T&S Disinformation Attribution Model

Author: Abhas Tripathi
Date: May 19, 2023
Last updated: Aug 4, 2023

Summary

Like in cybersecurity, ascribing a disinformation campaign to an entity, too, is an imperfect science. The Wikimedia Foundation's T&S Disinformation team investigates disinformation threats and, where credible, determines their origin. This document outlines the general framework that the team uses for determining the origin of disinformation campaigns. In professional terminology, this is referred to as "attribution modeling."

Disinformation campaigns are attributed using four categories of evidence, collected largely from publicly available data (with the exception of data gathered using the CheckUser tool and other non-public sources), and we express the likely association, or attribution, as an estimate, using a confidence interval ranging from low to high confidence.

What does T&S Disinformation do?

The Trust & Safety Disinformation team investigates threats made as part of a disinformation campaign and disinformation networks active on our projects.

These investigations focus on identifying two important but challenging attributes-

- the entity (individual or group) behind the disinformation attempt (the "who")
- the motives of the entity in question (the "why")

To this end, T&S Disinformation specialists gather largely publicly available data as evidence, both on-wiki and from other publicly available sources. This evidence is then combined with

- relevant socio-political context,
- knowledge of the language project and its existing administrative issue (if any),
- organizational memory, in the form of historical T&S investigations, and
- the investigators' own experience dealing with similar disinformation attempts

The action of using all available evidence to assess with the best possible confidence that the disinformation campaign was caused by a particular entity is called ***attribution****.*
In essence, T&S Disinformation investigates a disinformation threat, attributes (or associates) it to a particular entity and recommends appropriate actions.

A general framework for attribution

Analysts, researchers and investigators working in the field of disinformation and information operations have different methodologies of collecting evidence and use different frameworks of attribution.

Publicly used attribution models vary in the factors they consider and remain inconsistent. Some recent studies, however, have attempted to provide a more comprehensive, useful and public framework for attribution.

A highly cited attribution model, popularized by Pammet and Smith[1], considers

- three categories of evidence (technical, behavioral, and contextual) and
- three types of sources (open, proprietary and classified)

Attribution in Foundation Disinformation investigations

The Pammet and Smith framework is best used for military operations work, where the availability of proprietary and classified information means an analyst has more than one source to consider for attribution.

For attribution in the Wikipedia ecosystem, almost all data is essentially public, available for anyone to view and access. There are some restrictions based on an editor's user access level, such as data that can be obtained via the CheckUser tool or material that has been deleted may still be viewed by administrators. However, very little information is available only to users of Wikipedia (what would be the equivalent of 'proprietary') or officially-authorized people ('classified').

As such, T&S Disinformation has adopted a modified version of the Pammet and Smith framework, which considers only one type of source and four categories of evidence.

Evidence Categories

The T&S Disinformation Attribution Model considers the following categories of evidence

- Technical
- Behavioral
- Contextual
- Circumstantial[2]

[1] https://stratcomcoe.org/pdfjs/?file=/publications/download/Nato-Attributing-Information-Influence-Operations-DIGITAL-v4.pdf

[2] Even though the terms *contextual* and *circumstantial* are semantically related, when referring to evidence of a disinformation campaign within the Wikimedia ecosystem, the two are distinct categories and provide distinctly valuable evidentiary information

Evidence Category	Examples
Technical	IP addresses, user access levels, sockpuppets, email addresses, CheckUser actions
Behavioral	Account activity including on talk pages, editing pattern, interaction with other editors, social network analysis
Contextual	Social media activity and account information, political context, possible beneficiaries, linguistic markers, geo-political events
Circumstantial	Affiliation and links to institutions, articles of interest, conflict of interest (CoI), biased editing

Estimative Language

Assessments, and related judgements on the *who* and *why* of any IO campaign, are most often made using incomplete information, and as such cannot be certain

As such, after considering all available evidence, any attribution is expressed in estimative language, showing varying degrees of confidence.

- High confidence generally denotes judgments based on high-quality information where 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 are drawn 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 only questionable or implausible information is available to judge. The information is either too disjointed or too poorly corroborated to make solid analytic deductions, or there were significant concerns as to the credibility of the sources used.

The final report uses probabilistic terms such as *probably* and *likely* and verbs such as *judge, assess,* and *estimate* to convey analytical assessments and explicitly defines our level of confidence in our conclusions.