DRAFT WHITE PAPER: An attribution model for influence operations

MICROSOFT'S DIGITAL THREAT ANALYSIS CENTER (DTAC)

January 31, 2023

Executive Summary

- Publicly available standards for attributing influence campaigns remain inconsistent.
- However, some recent studies have provided useful heuristic and normative models which may augur toward a more standardized framework for attributing influence operations.
- DTAC has adopted a modified version of the Pamment and Smith (2022) framework, which considers three categories of evidence—technical, behavioral, and contextual—either in the open source alone or enriched with proprietary (i.e., platform-held) data.
- While the majority of DTAC's attributions will be based solely on publicly available open-source datasets and techniques, Microsoft's telemetry may at times be used to leaven our attributions.
- Once all available evidence from open and proprietary sources are considered, attributions are then expressed in estimative language consistent with US intelligence community (IC) standards.
- Estimative language should lead to a confidence assessment from *Low Confidence* to *High Confidence* and should make use of terms such as *probably* and *likely* and verbs such as *judge*, *assess*, and *estimate*.
- Such judgments are necessarily made using incomplete information, and as such are not meant to imply proof, or certainty.

Context

Over the past several years, academic institutions, think tanks, and social media companies have offered a variety of heuristic and normative models for influence operation attribution. Attributing such operations is a notoriously challenging task which has historically required a combination of technical expertise and geopolitical context. The attribution of influence operations is different in kind from attributions focused on cyber actors. While attributions of the latter are necessarily dependent upon technical signatures, influence attributions must pair those signatures (where they exist) with behavioral and contextual clues and evidence.

On the technical side, researchers may analyze the digital infrastructure used in such operations, including domain and IP address ranges used, as well as the tactics, techniques and procedures (TTPs) which characterize Advanced Persistent Manipulators (APMs). This is complicated, however, by APMs' routine use of sophisticated masking techniques, including the use of proxy servers, virtual private networks (VPNs), and encrypted or obfuscated messaging channels.

The information offered by these technical indicators can be bolstered by geopolitical knowledge and an analysis of (usually) nation-state priorities. This is an imperfect science, as multiple countries often have similar priority sets, and have been observed employing so-called "false flag" influence efforts, whereby a campaign or persona is designed by one nation to appear as though it shares the motivations and background of another.

With the above limitations in mind, this white paper defines an initial framework by which influence operations—sometimes but not always paired with Advanced Persistent Threat (APT) group cyber actions—may be attributed along a confidence interval ranging from low to high confidence.

The Framework

DTAC's framework for attributing influence operations with confidence is based on a July 2022 report by James Pamment and Victoria Smith at the NATO Strategic Communications Centre of Excellence and the European Centre of Excellence for Countering Hybrid Threats.[1] The advantage of this framework is its applicability across both open-source data—which is most often employed by DTAC analysts in their determinations—as well as proprietary "platform" telemetry, which may be used to enrich attributions in limited circumstances.

The DTAC framework borrows both the three categories of evidentiary basis (technical evidence, behavioral evidence, contextual evidence) as well as the two types of sources (open, proprietary) from the NATO paper. A third category of source—namely, classified sources—are not considered in determinations of attribution.

	Technical evidence	Behavioral evidence	Contextual evidence
Open source	Domain ownership, IP ranges, documented financial relationships, etc.	Account or page activity, posting patterns, cross-posting, sharing patterns, social network analysis	Political context, narrative analysis, analysis of media, linguistic markers, possible beneficiaries
Proprietary source	Data sourced through proprietary telemetry or platform backend	As with open source, enriched by proprietary platform data	As with open source, enriched by proprietary data from previous attributions and disclosures

Figure 1: DTAC influence attribution matrix. Adapted from Pamment and Smith (2022)

Borrowing from the NATO paper, these categories of evidence and source are as follows:

- Open-source technical evidence is derived from open-source intelligence investigations (OSINT). This includes publicly available domain registration information, corporate registry or other beneficial ownership ties, information gathered via platform APIs (application programming interfaces), and actor IP addresses (where available and not masked by a VPN).
- Open-source behavioral evidence focuses on the activities and techniques used by accounts suspected to be part of an influence campaign. This includes, for example,

[1] Pamment, James, and Victoria Smith. "Attributing Information Influence Operations: Identifying those Responsible for Malicious Behaviour Online." (2022). https://stratcomcoe.org/pdfjs/?file=/publications/download/Nato-Attributing-Information-Influence-Operations-DIGITAL-v4.pdf

patterns of posting, reposting, and cross-posting, and other amplification methods such as following relationships, and liking and sharing of posts. This may also include more advanced social network analysis techniques.

- Open-source contextual evidence focuses on both the content of influence operations as well as the geopolitical environment in which the campaign or actor operates. This also includes the type of language used and the tenor of that language. Together with behavioral evidence, contextual evidence helps determine which actor might benefit from such an operation.
- Proprietary source technical evidence covers information collected by platforms, and can include account creation date, other technical markers (such as email addresses and egress IPs) used to access the account, and potentially patterns derived from other "off-platform" online activity.
- Proprietary source behavioral evidence includes behavior and indicators not available in the open source, including activity in closed or private groups or countermeasures the actor takes to access or make use of the account.
- Proprietary source contextual evidence primarily consists of collecting patterns based on previous platform-level attributions, which may exist in aggregate in proprietary datasets managed by the platforms themselves.

Estimative Language

Once all available evidence from open and proprietary sources are considered, attributions are then expressed in estimative language as detailed, *inter alia,* by the Office of the Director of National Intelligence.[2] As such:

- High confidence generally denotes judgments based on high-quality information and/or the nature of the issue at hand makes it possible to render a solid judgment. "High confidence" does not indicate a fact or a certainty, and still carries a risk of being wrong.
- Moderate confidence, in general, results from credibly sourced and plausible information. However, such information is either not of sufficient quality, or there is not sufficient corroboration to warrant a higher level of confidence.
- Low confidence generally means questionable or implausible information was used to arrive at a judgment. The information is either too disjointed or too poorly corroborated to make solid analytic deductions, or that there were significant concerns as to the credibility of the sources used.

Attributions should use probabilistic terms such as *probably* and *likely* and verbs such as *judge*, *assess*, and *estimate* to convey analytical assessments. Judgments are necessarily made using incomplete information, and as such are not intended to imply proof or certainty.

[2] National Intelligence Estimate. "Iran: Nuclear Intentions and Capabilities," November 2007. https://www.dni.gov/files/documents/Newsroom/Reports%20and%20Pubs/20071203_release.pdf