Towards Automatic Veracity Assessment of Open Source Information

Marianela García Lozano, Ulrik Franke, Magnus Rosell
DSS, Department of Decision Support Systems
FOI, Swedish Defence Research Agency
164 90 Stockholm, Sweden
Email: {garcia, ulrfra, magros}@foi.se

Vladimir Vlassov
SCS, Department of Software and Computer Systems
KTH, Royal Institute of Technology
164 40 Kista, Sweden
Email: {vladv}@kth.se

***Abstract*—Intelligence analysis is dependent on veracity assessment of Open Source Information (OSINF) which includes assessment of the reliability of sources and credibility of information. Traditionally, OSINF veracity assessment is done by intelligence analysts *manually*, but the large volumes, high velocity, and variety make it infeasible to continue doing so, and calls for automation.**

Based on meetings, interviews and questionnaires with military personnel, analysis of related work and state of the art, we identify the challenges and propose an approach and a corresponding framework for automated veracity assessment of OSINF. The framework provides a basis for new tools which will give the intelligence analysts the ability to automatically or semi-automatically assess veracity of larger amounts of data in a shorter amount of time. Instead of spending their time working with irrelevant, ambiguous, contradicting, biased, or plain wrong data, they can spend more time on analysis.

***Keywords*-big data; data veracity; veracity assessment; reliability and credibility; trust; OSINF; NATO STANAG 2511**

I. Introduction

Big data is often used to describe data and problems characterized by the traditional three Vs, namely, *Volume*, *Velocity*, and *Variety* [1]. This has in more recent times been expanded by a fourth big data dimension, namely *Veracity*, which refers to the *quality or trustworthiness* of the data. There is no clear definition of veracity within the big data community but other words and descriptions that also are used in relation to veracity are "uncertain or imprecise data", reliability, credibility, fidelity, "biases, noise and abnormality in data". Assessing veracity is a challenging problem and, so far, not much work has been done within this area.

Within the military intelligence domain there is a long tradition of using, managing and analyzing uncertain data. An important part of this is to *manually* assess the *reliability* of sources and the *credibility* of information, in essence making an assessment of the trustworthiness and quality, i.e., veracity, of the source and information. In this paper, we use Veracity interchangeably with trust, reliability and credibility.

In the intelligence domain, Open Source Information (OSINF) consists of all information which is *publicly* available, e.g., books, news media, radio, public databases, reports and everything found on the Internet, see Figure 1. There are huge volumes of data to explore and the potential benefits of harnessing it are vast. Unfortunately, the major part of OSINF is unstructured and there are few producers which are trusted. Hence, OSINF is of very diverse quality, and comes in all shades of being incorrect, biased, outdated, incomplete, inconsistent and contradicting, making the challenge of assessing veracity demanding.

Fig. 1: Examples of types of information found in OSINF

A. Goal, Contributions and Paper Outline

The core question of this paper is: Which are the challenges we need to address, and the approaches we may take, to *automatically* assess the veracity of OSINF in order to be able to analyse large amounts of data within a short amount of time? We consider situations characterized by: Unstructured data, e.g., plain text; Large data volumes; Continuous streams of data; Data that can be (intentionally) misleading, e.g., incomplete, biased, contradicting, wrong, and, outdated.

The contributions of this paper are:

- an empirical study of veracity assessment within the military domain;
- a framework for the veracity assessment challenges;
- an approach for how to handle the challenges and move towards automating veracity assessment of OSINF.

The remainder of this paper is structured as follows. In Section II we review related work and analyse the state of

978-1-4673-7278-7/15 $31.00 © 2015 IEEE
DOI 10.1109/BigDataCongress.2015.36

the art in automation of veracity asessment of OSINF within the military domain. Section III, describes the theory for how veracity should be manually rated within the military intelligence and the main issues identified with this. The second part describes our empirical investigation into how assessment is de facto done within the military domain. Based on our findings we introduce a theoretical framework described in Section IV, that is used to outline, break down and reason about the veracity assessment challenges. Given the presented challenges, in Section V, we propose a veracity assessment automation approach based on probabilistic networks, trust propagation, and semantic similarity. Finally, in Section VI, we conclude and summarize our work, outlining future work.

II. Related Work and State of the Art

This section is divided into two subsections. The first discusses related work in different domains such as trust in social networks and data quality assessment. The second subsection focuses on approaches to automation of veracity assessment within the intelligence analysis domain.

A. Related Work

Some researches have differentiated between trust and reputation. [2] give three definitions of different types of trust: Reliability Trust, Decision Trust and Reputation. These differences are interesting if we view veracity assessment from a trust or a recommender perspective. The latter would be synonymous with e-commerce systems that give recommendations of the type ”buyers who liked / bought this book also liked / bought...”. Would intelligence analysts accept a system saying ”Analysts who viewed / trusted this information also viewed /trusted...”? If we view veracity assessment from a trust perspective there are several algorithms, e.g., SUNNY [3] and MoleTrust [4] which provide trust metrics for transitive trust relationships in social networks.

Some researchers have explored topical trust [5], [6] and also homophily in trust, e.g., in network analysis [7] and by exploiting taste distances using the Pearson coefficient [3].

Another useful approach [8] is to use a network centrality measure. In this case it was based on TunkRank[1], a Twitter analogy to PageRank [9]. The Network centrality measure can be seen as an indirect measure of trust where nodes that are more retweeted, referred or linked to are more trusted. [10] have proposed TwitterRank a method to measure a tweeter's influence taking both the topical similarity between users and the link structure into account. They use Latent Dirichlet Allocation (LDA) [11] to explore topic sets. [12] uses LDA to mine for opinion distances in Twitter based social networks. This can then be used to find similar minded groups within the network and also view the opinion distances between groups. According to [12] the more utilized methods for topic exploring are Probabilistic Latent Semantic Analysis (PLSA) [13] and LDA.

[1]TunkRank http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank

Provenance which is an important factor in assessing trust is a fairly new area. [14] stated (2010) that ”15 years ago the term *data provenance* was not in use”. There have been some interesting works done in this field like [15] that explore the why and where of provenance, i.e., why was the data created and where does it come from.

Data and information quality are also used in connotation with veracity. The most widely used definition is the one proposed by [16] where quality is defined as ”fitness for use”. There are two main implications of this definition. The first being that quality is task-dependent and that a user may consider information appropriate for one task but not for another task. The second is that information quality is subjective as users may perceive the quality of the same information differently. The first issue is in line with the previously stated context awareness challenge and the second issue is in line with the subjectivity issue present in veracity assessments. In an effort to capture the aspects of data quality that are important to data consumers [17] did a two stage survey that resulted in a set of data quality dimensions (Intrinsic Data Quality, Contextual Data Quality, Representational Data Quality and Accessibility Data Quality). Many have tried to use this to create quality aware systems that filter out unwanted or low quality information, e.g., [18].

Our inability to give exact estimates of world states is reflected in the subjectivity issue in assessing sources and information. A method to deal with this is to use probabilistic methods [19]. Many of the approaches mentioned in this section use probabilistic methods such as Baysian Networks, Dempster-Shafer theory and LDA to calculate, among other things, transitive trust and topic distances.

B. State of the Art in Veracity Automation within the Intelligence Analysis Domain

In a semi-automatic evaluation process that uses an ontology and Natural Language Processing (NLP) to detect similar items of information and [20] introduce a user-centric semantic based model to assess information. [21] also use semantic analysis to estimate the correlation of HUMINT data, but the proposed Shallow Semantic Analysis (SSA) approach is unsupervised and automatic.

In [22] the authors survey approaches for automatic information evaluation and assess their applicability to answering Priority Intelligence Requirements (PIR).To follow up their survey [23] the authors present a proof-of-concept Semantic Wiki Alerting Environment (SWAE). Much in the same way as previous authors and approaches streamed reports (Twitter, Blogs, Flickr) are processed by entity extraction and semantic analysis systems. In an improved approach to apply assessment standards, i.e., STANAG 2022, to Twitter [8] they make use of network centrality measures. The authors also deal with the possibility of users making false retweets where users attribute tweets to other users. If a user is found to have done such a thing the user is then marked as unreliable.

III. Current Theory and Application of Veracity Assessment within the Military Domain

In this section we will review the dominating veracity assessment system used within the military, the issues that have been pointed out with it and how some countries have added to it. We will also review how veracity assessment is done in practice.

A. NATO STANAG 2511

The North Atlantic Treaty Organization (NATO) Standardization Agreements (STANAG) 2022 and its updated version 2511 [24] outline a ranking system[2] for assessing intelligence reports. The two main concepts are the *reliability* of the source and the *credibility* of the information.

The source reliability rating notation uses an alphabetic coding, A-F, where F is actually not an evaluation of a source, but an indication that its reliability cannot be judged. In short it is; *A – Completely reliable, B – Usually reliable, C – Fairly reliable, D – Not usually reliable, E – Unreliable, F – Reliability cannot be judged.*

An information item's credibility is classified using a numeric code ranging from 1-6. The assessment is based on likelihood and levels of corroboration by other sources. When an item cannot be classified it is given the rating of 6. In short it is; *1 – Confirmed by other sources, 2 – Probably true, 3 – Possibly true, 4 – Doubtful, 5 – Improbable, 6 – Truth cannot be judged.*

Reliability and credibility are to be assessed *independently* and all combinations (A-F, 1-6) are possible. In the NATO System there is no methodology on how to proceed with the rating and further more, the system is open to interpretation resulting in the ratings being highly *subjective*.

B. Assessment Issues with NATO STANAG 2511

A number of authors have on different occasions presented issues that they have identified with the NATO STANAG 2511 and its predecessor [25]–[29]. Summarizing the challenges with the NATO STANAG 2511 recommendation we note that the issues pointed out, namely *ambiguous, missing and imprecise* definitions of the core concepts, combined with *undefined situations*, lead to different interpretations of how the recommendation should be applied.

Why is this a concern? The differences in interpretation of STANAG 2511 lead to a great variance in the ranking. The analysts are left to depend on their own experience, domain expertise and their own biases. From a system perspective this could almost be equated to a random behaviour in the assessment. Making it highly difficult to achieve a notion of *quality* or *goodness*. An automation and systematization of the veracity assessment would lead to a predictable and consistent behaviour. Which in turn would make the comparison between different pieces of assessed information and sources easier.

To combat some of these issues certain countries, e.g., USA and Sweden, have made additions to their assessment frameworks by adding requirements of likelihood and and analytical confidence.

In 2007 the American National Intelligence Council (NIC) released a document[3] outlining how likelihood and analytic confidence should be used by American intelligence analysts. Likelihood is meant to reflect a sense of the probability of a development or an event. As seen in Figure 2 likelihood is depicted as a sliding scale with terms ranging from *Remote* to *Almost certainly*. The NIC text is unfortunately somewhat confusing since the description of the figure also states that words such as "we cannot dismiss", "we cannot rule out", and "we cannot discount" are used to reflect an unlikely or even remote event. Further on, words such as "may be" and "suggest" are to be used in situations where likelihood cannot be assessed.

Fig. 2: Event Likelihood figure from http://goo.gl/rPdZk7

In addition a judgment confidence level ranging between "high", "moderate" and "low" is also to be provided. No clues or method on how to assess likelihood or confidence is given in the NIC document.

C. Study of Veracity Assessment Implemented in Practice

Due to the found issues with the NATO standard and the need for introducing auxiliary concepts (likelihood and confidence), we wished to see veracity assessment implemented in practice. We were especially interested in seeing how OSINF is used and assessed. From what we had gathered there are large gaps between the recommended assessment standard and the feasibility of using it on OSINF. Hence, we needed to find out any methods or heuristics developed and employed by the analysts. We approached this task in three ways, i) questionnaires, ii) interviews, and iii) meetings.

1) Combined Joint Staff Exercise 2013 - Questionnaire: The aim of the Combined Joint Staff Exercise (CJSE) is to prepare the individual participants for work in an international staff within a multi-functional / multinational Crisis Response Operation (CRO), both on tactical and operational levels. The CJSE Exercise series is multinational and CJSE 13 was supported by participation from several defence colleges[4]. The Training Audience (TA) consisted of HQs representing Operational, Land, Maritime and Air Components. CJSE 13 took place 16-26th of April 2013 at four different locations in Sweden and in total the exercise included around 1000 people from both military and civilian organizations.

At the CJSE 13 exercise we had the opportunity to ask the exercise intelligence participants to answer a questionnaire

[2]Sometimes NATO STANAG 2511 is also referred to as the *Admiralty System, Admiralty Scale* or the *NATO System*.

[3]See http://goo.gl/rPdZk7

[4]The Swedish Defence College, the Baltic Defence College, the Finnish National Defence University, the Austrian National Defence Academy, the Norwegian Defence Command Staff College, and the Swiss Armed Forces Headquarters.

focusing on the some of the issues we had identified with NATO STANAG 2511. We were mostly interested in either confirming, rejecting or adding to the list of issues that we had found in the literature or thought of ourselves.

On the intelligence side there were a total of 145 intelligence positions which were seen as potential respondents, see Figure 3. Among these, there were both vacancies and executives with no previous intelligence experience. A link to the questionnaire was sent out on one of the final days of the exercise and 45 answers were obtained in total.

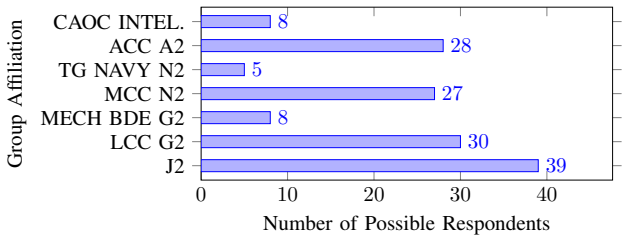


Fig. 3: Total number (145) of Possible Respondents divided into group affiliations

The questionnaire consisted of four main parts, it can be seen in full at http://goo.gl/MOyzW8.

a) Part 1, Experience: – 4 questions regarding the respondents' experience with NATO STANAG 2511, years of service and their opinion on what OSINF is.

Here we found that the respondents were a mixed group with a wide range of experience from 2 to 30 years and most seemed to have a good, but perhaps sometimes narrow, grasp of what OSINF is.

b) Part 2, Definitions and Attributes: – 6 questions about the respondents' perception of the basic NATO STANAG 2511 definitions, attributes and quality of assessment.

Here we asked about the main attributes that were used to assess a source's reliability and information's credibility. The four attributes most respondents stated they used to assess a source and information were, in falling order, "reliability", "Accuracy and correctness", "objectivity", "traceability and provenience". Worth noting is that the same attributes were given for assessing *both* reliability and credibility. The final question pertained to how the respondents would go about to evaluate and give feedback on an assessment. The answers highlighted two things that were recurring, the first being the need for traceability and the second being the use of the so called "gut feeling".

c) Part 3, Source Judgment: – 5 questions regarding the respondents' approach to judge a source's reliability, independence between sources and how the respondents perceive the basic concepts.

For example, the respondents were asked in one question who they considered to be the source in a newspaper article reporting an eye witness statement. The alternatives were 1) the newspaper, 2) the article's author, 3) the eye witness or, 4) other. Noteworthy is that the answers were *equally* distributed between the first three choices. Highlighting the fact that there seems to be confusion on who is to be regarded as the source. Another issue we asked the respondents about in this section was to describe how they would "... proceed about verifying the independence of Open Source Information (OSINF) sources?". Many of the answers demonstrated a consciousness of 1) the need to assert whether a source really is the originator of the information or only relying it, 2) check the history and associations of a source and, 3) also follow it over time. However, one of the answers was perhaps the most honest about how reality is stating that they "usually don't verify independence".

d) Part 4, Information Judgment: – 10 short questions on the respondents' use of and approach to judging information and its credibility.

For example, two different questions asked which information the respondents would i) *trust more* or they ii) *would judge to be more probable* - information rated A5, E1 or both the same? Interesting was that the majority of the respondents answered that they would *trust* information rated A5 more. In other words, they trusted the source not the information. The answers to which they would *judge more probable* were almost equally distributed between the three alternatives showing that there is a lot of room for interpretation of the NATO scales.

2) CJSE 13 - Force Headquarters Interviews: At the Swedish Combined Joint Staff Exercise 2013 (CJSE13) we also conducted interviews with 9 members of the Force Headquarters (FHQ) to get a view of their perception of information and source assessment. The interviewees came from a varied background ranging from analysts with many years of experience within intelligence to company officers with specialist officers under them. At the interviews we not only asked them on their use of OSINF in the exercise but also in their normal day to day positions in the SwAF.

Some of their opinions and experiences were: The NATO scales are connected and not completely independent, usually a source rated A provides information rated 1-3; Open Sources are seldom rated higher than C3; The Web is viewed as a Single Source by some, others see different web sites as different sources; The constant lack of time is a large problem in the veracity assessment of sources and information; It is perceived as difficult / time consuming to have a continuous assessment of sources and information, follow ups of assessments are seldomly done.

3) Swedish OSINF Intelligence Analysts: We interviewed two Swedish intelligence analysts who focus on OSINF and discussed information and source assessment. They were aware of the shortcomings of the NATO STANAG 2511 scale and were of the opinion that it is not really applicable to OSINF. Instead they have developed a straightforward model to judge source reliability. This model is applicable to sources who have given similar type of information during a long time, e.g., news media. The model consists of three parts which aim to answer what the sources possibilities of producing its own news material are: i) Access, e.g., How did the source get the information? ii) Motivation, e.g., what is their affiliation / bias? iii) Precondition & Capacity, e.g., do they have boots

on the ground? The required language skills to obtain the information? Number of reporters? Number of offices? What is their monetary situation? The answers to these questions are combined with the analysts' gut feelings, information from the source itself, the press, wikipedia and other web tools such as sourcewatch.org, domaintools.org, etc. An important part of the method was the need for continuous follow up and comparison with other sources to get a feel for the reliability of the source.

The result of these questions and approach is not meant to be a compact code like the NATO scale but rather a motivation to the source evaluation.

IV. Veracity Assessment Framework

In order to reason about veracity assessment and the related challenges we introduce here a theoretical framework. Assume our world consists of a large number of OSINF components, see Figure 4. These components are divided into *assessed* components and *not assessed* components. The dots (or nodes) in the figure represent producers / sources or consumers of information, denoted by n_i where $i \in \{0, ..., N\}$ if assessed (blue dots), or $\hat{n}_i$ if not assessed (red dots). Similarly the rectangles represent information elements, denoted by e_l where $l \in \{0, ..., M\}$ if assessed (green rectangles), or $\hat{e}_l$ if not assessed (light red rectangles).

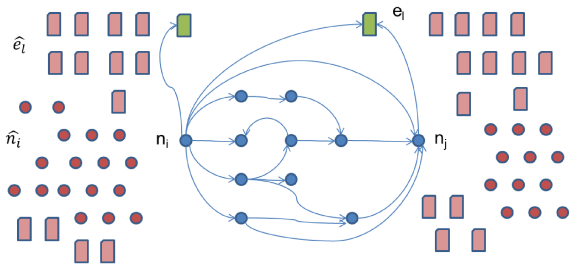


Fig. 4: OSINF Components

Reliability of n_j where $j \in \{0, ..., N\}$ can be assessed from the perspective of n_i: $R(n_i, n_j)$. The value can be arrived at either directly or transitively. Note that assessing a node does not imply that we "trust" them.

In the case that a source n_i has manually assessed n_j we call it a direct assessment, denoted by the subscript D: $R(n_i, n_j) = R_D(n_i, n_j)$, see the blue links in Figure 5a. R_D gives rise to a directed network W_{R_D} of direct reliability assessments. This could also be compared to a social, trust or probabilistic network.

In the case that the node n_j has not been directly assessed by n_i, n_i could decide to use the assessments of its peers and calculate a transitive reliability value from theirs. See examples of transitive reliability links in Figure 5b, marked by orange links. We could also use the name "transitive trust" for this since it is used in social networking. Given that there exists at least one path in W_{R_D} from n_i to n_j:

$$R(n_i, n_j) = R_T(n_i, n_j), \tag{1}$$

where the subscript T denotes *transitive* reliability. R_D combined with R_T gives rise to a new directed network $W_{R_{D+T}}$ of direct and transitive reliability assessments.

Analogously the direct credidibility assessment $C_D(n_i, e_l)$ is the manual assessment of an information element e_l by node n_i, see Figure 5c. Similarily this also forms a directed network of direct assessments, W_{C_D}. If there is no direct credibility $C_D(n_i, e_l)$ between n_i and e_l, but there is a reliability assessment $R(n_i, n_j)$ and a direct credibility assessment $C_D(n_j, e_l)$, see Figure 5d, we can calculate the transitive credibility:

$$C_T(n_i, e_l) = R(n_i, n_j) \otimes C_D(n_j, e_l), \tag{2}$$

where $\otimes$ is a suitable operator.

C_D combined with C_T gives rise to a new directed network $W_{C_{D+T}}$ of direct and transitive credibility assessments.

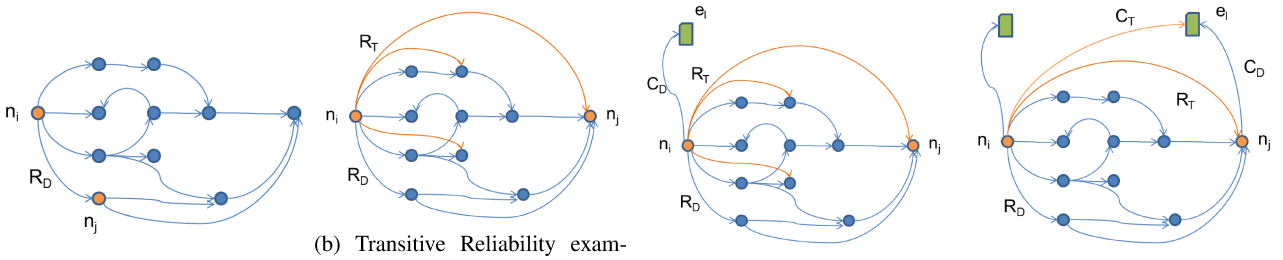
A challenge is to decide which *assessment scale* to use to represent a veracity assessment. Remembering the issues brought up earlier with the NATO scale, the main thing is that the user understands *how* the scale should be used and *what* it refers to. This includes the issues with basic concepts and undefined situations. For example, a reliability and credibility value could be given using the NATO scale (A-F, 1-6), or a set of numerical values in $[0, 1]$, as long as it is clear to end user what is meant. Also, to evaluate a veracity assessment, automatic or not, a notion of quality is needed. When has a good job been done? We have looked at how it is done today within the intelligence but the challenge is that we are dealing with *subjective* assessments.

Another challenge is achieving *Context Awareness*. This 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* and *confidence* level where an analyst may be more or less inclined to use or discard a piece of information. We believe that for a future automatic veracity assessment / recommendation system, context awareness will need to be included.

However, the main challenge for automation is to obtain assessments for as many of the sources and information elements as possible and as quickly as possible. Also, these assessments should be related to the analyst's direct assessments and be of as high a quality as possible. The biggest difficulty lies in achieving automatic veracity assessments for nodes, $\hat{n}_i$ and information elements $\hat{e}_l$ that no one in the trust network has assessed, see Figure 4.

V. Exploiting Similarity for Veracity Assessment of Unassessed Nodes and Information Elements

Our end goal is to achieve automatic or semi-automatic veracity assessments of previously unknown sources and information elements. As we know OSINF is being created at a high velocity making it virtually impossible to keep track of everyone and everything. We are bound to have huge volumes



(a) Direct Reliability (b) Transitive Reliability examples (c) Direct credibility (d) Transitive Credibility example

Fig. 5: Veracity Assessments of nodes and information elements, described in Section IV.

of previously unseen sources and information. An approach which we believe may be fruitful is exploiting different types of similarities to get a veracity assessment.

We define similarities to be: $S_n(n_i, n_j)$ for nodes and $S_e(e_l, e_k)$ for information elements, see Figure 6. Another term one could use in place of similarity is *distance*.

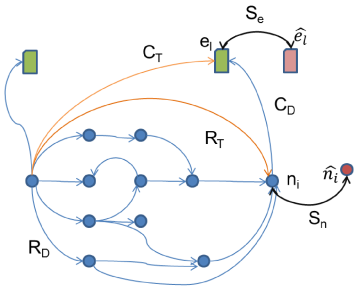


Fig. 6: Similarity

Similarities may be calculated between all configurations of assessed and unassessed nodes and information elements, for examples, see the black links in Figure 7. The similarities can then be used to extend the networks $W_{R_{D+T}}$ and $W_{C_{D+T}}$ to $W^{ext}_{R_{D+T}}$ and $W^{ext}_{c_{D+T}}$. With these networks more transitive assessments, R_T and C_T, can be calculated using Equations 1 and 2 weighted by the relevant similarity values. Which similarity values that are relevant has to be decided both by the context and suitable thresholds (since it would not be feasible or desirable to have all W^{ext} values in a calculation).

The thresholds are also necessary when calculating transitive veracity assessments since the paths from n_i to n_j might be too many, too long, have too large variance or be too old. This also emphasizes the need for adding *confidence* values to the veracity assessments. Which is in line with how the NATO assessment scales are used today.

A. Approaches to Similarity

To assess the similarities between nodes or information items we believe that there are three main types of metrics that can be used, i.e., content, meta-information and rating based.

The first type, i.e., the *Content based* metrics, exploit the information itself. The information can be compared with other

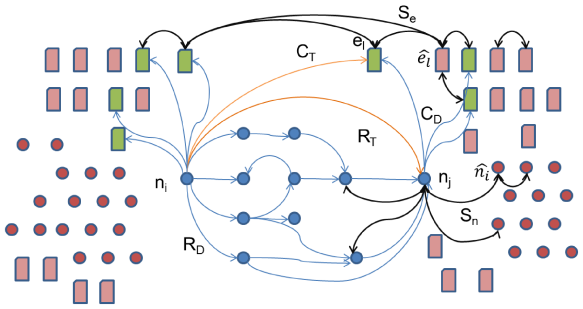


Fig. 7: Start of Similarity Network

information to compare or verify facts. It can also be compared to itself to see whether it is consistent. Examples of use: topic and opinion detection; sentiment and affect analysis; style.

The second type, i.e., *Meta-information based* metrics, use meta-information about the information and circumstances in which the information was created. Examples of things to use: author or publisher of the information; provenance; sources used to produce the information.

The third type, *Rating based* metrics, relies on explicit and implicit ratings about the information itself, information sources, or information providers. Examples are: grade; influence and popularity measure; network centrality. These type of metrics give rise to both direct and transitive veracity assessments. For example, the explicit rating is the equivalent of having a direct veracity assessment. The implicit ratings are the equivalent of having either transitive veracity assessments or making use of network centrality measures, e.g., the amount of followers and retweeters a source has alternatively in the information element case, the amount of links going to and from other information elements.

In related work we saw several approaches, e.g., topical trust, homophily in trust, network centrality and opinion mining that can be used to calculate similarity. Several of the state of the art approaches utilized semantic similarity to find correlation between information elements. Some, like [20], [21], [23] use content based metrics to try to assess similarities between already gathered information. Semantic methods and NLP are vital in this. In [8] they use rating based metrics by

employing network centrality measures. Provenance and meta-information based metrics are in turn used very little and only indirectly where some sources may be tagged as reliable or unreliable based on previous behaviour.

We believe that automatically calculated similarities can be used as a starting point to assess source credibility and information reliability. We base this on two reasons: the first is that the manual assessments are *subjective* in nature. The second reason is found in previous work, state of the art and results from our empirical studies. It has also been shown in studies that similarity is a successful method for pinpointing trusted links since individuals tend to associate with similar others [7]. It has also been argued that we humans use the recognition primed decision model [30] which is conceptually very similar to Case-Based Reasoning (CBR) heuristic, i.e., solve new problems based on our experience of the solutions of similar past problems [31]. If we follow the analogy of CBR in intelligence analysis this means that we tend to rate similar things in the same way.

B. Veracity Assessment Example

Let us consider the following example, assume that we, i.e., n_0, have a database of information elements $\{e_0, e_1\}$ that we have assessed and rated $\{C(n_0, e_0), C(n_0, e_1)\}$ and we have also assessed their sources $\{n_1, n_2\}$ with $\{R(n_0, n_1), R(n_0, n_2)\}$. We are presented with a new information element $\hat{e}_2$ which we have not yet assessed. This information element might in the first case have been produced by a source we know, e.g., n_1 and in the second case have been produced by a source that is completely unknown $\hat{n}_3$.

The question becomes – What kind of trust can we put on this information item and / or its source?

If we divide this example into two cases: the first being transitive veracity and the second being unassessed veracity. Two main questions arise:

- Has anyone else previously assessed the information and if so, how much do we trust them?
- Is the information element similar to something we, or someone else, have already assessed?

Depending on the answer to these questions we may approach the example from the source trust point of view or the similarity point of view. In this example we choose 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 $\{S_e(e_0, \hat{e}_2), S_e(e_1, \hat{e}_2)\}$. These similarity values are then combined with the reliability values that we have for the other elements' sources (given that we do not already have a reliability assessment for the new element's source).

More concretely, how to approach the described example we suggest using a combination of topic detection methods and transitive trust. We begin by utilizing LDA [11] to discover topic similarity between the information elements. Assuming that these values fall above some given threshold we may then combine them with the trust that we have for the sources of the other information elements. Assuming that we have not directly assessed the other element's source the trust value for them is calculated using a transitive trust algorithm, e.g., SUNNY [3]. Depending on the range within which the reliability and credibility values would fall they would then be mapped to NATO System ratings or an alternative assessment scale together with a confidence value.

A simple starting strategy for fusing veracity assessments could be to use a weighted average of the most similar sources and information elements.

VI. Summary and Conclusions

As we have seen OSINF fulfill the 3V of big data and the advantages of harnessing it are vast. But, before it can be used a veracity assessment of its quality and trustworthiness needs to be done. Within the military domain there is a long tradition of dealing with uncertain information and manually assessing it. We reviewed the leading assessment scale (NATO STANAG 2511), how intelligence analysts view it, and their approaches to using it in combination with OSINF. The dominating issues brought up were: the lack of time to do any type of veracity assessment, the subjective nature of the assessments and, the ambiguity and fuzziness of the assessment scale. We can conclude that an automation and systematization of the veracity assessment would be highly beneficial.

For a veracity assessment automation approach to be qualitative, trustworthy, and accepted there are some things that have stood out, in our questionnaires and interviews, as necessary. An assessment needs, for example, to be traceable, i.e., used sources, information, methods and other assessments need to be accessible to the end user, also the ranking scale used in an assessment needs to be well defined and unambiguous so it is clear to the end user exactly what has been assessed and how the assessment should be interpreted. Also for a source to be seen as reliable it needs to be perceived as objective and it needs to have produced similar type of information during a long time. The source's access and motivation to produce information coupled with their capacity to do so are important factors to take into consideration in an assessment. Further, an assessment should also have a confidence value outlining the "goodness" of it and perhaps even a time stamp, i.e., best before date.

In order to reason about veracity assessment and the related challenges we introduced a theoretical framework. The framework describes necessary components and shows how a veracity assessment network is gradually built up and expanded from direct and transitive veracity assessments.

In this paper we have argued that a similarity based approach is the way to achieve an automatic or semi-automatic veracity assessment of unassessed nodes and information elements. We suggest that there are three main types of metrics that can be used to obtain indicators to help us estimate similarities, i.e., content, meta-information and rating based metrics.

Working towards automatic veracity assessment of OSINF will provide new tools which will give the stakeholders access to more qualitative data and more time to spend on the analysis

of the data instead of spending it working with irrelevant, ambiguous, contradicting, biased or plain wrong data.

A. Future Work

To continue our work we will implement the framework and try it out on varying scenarios with the aim to perform experimental test. Hence, obtaining quantitative validation and more data backed results.

REFERENCES

[1] D. Laney, "3d data management: Controlling data volume, velocity and variety," *Application delivery strategies, File*, vol. 949, 2001.

[2] A. Jøsang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decision support systems*, vol. 43, no. 2, pp. 618–644, 2007.

[3] U. Kuter and J. Golbeck, "Sunny: A new algorithm for trust inference in social networks using probabilistic confidence models," in *Proceedings of the national conference on artificial intelligence*, vol. 22, no. 2. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007, p. 1377.

[4] P. Avesani, P. Massa, and R. Tiella, "A trust-enhanced recommender system application: Moleskiing," in *Proceedings of the 2005 ACM symposium on Applied computing*. ACM, 2005, pp. 1589–1593.

[5] J. Golbeck and J. Hendler, "Accuracy of metrics for inferring trust and reputation in semantic web-based social networks," in *Engineering knowledge in the age of the semantic web*. Springer, 2004, pp. 116–131.

[6] T. Knap and I. Mlỳnková, "Towards topic-based trust in social networks," in *Ubiquitous Intelligence and Computing*. Springer, 2010, pp. 635–649.

[7] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, pp. 415–444, 2001.

[8] B. Ulicny and M. Kokar, "Toward formal reasoning with epistemic policies about information quality in the twittersphere," in *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*. IEEE, 2011, pp. 1–8.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web." 1999.

[10] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 261–270.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[12] N. Dokoohaki and M. Matskin, "Mining divergent opinion trust networks through latent dirichlet allocation," in *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*. IEEE, 2012, pp. 879–886.

[13] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.

[14] R. Blake, "Identifying the core topics and themes of data and information quality research," *AMCIS 2010 Proceedings*, 2010.

[15] P. Buneman, S. Khanna, and T. Wang-Chiew, "Why and where: A characterization of data provenance," *Database TheoryâĂŤICDT 2001*, pp. 316–330, 2001.

[16] J. Juran, *Quality control handbook*, 3rd ed. McGraw-Hill, 1974.

[17] R. Wang and D. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of management information systems*, pp. 5–33, 1996.

[18] C. Bizer, "Quality-driven information filtering in the context of web-based information systems," Ph.D. dissertation, Freie Universität Berlin, Mar. 2007. [Online]. Available: http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000002736

[19] R. Jeffrey, *Subjective probability: The real thing*. Cambridge University Press, 2004.

[20] J. Besombes, L. Cholvy, and V. Dragos, "A semantic-based model to assess information for intelligence," *AerospaceLab*, 2012.

[21] V. Dragos, "Shallow semantic analysis to estimate humint correlation," in *Information Fusion (FUSION), 2012 15th International Conference on*. IEEE, 2012, pp. 2293–2300.

[22] B. Ulicny, G. Powell, C. Matheus, M. Coombs, and M. Kokar, "Priority intelligence requirement answering and commercial question-answering: Identifying the gaps," in *Proceedings of the 15th International Command and Control Research and Technology Symposium (ICCRTS '10)*, 2010.

[23] B. Ulicny, C. Matheus, and M. Kokar, "A semantic wiki alerting environment incorporating credibility and reliability evaluation," in *Proceedings of the 5th International Conference on Semantic Technologies for Intelligence, Defense, and Security (STIDS 2010), Fairfax, VA*, 2010.

[24] N. A. T. Organization, "Standard: Nato - stanag 2511, intelligence reports - ed 1."

[25] L. Cholvy, "Information evaluation in fusion: a case study," in *Proceedings of the conference IPMU 2004*. Citeseer, 2004.

[26] L. Cholvy and V. Nimier, "Information evaluation: discussion about stanag 2022 recommendations," DTIC Document, Tech. Rep., 2004.

[27] V. Nimier, "Information evaluation: a formalisation of operational recommendations," in *Fusion 2004: Seventh International Conference on Information Fusion*, 2004.

[28] J. Besombes and A. d'Allonnes, "An extension of stanag2022 for information scoring," in *Information Fusion, 2008 11th International Conference on*. IEEE, 2008, pp. 1–7.

[29] T. Delavallade and P. Capet, "Information evaluation as a decision support for counter-terrorism," in *NATO symposium on C3I in Crisis, Emergency and Consequence Management, IST*, vol. 86, 2009.

[30] G. A. Klein, *Sources of power: How people make decisions*. The MIT Press, 1998.

[31] R. C. Schank, *Dynamic memory: A theory of reminding and learning in computers and people*. Cambridge University Press, 1983.