ADAC.iO
Attribution Data Analysis
Countermeasures Interoperability

DELIVERABLE 1.2

JANUARY 2025

Attribution and Information Influence Operations: A 'Field Guide' for Open-Source Investigators and Researchers

Co-funded by
the European Union

About ADAC.io: Attribution, Data, Analysis, Countermeasures and Interoperability

ADAC.io is a Horizon project funded by the European Union and coordinated by the Psychological Defence Research Institute at Lund University. It engages seven partners and has a three-year duration ranging from February 1, 2024 to January 31, 2027.

Based on the concept of Foreign Information Manipulation & Interference (FIMI) as elaborated by the EU EEAS, the purpose of ADAC.io is to protect democracy in the EU by strengthening the ability to deny the intended effects of FIMI on society. ADAC.io hence aims to significantly develop upon current knowledge of how FIMI can be detected, categorised, analysed, shared, and countered.

The project engages the following partners: Alliance4Europe (DE), Debunk EU (LT), Dortmund University - Institution of Journalism (DE), Cardiff University - Security, Crime and Intelligence Innovation Institute (UK); University of Social Sciences and Humanities (PL), Leiden University - The Hague Program for Cyber Norms (NL), Lund University - Psychological Defence Research Institute (SE).

Author(s): Martin Innes; Security, Crime and Intelligence Innovation Institute; Cardiff University
Anneli Ahonen; Security, Crime and Intelligence Innovation Institute; Cardiff University

Cover image: Designed by Freepik

This work was funded by the European Union Horizon Europe research and innovation program [grant number 101132444 – ADAC.io] and the UKRI under the UK government's Horizon Europe funding guarantee [grant number 10105669]. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union, the European Commission or UKRI. Neither the European Union, the European Commission, nor the UKRI can be held responsible for them.

Contents

1 Introduction and purpose

Attribution can be defined as a process of data collection and interpretation used to infer responsibility for authoring or amplifying text and/or imagery, or a series of these orchestrated as a campaign, where the author's identity is obfuscated to some extent. This report describes some of the key challenges when attempting to attribute responsibility for the organization and conduct of information influence operations and disinformation campaigns, and some methods to manage and mitigate these challenges. As guidance it is narrowly focused upon the attribution task, on the grounds that it is routinely a key moment for researchers and investigators as they work with public open-source materials, but also one where mistakes and errors are often made.

In adopting this focus, the report was designed to be read in conjunction with a companion piece authored by researchers at Lund University.[1] The focus of the Lund report is an attribution framework - how data derived from open, proprietary and classified sources can be structured, integrated and used by professional analysts, alongside the other modes of intelligence and information they have access to. Noting in particular how the ability to triangulate across different types of data affords particular opportunities when it comes to attributing particular operations and campaigns. A key discussion point in the Lund work is to set out and illustrate the utility of the NATO Stratcom Centre of Excellence and Hybrid Centre of Excellence Information Influence Operations Attribution Framework (IIO Attribution Framework). In contrast to which, the target audience for this report is the research and investigation community that relies principally upon material derived from public open sources.

Whilst recognizing these distinguishing features, it is equally important to clarify the main points of convergence between the two reports. For example, both adopt a common definition of attribution, and focus upon how attribution works when it comes to 'information influence operations'. The latter term is important as it is commonplace in the literature to deploy concepts of 'information operation' and 'influence operation' quite interchangeably. However, we prefer to blend these terms of the grounds that: (i) Not all influence operations are organized and conducted via the information environment and involve digital data, which is the focus herein; (ii) Moreover, juxtaposing the concepts of 'information' and 'influence' together serves to clarify what the aims of manipulating the data are in terms of inducing a shift in the state of the targeted entity; (iii) The accent upon 'operations' signals how our interest is not in

[1] Palmertz, B., Isaksson, E. & Pamment, J. (2025). A Framework for Attribution of Information Influence Operations. ADAC.io EU Horizon Project Deliverable 1.1. Psychological Defence Research Institute; Lund University.

single episodes of disinforming communication, but linked series of actions performed by operatives linked to a state apparatus in some fashion, that are sustained over a period of time, oriented by a defined guiding objective. Sometimes in the discussion we also include references to 'disinformation campaigns' as a way of acknowledging how non-state third-party and proxy organizations are increasingly enlisted either, wittingly or unwittingly, into state-backed efforts.

Consistent with this framing, this document is also not intended to provide a guide to open-source investigations more generally, as there are plenty of texts doing this already. Nor can it provide a highly detailed account of all the nuances and intricacies of how specific attribution assessments and judgements are made across a range of settings and circumstances. Rather it is positioned so that the content provides a disciplined account of some of the generalizable principal issues that need to be thought about and factored in, when making decisions that impact upon attribution work performed using open-source data.

Accordingly, the guidance is pitched to be useful for researchers and investigators operating across a range of contexts and disciplines, rather than being fitted to the requirements of one specific type of role. We envisage this spanning open-source work taking place across: journalism; policing and public safety; commercial 'intelligence' providers; national security operatives; and civil society efforts to expose human rights abuses, amongst others. The need for common guidance when it comes to constructing attribution assessments and judgements is tied to how, over the past decade, there has been a rapid and significant increase in the use of what are frequently referred to as 'open-source intelligence techniques' (OSINT)[2] across all of the disciplines outlined above.

There have been multiple drivers of this overarching trajectory of development, including:

- The availability of vast volumes of public data generated by new technologies and the associated information environment, especially the internet and social media;
- Growing recognition of the fact that a range of actors, including foreign states and extremist groups in particular, are seeking to exploit and manipulate this information environment, and that their activities can be discovered and detected through careful collection, analysis and interpretation of public data.
- Because discovery of such actions has proven able to command significant political and public attention, functioning as a major source of media stories

[2] Other terms used alongside with OSINT are OOSI and OSI. Online open-source investigations (OOSI) refer to use of only online sources, while open-source investigations (OSI) also use offline sources. For simplicity and clarity, we refer to OSINT throughout.

and publicity for those responsible. In turn, this cycle has attracted considerable funding from a variety of sources, that then supports further work.

Whilst there is much to be celebrated about the growth and development of OSINT methods, there are also several areas for concern, that are growing in intensity and severity. First, the increased interest in the affordances of OSINT and the expansion of professions engaging in applying such methods, almost inevitably induces marked variations in quality. This matters because, as will be elaborated on below, as well as successes, there have also been notable failures in terms of misidentifications and inaccurate descriptions of events. In its most extreme forms this has generated misinformation about disinformation.

At the same time, many contemporary information influence operations and disinformation campaigns have been getting more sophisticated. For not only have those seeking to defend against various forms of information operation learned and evolved through their shared experiences, so too have their adversaries. For example, by increasingly outsourcing the conduct of their campaigns to third-party providers and utilising ordinary digital services to obscure their identities.

Reflecting and responding to such developments, attribution work is fundamentally about issues of identifiability, and ways to 'pierce' various forms of anonymity and pseudo-anonymity adopted by those engaged in crafting information influence operations of different kinds. This can involve various kinds of deception by the operatives implementing influencing campaigns, using a range of 'masks' and 'covers', including:

- Adopting fake personas to pose as 'ordinary' citizens;
- 'Spoofing' the identities of public figures and/or influencers;
- Bots imitating human social media users;
- Creating 'front' organizations, for example posing as a think-tank;
- Mimicking the identities of journalists and media organizations.

This is clearly not an exhaustive list; it is merely sufficient to convey the extent to which there are multiple modes of obfuscation available and utilized on a regular basis in the execution of information influence operations and disinformation campaigns.

Additional challenges are emanating from how the amount of open-source data freely available to researchers and investigators is being increasingly restricted. For example, at the time of writing, Meta has recently decommissioned its CrowdTangle platform, which was widely used in the OSINT community, replacing it with an inferior more closely regulated substitute. Previous to which, X (formerly Twitter) had withdrawn its free API access to researchers. Open-source data is still available, but at a reduced scale and requiring significantly more effort to access. Some investigators are able to partially offset these issues by procuring third-party data collection and analysis packages, but these are often costly and thus can be a barrier to entry. Either way, the

OSINT community is increasingly required to work on a restricted data 'diet', where they know they are making assessments and judgements based on partial material.

There are also increasing political-legal risks to those engaging in OSINT work. Most obviously, are those pertaining to monitoring in or on autocratic regimes. But there are related risks cropping up in democratic countries also. Notably a number of organizations who were involved in monitoring the 2020 US elections for interference have reported repeated threats of litigation and political interference. In sum, the potential 'costs' of making misattributions are increasing.

Framed by these challenges, the purpose of this report is to set out a framework of principles to help guide open-source researchers and investigators to improve the rigour, precision and accuracy of their attribution work for information influence operations and disinformation campaigns. Each of these concepts is crucial to the perspective that is set out in the following pages, and consequently warrant briefly unpacking:

- *Rigour* is about ensuring a comprehensive and systematic approach to collecting, analysing, interpreting empirical data relevant to the subject of an investigation. However, rigour on its own is a necessary but not sufficient for delivering a desired outcome. For a process of enquiry can be highly rigorous but still reach a misleading conclusion.
- Because of this we also need a concept of *accuracy*, framed in terms of precision. Mirroring the comments made in respect of rigour, it is possible for an enquiry to be accurate in terms of the inferences drawn, but to lack rigour. This matters because one purpose of open-source research is to persuade an audience of the claims made about who has done what to whom and why. The chances of this happening is very much enhanced where all claims made are accurate and precise, and supported by rigorously collected evidence.
- *Precision* is about rendering assessments and judgements about responsibility that are specific and evidenced, rather than generic. To put this in more concrete terms, this involves making sure that, whenever possible, responsibility is connected to specific information operations and organizations, rather than the oft used labels of 'the Kremlin' or 'Chinese state' for example. This matters because the key state threat actors support multiple information influence operations, and they vary in their methodologies and tactical objectives.

In engaging with these themes and topics, the rest of this report moves through several sections. First, we briefly lay out our credentials and experience for setting out this guidance. This is followed by a conceptual account of the main ways that attribution claims are constructed and the resources that typically inform and shape these. In turn, this sets up a discussion of the main mistakes and errors made when attributing responsibility for information influence operations and disinformation campaigns. These are categorized as problems of 'over-' and 'under-attribution'.

Having mapped out the principal challenges, the discussion describes some ways of resolving these, informed by the contents of the preceding sections. Interspersed with the more descriptive passages, are some empirical materials intended to illuminate what these more abstract issues look like in terms of 'real world' issues and contexts.

1.1 Credentials and positioning

In drawing this guidance together, we have been informed by three main sources of experience and evidence:

- A decade of experience working for various government departments and agencies doing OSINT investigations and research on a number of threats and topics, helping them to understand the causes and consequences of a number of different information operations and disinformation campaigns. This has included working on or in over forty countries, across problem sets including: wars and conflicts; election integrity and interference; public health protection; the purposive manipulation of scientific research; and public disorder and terrorism events.
- Empirical research conducted on professional OSINT police officers working for UK Counter-Terrorism policing and the National Crime Agency. This involved observing their work and interviewing them in-depth about how they do what they do.
- A history of studying the organization and conduct of police crime investigations, and the recursive interactions between knowledge and action that occur in such inquiries. The lead author was responsible for a series of pioneering academic studies 'investigating the investigators' in terms of how police construct their knowledge and understanding about how criminal responsibility for harmful acts is ascertained, and the patterns of action and interpretation that shape such judgements.

It is this blend of insights, triangulated from a range of sources and perspectives, that underpins the position that we have adopted. In addition to which, we draw in key ideas about how to do attribution from several relevant disciplines. For example: the investigative journalist group Bellingcat has several guides and tools available online[3]; Columbia Journalism Review has an extensive OSINT guide available;[4] the Global Investigative Journalism Network has published on how to detect and analyse troll campaigns; [5] and the European External Action Service has guidelines on conducting

3 https://www.bellingcat.com/category/resources/, https://www.bellingcat.com/resources/2020/12/14/navalny-fsb-methodology/

4 https://www.cjr.org/tow_center_reports/guide-to-osint-and-hostile-communities.php

5 https://gijn.org/resource/investigating-digital-threats-trolling-campaigns/

OSINT analysis on identity and gender-based disinformation.[6] Many media outlets have their own guidelines for how journalists should practice techniques like going undercover online and concealing their true identity,[7] or using anonymous sources.[8]

Compared with the current state-of-the-art for information influence operations, a more clearly developed position on the specifics of 'how to' attribute a digital campaign is to be found in the cyber-security domain, where key concepts have inspired and influenced the attribution debate with regards to information operations.[9] Notably, the discourse about detecting and evidencing the Tactics, Techniques and Procedures (TTPs) used by a defined threat actor when attributing cyberattacks, has had a direct and strong shaping effect on the information influence operations analysis community. However, Pamment and Smith (2022) have critiqued this direct 'translation' on the grounds that the attribution of information influence operations is fundamentally different from the cyber field. They contest that information influence is a communication problem, where behavioural and contextual evidence are more important. Moreover, interfering in public debate, is in most cases, not illegal. Following this line, Hedling and Ördén (2024) discuss how attribution, non-attribution and diffused attribution has been deployed by governments in politically charged situations.[10]

Such critiques notwithstanding, some elements of cyber-security's approaches to attribution may still be helpful in understanding the challenges of responding to information influence operations. For example, Healey (2012) establishes a 'spectrum' of state responsibility as a tool to help analysts to assign responsibility for cyberattacks. Ten categories are marked with a degree of responsibility, starting from a very distant, passive degree of responsibility, through to actively planning and executing an attack.[11]

6 https://www.eeas.europa.eu/sites/default/files/documents/2024/EEAS-DataTeam-OsintGuidelines-04-Digital.pdf

7 https://www.nytimes.com/editorial-standards/guidelines-on-integrity.html

8 https://blog.ap.org/behind-the-news/when-is-it-ok-to-use-anonymous-sources

9 https://stratcomcoe.org/pdfjs/?file=/publications/download/Nato-Attributing-Information-Influence-Operations-DIGITAL-v4.pdf?zoom=page-fit, https://www.hoover.org/sites/default/files/research/docs/lin_webready.pdf, https://cyberdefensereview.army.mil/Portals/6/Documents/CDR%20Journal%20Articles/Determinants%20of%20the%20Cyber_Kostyuk_Powell_Skach.pdf?ver=2018-07-31-093725-923, https://www.tandfonline.com/doi/abs/10.1080/01402390.2014.977382

10 Hedling, E. & Ördén, H. (2024). Disinformation, Deterrence and the Politics of Attribution. *International Affairs*. https://lup.lub.lu.se/search/publication/6681ef5c-3eab-4066-8eea-ab21206662b8

11 https://www.atlanticcouncil.org/wp-content/uploads/2012/02/022212_ACUS_NatlResponsibilityCyber.PDF

1.2 Case study #1: A Campaign Against Sweden's Childcare and Complex Attributions

Prior to engaging in a more analytic register about the challenges of and possible solutions for constructing persuasive attributions, it is perhaps first useful to start with a case study that articulates some of the complexities of attribution work. The case has been selected because of how it illuminates some of the dependencies that shape decisions about who can attribute and how, as well as some of the consequences that flow from these.

The case concerns an aggressive influence campaign targeting Sweden's childcare that evolved during late 2021 and peaked in the first months of 2022.[12] The narrative was that Sweden's social services were accused of forcibly taking immigrant children away from their parents – a message which had been previously authored and amplified by foreign state actors, including Russia, in their disinformation campaigns. As a case, it sheds light into how attributing works when facing a complex campaign, where both foreign and domestic actors play a role. In so doing, it foregrounds questions about which parts of a complex, multi-faceted campaign should be attributed and by whom?

There were multiple triggers that seeded the base narrative and public interest in it, including a report by Swedish radio stating there is a greater risk of children of immigrants being taken into care by social services.[13] Subsequently, the campaign came to be named after the Swedish law from 1990, "LVU". This act gives Swedish social services the power to act to protect children and young people under the age of 21. As noted in a study on the campaign commissioned by Sweden's Psychological Defence Agency,[14] the issue of the role of social services had been an emotionally charged topic among several communities in the Swedish population for a long time. In particular, rumours and conspiracy theories about the mandate and mission of social services had existed for years and circulated within Swedish Salafist and Salafi-jihadist movements. Before the viral LVU campaign, there were rumours and claims that social services take Muslim children away from their families without any valid reason, or even sell them. Claims that were reheated and resurfaced by the LVU campaign.

It appears that the origins of the campaign in 2021-22 were in interactions that took place between Swedish activists and Moustafa El-Sharqawy, better known for his online presence @Shuounislamiya with a large following. In early February 2022, the campaign gained international momentum when social media posts about the allegations made by El-Sharqawy and an Egyptian YouTube personality Abdullah El-Sherif went viral. Key campaign messages included, amongst others, that Sweden is a fascist state where social services deliberately kidnap Muslim children to forcibly turn

[12] https://www.fhs.se/download/18.32d29dd2187bd01d5e455265/1682576119173/LVU-kampanjen.pdf

[13] https://sverigesradio.se/artikel/barn-till-invandrare-tvangsomhandertas-oftare-av-socialtjansten

[14] https://www.fhs.se/download/18.32d29dd2187bd01d5e455265/1682576119173/LVU-kampanjen.pdf

them into Christians, or place them in paedophile families. Later on, prominent preachers with ties to Islamist and radical Islamist movements also started promoting the campaign's messages. This in turn encouraged international media outlets like Al Jazeera, TRT, and Al Arabiya to report on the campaign, amplifying its message. International Muslim organizations and scholars signed a petition called #InDefenseOfChildren. The petition demanded an end to child custody interventions. Some Swedish Muslim representatives condemned the campaign, while others acknowledged problems with LVU and the Swedish social services.

A growing sense of concern that the online campaign was gaining momentum was reinforced by physical demonstrations in Stockholm, Gothenburg, and Malmö, that gained attention from Swedish and international media, particularly Arabic and Turkish-speaking outlets.

This is an appropriate juncture at which to pause, in order to define the multiple challenges in play in such a situation. There are after all multiple actors, some domestic but others internationally based, operating across different media channels, making different claims. Thus, there are considerations about which components of the campaign warrant a public attribution and why, and equally who is positioned to do so? Not least because in most Western democracies there is a clear division of labour amongst state security agencies between those engaged in mitigating internal threats, and those with a foreign policy remit. The point being that attribution decisions are not just data and analysis problems; they frequently also have a political dimension as well. Given that prior to any attribution work being undertaken it is frequently highly uncertain about the point of origin for a threat, this can induce a degree of inertia in terms of who has investigative primacy, at least when it comes to state responses.

In terms of the response to the LVU case, the newly established Psychological Defence Agency's mandate was limited to countering foreign actors' undue influence campaigns, and it knew it had to effectively communicate its decisions when reacting publicly. In the mean-time, on social media, those who tried to mitigate the campaign's effects were targeted with hateful or negative comments. Online discussions were occurring as to how Swedish institutions and embassies can and should be targeted abroad. Sweden's Arabic language media house's communications were shut down on social media due to the campaign's activities, and they couldn't reach their audiences. Personal information of social workers in Sweden was published online and they were threatened.

The defensive response started by taking the topic to the national crisis coordination agency to co-ordinate across relevant stakeholders. Following on from which, the Psychological Defence Agency decided to attribute and expose the main threat actor via an interview to Swedish media.[15] The post event evaluation study commissioned by the Agency on the LVU campaign and the responses to it concluded the Agency's

[15] https://sverigesradio.se/artikel/hotkampanj-mot-sverige-uppmanar-till-terrordad

actions helped in providing a common situational awareness and priority actions across state and local actors.[16]

The same evaluative study also noted a lack of counterforces against the disinformation and narratives spread by the campaign, despite efforts by organisations like the Swedish Institute (SI) and Alkompis, a Swedish Arabic-language media house, to address the disinformation online. Critically, the study noted how widespread distrust towards Swedish institutions posed a fundamental problem for many aspects of the control responses, including the willingness of key communities to believe the Agency's attribution of responsibility.

In this case, the involvement of a foreign actor made it legally permissible for the Swedish Psychological Defence Agency to intervene and coordinate responses towards those they identified as responsible for the harmful campaign. However, attributing only one part of a complex campaign raises questions about if it is proportionate to call out a foreign actor's role, while much of the harmful activity is conducted by domestic actors? At a conceptual level this moves us towards differentiating between 'single attributions' and 'complex attributions.' The former referring to tasks where the focus is on assigning responsibility for a specific operation to a single entity. This is contrasted with more complex attributions, where the harmful communications are being propagated by multiple actors, some of whom may be domestically situated and others internationally, and unconnected from each other. As a consequence of which, decisions have to be taken about what elements of the overarching campaign can and should be attributed and by whom, alongside the more routine challenges.

[16] https://www.fhs.se/download/18.32d29dd2187bd01d5e455265/1682576119173/LVU-kampanjen.pdf

2 The social organization of open-source attribution work

Open-source intelligence or OSINT is something of a 'plastic' catch-all term used to describe a range of similar but distinct activities, performed by different organizations with a range of aims and purposes. This matters because not all of what gets labelled as 'OSINT' is the same. That said, typically, there is a shared commitment to using public data, most often derived from digital sources. Although this is sometimes blended with non-public sources.

More contestable, is whether most OSINT properly constitutes 'intelligence' or not. For as a concept intelligence is generally held to result from information being processed and interpreted to afford foresight. Much 'open-source intelligence' work does not possess this future-facing orientation, instead describing and explaining the causes and consequences of events that have happened. Consequently, such cases are more accurately thought of as forms of open-source research or investigation. Herein, we use the terms interchangeably.

Understanding this plurality and diversity of applications is important in starting to frame a position about the key knowledge and skills attributes of effective open-source researchers and investigators. In many accounts of and 'training manuals' for OSINT, the discussion routinely becomes focused upon the need for specific kinds of subject matter expertise, and the technical skills needed to perform particular kinds of analyses, whether that be using images, text or quantitative data. For our purposes here however, replicating such discussions is unhelpful. For what we are seeking to identify are some generalizable and transferable principles to aid the task of attribution, whatever modes of data are being utilized.

Positioned in this way, when it comes to OSINT work, it is potentially helpful to differentiate between the need to possess 'technical-digital' and 'investigative' skills. The former captures how operators need to understand the particular affordances of different social media platforms, and various web technologies, if they to be able to manipulate data from these sources on a regular basis and at scale. This is different knowledge and expertise from that which is required to conduct an effective investigation. Research on investigations has highlighted the importance of a number of qualities and attributes of effective investigators, including:

- A methodical and systematic approach to identifying, collecting, processing and interpreting data from a range of appropriate and relevant sources;

- An ability to construct plausible 'abductive' inferences. Abduction is reasoning to the best explanation to explain the causes of an event or situation, based on limited and incomplete information.
- A familiarity with the legal and regulatory structures within which one is working, in order to know how and when particular boundaries can be pushed and stretched, but not necessarily breached.

Consistent with which, albeit describing it in a different way, OSINT investigations are typically blending **art** (intuition and 'feel for the game'), **craft** (skills in understanding and manipulating different data) and **science** (uses robust methods, hypotheses and falsification).

Reflecting the range of applications of OSINT and the variety of backgrounds that OSINT analysts increasingly come from, it is quite usual to find that some such workers are more skilled and competent in the data handling elements of their role, where others are more oriented to the investigative skills involved in identifying leads and anomalies, and case building. Distinguishing between the two different principal skill sets in this way, and the fact that individual open-source researchers and investigators may have higher and lower capabilities, helps us to break down and map out the work into four main profiles, as summarized in Table 1 below:

	High Investigative Skills	Low Investigative Skills
High Data Skills	**OSINT Operators** – seamlessly blend art, craft and science.	**Data Crunchers**; academic researchers – often strong methods but miss practical exploits.
Low Data Skills	**Online Investigators**; practitioner backgrounds – lack technical data skills, but have specialist knowledge.	**Impressionists** – mimic the discourse, but don't deliver insight and understanding.

Table 1: Profiling Open-Source Data and Investigation Skills

The salience of this approach is it helps to clarify aspects of the challenge of attribution using open-source data. The point being that, to some degree, the nature of the challenge is shaded by the skills that different researchers bring to their work.

A key point is that only a minority of people are able to combine high technical-data knowledge and skills, with the ability to creatively exploit these to build compellingly evidenced cases that clearly articulate the involvement of specific entities in authoring and/or amplifying harmful online behaviours. More numerous within the open-source investigation community, are individuals and groups who have a particular predilection for either investigative work, or for processing large amounts of data using data science methodologies, but are perhaps less adept in understanding how the results can (or cannot) be practically exploited. Equally, it needs to be recognized that as OSINT has

increased its profile and status, it has also attracted the involvement of some individuals who are neither trained, nor skilled, but who invoke the language and terminology of OSINT, in an attempt to further their own idiosyncratic agendas. At a minimum, the take-home point of the preceding discussion is that there are different ways of doing open-source investigations, and hence different ways of making attributions.

It is also worth pointing out that the risks for individual journalists, media organizations or non-governmental organizations are significantly higher than for state entities making attributions. Impartiality and balanced and unbiased coverage are core rules in journalism, and the stakes when attributing information influence operations, like any other malign activity, can be high.[17]

For example, three Russian oligarchs initiated libel proceedings against journalist Catherine Belton and Harper Collins the publisher of her book "Putin's People". In the end, the defendants settled or withdrew their claims. The publisher agreed to amendments related to the Russian oligarch Roman Abramovich. Carole Cadwalladr was sued following her revelations about Cambridge Analytica, with international journalist associations deeming the lawsuits as "intended to silence Cadwalladr's courageous investigative journalism".[18] Some Russian oligarchs have managed to get themselves removed from the sanctions list after challenging the decisions in court,[19] but the EU's general court backed the ban on RT which appealed the designation.[20] One of the Russian information operation contractors, ANO Dialog, announced in October 2024 it will sue the United States' Federal Bureau of Investigation for "false allegations."[21]

The point being that there is a clear trend for individuals and organizations associated with the Russian state, who are publicly labelled as responsible for harmful activities, to increasingly resort to 'lawfare' methods to try and suppress such claims. It is a trend that open-source researchers need to be aware of, and reinforces the important of ensuring rigour, accuracy and precision, when making attribution statements.

[17] https://reutersinstitute.politics.ox.ac.uk/risj-review/how-to-rethink-impartiality-digital-age, https://www.reutersagency.com/en/about/standards-values/#:~:text=A%20Reuters%20journalist%20shows%20integrity,our%20sources%20and%20our%20readers.

[18] https://stratcomcoe.org/pdfjs/?file=/publications/download/Information-Conflict-DIGITAL-FINAL.pdf?zoom=page-fit

[19] https://www.dw.com/en/russian-oligarchs-win-appeal-to-partially-annul-eu-sanctions/a-68788005

[20] https://www.dw.com/en/eu-court-upholds-ban-on-russian-broadcaster-rt/a-62618278

[21] https://t.me/uranews/102869)

3 Over-attribution and under attribution

Prior to setting out more specific details about how constructing plausible and evidenced attributions can be conducted, it is first worth outlining the main ways that things go wrong. Broadly speaking, the main errors and mistakes that arise can be traced to one of two principal conditions, that can be defined as follows:

- Over-attribution – occurs where weak signals are interpreted to make strong claims about involvement and responsibility, even though the available evidence is ambiguous.
- Under-attribution – is a condition that results from failing to detect and/or make use of signals in data, so that an accurate assignment of responsibility to an actor or entity is not made by an analyst.

In some ways the latter condition is the easier to explain, in light of the previously outlined trend, wherein state actors are becoming more accomplished in terms of their ability to obfuscate their involvement. This can trace back to their increasing willingness to 'out-source' aspects of their campaign, or to make use of a range of 'paid for' digital services that effectively provide a 'privacy shield' for the operators. Two case studies further illuminate the make-up of over- and under-attribution.

3.1 Case study #2: Over-attribution in the UK General Election 2024

Leading up to 2024 there was considerable political and public consternation being expressed about the large number of elections taking place globally, and the potential for malign influence and interference operations to be run in ways that would subvert democratic processes. This included the United Kingdom, where it was known that a General Election would have to be called, albeit the precise timing for this was uncertain. As it happens, the decision was taken in May to call a snap election for July, timing that caught many political commentators off-guard, as most expectations were that it would take place in the Autumn.

Given the wider geopolitical environment, there were widely shared assumptions that the UK election would be a prime target for covert foreign state interventions, particularly by Russia, but also possibly China and Iran. Consequently, there were lots of briefings issued by security agencies, many of which received substantive media coverage. As it transpired, little such activity was publicly detected and attributed. There were though several instances of 'over-attribution'.

One such was a claim mounted by Australian journalists using 'open-source' techniques that they had found a small network of 6 accounts operating across Meta's surfaces, that they attributed to Russia. It appears that this connection was based on two grounds. First the nature of the content that was being distributed by the accounts in a loosely co-ordinated way, resonated with pro-Russian narratives around Ukraine and other issues. The second, was that the network's activities were being amplified by inauthentic accounts operating out of Nigeria. The journalists involved identified that previous Russian information operations had been supported and boosted by accounts geo-located to Africa. It may be relevant that roughly two weeks prior to their story appearing, Cardiff University researchers working with ITV News had publicly attributed networks of Nigerian bot accounts being used to boost political communications from the Reform Party on TikTok. Importantly however, Cardiff made no indication of a Russian connection.

In terms of the journalist's attribution, there were no other technical or behavioural signals that could be found to support the assertion made of direct Russian involvement. An explanation for this is that maybe there was such an expectation being built up that influence operations and interference measures would be initiated, that it 'primed' investigators to go out there and find them, in such a way that they 'wanted to believe' that possible indicators were stronger than they actually were, thereby exemplifying the socio-technical dynamics of over-attribution.

As an aside, it is worth noting that in addition to operations where some attempt is made to hide or misdirect authorship, there are some information operations, where no such efforts are made and the activities are effectively 'avowed'. For example, the Russian information operations Doppelganger and Portal Kombat both possess 'signature' features that are well known, and clearly visible.

3.2 Case Study #3: The 2024 Romanian Presidential Election

A second election-based case study can be used to further illuminate how problems of over- and under-attribution can manifest themselves. Voters went to the polls in Romania in November 2024 to elect a new President, and somewhat unexpectedly the relatively unknown ultranationalist independent candidate Călin Georgescu won the first round, with his pro-Russia and anti-NATO messaging and sovereigntist approach to agriculture and economy.[22] Shortly after the result was announced, the Romanian intelligence community declassified documents claiming Georgescu had significantly benefited from a social media campaign, especially on TikTok, that was highly similar to some of the Kremlin's influence operations targeting Ukraine and Moldova.[23] The documents claimed that paid influencers as well as extremist, right-wing group members and people with ties to organized crime were behind Georgescu's social

[22] https://disinfo-prompt.eu/posts/6gVQHsgN02LeYCnVzI6wGx

[23] https://www.presidency.ro/ro/media/comunicat-de-presa1733327193

media promotion. The declassified documents did not directly blame Russia for election interference, but they did suggest it strongly.

One element of the campaign described in the declassified documents, introduced[24] a hashtag "Equilibrium and Verticality" via a script allegedly shared with groups of Romanian social media influencers. The script recycled messages by Călin Georgescu, but didn't mention his name. The Romanian Ministry of Interior interpreted this as directly emulating the methodology of a Russian campaign targeting Ukraine called "Brother near Brother".

Significantly, whilst the Romanian governmental authorities flagged up resemblances and similarities to prior Russian campaigns, they did not publish any evidence to substantiate these claims This was picked up by local media who were quick to express concern that the authorities hadn't provided evidence of foreign interference[25], despite this being a very serious allegation.

In effect, the journalists involved were suggesting that, at least in terms of what is currently in the public domain, the Romanian intelligence community had clearly engaged in 'over-attribution.'

In response to this situation, in December 2024, the Constitutional Court of Romania annulled the results of the first election round on the grounds that a Russian influence campaign had impacted the vote. Unsurprisingly, the government has been coming under increasing pressure to back up their allegations with proof. The European Commission has ordered TikTok to preserve data linked to the Romanian election. TikTok has told the Romanian authorities it detected a network of accounts linked to the Russian state news agency Sputnik targeting users in Romania and Moldova, according to the intelligence documents[26]. However, the numbers being referred to by TikTok, were by no means at the scale needed to shift the overall vote in the ways that occurred. In addition, the documents said that user access data for electoral websites was stolen and then posted to "cyber-crime platforms that originated in Russia".

At the time of writing, the situation in Romania is still evolving and has yet to resolve itself. Intriguingly, it does appear that whether or not his initial campaign was boosted by Russian efforts, the publicity surrounding his case, may yield political capital for Georgescu. Recent public opinion polling about voter intentions suggests he may well be placed first in the re-run vote.

---

[24] https://disinfo-prompt.eu/posts/6gVQHsgN02LeYCnVzI6wGx

[25] https://www.g4media.ro/ce-stim-pana-acum-despre-presupusa-implicare-a-rusiei-in-alegerile-prezidentiale-sie-a-lansat-tema-sri-nu-are-date-certe-cu-privire-la-atacatorul-care-a-vizat-sistemele-it-parchetele-nu-au-lansat-p.html?uord=Qo3hkCczib10Zm5cy6EnQsSFunQumCIdLP/B7pT7i4Icsr4SYZKomI2LXcNym2xLtkIUSQjJQFUWiDldl6EoU2Kon6gATsGEiXMukifxK1/69/MihVEbhJsOMSWum/sz7vSAj8E3

[26] https://www.rferl.org/a/romania-election-scandal-tiktok-bogdan-peschir-georgescu/33229674.html

4 Attribution fundamentals

When training police detectives they are sometimes encouraged to think in terms of 'the A, B, C of investigation'. This stands for: 'Assume nothing, believe no-one, challenge everything'. Allowing for the need to adapt these somewhat for them to be relevant for the digital world, they provide a good starting point for the basic mindset that should be adopted by open-source investigators and researchers. The idea of focusing upon what can be directly observed, and clearly differentiating these elements from inferences and assumptions used to construct claims and judgements, is vital. Indeed, it appears that one of the ways that open-source investigations most frequently go wrong in terms of making 'over-attributions' is by mixing inferences and observations.

4.1 Four types of signals

In their account of open-source work, Pamment and Smith (2022) draw a helpful distinction between three main types of 'raw materials' used to inform attributions. To which we have added a fourth:

(1) Content Signals: are where text, images or video data contained within messages and/or websites are the focus of analysis. The idea being that these are being used by an actor of interest to pursue their interests and objectives in some way, and can thus be used by an open-source researcher to inform their assessment of who is responsible for authoring them.

(2) Behavioural Signals: rather than attending to what is communicated, behavioural analysis is concerned with how the distorting, distorting or deceptive activities are organized and conducted. The idea being that particular threat actors or information operations possess 'behavioural signatures' that are fairly common features in their activities. Therefore, by detecting and recognizing these patterns of behaviour, it is possible to make plausible attribution assessments.

(3) Technical Signals: a similar but different logic applies to 'technical' attributions. The focus here is upon the technological infrastructure used to mount an influence operation, in terms of IP addresses, particular servers and so forth. Where these elements have previously been linked to an operation, then this might be an indicator of a common author.

(4) Financial Signals: to the above three, we would accent the increasing importance of financial data in making open-source attribution judgements.

> ‘Following the money’ can be useful both in terms of discriminating between different threat actors, but it can also be uniquely important in terms of establishing ‘direction and control’ back to a state authority.

The obvious point to make is that a more confident attribution is more likely to occur where a combination of content + behavioural + technical + financial signals all align to suggest a particular author. The practical reality, however, is that it is rare to have all of these available and blended together. Frequently, open-source workers are making their assessments and judgements based on one or two of these types of data that they are ‘mosaicing’ together. As a general rule of thumb however, attributions based on only one type of data, can command only limited degrees of confidence in terms of their accuracy and precision. For our purposes here, the value of this framework is in encouraging researchers and investigators to think about the types of data they are relying upon when constructing their assessments and judgements.

4.2 The “A2E” framework

An alternative, but related approach to that outlined above is the ‘A2E’ framework. The original formulation of which was devised by Camille Francois of Graphika as ‘A to C’ but was then extended and augmented by Innes (2019). It is important to be clear about what ‘A2E’ can and cannot be used to do. It is, after all, essentially descriptive. However, this in itself can be useful in keeping an analyst’s attention fixed upon only what can be directly observed the available data, thus avoiding making leaps of intuition and unintentionally weaving in assumptions about what the ultimate author intended.

The key components of the A2E framework are defined as follows:

> A = Accounts, Authors, Amplifiers and Audiences – this vector covers the roles performed by the key actors engaged in the transmission and reception of disinforming messages.
>
> B = Behaviour – a critical component of the assessment and evaluation of accounts authoring and amplifying false information is a focus upon their behaviours and interactions. Put another way, this analytic dimension focuses upon how communication is conducted, rather than the content of what is communicated. The idea is to detect behavioural patterns and ‘signatures’ as outlined above.
>
> C = Content – this involves a close reading of the material to determine how this has been constructed and the motivations that might underpin why someone would seek to communicate it. This can encompass both textual and visual analysis, undertaken using a variety of qualitative and quantitative content analysis techniques.
>
> D = Distribution – attends to the pathways via which a particular message or group of messages is disseminated. For example, does it achieve high traction on a single

platform, or does it develop more 'cross platform' dynamics? What role is played by established 'influencers' in terms of boosting the signal?

E = Effect – at the current time, the state-of-the-art in terms of ability to measure the influence effects of disinformation campaigns and propaganda precisely and accurately are limited. This limitation notwithstanding, knowing that a particular audience segment is being targeted, can sometimes provide a clue to the guiding operational hand, if it is consistent with prior attributions.

Depicted in this way, the A2E framework provides a systematic approach to breaking down an operation (or part thereof) into its component parts. It has perhaps received most attention and interest from those open-source researchers coming from academic backgrounds.

4.3 TTPs

A similar logic of breaking down communicative actions and operational sequences into their constituent parts, also underpins approaches such as the DISARM framework. This starts from the premise that by developing a detailed taxonomy of the variety of tactics, techniques and procedures (TTPs) that those transmitting disinforming, distorting and deceptive messages can use, a systematic and replicable analysis can be recorded. The perceived value for attribution work is that by looking at the particular combinations of TTPs featuring in a current operation and comparing these with prior analyses of past attributed operations, can provide indicators of responsibility.

Whilst this approach sounds reasonable in principle, there are a number of 'real world' practical bedevilments. First, reflecting the diversity of ways in which information operations can be configured, the DISARM framework has grown into an intricate and elaborate list of TTPs. This contributes to a second challenge. Because there are so many categories it is not clear that different researchers understand and apply them all in a similar way when recording observable features of information manipulation efforts. A further complexifying issue concerns the presence of social learning amongst online groups and communities constructing information based influencing efforts. Stated simply, the TTPs associated with a successful information operation, will be copied and imitated by others. For example, following the significant media publicity about the digital activities of the St Petersburg based Internet Research Agency, many other actors adopted and adapted some of their methods. This included other state actors, but also those engaging in clickbait monetization schemes. The take-home point being that where this methodological proliferation occurs, identifying specific TTPs has less discriminatory power in terms of being able to attribute to a specific authoring hand.

Analysing a known threat actor's TTPs can be beneficial in understanding its patterns of behaviour and the capabilities it is able to exploit. This can be especially insightful

when looking at cyber-enabled influence operations combining cyber-attacks with information operations. However, TTP analysis has less usability in attributing unknown, new influence operations for the reasons outlined above. However, there may be considerable opportunities to enhance TTP analysis going forward, by harnessing the power of AI based Large Language Models and their ability to process large volumes of data with unprecedented speed and scale.

4.4 'Acronym Soup': FIMI vs CIB

For modes of open-source research and investigation into information influence operations where attribution is an important precursor for some kind of response or reaction, there are at least two other frameworks that have been more influential. Social media platforms, in particular Meta, have framed the issues in terms of 'co-ordinated, inauthentic behaviour' (CIB). Counter-pointed with which, many governments, led by the European Union, have adopted a different concept of 'Foreign Information Manipulation and Interference' (FIMI). These are sometimes used together and in ways that present them as inter-changeable, but they are not.

Before detailing these approaches, it is first useful to highlight how, in different ways, they raise questions about the purpose underpinning an attempt to make an attribution and the standards of proof that follow on from this. The CIB approach is typically focused upon interventions that de-platform or reduce the visibility of material from accounts whose activity are viewed as problematic and in breach of platform 'Terms of Service'. It quite explicitly eschews any consideration of the content of what is being transmitted, largely as a way for platforms to avoid allegations that they are 'policing' freedom of expression.

Critically however, platforms have very different interests from governments. Most obviously governments tend to engage when they suspect other states are seeking to use disinforming, distorting or deceptive communications to influence public understanding or political decision-making in some manner. Proponents of the FIMI approach argue that it is preferable to either the CIB framing, or indeed other terms, such as disinformation. This is on the grounds that it is sufficiently flexible and wide-ranging so as to be able to capture the wide variety of sources and methods used as operators seek to leverage shaping effects upon both their adversaries and allies.

However, viewed through an attribution lens, FIMI is positioned at a level that is not especially helpful in terms of encouraging accuracy and precision in assigning responsibility. This is because notions of information manipulation and interference are terms that cover a wide variety of actions, that stretch far beyond the methods used in the organization and conduct of information influence operations. If one is looking to assign blame for general activities of manipulation and interference, then it is easy to resolve this to a general level of 'the Kremlin' or the Chinese Communist Party, as opposed to steering analytic effort to more specific entities and organizations.

There are also issues in terms of what is the appropriate threshold for defining something as 'foreign' interference, given how there are often complex transmission pathways in terms of how disinformation travels and the ways that information gets manipulated. For example, given that many established and persistent foreign state information operations seek to identify and amplify 'organic' domestically originating disaffection narratives, how much of this boosting activity is needed to qualify as FIMI. Relatedly, similar concerns arise out of a key trajectory of development in the organization and conduct of contemporary information influence operations, whereby states are out-sourcing the delivery of such efforts to 'in-country' proxies. Regularly resulting in situations where appearances are suggesting a domestically originating campaign, whilst the analyst suspects foreign interference, but really struggles to substantiate these suspicions to 'get them over the line.'.

Potentially one method for offsetting these issues, is adopting a technique used by the police when investigating crimes. Police investigators are encouraged to think in terms of there being specific 'points to prove' for different criminal offence types. This involves deconstructing a crime down into the constituent parts. This then enables them to identify all the areas where they will require evidence of actions or intentions, if they are going to be able to plausibly demonstrate 'beyond reasonable doubt' that an individual or individuals were responsible for an illegal act. Defining the specific points needed to attribute a particular style of information manipulation or interference, prior to commencing an enquiry, seems to have some utility in rendering FIMI more analytically useful. Moreover, the points to prove approach might also help open-source researchers and investigators to arbitrate between whether they are searching for indicators of CIB or FIMI.

For example, Facebook defines CIB as "coordinated efforts to manipulate public debate for a strategic goal, in which fake accounts are central to the operation." This means that "people coordinate with one another and use fake accounts to mislead others about who they are and what they are doing." Translating this into a 'points to prove' approach, could involve the following:

- Identifying indicators of 'co-ordination' – this might evidence 'temporal co-ordination' in terms of groups of accounts issuing similar messages in patterned 'bursts' of activity at similar points in time; or engaging in 'content co-ordination' where multiple accounts all share the same textual or image-based material; or 'follower co-ordination' where multiple accounts are all boosted by the same sets of accounts.
- Establishing the 'inauthenticity' of accounts – aspects of the analysis would focus upon identifying that the account profiles propagating the messaging are not who they claim to be in some respect. At one extreme this can be because they are fully automated 'bots' or 'cyborg' accounts, or alternatively because the operators are choosing to imitate some assumed or falsified identity.

- The final component is that there is some attempt to mislead or manipulate audience members engaging with the content being shared. Various options to monitor for here would be the sharing of disinformation narratives, making use of hack and leak materials, or the 'laundering' of ideas and narratives from state media sources.

In one takedown in 2023, Facebook removed 33 accounts, six pages, six groups and four Instagram accounts originating in China and targeting US audiences.[27] Fake accounts were used to pose as members of US military families and anti-war activists. Some of the pages and groups were also focused on military themes, especially US aircraft carriers. The accounts posted links to news articles and memes in English, copy-pasted content from elsewhere, about the US military and criticized US policy towards Taiwan, including via a petition which attracted 300 signatures. It was also active on YouTube and Medium.

Whether the aim is to demonstrate the presence of co-ordinated inauthentic behaviour, or information manipulation by foreign actors, a cross-cutting method that is critical to constructing solid attributions is the ability to 'mosaic' together information of different types and from a range of sources. Most attributions, most of the time, result from the ability of an analyst to link together 'bits' of information, where the whole is greater than the sum of the parts. This may well include materials collated from across multiple social media platforms, and from across the content, behavioural and technical layers, outlined previously.

One further piece of learning from the conduct of police investigations, is a disciplined and repeating willingness to 'test' the robustness of key assumptions and inferences, especially as new bits of relevant information and data is uncovered. To put it another way, continually posing 'counter-factuals' in terms of alternative explanations for a given constellation of evidential material, can be an important way of advancing an enquiry. After all, sometimes being able to discount a plausible scenario and determine that it did not happen, is as important as being able to confirm aspects of a preferred hypothesis or explanation. It is for this reason that police often make use of deliberately designed external review functions to test the integrity of their case-building on high profile enquiries, and in science such an accent is placed upon the role of 'peer-review'.

By way of summary, the discussion to this point has illustrated how there are a range of different frameworks that can be and are used by researchers to inform their attempts to construct plausible and persuasive attributions. These alternatives have different strengths and weaknesses, and their relative perceived values are shaped by the position of the researcher / investigator, and the nature of their principal tasks. To some degree,

[27] https://scontent.fmmx4-1.fna.fbcdn.net/v/t39.8562-6/428391559_3726563917669737_441724410168993669_n.pdf?_nc_cat=102&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=lznahHVGhuEQ7kNvgGd6oUW&_nc_zt=14&_nc_ht=scontent.fmmx4-1.fna&_nc_gid=A4k1aUpi06q-UHmpBFbJxP_&oh=00_AYDOF25MSWYh0zVEH-HcY_7GZ9q_t2OpNboCovPjknwL0w&oe=674CCB77

all of the approaches outlined above, perhaps underplay the significance of covertness and deception. After all, if no attempt is made to obfuscate the identity of those transmitting disinforming, distorting or deceptive messaging, then the challenge of identifiability does not arise.

Piercing this veil of anonymity and pseudo-anonymity has often depended upon the work of investigative journalists. While governments have access to different types of data than media outlets and might also rely on classified sources, it has frequently been journalists who have demonstrated a willingness to try and track down specific individuals and groups in-person and 'test' the strength of evidence against them. Journalists have also tended to 'be on the front foot' in terms of obtaining and publicizing information based on leaked documents, often offering invaluable insights into the work of the information influence operatives. This is a role that should not be under-estimated. For it is often only taking this step that a greater degree of certainty about a claim of responsibility for an information influence operation can be secured. In addition, where doubt may exist on significant points of fact, basic principles in journalism require that information must be based on at least two independent sources.[28] This type of verification is also vital for OSINT work, where analysis of primary sources and their independent verification form the basis of the job There are detailed guides available for journalists investigating online manipulation, which can also be useful for OSINT analysts coming from different professional backgrounds.[29]

4.5 Time and attribution

One final dimension that frequently contributes to attribution is the often under-appreciated role of 'time'. Insights gleaned by collecting data on a suspected operation over an extended period, can often cumulatively build. Many (but not all) analyses of specific information influence operations tend to study their activities only across a limited time-window, before moving to report on and expose them in some fashion. Longitudinal analyses that track operations over time and trace how they evolve and adapt as they engage across multiple incidents and events, remain relatively rare.

There are parallels here with the methods and challenges sometimes encountered by crime analysts when working on serial offenders. In an early academic contribution on the challenges posed by serial murders, the criminologist Stephen Egger identified a problem he dubbed 'linkage blindness'. As he described it, it was only when police connected incidents and then composited together their various bits of intelligence and evidence that they had individually collected, that they had enough information to start to identify specific offenders.

[28] https://docs.rferl.org/en-Press/2022/04/27/fe28a8b6-624f-4047-badb-ca71c617b8bf.pdf

[29] https://datajournalism.com/read/handbook/verification-3

A similar argument can be constructed with digital information influence operations. Focusing upon a narrow time window might well delimit the amount of exploitable information, where tracking and linking activities over time enables more and more details to be mosaiced together. To illustrate this a brief case study of the Ghostwriter information operation is presented below.

4.6 Case study #4: Ghostwriter and Ambiguous Attribution

Ghostwriter is a cyber-enabled influence campaign named by cybersecurity firm Mandiant in 2020. The Ghostwriter campaign is especially interesting because it has been active at least since 2016, and has been able to evolve and adapt mainly due to two reasons. First, it was initially understood as separate incidents of cyber-attacks and disinformation postings, rather than as a persistent operation involving multiple linked actions. As a result, it was subject to multiple responses from governments, private cyber firms, social media platforms, media outlets and civil society. But their respective interventions were, until 2020, targeted at the individual incidents, rather than the actor or activity behind the series.

Second, according to the multiple entities that have been involved in attributing the campaign, Ghostwriter's cyber activity is supported by a foreign state threat actor, either Russia, Belarus, or both. This ambiguity in attribution has likely hindered the response. It is unknown who is responsible for producing the content of the influence campaign.

This resonates with the aforementioned criminological concept of "linkage blindness". In Ghostwriter's case, a similar phenomenon has occurred, resulting in a lack of intelligence sharing and coordination, that has frustrated understanding who is ultimately behind the campaign. Based on open-source data, Ghostwriter has impacted thousands of email users.[30] It has hacked dozens of social media accounts and media websites, published hundreds of false blogposts and other falsified content, and impersonated multiple government officials, NATO representatives and journalists in Europe.

Critically for our purposes in this report, different governmental, commercial and civil society groups have been looking at different elements of the overall operation, and have attributed these to different responsible actors. For example, Mandiant attributed the cyber incidents to UNC1151 an Advanced Persistent Threat actor known to operate out of Belarus, but others have reached different conclusions. Potentially, this may be an artefact of how Microsoft describes UNC1151 as a "group in development".[31]

[30] https://www.cardiff.ac.uk/__data/assets/pdf_file/0005/2699483/Ghostwriter-Report-Final.pdf

[31] https://learn.microsoft.com/en-us/defender-xdr/microsoft-threat-actor-naming

The attribution and linking of individual incidents to Ghostwriter has developed slowly over time. Parts of Ghostwriter's cyber activity have been attributed to Russia's military intelligence (by Germany) and to the Russian state (by the EU and Poland), as well as to Belarus (Mandiant and Google). In terms of being able to arbitrate between such claims and counter-claims, it is difficult owing to how the state-led attributions frequently depend on classified sources. Overall, Ghostwriter has been quite successful in obscuring the origins of its malign activity. It has been openly producing content in Russian language and using Russian blogging platforms. Relatedly, Russian state media has been careful to avoid amplifying its fakes directly. Instead, it has focused upon amplifying the rebuttals and ridiculing Western governments, accusing them of Russophobia. It is a positioning strategy that is clear in terms of its ideological alignment, but creates a certain amount of ambiguity about the actual degree of state involvement. For instance, in Poland, public distrust in the governmental communications led some to question if a foreign actor was involved at all.

Ultimately, one is left wondering whether, in this particular example, it is mistaken to seek a single guiding hand. It seems plausible that maybe the different parts of the operation, in terms of the cyber-attacks, the targeting of surveillance and engagement in influence engineering through social media content creation, are being organized and conducted by different networks. It is possible that these components are not fully aware of each other's activities. More certain is that the failure to secure an agreed upon attribution has been an impediment to imposing control. So far, no sanctions have been designated to those who are behind Ghostwriter, and the EU's cyber sanctions toolbox has not been put to use.

5 Guarding against unethical practices

5.1 Perception hacking

Although not directly about how to construct attributions, it is also worth thinking about some of the downstream consequences that can flow from them, especially in terms of how they manifest in terms of the risks of 'perception hacking'. The latter is a term used by Meta (and others) to capture how it is sometimes sufficient, in terms of triggering harmful social reactions, to manipulate public perceptions to believe that an information influence operation has taken place, even when it has not.

Especially when people are 'primed' to expect some kind of information-based interference, as often happens in the lead-up to elections, it can sometimes lead open-source researchers to confirm these expectations, even though they are in actuality dealing with ambiguous and uncertain information. It is therefore important that, when in such situations, analysts are mindful to retain their critical faculties and apply consistent standards. They should actively try to avoid 'giving the benefit of the doubt' towards making a positive attribution, just because of the prevailing atmospherics.

This matters because when open-source research makes claims in error about who is responsible for information circulating online, this can damage public trust and confidence going forward. The upshot of which is that it can render it more difficult to persuade the public or politicians of the accuracy of future attributions.

5.2 Ethics and 'collection creep'

Given the mounting challenges, outlined above, in terms of being able to construct solid attributions it is perhaps understandable that analysts and investigators sometimes become frustrated and seek ways to circumvent the impediments that are limiting their understanding. This is particularly the case in an 'information rich' environment where there are a myriad sources of information, albeit of varying provenances and origins. This can create an 'invitation' to unethical practices, particularly collateral intrusion, or what be dubbed 'collection creep'.

Collateral intrusion refers to when trying to collect information on a subject of interest, one way of doing so is to deeply research their network of contacts, as an indirect route to obtain material on the subject concerned. Sometimes this can be a legitimate pathway for progressing a line of enquiry, but it can easily expand in terms of its remit, to become profoundly unethical. For example, by collating material from family and friends of the subject, even though they are wholly unconnected with and unaware of

the activities of interest. Or, drawing in data that was put into the public domain through illegal means, where it might well have been manipulated.

This constitutes a form of ‘collection creep’ that needs to be guarded against. In pursuit of making an attribution it is easy to be lured into thinking ‘if I just draw in a little bit more data to make use of this particular information source’, even though I ideally would not rely upon it, then the ends justify the means. One way to help insure against the influence of such forms of collection creep is to prepare a collection plan, prior to commencing any research or investigation. This defines what kinds of data and question are legitimate and ‘in scope’. As with the previous point about perception hacking, guarding against the invitation to collateral intrusion and other forms of unethical practice, are important for maintaining overarching public trust and confidence in the attribution work that is completed using open-source data. This is especially imperative in a historical moment where public trust in key public institutions in many Western countries is being rendered more ‘brittle’, or in many cases declining.

6 Conclusion

Attribution is a critical task in open-source research and investigation. It functions as a predicate for many actions to counter information manipulation and its effects, but also for understanding how and why the digital influence engineering is occurring in the first place. For a number of overlapping and interacting reasons, constructing plausible and solidly evidenced attributions for covert information influence operations is being rendered increasingly difficult. That said, at the same time, especially following Russia's invasion of Ukraine, a number of operations are seemingly less concerned to obfuscate their origins.

In response to these developments, there are two emerging trends in attribution work that are worth highlighting in case they become increasingly important in the future. These we dub 'participative attribution' and 'pre-emptive attribution'. The trend to multi-actor participation and collaboration is highlighted in Open AI's recent takedown reports[32]. Open AI's attribution to an Iranian threat actor followed this logic. It used Microsoft's reporting and a set of domains they had published earlier. Open AI identified further activity on X and Instagram and shared it with those platforms. Meta then confirmed a link to the Iranian campaign, which targeted Scottish users. This type of collective way of advancing attribution requires trust between industry partners and adequate resources to invest the time and skills in work to disrupt malign operations. Not all the social media platforms are equally engaged in such efforts.

Pre-emptive attribution is more loosely formulated and is connected to the increasing use of public disclosures and pre-bunking initiatives in anticipation of potentially harmful information operations being launched. Approximately two months prior to the US Presidential Election, and informed by experiences gleaned from the previous two democratic cycles, the American intelligence community issued repeated pre-warnings of foreign interference threats. This included publishing details of how Tehran was behind attempts to compromise candidate Trump through hack-and-leak campaigns of his staff members.[33] Additionally, several Russian entities and individuals were sanctioned ahead of the election and extensive information about the details of their operations were publicized. Russian government media networks RT and Sputnik were exposed as having acquired cyber capabilities and organizing support for Russia's war efforts in Ukraine. The aim being to create sufficient public awareness and provide sufficient detail about the hostile influencing methodologies, that a wider

[32] https://cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update_October-2024.pdf

[33] https://www.cisa.gov/news-events/news/joint-odni-fbi-and-cisa-statement-iranian-election-influence-efforts

community of open-source analysts and the public at large, would be better equipped to detect the assets and claims emanating from both covert and overt influence operations.

Set against this backdrop, in seeking to formulate guidance to inform attribution work, we have sought to draw in ideas from and analogies with other modes of investigation. These do not always provide a perfect fit for the complexities of the flows of the vast streams of digital data that are routinely publicly available, but we think they are helpful in terms of providing a toolkit for thinking about the key tasks engaged in assigning and attributing responsibility for authoring disinforming, distorting and deceptive public messaging.