 2023 uses data queried from the Twitter V2 API, querying the hashtag #DemocracySummit during March 11 - April 20, 2023. This more recent dataset, which includes 90,412 tweets from 63,990 accounts, provided a broader perspective on the tactics currently employed by unattributed networks operating on behalf of China. We specifically analyzed a subnetwork that used the hashtag #USA and performed a two-hop query of agents who used this hashtag, resulting in a small campaign with 571 accounts and 1,237 tweets. Based on previous research on how Chinese diplomatic accounts used the hashtag #WhoDefinesDemocracy to counter the Democracy Summit in 2021, we analyzed this data set for a similar hashtag (Jacobs and Carley 2023).

Secondary Dataset and Bot Detection To further enrich our analysis, we incorporated a secondary data set containing Twitter handles from more than 350 government- and state-sponsored Chinese media outlets (Jacobs and Carley 2023). Additionally, we use a tier-based random forest model Bothunter to distinguish inauthentic actors within this data set, specifically accounts that pass a bot threshold of .70 (Beskow and Carley 2018; Ng, Robertson, and Carley 2022). We were unable to use bot detection on the benchmark dataset due to the data set missing certain key values used for bot prediction to include the account screen names. Lastly, we used ORA network analysis software to generate network mappings and open-source Python tools to statistically analyze the data. In this study, we omit the use of account screen names, adhering to strict ethical guidelines that prioritize privacy and responsible use of digital data.

Identifying Detection Evasion Techniques Our study uses methods to analyze the evolution of detection evasion techniques (DET) in different social media campaigns. We focus on two primary analytical dimensions: Tweet or Message Similarity and Account Network Measures.

- **Tweet or Message Similarity:** This approach addresses the tactics of REPEATER BOTS networks, which disseminate identical or slightly modified messages to evade detection. We assess the similarity of messages using a combination of text analysis techniques, identifying common patterns and artifacts within the tweets.
- **Account Network Measures:** We examine the structure and dynamics of account interactions, particularly through retweets, to understand how networks of accounts are engineered to maximize dissemination while avoiding detection.

Analytical Techniques

Coordinated Tweet Groups For the benchmark data, we used a combination of regular expressions and tokenization comparison to all artifacts except for the *Chinese Phrase* artifact, which we had to find manually due to the obfuscation with other Chinese language words. In determining whether a tweet was part of a specific repeated sequence, we alphabetically sorted the messages, tokenized the tweet text, and

performed a rolling comparison of tweets within a 48-hour period to denote if a tweet had the same text as the tweet before it. We limited tweets that were repeated at least three times and only compared the first three tokens since this bypassed any artifacts that were added to the text.

For the 2023 dataset, we grouped a given original tweet and all subsequent retweets, regardless of time constraints. Tweet IDs were used instead of the text of the tweet due to some accounts retweeting the exact same message at a later date that was then amplified by a sequence of bot accounts.

K-Means Clustering We utilize K-means clustering to partition the data set into distinct groups based on latent patterns within our networks. The assignment of cluster labels to each sender enabled the identification of coherent user segments, characterized by similar behaviors, network positions, and content preferences. We calculated the average tweeting per hour and the total tweets per sender, which served as a proxy for user activity patterns. For both campaigns, we used NLP and basic regular expression parsing methods to group tweets either as a repeated tweet for the 2021 dataset or all tweets and their respective retweets for the 2022 dataset. Additionally, we quantified network characteristics through metrics such as the number of unique groups a sender interacts with and the average number of tweets per tweet group, providing insight into the sender’s engagement diversity and intensity. We assessed the social network structure by deriving In-Degree and Out-Degree centrality measures, indicative of a sender’s popularity and outreach, respectively. The “Number of Followers” metric further enriched our understanding of the influence and reach of an agent within the network.

Entropy Analysis For the benchmark dataset, we calculate an entropy score to measure the diversity of artifact usage in tweets, offering insight into the sophistication of evasion strategies using artifacts to differentiate tweets. We define an artifact within a tweet as a unique string element (e.g., word, number, symbol) that differentiates it from others. The entropy score, $E(a_k)$, measures the diversity of artifact usage by an actor a_k , indicating their adaptability and sophistication in evasion tactics:

$$E(a_k) = - \sum_{T_k \in \mathcal{T}} p(T_k) \log_2(p(T_k))$$

where $p(T_k)$ is the proportion of tweets by actor a_k that contain a specific type of artifact T_k , illustrating the variety of artifacts used to evade detection.

$$p(T_k) = \frac{1}{|M(a_k)|} \sum_{m_i \in M(a_k)} A_{T_k}(m_i)$$

This measure, rooted in information theory, captures the degree of spread in the use of artifact types by an actor, with higher scores indicating a greater spread or variety.

Account Similarity: Analyzing Naming Patterns and Network Roles In our analysis of account similarity, we grouped accounts based on the similarity of their naming conventions, using regular expressions to find user accounts

with the same letter and digit ration. This approach enables cluster analysis of bot accounts or coordinated campaigns based on naming patterns that could indicate automated account creation processes. We also used in-degree and out-degree centrality measures to distinguish between content creators and amplifiers within the network. Specifically:

- In-degree centrality was used to identify key content creators, reflecting accounts that receive a high volume of retweets, mentions, or replies, thus acting as primary sources of information or disinformation.
- Out-degree centrality helped pinpoint the amplifiers, accounts that actively spread the content through retweets, mentions, and replies, thereby amplifying the reach of the content creators’ messages.

To gauge the connectedness among the amplifier accounts, we employed the Jaccard similarity measure:

$$J(i, j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|}$$

where N_i and N_j represent the sets of neighbors for accounts i and j , respectively. This metric provides insight into the extent of overlap in the audiences or networks that different accounts engage with, indicating potential coordinated activity among amplifiers.

Results

2021: Xinjiang Repeater Network

Our investigation into the Xinjiang benchmark dataset revealed significant findings about the strategies employed by covert networks. The dataset, comprising 16,009 tweets from 1,978 agents between January and April 2021, offered insights into the tactics used to refute allegations of human rights abuses in the Xinjiang Uyghur Autonomous Region. We primarily found narratives and themes that a) refuted claims of human rights abuse against Uyghurs, b) spread positive messaging about China’s Xinjiang Autonomous Region, and c) targeted Uyghur activists, groups, and key US politicians such as Mike Pompeo.

Key Account Identification After recovering and merging account names from retweets and mentions, we identified 49 PRC-affiliated accounts and 65 accounts associated with Uyghur activists. The primary narratives promoted by these accounts refuted human rights abuse claims, spread positive messages about Xinjiang, and targeted figures such as Mike Pompeo and Uyghur activist groups. Additionally, we found 9 US Government accounts and 11 accounts associated with western media outlets, almost all of which were affiliated with the BBC outlet.

Detection Evasion Techniques using Artifacts Our analysis revealed various detection evasion techniques (DETs) used by the network to obfuscate its activities. Table 1 categorizes these evasion tactics, highlighting their prevalence and nature. We found that the network changed tactics from January to April 2021, likely in response to Twitter network detection. We found four methods that the network used over time: a) random capitalized 4-letter blocks, b) random Chinese words and characters, c) random punctuation

Table 1: *Detection Evasion Types*: Tweets from covert networks contain character artifacts to differentiate from other similar tweets.

Type	Count	Example
4-Letter	7,168	#Xinjiang Situation: Report finds ‘forced labor’ accusation false url EHTK
Mentions	1,412	Xinjiang’s counter-terrorism measures protect human rights #Xinjiang @fuck_next
Punctuation	670	#Xinjiang Online Vocational education and training is key measure to protect human rights)(
Chinese Phrases	210	#Xinjiang Xinjiang official refutes allegations, welcomes visits [URL] 饜 伧 褛

characters, and d) engagement around a single account to build a network structure. The purpose of these artifacts is likely to differentiate messages and evade detection algorithms for repeatable, bot-like behavior.

- **4-Letter Block** This method was used from February 4 to April 5, 2021 and represented the largest evasion method, with 7,168 tweets. The character block was a random, unique combination of four capitalized letters of the Latin alphabet. The block was primarily at the end of the text, but it was also found in the middle and for Arabic tweets on the left-hand side of the tweet text.
- **Chinese Phrases** This artifact was used in tweets from January 25 to February 3, with 210 tweets. Tweets include a 5-character block of simplified Chinese text. Many of the text blocks are words or phrases, such as 糙反。 (rough) or (Min Zhen stands in front of the oyster) in Figure 1. However, we found many instances of random characters without translation, indicating either a random arrangement or the use of more obscure characters that have not been integrated into the machine translation capabilities of Google Translate.
- **Punctuation** Punctuation artifacts are either at the end of a tweet or buried towards the middle. For this category, there were 670 tweets between January 18-26. An example of this may be where three accounts all tweet the same message, where only the email pointer and punctuation artifact differ.
- **Mentions** This network was different from the other methods in using engagement internally to emulate the network structure, using the Twitter mention and Retweet functions to engage internally. All accounts were related to a single account within a two-hop radius.

Temporal Analysis Figure 1 demonstrates the temporal nature of how the repeater bots used the artifacts. The dotted

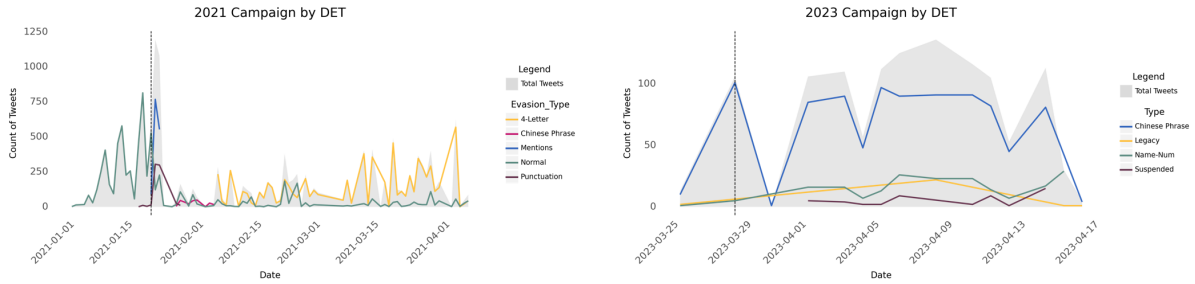


Figure 1: (Left) Timeline for the 2021 covert campaign. The dotted line on January 19, 2021 indicates when US Mike Pompeo made a press release against China’s treatment of ethnic Uyghurs in the Xinjiang autonomous region. After which, we see a peak in tweet messaging and different types of DETs. The 4-letter method endured until the end of the campaign. (Right) 2023 Democracy Summit Campaign by account type: The dotted line indicates the occurrence of Democracy Summit, at the peak of network message traffic.

line in this figure is on January 19, 2021, when then US Secretary of State Mike Pompeo conducted an interview with Fox News and published a press release condemning China’s treatment of ethnic Uyghurs within the Xinjiang Uyghur Autonomous Region in China (DOS 2021). The temporal analysis indicated a significant spike in tweet volume following then US Secretary of State Mike Pompeo’s public condemnation of China’s treatment of the Uyghur population on January 19, 2021. After the Pompeo press release, we see the beginning of tweets made with detection evasion methods, specifically, the added punctuation, 4-letter blocks, Chinese characters, and the use of mentioning other accounts to emulate engagement. We used a nonparametric Mann-Whitney U test to determine whether there was a difference in tweet counts on weekends versus weekdays, as cited in previous studies (Dube 2021). Our findings with a p-value of .475 indicated that there was no difference in the distribution of tweets between weekdays and weekends, indicating that bots probably contributed to most of the content of tweets.

Table 2: Influence of Evasion Tactics on Twitter Activity for 2021 Xinjiang Repeater Network

Var.	coef	std err	P>	z	CI (95%)
Int.	2.890	0.054	0.000		[2.78, 2.99]
4-Let.	0.796	0.095	0.000		[0.61, 0.98]
Chinese	-0.395	0.266	0.137		[-0.91, 0.125]
Mentions	-3.322	0.077	0.000		[-3.47, -3.17]
Punct.	-1.388	0.114	0.000		[-1.61, -1.16]

Logistic Regression Analysis We conducted a logistic regression analysis to understand how various evasion tactics affect the likelihood that a given agent will continue tweeting in the future after a specific evasion tactic ends. We quantified the impact of these tactics on the probability of an account’s continued tweet engagement, utilizing a dummy variable to capture whether an agent persists in tweeting for each type of evasion detection type. The results, presented in Table 2 show varying effectiveness between different tactics. Our findings reveal that our *Mentions* network had the highest probability of not tweeting again, indicating that it

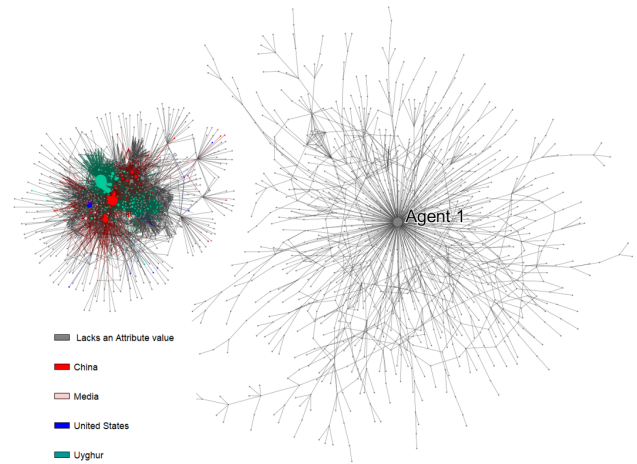


Figure 2: 2021 Xinjiang Campaign agents using an Agent x Agent All Communication network entailing all tweets, retweets, quotes, and mentions. Nodes are sized by *in-degree* centrality, indicating they were retweeted or mentioned. The network has two completely isolated sub-components, one of which conducted amplification of Chinese state media and conducted a doxing campaign against Uyghur activists.

was easily detectable. The *punctuation* category also indicated that the chances that these accounts continued to tweet were approximately 0.25 times those of accounts that did not use punctuation anomalies, suggesting a decreased probability of continued activity, possibly due to effective detection and action by Twitter. This subnetwork tweeted remained isolated, tweeted for three days, and then disappeared from our data set. Lastly, the odds of *4-Letter* accounts continuing to tweet were about 2.22 times higher than those of accounts that did not use this tactic, keeping other variables constant. These findings suggest that Twitter was not as effective in detecting *4-letter* evasion tactics, allowing these accounts a higher likelihood of remaining active. This finding is also supported by the temporal trend of this particular tactic which persists until the end of the data set.

Analyzing the Meta-Network: A Tale of Two Distinct Components Our comprehensive analysis of the Agent x Agent - All Communications network, encompassing all forms of Twitter interactions, showed two distinct component networks within the meta-network (see Fig 2). Each served unique purposes: one contained internal social engagement and interaction, while the other focused on targeting specific individuals associated with Uyghur human rights and amplifying favorable content for the Chinese government. The division between these networks underscores the multifaceted strategies employed in the digital influence campaign.

Component 1: The Mentions Evasion Tactic (Agent 1) Component 1 is characterized by mentions of a specific account, Agent 1, and fosters positive engagement and support for Xinjiang. The Agent 1 network, which comprises 571 agents and spans just three days (January 19-21, 2021), remained completely isolated from the broader dataset. Its primary function appeared to be to promote positive narratives about Xinjiang without directly engaging with state-sponsored content or targeting Uyghur activists. The central figure, Agent 1, despite its inactivity, became a focal point for the simulated participation within this network. The analysis highlighted the use of mentions as a method for spammy accounts to create an illusion of interaction and engagement. The five URLs were linked to state-run Chinese media or government outlets with the following headlines in the China Daily or China.org.cn.

Component 2: Uyghur Targeting & Chinese Content Amplification The more extensive component of the network, consisting of 729 agents and 14,066 tweets, was actively involved in targeting Uyghur activists and amplifying content from Chinese government and media sources. We found 49 PRC accounts, 9 US accounts of senior-level politicians, and 72 accounts of Uyghur Activists and Uyghur Human Rights groups. Tweets that mentioned Uyghur accounts used verbiage that negatively targeted the individuals. Engagement with Uyghur activist accounts and senior US politicians through derogatory and confrontational language points to a deliberate strategy to discredit and intimidate voices critical of China’s policies in Xinjiang.

- “please read, you terrorists and separatists ! @uyghuraccount1 @uyghuraccount2 @uyghuraccount3 @uyghuraccount4 @uyghuraccount5 @uyghuraccount6”
- “UK ‘ban’ on China imports based on untenable accusations by anti-China think tanks, Xinjiang separatists @uyghuraccount7 @uyghuraccount8 @uyghuraccount9 @uyghuraccount10 @uyghuraccount11 <https://t.co/7dXow89UDv> <https://t.co/163zRfoqJx>”

Finding Key Actors that Adapt Through Entropy A deeper dive into the network structure using K-Means clustering and t-SNE visualization revealed clear divisions between the two main components and identified key actors heavily involved in targeting and doxing of Uyghur activists (Fig 3). The analysis of artifact entropy indicated that the top 5% of nodes, characterized by their diverse use of evasion tactics, played a significant role in the network’s more

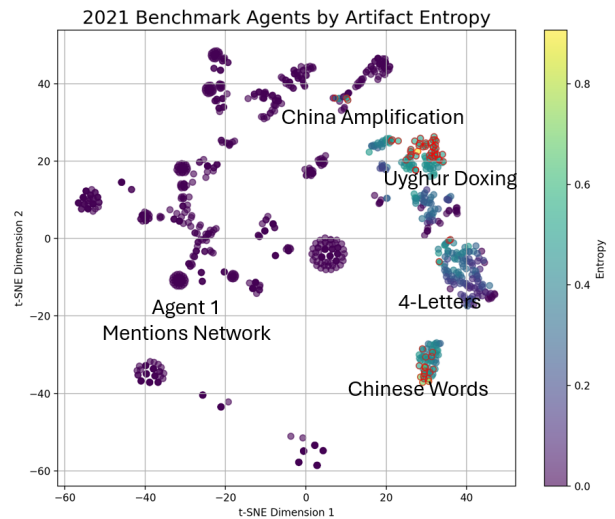


Figure 3: K-Means t-SNE visualization of agent nodes. Agent nodes are colored by entropy score that measures how diverse agents are in using tweets that use more than one type of artifact in their tweet. Key agents were involved primarily in the doxing of Uyghur activists and, to a lesser extent, the amplification of official Chinese government tweets. Top 5% Quantile nodes are outlined in red.

aggressive activities, including the targeting and doxing of Uyghur activists. These actors exhibited higher levels of in-degree and out-degree centrality, indicating their prominence in the dissemination and amplification of the targeted content (see Table 3).

Table 3: Network Characteristics of top Artifact Entropy Nodes and PRC-affiliated accounts. In-Degree indicates tweet generation whereas Out-Degree shows message amplification.

	PRC	Bottom 95%	Top 5%
In-Degree	0.0000	0.0001	0.0002
Out-Degree	0.0007	0.0001	0.0000
Degree	18.08	1.27	5.46
Followers	1,188,282	2.62	3.31

*Note: Centrality values for directional degree measures; Mean values for account type attributes.

2023: Summit on Democracy

In the period leading up to the Summit on Democracy, hosted by the United States (US) on March 29, 2023, we identified an organized network that disseminated negative discourse about the US. Unlike traditional text-based tweets, this network primarily engaged in retweeting messages accompanied by images that contain text, thereby circumventing character limits and complicating content analysis (see example in Fig 6). Originating from a two-hop query centered on the hashtag #USA, this subset revealed a com-

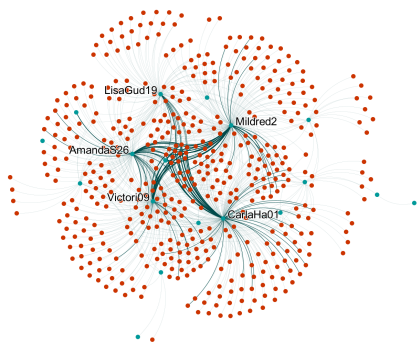


Figure 4: 2023 Bot Network using Agent x Agent Retweeted Network: Labeled green nodes indicate non-bot accounts amplified by red bot accounts with the highest Jaccard similarity score. Link width is relative to the number of connections with each account. All named accounts were within the *Name-Number* account type, while the vast majority of amplifying accounts were *Chinese Phrase* bot-like accounts.

position dominated by bot-like accounts, focusing on the retweeting activity of specific accounts. Although no direct connections were found with official Chinese government accounts, the characteristics of the campaign, such as the use of similar visual graphics, rhetoric, and themes focused on the US-China dynamic, point toward a link with China. Furthermore, key amplified accounts promoted PRC diplomatic activity with multilingual messages, although these posts did not link directly to PRC government or media accounts. This assessment is supported by the alignment with the tactics outlined in cross-platform disinformation campaigns (Nimmo, Eib, and Tamora 2019).

Network Composition and Tactics The network’s core consisted of 15 accounts responsible for generating “original” content, which was then amplified by a larger group of 561 accounts, resulting in 1,237 tweets (Fig 4). Our bot prediction indicated that 93% of these accounts passed the bot threshold for bot-like automated behavior. Despite the extensive activity, the platform’s intervention was minimal, with only 66 out of 571 accounts facing suspension.

Account Templates within the Network This network likely used specific account creation templates for types of accounts, each playing a specific role within the campaign. We found two types of account naming conventions and three overall types of accounts within this network: a) Type *Name-Number* or basic human personas with an alpha numeric ID and screen name depicting a first and last name, b) Type *Chinese Phrase* bot accounts with user IDs that are phonetically based on Chinese screen names and ending in primarily a single digit, and c) *Legacy* accounts that are much older and well-developed personas.

- **Name-Number** (89 accounts): These accounts, predominantly labeled as bots (excluding the 10 original content creators), featured usernames combining female first names with surnames, following a specific alphanumeric pattern. A significant portion displayed a 7-letter and 8-digit combination, with a smaller subset adhering to a 10-

letter and 5-digit format. The 10 original accounts had jaccard similarity scores in the top 5% of the network nodes, indicating that there was substantial overlap in accounts that were retweeting them (Fig 4).

- **Chinese Phrase** (448 accounts): Exhibiting near-universal bot behavior (99.5%), these accounts used screen names derived from Chinese phrases, hinting at their creation via Chinese character keyboards. They primarily functioned as amplifiers within the retweet network, characterized by limited interactions and underdeveloped profiles. These accounts had screen names using Chinese phrases; an example is the account screen name “在厕所卖姨妈巾致富的仙女which translates to “The fairy who got rich by selling towels in the toilet”. The user ID is a phonetic approximation of the Cantonese pronunciation “zicsumiyjnzhf1”, indicating that these accounts were probably created with a Chinese character keyboard.
- **Legacy** (23 accounts): Accounts that were created in 2021 or earlier with typically around 4-5,000 followers. Many of these accounts contained user names related to the words *sexy*, *love* and *girl* or had user names with ethnically Indian names. The three accounts that created original content for this category used video context, one of which was an edited snippet from the Chinese Foreign Ministry.

Analysis of Content and Engagement This network mainly shared articles with embedded images that were seemingly drafted in text editors, often containing grammatical and spelling errors, denoted by underlined wavy red lines. This approach suggests a manual creation process, potentially in a language other than English.

The top retweet and URL shared within this network is a well-developed *Name-Number* account *Mildred2* sharing what appears to be an article created in a text editor (Fig 6). We analyzed the timelines of the 15 accounts that are original tweeters and found additional tweets about the Democracy Summit that were not pulled within the query search. An example of this is the image to the right of Fig 6, which shows a tweet with an image likely created by a text-to-image generator, where the words “Summit for Democracy” are within the embedded text image.

Temporal Activity Our analysis of account creation dates and the account’s last activity revealed a pattern of synchronized tweets on June 12, 2023, primarily from accounts with Chinese naming conventions (see Fig 7). This synchronized activity, possibly using “scheduled tweets,” suggests an algorithmic effort to maintain account visibility. Legacy accounts also had a scheduled tweet on November 15th using the same type of tweet generation (e.g. random content, news article content, etc.). Approximately 26% or 152 accounts last tweeted in April 2023, when most of the campaign occurred around the Democracy Summit.

t-SNE Cluster Map Analysis The t-SNE cluster map analysis (Figure 5) presents a clear delimitation between the different types of accounts involved in the campaign. The majority of the 15 content creators formed a tightly knit cluster, underlining their pivotal role in initiating the cam-

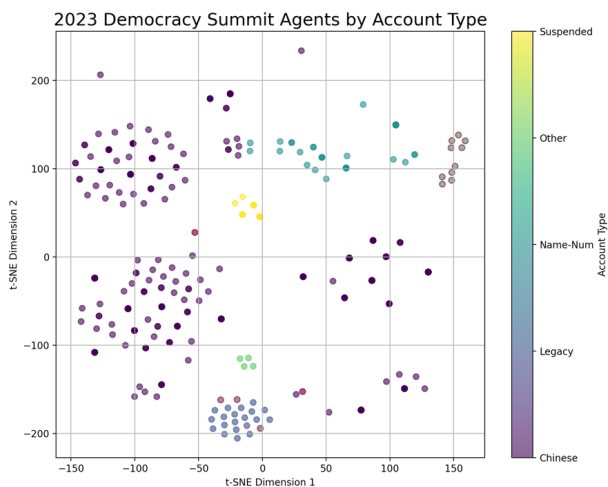


Figure 5: K-Means t-SNE visualization of agent nodes. Agent nodes are colored by their account type: *Name-Number*, *Chinese Phrase*, *Legacy* or other. Key agents were 10 name-number, 4 Legacy and 1 Chinese Phrase account which created tweets that were then amplified. These nodes outlined in red.

paig’s narrative against the Summit on Democracy. Table 4 also shows that these accounts were more connected and had higher follower rates. In contrast, the legacy accounts, despite their well-developed personas and longer presence on the platform, appeared as outliers with minimal impact on the campaign’s overall direction. The most pronounced cluster was formed by the Chinese Phrase accounts, identified as the campaign’s main amplifiers. These accounts, predominantly bots, were instrumental in spreading the campaign’s message across the network, showcasing a high degree of likely algorithmic coordination. The boundaries observed on the t-SNE cluster map underscore the division of labor within the network, where account type played a specific role in improving the campaign’s reach and impact.

Discussion

The transition from text-based to image-based detection evasion techniques marks a significant evolution in the strategies employed by state-sponsored social media campaigns. This shift not only complicates the task of identifying and mitigating these campaigns, but also reflects a broader trend of technological adaptation among actors engaged in digital disinformation. Our analysis of state-sponsored Chinese campaigns during two pivotal events in 2021 and 2023 clearly illustrates this trend. While 2021 saw the use of relatively straightforward text-based DETs to share PRC government content and attack perceived foes, by 2023, these campaigns, while using old retweet schemes, avoided sharing content that could be easily linked to Chinese state-sponsored sites. We found an algorithmically designed system of accounts that sought to mimic grassroots movements more effectively, using hashtags that were general enough to not warrant attention, in addition to image- and video-based content that is not easily analyzed.

Table 4: Comparative Analysis of Tweeters and Retweeters

Account Type	Tweeters	Retweeters
Chinese Phrase		
In-Degree	0.000	0.0007
Out-Degree	0.0031	0.000
Num Followers	805.0	0.12
Degree	9.0	1.8
Legacy		
In-Degree	0.000	0.0003
Out-Degree	0.0099	0.0001
Num Followers	4,287	5908.7
Degree	29	1.31
Name-Num		
In-Degree	0.000	0.0007
Out-Degree	0.0609	0.000
Num Followers	361	99.93
Degree	148	1.61
Suspended		
In-Degree	–	0.0006
Out-Degree	–	0.0
Num Followers	–	489.7
Degree	–	1.65

*Note: The table presents mean values for various characteristics of the network by account type.

These evolving tactics pose a direct challenge to social media platforms and policymakers striving to uphold the integrity of the online public sphere, amid uncertainty regarding what role if any social platforms should play with moderation. The adaptability of these campaigns, as evidenced by their ongoing presence and the adoption of new evasion methods, underscore the limitations of current moderation technologies and policies. The ability of state actors to stay ahead of detection algorithms by experimenting with different techniques highlights an urgent need for continuous innovation in detection methods, as well as a need for more nuanced policy frameworks that can adapt to the changing landscape of digital disinformation. Campaigns are becoming more difficult to detect and less attributed to state-sponsored digital fingerprints. From using simplistic methods such as high-frequency commenting and posting (Schliebs et al. 2021), strategic distraction (King, Pan, and Roberts 2017; Jacobs, Uyheng, and Carley 2023), and AI-generated content creation (Fredheim 2023), the capabilities of state-sponsored actors are growing with the available technology. Tools such as generative AI will only better enable state actors to create more persuasive context that is specifically geared toward an audience, bypassing current language and cultural barriers.

Although our study sheds light on adaptive strategies of state-sponsored campaigns, it is important to acknowledge the limitations in our data set and analytical approach. The snapshot provided by our analysis does not capture the full spectrum of state-sponsored activities on social media platforms. Furthermore, the content that we were unable to analyze due to its multimodal nature, such as video clips and AI-



Figure 6: (Left): The top tweet from the 2023 network. All amplified tweets have attachments featuring an article with an image. (Right): Example tweet not captured by API pull due to image and video content, likely created through text-to-image generation.

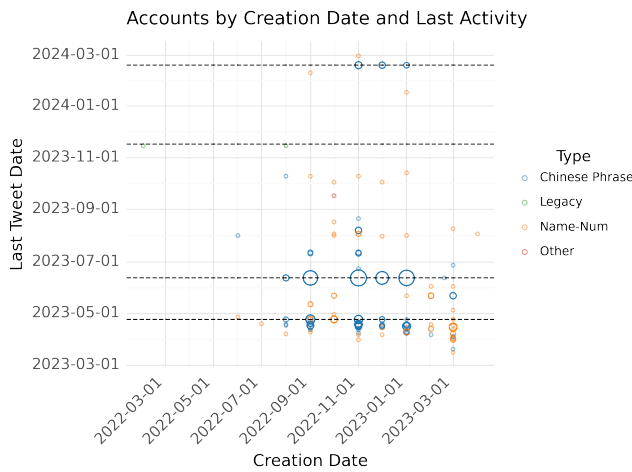


Figure 7: Account types by creation date and last account activity. Nearly half of the network, all Chinese-named accounts, tweeted random unique content on June 12, 2023 in what is likely a scheduled network tweet.

generated images, underscores the need for advanced machine learning techniques capable of detecting these evolving tactics amidst the vast amount of digital content. Future research should focus on developing these methodologies and exploring the broader sociopolitical contexts within which these campaigns operate.

Related Works

The exploration of state-sponsored campaigns on social media platforms demonstrates the multifaceted approach adopted by state actors to manipulate public opinion and shape geopolitical narratives. From Russia and Iran's integration of text and imagery to China's sophisticated use of platforms like Weibo for nationalistic propaganda, the strate-

gies deployed underscore the complexity of modern digital influence operations (Broniatowski et al. 2020; Danaditya, Ng, and Carley 2022).

The digital domain offers state actors the cloak of anonymity, allowing them to conduct covert operations while maintaining plausible deniability (Stout 2017). This aspect of digital geopolitics has been explored in the context of covert Chinese networks, which have been shown to utilize low-effort spammy bots for isolated information campaigns that are easily detectable (Jacobs, Ng, and Carley 2023). China's notorious Spamuflage network has been primarily linked to China's domestic issues, but is reported to be pivoting to target US domestic issues (Nimmo, Eib, and Tamora 2019; Warren et al. 2023; Thomas 2024). Another Chinese covert network, DRAGONBRIDGE, adopted American personas to persuade a local town in Texas to abandon rare-earth mining (Intelligence 2022). As we delve into the nuanced strategies of state-sponsored campaigns, particularly those orchestrated by China, our study seeks to contribute to the research community on digital diplomacy and information warfare, offering new insights for policymakers, social media platforms, and civil society organizations engaged in the defense of a free, open, and transparent information space.

Conclusion

In conclusion, the sophisticated and evolving tactics of state-sponsored social media campaigns represent a significant challenge to the integrity of global information ecosystems. By highlighting the past use of basic DETs and the challenges of emerging tactics that bypass current detection norms, this study contributes to a deeper understanding of modern information warfare. It underscores the importance of continuous innovation in detection and moderation technologies, collaborative international policy efforts, and ongoing prioritization of maintaining a transparent information domain.

Acknowledgements

This work was supported in part by the Knight Foundation and the Office of Naval Research grant Minerva-Multi-Level Models of Covert Online Information Campaigns (N00014-21-1-2765). Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, Office of Naval Research, or the U.S. government.

References

- Beskow, D. M.; and Carley, K. M. 2018. Bot-hunter: a tiered approach to detecting & characterizing automated activity on twitter. In *Conference paper. SBP-BRiMS: International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, volume 3.
- Blog, X. 2021. Disclosing state-linked information operations we've removed.
- Broniatowski, D. A.; Kerchner, D.; Farooq, F.; Huang, X.; Jamison, A. M.; Dredze, M.; and Quinn, S. C. 2020. The covid-19 social media infodemic reflects uncertainty and state-sponsored propaganda. *arXiv preprint arXiv:2007.09682*, 3(2).
- Danaditya, A.; Ng, L. H. X.; and Carley, K. M. 2022. From curious hashtags to polarized effect: profiling coordinated actions in Indonesian twitter discourse. *Social Network Analysis and Mining*, 12(1): 105.
- Davidson, H. 2021. Xinjiang: Twitter closes thousands of China state-linked accounts spreading propaganda.
- DiResta, R.; Miller, C.; Molter, V.; Pomfret, J.; and Tiffert, G. 2020. *Telling China's story: the Chinese communist party's campaign to shape global narratives*. Stanford Internet Observatory Stanford, CA.
- DOS. 2021. U.S. Department of State: Determination of the Secretary of State on Atrocities in Xinjiang.
- Dube, R. 2021. Understanding the Communist Party of China's Information Operations. *arXiv preprint arXiv:2107.05602*.
- Feng, S.; Tan, Z.; Wan, H.; Wang, N.; Chen, Z.; Zhang, B.; Zheng, Q.; Zhang, W.; Lei, Z.; Yang, S.; et al. 2022. Twibot-22: Towards graph-based twitter bot detection. *Advances in Neural Information Processing Systems*, 35: 35254–35269.
- Fredheim, R. 2023. Virtual Manipulation Brief 2023/1: Generative AI and its Implications for Social Media Analysis.
- Intelligence, M. T. 2022. Pro-PRC DRAGONBRIDGE Influence Campaign Targets Rare Earths Mining Companies in Attempt to Thwart Rivalry to PRC Market Dominance.
- Jacobs, C.; Uyheng, J.; and Carley, K. 2023. How China Uses Social Media in Grey Zone Operations toward Taiwan. *Journal of Information Warfare*, 22(4).
- Jacobs, C. S.; and Carley, K. M. 2023. # WhatIsDemocracy: finding key actors in a Chinese influence campaign. *Computational and Mathematical Organization Theory*, 1–21.
- Jacobs, C. S.; Ng, L. H. X.; and Carley, K. M. 2023. Tracking China's Cross-Strait Bot Networks Against Taiwan. In Thomson, R.; Al-khateeb, S.; Burger, A.; Park, P.; and A. Pyke, A., eds., *Social, Cultural, and Behavioral Modeling*, 115–125. Cham: Springer Nature Switzerland. ISBN 978-3-031-43129-6.
- King, G.; Pan, J.; and Roberts, M. E. 2017. How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American political science review*, 111(3): 484–501.
- Kofman, A. 2024. YouTube Promised to Label State-Sponsored Videos But Doesn't Always Do So.
- Mac, R.; Isaac, M.; and Frenkel, S. 2022. How War in Ukraine Roiled Facebook and Instagram.
- Moravec, P. L.; Collis, A.; and Wolczynski, N. 2022. Countering State-Controlled Media Propaganda Through Labeling: Evidence from Facebook. *isre.2022.0305*.
- Ng, L. H. X.; and Carley, K. M. 2023a. Botbuster: Multi-platform bot detection using a mixture of experts. In *Proceedings of the international AAAI conference on web and social media*, volume 17, 686–697.
- Ng, L. H. X.; and Carley, K. M. 2023b. Deflating the Chinese balloon: types of Twitter bots in US-China balloon incident. *EPJ Data Science*, 12(1): 63.
- Ng, L. H. X.; Robertson, D. C.; and Carley, K. M. 2022. Stabilizing a supervised bot detection algorithm: How much data is needed for consistent predictions? *Online Social Networks and Media*, 28: 100198.
- Nimmo, B.; Eib, C. S.; and Tamora, L. 2019. Cross-Platform Spam Network Targeted Hong Kong Protests. *Graphika*, September.
- Perez, S. 2024. It sure looks like X (Twitter) has a Verified bot problem — [finance.yahoo.com](https://finance.yahoo.com/news/sure-looks-x-twitter-verified-183718322.html). <https://finance.yahoo.com/news/sure-looks-x-twitter-verified-183718322.html>. [Accessed 19-03-2024].
- Reuters. 2023. Twitter drops 'government-funded' label on media accounts, including in China.
- Rocha, A.; Scheirer, W. J.; Forstall, C. W.; Cavalcante, T.; Theophilo, A.; Shen, B.; Carvalho, A. R.; and Stamatatos, E. 2016. Authorship attribution for social media forensics. *IEEE transactions on information forensics and security*, 12(1): 5–33.
- Schliebs, M.; Bailey, H.; Bright, J.; and Howard, P. N. 2021. China's Public Diplomacy Operations: Understanding Engagement and Inauthentic Amplification of PRC Diplomats on Facebook and Twitter.
- Sharma, K.; Zhang, Y.; Ferrara, E.; and Liu, Y. 2021. Identifying coordinated accounts on social media through hidden influence and group behaviours. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1441–1451.
- Stout, M. 2017. Covert action in the age of social media. *Georgetown journal of international affairs*, 94–103.

Thomas, E. 2024. Pro-CCP 'Spamouflage' network pivoting to focus on US Presidential Election.

Warren, P.; Linvill, D.; Fecher, L.; Warren, J.; Sheffield, S.; Taylor, J.; Gubanich, A.; Hundley, P.; Lamont, J.; Meadows, S.; et al. 2023. The 5-year Spam: Tracking a Persistent Chinese Influence Operation.