Trusting Open Source Information

Marianela García Lozano
FOI, Swedish Defence Research Agency
164 90 Stockholm, Sweden
Email: garcia@foi.se

Intelligence analysis is dependent on credible input data that comes from trusted and reliable sources. Open Source Information (OSINF) provides an abundance of data, but it comes with the price of noise, i.e., a lot of the data is irrelevant, ambiguous, contradicting, biased or plain wrong. Despite this, making full use of the wealth of data that OSINF encompasses would improve the quality of intelligence analysis. Traditionally the reliability of sources and credibility of information are *manually* assessed by intelligence analysts, but the large volumes and velocity by which OSINF is created make it unfeasible to continue doing so. Hence, automatic support in the form of methods, techniques and tools, are needed. The core question of the research described in this paper is: How can we *automatically* evaluate information and sources, and assess their veracity? Or, in other words, how can we automatically assess the credibility and reliability of a source and the information provided by it?

In order to reason about veracity assessment and the related challenges we introduce a theoretical framework. Assume our world consists of a large number of OSINF elements, as shown in Figure 1. These elements are divided into known elements, depicted by green dots or rectangles, and unknown elements, depicted by red dots and rectangles. The dots (or nodes) in the figure represent producers or consumers of information, denoted by n_k where $k = 0, ..., N$ if known, or n_u if unknown. Similarly the rectangles represent information elements, denoted by e_k if known or e_u if unknown . We define a node to be "known" if we have either directly assessed its reliability, or if an assessed node in turn has assessed another node's reliability, making the term "known" transitive. Hence, a directed network of known nodes is formed by following the reliability values that nodes have for each other. This could also be compared to a social or trust network. Note that assessing a node does not imply that we "trust" them. In the same way we define an information element to be "known" if we have either directly assessed its credibility or transitively if a "known" node has assessed its credibility.

Suppose that we have a database of information elements e_1, e_2 that we have assessed and rated $C(n_i, e_1) = c_1$, $C(n_i, e_2) = c_2$. We are presented with a new information element e_3 which we have not yet assessed. This information element might in the first case have been produced by a source we know, and in the second case have been produced by a source that is completely unknown. The question becomes – What kind of trust can we put on this information item and / or its source? Four main types of challenges can be identified in Figure 1: *transitive trust* (to achieve an automatic reliability assessment of a node we do not directly know, but that is known by someone else in the network); *trust of unknown nodes* (to achieve an automatic reliability assessment for a node that no one in the trust network knows); *transitive credibility* (to achieve automatic credibility assessment of a known information item); and *unknown credibility* (to achieve automatic credibility assessment of an unknown information item). An addition challenge related to veracity assessment has to do with the *context* in which an analyst is working. In one situation sources and information may be totally unacceptable and in another the analyst's operational frame and task may allow for some leniency. The context does not in practice change the credibility or reliability assessment but it does change the *acceptance* level where an analyst may be more or less inclined to use or discard a piece of information. We use the term *context awareness* to name this challenge. We believe that for a future automatic veracity assessment / recommendation system context awareness will need to be included.

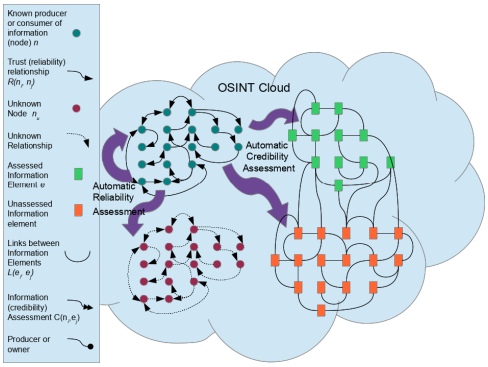

Fig. 1. Known and Unknown Producers and Consumers of Open Source Information Elements. Credibility and Reliability Challenges

Returning to the example database, we have two cases: transitive credibility and unknown credibility. We are interested in whether someone else has assessed the information or if the information item is similar to something already assessed? Depending on the answer to these questions we may approach the example from the source trust point of view or the information similarity point of view. In this example we chose the latter and by using our similarity metrics we calculate the similarity value that the new element has to the elements already stored in our database . These similarity values are then weighted with the trust values that we have for the other elements' sources (given that we do not already have a trust value for the new element's source).

To continue our work we will begin by implementing the example that we have used in the paper. We will evaluate it by comparing the results with manually assessed information items.

978-0-7695-5062-6/13 $26.00 © 2013 IEEE
DOI 10.1109/EISIC.2013.77