Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
May 3, 2018
A Regional Conference of the Society for Industrial and Systems Engineering

Origins of Intelligence: Discerning Corroboration vs. Replication in Open Source Intelligence Production

Joseph Bosse, Conor Glancy, Justin McDaniels, Daniel Provaznik, and Jillian Wisniewski

United States Military Academy, West Point, NY 10996, USA

Corresponding author's Email: Daniel.Provaznik@usma.edu

Author Note: Cadets Joseph Bosse, Conor Glancy, Justin McDaniels, and Daniel Provaznik are seniors in the United States Military Academy's Class of 2018. CDT Bosse will graduate with a Bachelor of Science in Engineering Management and commission into the Infantry. CDT Glancy will graduate with a Bachelor of Science in Systems Design and Management and also commission into the Infantry. CDT McDaniels will graduate with a Bachelor of Science in Operations Research and commission into the air defense artillery. CDT Provaznik will graduate in May of 2018 with a Bachelor of Science in Systems Engineering with a minor in terrorism studies and commission into Army aviation. The capstone team would like to thank their advisor, MAJ Jillian Wisniewski, for her support, coordination, and mentorship throughout the capstone process as well as that of LTC Matthew Dabkowski who, on multiple occasions, offered technical expertise.

Abstract: A well-documented fallacy across the intelligence community (IC) is the evaluation and processing of information through existing analytic products, which too often replicates previous findings instead of contributing an original assessment. The fallacy, which stems from myriad combinations of errors and biases in reporting, allows information to gain unwarranted legitimacy as it proliferates: its replication is often misconstrued as corroboration. Open Source Intelligence (OSINT) is especially vulnerable to such misinterpretation because it is produced from publicly available information (PAI), which, under its massive volume, buries source origins. This study computes a bipartite network of citations and derives an interdependence measure of its sources, where high or low measures indicate replication or corroboration, respectively. The application of bibliographic coupling, distance measures, and network centrality to a set of 450 scholarly sources, with over 9,100 citations, provides a foundation on which to build an objective measure of originality for a source document.

Keywords: Echo Effect, Open Source Intelligence, Bias, Corroboration, Bibliographic Coupling

1. Introduction

1.1 Background

In his article "Why Intelligence Failures Are Inevitable," Richard Betts claims, "in the best known cases of intelligence failure, the most crucial mistakes have seldom been made by collectors of raw information...but most often by the decision makers who consume the products of intelligence services." For this reason, it is vital that analysts create intelligence products with the most pertinent, unbiased, and accurate information to aid policy makers in their decision making. In one of the best-known intelligence failures, President George W. Bush and his staff went to war with Saddam Hussein's regime based on false intelligence that Iraq possessed Weapons of Mass Destruction. Reports of Iraq pursuing Yellowcake from Niger and specialized aluminum tubes from Hong Kong gained unwarranted legitimacy from the constant repetition of these stories throughout the intelligence community. Although each report seemed independently derived, an examination of these reports not only revealed a high degree of interdependence among these reports but also that a single inaccurate report provided the premise for their findings (Barstow 2004, Pfiffner 2007).

The Yellowcake case provides a cautionary tale of the prevalence of biases in intelligence, including: oversensitivity to consistency, overconfidence, law of small numbers, distortion in replication, and confirmation bias. In particular, Heuer warns the IC about the effects of these biases in his seminal work for the CIA, *Psychology of Intelligence Analysis*, because "impressions tend to persist even after the evidence that created those impressions has been fully discredited" (Heuer, 1999). Ultimately, President Bush and his staff decided to invade Iraq on the basis of fundamentally invalid evidence. This study offers a method to reduce illegitimate replication in intelligence reporting, also referred to as the "echo effect." With a means to discern corroboration from replication, intelligence analysts and policy makers will be made aware of only the most accurate and unbiased information, facilitating optimal decision making.

ISBN: 97819384961-2-7

1.2 Challenges in OSINT

OSINT is especially vulnerable to biases in evaluation of evidence because it is produced from PAI, which, under its massive volume, buries source origins. Army Directive 2016-37 defines OSINT as: "intelligence that is produced from publicly available information and is collected, exploited and disseminated in a timely manner to an appropriate audience for the purpose of addressing a specific intelligence requirement" and outlines its use in the Army (2016). Historically, intelligence consumers have undervalued OSINT, favoring instead classified intelligence products; however, the modern information landscape has called attention to the need to exploit the open source environment.

Exploiting OSINT is nontrivial. Since OSINT constitutes every unclassified publication by the government, academia, news reporting, and the private sector, its scope is enormous. The National Security Agency is only able to "touch" two petabytes of the thousands created every day in the public sphere, underscoring the qualification of OSINT as "big data" (Lopez n.d., Stoddard n.d.). In order to collect and process this big data effectively, businesses and government agencies need to have a clear data strategy, an efficient data model, and the right integration tools to effectively utilize OSINT and exploit its full potential (Lopez, n.d.). Further compounding the challenges, data is often stored in terms of how it is described, not in terms of what it can help us achieve, a characteristic that was first highlighted by Dr. Marvin Minsky in his book on artificial intelligence, *The Emotion Machine.* Minsky also explains that analysts know more about the goal they want to accomplish than about the data they must gather and analyze in pursuit of that goal. Because the structure in which data is stored is at odds with how analysts search and access it, errors in judgement are more likely to occur (Heuer 1999, Minsky 2006).

1.3 The Echo Effect and Other Utilization Problems

In today's environment, information spreads rapidly in a fashion very similar to the Bass Diffusion Model, gaining initial traction before rapidly proliferating (Norton and Bass, 1987). With news reporting, as an outlet catches a story, it receives the raw original information and tailors it to its inherent bias and audience. It then republishes it, causing it to gain distance from the original source. This republication accelerates distribution and as the distance increases from the original information, it gains an element of unwarranted legitimacy that is compounded as more outlets pick up the story. Academic publications function similarly within a citation network. Flawed or biased publications--when cited--gain unwarranted credibility.

OSINT cannot reach its full capacity for the following reasons: an "echo effect" inherent in big data, a bias preferring classified to open source, a lack of subject matter expertise in regards to its associated training programs, and a limitation on effective language processing. Today, the "echo effect" is a concern because many sources produce the same information (Best & Cumming, 2007). When information is repeated across multiple different sources, it may gain unwarranted credibility which can cause users to only "reinforce their current understandings and opinions" (Best & Cumming, 2007). This echo effect allows the user to become subject to confirmation bias, seeking and responding only to the information that supports the beliefs they already hold. The transfer of this repetitive and subsequently biased information only adds to the total amount of data, straining data collection, analysis, and integration. Additionally, in the intelligence community, there is bias against Open Source collection and analysts. Because it is less resource intensive and lower risk to access, many dismiss it as unimportant and consider information collected from classified sources better because it is harder to acquire. Analysts also often lack proper training and subject matter expertise to utilize OSINT, limiting the population that can utilize its advantages (Mercado, 2009).

1.4 Problem Statement

The goal of this research is to aid intelligence analysts in overcoming the echo effect by using open source network analysis tools to create an objective filtering and assessment methodology that assigns an originality metric to an individual report that ultimately allows analysts to minimize bias in their intelligence products.

2. Methodology

2.1 Data Collection

The initial data set was generated by querying the Web of Science database for the key term, "network science." The results were filtered for peer reviewed status and then ordered for relevance. The first 450 results, depicted in Figure 1, were exported in Bibtex format and categorized using the following terms (listed with associated color scheme): technology (orange), natural science (blue), and social networks (grey).

ISBN: 97819384961-2-7

Figure 1. Keyword Analysis of 450 Data Entries

2.2 Matrix and Network Creation

The 450 sources and their 9,121 associated citations were read into the RStudio programming engine and converted into a dataframe to facilitate analysis using the R bibliometrix package, which was used to generate the bipartite network, which mapped each source to the full set of citations (Aria & Cuccurullo, 2018). The bipartite network was then read into the ORA-Lite Software to create the visual representation of the network's structure, depicted in Figure 2 (CASOS, 2017).

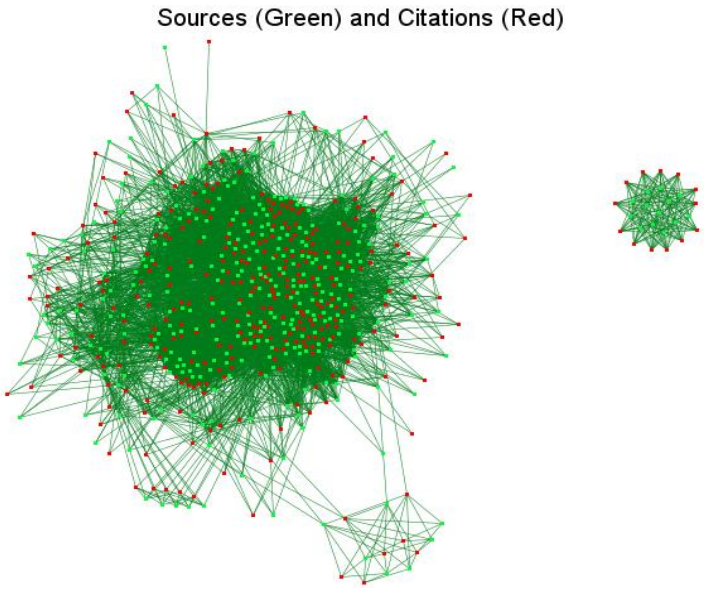

Figure 2. Network of Sources and Citations

2.3 Interdependence Measures

Classic methods, including bibliographic coupling, distance measures, and network centrality measures, provide the foundation to build an objective measure of originality for a source document. Bibliographic coupling provides a numerical measure of how related, or coupled, two articles are through how often "one cited source appears in the bibliographies or reference lists of both articles" (Aria & Cuccurullo, 2018). The Jaccard Similarity Coefficient provides a distance measure that conveys a pair-wise comparison of each article by measuring the proportion of shared citations over instances where at least

ISBN: 97819384961-2-7

one of the citations in either article is present (StataCorp, n.d.). Measurement degree centrality creates an objective value that is based on the number of direct connections held by each respective node in the network (Disney, 2014). Direct connections are the one-to-one relationships shared between two nodes. On the other hand, betweenness centrality measures "the number of times a node lies on the shortest path between other nodes" (Disney, 2014). The significance of betweenness centrality is that it identifies specific nodes that are shared between two other nodes; calculation is determined by identifying the shortest paths within the network and counting them accordingly. The initial examination of the citation network through each of these methods provided an initial understanding of the network's structure and enabled the generation of the data formats that would be required in developing the desired objective measure of originality for each source within the network.

3. Findings

3.1 Applying Interdependence Measures

To provide an example for the application of these interdependence measures, the first four of the 450 sources were presented for analysis in matrix format. These matrices provide a way to illustrate the quantitative relationship between one source and the other sources within the network.

3.1.1 Bibliographic Coupling Application to the Network

In Table 1, the bibliographic coupling illustrates two concepts: total number of connections and total number of shared connections. The matrix diagonal values represents the former. For example, the first source, "Roweis St, 2000," has a nineteen total sources cited within its article; "Newman MEJ, 2006," has 32 total sources cited within its article. On the other hand, non-diagonal values represent the later, the shared connections. For example, "Milo R, 2002" shares one source with "Newman MEJ, 2006" and three sources with "Newman MEJ, 2001." One thing to note is that the total number of shared citations will aggregate into the diagonal value in this 450 source sample; however, this observation may not be the case for all samples that could be taken by the IC. There is a possibility that a sample of sources could have one citation not shared among any other sources within the network except one, specific source. Another point to note is that the number of shared connections is mirrored along the diagonal of the matrix mirrors.

Table 1. Snapshot of First Four Sources - Bibliographic Coupling

	ROWEIS ST, 2000, SCIENCE	MILO R, 2002, SCIENCE	NEWMAN MEJ, 2006, PROC NATL ACAD SCI U S A	NEWMAN MEJ, 2001, PROC NATL ACAD SCI U S A
ROWEIS ST, 2000, SCIENCE	19	0	0	0
MILO R, 2002, SCIENCE	0	11	1	3
NEWMAN MEJ, 2006, PROC NATL ACAD SCI U S A	0	1	32	2
NEWMAN MEJ, 2001, PROC NATL ACAD SCI U S A	0	3	2	30

3.1.2 Jaccard Similarity Coefficient Application to the Network

When applying the Jaccard Similarity Coefficient, the group could identify a ratio that can roughly calculate originality relative to the sample of sources collected. The Jaccard Similarity Coefficient takes any particular node's total number of connections and divides it by the same total number plus the number of shared connections. If a number calculated is small, then there is more shared connections, which increases the denominator within the calculation. In Table 2, "Milo R, 2002" relative to "Newman MEJ, 2006" yields a value of 0.0323. This particular value says that "Milo R, 2002" shares only 3.23% of all the sources across the entire sample of sources with "Newman MEJ, 2006." Notice that the diagonal values in Table 2 are zero; this value makes sense because it takes into account all the sources, which essentially disregards total number of sources within the calculation. Thus, the Jaccard similarity coefficient is a good measure based on the objections initially defined.

ISBN: 97819384961-2-7

Table 2. Snapshot of First Four Sources - Jaccard Similarity Coefficient

	ROWEIS ST, 2000, SCIENCE	MILO R, 2002, SCIENCE	NEWMAN MEJ, 2006, PROC NATL ACAD SCI U S A	NEWMAN MEJ, 2001, PROC NATL ACAD SCI U S A
ROWEIS ST, 2000, SCIENCE	0.00000000	0.00000000	0.00000000	0.00000000
MILO R, 2002, SCIENCE	0.00000000	0.00000000	0.03225806	0.09090909
NEWMAN MEJ, 2006, PROC NATL ACAD SCI U S A	0.00000000	0.03225806	0.00000000	0.05882353
NEWMAN MEJ, 2001, PROC NATL ACAD SCI U S A	0.00000000	0.09090909	0.05882353	0.00000000

3.1.3 Network Centrality Measures

When applying network centrality measures, the group could identify notable groups within the sample. Specifically, the network centrality measures used are Authority Centrality and Hub Centrality, depicted in Figure 3, respectively. Authority Centrality represents how much information is held by a given node, while Hub Centrality represents the connection between any two nodes given the initial framework of information (Dodds, 2011). The significance of these measures are that they may identify authoritative sources on a given subject and their inclusion in any measure of interdependency among sources would not necessarily imply replication. In deriving an objective measure of originality for a given source, a list of related authoritative sources may provide a means to identify sources that are strengthened by their consideration of appropriate background literature. Furthermore, the list of authoritative sources may also assist in calibrating the measure of interdependence, accounting for any magnification of its value due to the inclusion of these types of references.

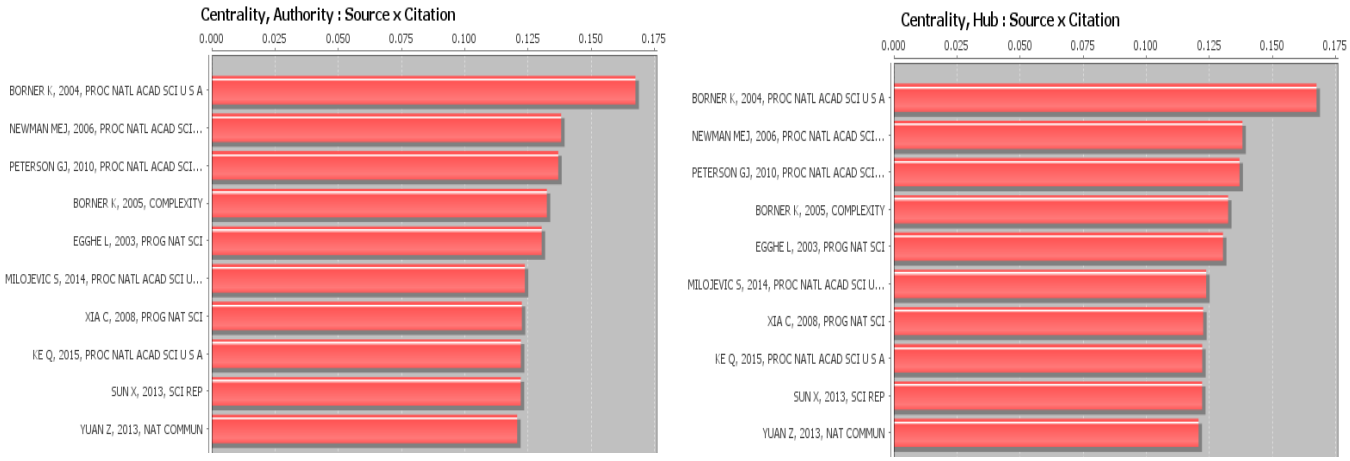

Figure 3. Histograms of the Top Ten Authority and Hub Centrality Measures of the 450 Source Sample

3.2 Conclusion

Ultimately, this data set of 450 sources provided a useful proof of concept that can be subsequently extended to intelligence production, as these sources were successfully manipulated into the matrix form necessary for data analysis. The bibliographic coupling and the Jaccard Similarity Coefficient distance measures produced two separate value measures depicting each source's degree of shared and unshared citations. The Jaccard Similarity Coefficient offers the most utility to OSINT analysts due to its intuitive ratio, and therefore, teachable nature. Moving forward, the Jaccard similarity coefficient will be compared to additional distance measures in order to provide analysts a more in-depth understanding of the relationships between future OSINT intelligence reporting.

Some limitations of these findings is that the Jaccard Similarity Coefficient cannot account for either reliability or credibility. For example, an article could be 100% original in nature but cite non-credible sources or false sources. Additionally, these findings do not account for the credibility of shared sources, which could latently skew our Jaccard metric. For example,

ISBN: 97819384961-2-7

one shared-source could be cited hundreds of times, which would decrease originality of any analyzed sources in theory; however, a source cited multiple times among an entire sample could just signify that over-cited sources as an accurate source. Further research would look to combat this possible problem and determine an objective metric to measure a given article's credibility in conjunction with its shared-sources and originality determined by the Jaccard similarity coefficient.

4. References

Alexander, K. B. (2004, August 20). Letter to the Honorable Robert R. Simmons [Letter]. Office of the Deputy Chief of Staff for Intelligence, Washington, D.C.

Aria, M., & Cuccurullo, C. (2018, January 01). A Brief Introduction to Bibliometrix. Retrieved from https://cran.r-project.org/web/packages/bibliometrix/vignettes/bibliometrix-vignette.html

Best, R. A., & Cumming, A. (2007). *Open Source Intelligence (OSINT): Issues for Congress* (Report No. RL34270). Washington, D.C.: Foreign Affairs, Defense, and Trade Division.

Betts, R. K. (1978). Analysis, War and Decision: Why Intelligence Failures Are Inevitable. *World Politics*, *31*(1), 61-89. http://dx.doi.org/10.2307/2009967.

Bosse, J. C. (2018, February 5). Re: Building a Data Frame and Matrix for Citation Network [Online discussion group]. Retrieved from https://datascience.stackexchange.com.

CASOS, Carnegie Mellon. (2017). **ORA-Lite: Overview*. Retrieved from http://www.casos.cs.cmu.edu/projects/ora/index.php.

Disney, A, (2014, December 3). *KeyLines FAQs: Social Network Analysis*. Retrieved from https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/

Dodds, P, (2011). *Measures of Centrality: Complex Networks.* Retrieved from http://www.uvm.edu/pdodds/teaching/courses/2011-01UVM-303/docs/2011-01UVM-303centrality-flat.pdf.

Lopez, J. A. (2012). Best Practices for Turning Big Data into Big Insights. *Business Intelligence Journal*, *17*, 17-21.

Mercado, S. C. (2009). Sailing the Sea of OSINT in the Information Age: A Venerable Source in a New Era. Retrieved from the Central Intelligence Agency website: https://www.cia.gov.

Minsky, M. L. (2006) *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York, NY: Simon and Schuster. 182-185.

Norton, J. A., & Bass, F. M. (1987). A Diffusion Theory Model of Adoption and Substitution for Successive Generations of High-Technology Products. *Management Science, 33*(9), 1069-1086. doi:10.1287/mnsc.33.9.1069.

Schaurer, F., & Storger, J. (2013). The Evolution of Open Source Intelligence. *Journal of U.S. Intelligence Studies*, *10*(3), 53-55. Retrieved from https://www.afio.com/publications/Schauer_Storger_Evo_of_OSINT_WINTERSPRING2013.pdf.

StataCorp (n.d.). *Manual: mvmeasure_option.* Retrieved from https://www.stata.com/manuals13/mvmeasure_option.pdf.

Stodder, D. (2013) Data Variety: The Spice of Insight. *Business Intelligence Journal*, *18*(4). 53-55. Retrieved from file:///C:/Users/x85856/Downloads/TDWI_BIJV18N4_Web.pdf

Verton, D. (2016, May 2). *The 'Googlization' of Intelligence and Counterterrorism Analysis*. Retrieved from https://www.fedscoop.com/the-googlization-of-intelligence-and-counterterrorism-analysis.

Washington, D.C. United States Army. (2016). *U.S. Army Open-Source Intelligence Activities: Army Directive 2016-37*. Washington D.C.: Government Printing Office.

ISBN: 97819384961-2-7