An Overview of Attribution in Cyberspace Influence Operations

Takamichi Saito
(Researcher at Cybersecurity Laboratory and Professor at School of Science and Technology, Meiji University)
March 4, 2024

Abstract

Who is behind the influence operations that are extending the battlefield to social media, and what are their intentions? This paper outlines attribution in cyberspace influence operations and discusses ways to uncover it. As well as presenting concepts, analysis cycles, and models of attribution, the paper puts the models into practice in a campaign conducted just prior to the 2023 G7 Foreign Ministers Meeting. It also discusses the limitations of attribution using information sources closed to cyberspace in terms of estimating intent, obtaining data, and identifying true senders.

Introduction

Assuming readers have backgrounds in IT and data analysis, this paper attempts to outline attribution in cyberspace influence operations (hereafter, attribution) based on the reference [1-4] and previous analyses conducted by the author. The terminology used in this paper is based on these references. To avoid identification, please note that some ambiguous terms are used.

Influence Operations and Attribution

Influence operations are defined here as "a type of information warfare in the competition (conflicts) between states, consisting of a series of actions to influence the decision-making of a competing country and to promote changes in the target's behavior." Its distinctive feature is that it leads to behavioral changes, such as behavioral inhibition. Influence operations have the following objectives. More details are found in the reference [1].

(i) Causing social discord
(ii) Increased support for the ruling party
(iii) Electoral intervention and erosion of legitimacy
(iv) Recruitment

In particular, "causing social discord" is intended to weaken the cohesiveness of a country's leadership and make political decisions on important issues more difficult by, for example, fragmenting public opinion in a hostile country. Influence operations became more sophisticated and systematized with the advent of radio in the early 20th century and the subsequent spread of television. The spread of the Internet has expanded the battlefield to social media. In this paper, the discussion will focus on attribution in cases where actors are covert.

Next, based on the reference [3], attribution is defined as "a process of analysis that attempts to answer who was behind the cyber activity and why they conduct such operations." In general, from a security perspective, the threat posed by a country in a competitive (hostile) relationship is derived as a multiplication of the competing country's capability and that of intention. Therefore, attribution in the context of threat analysis involves, in

part, estimating the intentions of competing countries (hereafter, Estimating Intentions).

Intelligence Cycle

The process of identifying an attribution and providing it to the requester is considered in this paper as an intelligence cycle and is defined by the following steps. Due to space limitations, a general explanation is left to the reference [2] and other sources. This paper supplements such explanations by focusing on the key points.

(1) **Planning & Direction**: Identify objectives and establish information collection strategies.
(2) **Collection**: Collect necessary information, especially Indicator of Compromise (IoC),[1] from specific information sources. It is desirable that information be collected across multiple media.
(3) **Processing**: Consists of sorting, classifying, assessing, and storing. It also includes profiling and chronology.[2]
(4) **Analysis & Production**: Analyze information and produce the required intelligence.
(5) **Dissemination**: Distribute the analyzed intelligence to relevant stakeholders.
(6) **Evaluation**: Evaluate the results of the use and quality of intelligence and undertake the cycle again if necessary.

The “Planning & Direction” step identifies the campaign that is the target of the attribution analysis.[3] It then determines what data will be collected and to what extent, what methods will be used for analysis, and whether immediacy will be required. The tools to be used, the division of roles within the organization, and the cost of inputs should also be determined. In general, attribution for cyber attacks on IT infrastructure is initiated after recognition of the fact that an attack has occurred. In contrast, attribution for influence operations must distinguish between an operation and a spontaneous viral event. Attribution identifies the campaign through narratives and memes.

[1] IoC generally refers to information used to identify and detect system breaches and attacks. In addition to elementary data, memes and narratives are assumed in this paper.
[2] Chronology is defined as a list of events organized according to a temporal order.
[3] A campaign is an organized activity or action with a series of objectives.

In the “Collection” phase, the data needed for the “Analysis” phase is collected from social media, blog sites, and news sites related to the campaign in question. For X (formerly Twitter), the data to be collected includes the volume and frequency of tweets about the campaign, the time of posting (considering time zone, vacation periods, etc.), when the account was created, and the language used. Posts that only use copy and paste or hashtags are also collected. In addition, if machine learning is used to identify bots or estimate location information, the data necessary for these purposes is also collected. Further, information is collected to establish chronology. Where possible, internal information from relevant parties is also collected as non-digital information.

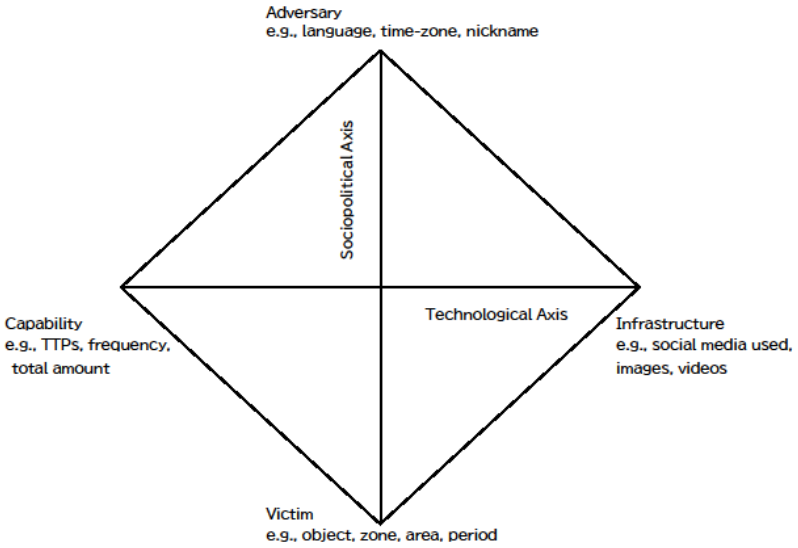
The “Processing” step identifies Intrusion Sets by, for example, grouping similar submissions.[4] It also creates a database (hereafter, DB) of actors and sorts out disinformation. The more data is gathered from similar campaigns, the higher the probability of human error on the part of the actors involved, and the more effective the intrusion indicators can be. Increased DB size also increases statistical accuracy. However, since collecting all the data would be enormous, it is necessary to select what data to collect. In addition, the data must be formatted and stored in a reusable format so that it can be used for analysis in another campaign.

Diamond Model Analysis

This section outlines attribution analysis using the Diamond Model and examples of such analysis. The Diamond Model is an analytical method that generally identifies campaigns along two orthogonal axes: a sociopolitical axis and a technological axis. More details are found in the reference [3]. In attribution, these two axes are used to classify campaigns as “Victims or Adversaries” and “Capabilities or Infrastructure” (see Figure 1).

[4] Intrusion Sets refers to a set of attack patterns or methods used by a particular group of attackers or campaign. This paper assumes news sites and other sites used in a campaign.

Figure 1: Diamond Model applied to Attribution.



Source: Partially modified by the author from the reference [3], p. 32.

Adversaries: Refers to information about the influencer's attributes (e.g., when the account was created, language used, etc.). In addition to account personas, it can take the form of media outlets, such as national and international media, journalists, and foreign embassies.

Capability: Refers to strategies such as hard or soft power,[5] or a combination of tactical techniques, i.e., TTPs (Tactics, Techniques, and Procedures). Capability can include clickbaiting, botting, sock puppetry, trolling, spreading disinformation, doxing, or any combination of these (Intrusion Sets), account linking, and frequency or total volume of posts. For more information, see the reference [1] and [4].

Infrastructure: Refers to the physical or virtual resources used by adversaries to conduct an attack. This includes the social media, blog sites, news sites, and forms of media used.

Victims: Refers to the individual, gender, race, ethnicity, organization, area, or event that is the target of the attack.

In addition to analyzing the campaign in question using the Diamond Model, a DB of past campaigns is created in advance using these two axes, and comparisons are made between past campaigns and the campaign in question. The more consistent the pattern of the Diamond, the clearer the

[5] Soft power refers to methods of exerting influence through culture and values. Sharp power refers to strategic methods of changing the decision-making of other countries through information operations.

attribution to a known group. Then compare with chronology and check for consistency with others, derive insights, and determine attribution for the campaign in question.

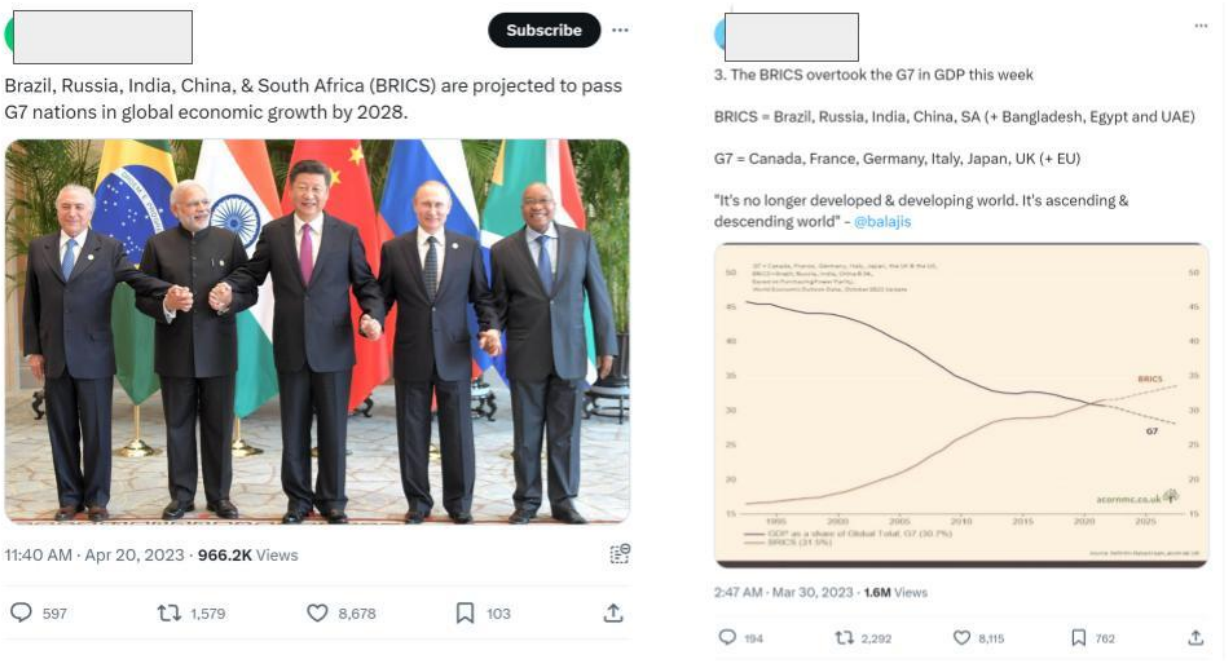
Next, this paper presents an example of analysis using the Diamond Model for a G7 campaign developed just before the 2023 G7 Foreign Ministers Meeting (April 16–18 in Nagano, Japan). The campaign was launched just before the G7 event, and at its peak, more than 10,000 related posts were made repeatedly from multiple accounts per day. Narratives such as "G7's economic strength is declining and being overtaken by BRICS" and the image shown in Figure 2 were "cooperatively spread"[6] as memes on X (formerly Twitter). There did not seem to be much lead to go to specific news sites. As for the "Adversaries," there were signs of organized activity, such as the observation of a certain number of BRICS-related accounts and a certain number of accounts that were suspected to be bots. The "Victims" were English-speaking, with some targeting specific regions of the Global South.

Looking at the chronology, there have been several events since the G7 Foreign Ministers Meeting. For example, the UN Security Council elections were held in June, and two countries from the Global South (Country A and Country S) were elected as non-permanent members. In addition, according to media reports, Country A agreed in July to "strengthen cooperation with China in areas such as security and defense." In Country S, it was reported that the Chinese president sent a congratulatory telegram to the re-elected president of Country S in the presidential election (held in July). In addition, six countries, including Egypt, will become new members at the BRICS meeting in August.

From the above, it can be inferred that the campaign in question was a sharp power-leaning influence campaign, and also that it may have been a campaign centered on an external propaganda organization, based on past examples.

[6] By "cooperatively spread," this paper assumes that multiple actors post the same meme or narrative in a short period of time. For example, posting the same content sentence or only the same hashtag.

Figure 2: X (formerly Twitter) postings from the campaign prior to the 2023 G7 Foreign Ministers Meeting



Source: Collected in October 20, 2023 on the translator's X account (partially masked)[7]

Limitations of Attribution

The limitations of attribution obviously vary between cases where the information source is closed to cyberspace and those where other information sources are used. This section outlines the limitations of methods that are closed to cyberspace.

Regarding Estimating Intentions, the intention is primarily known only to the person concerned. In addition, intent can be changed by different circumstances. Therefore, Estimating Intention is an estimate based on an observation at a particular time. If there is a lack of information or bias, it is impossible to conduct Estimating Intention correctly. Furthermore, it is not always possible to ultimately get the "right" attribution. In the past, some

[7] Translator's note: In the original version of this paper written in Japanese, the source is described as "collected in August 2023 on the author's X account (partially masked and using automatic translation)." To make this figure easy to understand in English, the translator re-collected the tweets from X. Their contents are identical to those used in the Japanese version.

cases have been uncovered by intelligence revelations, but these seem to be rare. The relationship between the campaign and an associated conspiracy theory can also make attribution difficult.

In a typical cyber attack, a relatively large number of traces, such as malware and C2 server information, can be obtained as IoC, but there are limits to the data that can be obtained in an influence operations campaign. In addition, some social media have restrictions on data collection, and in some cases, posters later delete what they post, which can make it difficult to extract all the data. It takes time and effort to find valid and meaningful information from the vast amount of data. This can sometimes be covered by tools.

False flags that intentionally mimic others to mislead the analysis are also assumed.[8] In addition, campaigns using proxies (substitutes) are also envisioned. This is an activity in which the attacking actor does not act directly, but indirectly manipulates or influences through a third party or organization. For example, if an organization launches a campaign, it may use political organizations within the target country through which it can conceal its involvement while conducting the campaign. In addition to receiving direct support in the form of funds, tools, and information, the organization may also use "ideological empathy" to develop a long-term campaign. Especially in the latter case, the original actors may not be traced.

Conclusion

This paper provides an overview of attribution in cyberspace influence operations based on the author's methodology, assuming readers have backgrounds in IT and data analysis. The application of the Intelligence Cycle to attribution is presented, and the campaign for the G7 Foreign Ministers Meeting is analyzed using the Diamond Model. The limitations of attribution in influence operations are also discussed. Although the space limitations of this paper will inevitably engender a number of shortcomings, it can be hoped that readers will gain useful insights.

[8] False flags are hostile actors who intentionally disguise information or imitate other TTPs, thereby failing in attribution.

Acknowledgements

I would like to thank Professor Maiko Ichihara for giving me the opportunity to write this paper.

【translated by】
Takahiro Nakajima (Bachelor's student, Faculty of law, Hitotsubashi University)

Reference

[1] Kazuki Ichida, Takamichi Saito, et al., *Online Public Opinion Manipulation and Digital Influence Operations: Making the "Invisible Hand" Visible* [*Netto Seronkosaku to Dejitarueikyokosaku – 'Miezarute' wo Kashikasuru* (in Japanese)], (Takamichi Saito, Chapter 2, Playbook of Digital Influence Operations, [Dainisho Dejitarueikyokosaku no Pureibukku (in Japanese)], Hara Shobo, 2023).
[2] Atsumori Ueda, *An Introduction to Strategic Intelligence* [*Senryakuteki Interijensu Nyumon* (in Japanese)] (Namiki Shobo, 2016).
[3] Timo Steffens, *Attribution of Advanced Persistent Threats: How to Identify the Actors Behind Cyber-Espionage* (Berlin and Heidelberg: Springer Vieweg, 2021).
[4] Takamichi Saito, A Segment of Public Opinion Manipulation in Information Warfare: Overview of Cyber Influence Operations and Domestic Situation [Johosen niokeru Seronyudokosaku no Henrin -Saiba-Inhuruensuoperesyon to Kokunai deno Gaikyo (in Japanese)], *Defense Technology Journal*, 43(1), 2023, 6-14.

Takamichi Saito Profile

Takamichi Saito is a researcher at Cybersecurity Laboratory and a professor at the School of Science and Technology at Meiji University. He is also Representative Director of Rangeforce, Inc. and holds a Ph.D. in Engineering. He specializes in cyber security and information security technologies. His research interests include web security, browser tracking (browser fingerprinting) technology, and cyberspace influence operations. His publications include *Mastering TCP/IP Information Security* (Second Edition) (Ohmu Sha) and *Online Public Opinion Manipulation and Digital Influence Operations: Making the "Invisible Hand" Visible* (Hara Shobo).