A Granular Approach to Source Trustworthiness for Negative Trust Assessment

Davide Ceolin, Giuseppe Primiero

A Granular Approach to Source Trustworthiness for Negative Trust Assessment

Davide Ceolin[1] and Giuseppe Primiero[2]

[1] Centrum Wiskunde & Informatica, Amsterdam, the Netherlands
Davide.Ceolin@cwi.nl
[2] Department of Philosophy, University of Milan, Italy
Giuseppe.Primiero@unimi.it

Abstract. The problem of determining what information to trust is crucial in many contexts that admit uncertainty and polarization. In this paper, we propose a method to systematically reason on the trustworthiness of sources. While not aiming at establishing their veracity, the method allows creating a relative reference system to determine the trustworthiness of information sources by reasoning on their knowledgeability, popularity, and reputation. We further propose a formal rule-based set of strategies to establish possibly negative trust on contradictory contents that use such source evaluation. The strategies answer to criteria of higher trustworthiness score, majority or consensus on the set of sources. We evaluate our model through a real-case scenario.

1 Introduction

Assessing information quality is a challenging task. Assuming a minimal definition of information as 'data + semantics', assessing its quality means to establish fitness for purpose for a given piece of information. Given the huge number of possible purposes and to make its computation feasible, information quality is often broken down into 'dimensions' [13], like accuracy, precision, completeness. Despite its complexity, humans deal with quality on a daily basis using heuristics to approximate ideal values and using them as a proxy for deciding whether to trust information or not. Notwithstanding the possibility of being deceived by our heuristics, a formalization of such strategies is a useful tool for understanding and prediction. We provide here a framework to mimic such strategies and a relative reference system of sources. When an oracle or fact-checking service is available, such a reference system can be turned into an absolute one, i.e., determining which sources are veracious and which not. Otherwise, our result will still provide a relative ranking of the importance of sources. This task relies on providing appropriate understandings of trust and trustworthiness.

Among the large number of its definitions in the literature, for our purpose trust on contents can be minimally identified with the result of a consistency assessment: a piece of information consistent with the agent's current set of beliefs or knowledge base is trusted when it allows to preserve other information considered truthful. This approach requires a methodology to deal with inconsistent information and it calls upon the problem of assessing source trustworthiness. The logic (`un`)`SecureND` [20] provides a mechanism to deal with this aspect through the introduction of separate protocols to deal with failing consistency. An agent A reading a piece of information ϕ from an agent B, where ϕ is inconsistent with A's knowledge base, has two possibilities: (1) *distrust*: to reject ϕ and preserve $\neg\phi$ and its consequences; and (2) *mistrust*: to remove $\neg\phi$ from her profile and to accept ϕ. (`un`)`SecureND` does not have a selection mechanism for either form of negated trust. In real case scenarios, the choice between distrust and mistrust will be determined by evaluating the source. While *trust* is the mechanism to establish admissible consistent information, we call *trustworthiness* the assessment quality on sources. We introduce an ordering function and several decision strategies aiming at providing computational mechanisms to mimic the subjective quality assessment process called *trustworthiness*. Through any of these mechanisms, A can decide whether the estimated trustworthiness of B is high enough to trust the new information ϕ. Consider a simplified scenario, with a finite set of sources sharing information on a common topic and referencing each other (to a lesser or greater degree): some of them will be in conflict and some will be consistent with one another. We identify three dimensions:

- *Knowledgeability*: the number of sources to whom a source B refers. This value is used as an indicator of B's knowledge of other views;
- *Popularity*: the number of sources referring to B. This counts the number of inbound links, and it does not involve their polarity. Citing a source, even to attack it, is seen as an indication of the popularity of the latter;
- *Reputation*: the proportion between positive and negative evaluations of B.

These dimensions are used for assessing the trustworthiness of B, to compare contradictory sources by a receiver, and to formulate decision strategies.

The paper continues as follows. Section 2 describes formal preliminaries, Section 3 describes the different strategies available to resolve the presence of contradictory contents, Section 4 translates these strategies in implementable rule-based protocols, Sections 5 and 6 present and discuss a use case implementation of the proposed logic. Section 7 surveys related work, and Section 8 concludes.

2 Formal Preliminaries

Consider a set of sources $\mathcal{S}$ and a (possibly partial) order relation $\leq_t$ over sources $\mathcal{S} \times \mathcal{S}$ expressing source trustworthiness; once defined, this is used as a proxy to establish trust in contents in the rule-based semantics presented in Section 4. We define the trustworthiness order $\leq_t$ as a function over three dimensions: reputation, popularity, and knowledgeability.

Reputation is an order relation $\leq_R$ over sources $\mathcal{S} \times \mathcal{S}$: intuitively, $S \leq_R S'$ means that source $S \in \mathcal{S}$ has at least the same reputation as $S' \in \mathcal{S}$. For simplicity, reputation is evaluated on the following criteria:

- we denote with $w(S)_{S'}$ a fixed weight of S received by S';

- $w = \{1, -1\}$, respectively for a positive and a negative assessment;
- we denote each $w(S)_{S'} = 1$ as *pos* and each $w(S)_{S'} = -1$ as *neg*;
- for any source $S \in \mathcal{S}$, a reputation assessment $r(S)$ by other sources in $\mathcal{S}$ is

$$r(S) = \frac{|pos| + 1}{|pos| + |neg| + 2}$$

We note that instead of computing the simple ratio of positive assessments over the total number of assessments, we add a smoothing factor like in Subjective Logic [15]. This allows us to represent assessment as performed in a 'semi-closed world': we base ourselves on the evidence at our disposal, but our sample is limited. The smaller our sample, the more the resulting reputation will be close to the neutral prior 0.5, since no prior knowledge is available to believe the source is fully trustworthy or untrustworthy. The larger our sample, the more the weight of the sample ratio will count on the reputation estimation. On the basis of the reputation assessment, we establish the corresponding order on $\mathcal{S}$:

Definition 1 (Reputation). *For any* $S, S' \in \mathcal{S}, S \leq_R S' \leftrightarrow r(S) \geq r(S')$

A second-order relation $\leq_P$ over sources $\mathcal{S} \times \mathcal{S}$ is defined: intuitively, $S \leq_P S'$ means source S has at least the same popularity as S', where popularity reflects the number of sources which refer to S. We denote the referenced sources as *outbound_links* and the referencing sources as *inbound_links*; non-referenced or non-referencing sources are denoted as *missing_links*. Note that $\forall S, S'$, if $S \in outbound\_links(S')$ and $S' \in outbound\_links(S)$, we can assume both sources have explicit knowledge of each other's information. We assume this fact and express that S' reads from S (or alternatively that S writes to S') as $S' \in outbound\_links(S)$. Note that in the calculus presented in Fig.1 these access operations are explicit. By our definition of reputation, we can assume that for every source S referenced by S', $w(S)_{S'} \in r(S)$. Hence, the popularity of S is

$$p(S) = \frac{|inbound\_links| + 1}{|inbound\_links| + |missing\_links| + 2}$$

On its basis, we establish the corresponding order on $\mathcal{S}$:

Definition 2 (Popularity). *For any* $S, S' \in \mathcal{S}, S \leq_P S' \leftrightarrow p(S) \geq p(S')$.

Finally, we define a third order relation $\leq_K$ over sources $\mathcal{S} \times \mathcal{S}$: intuitively, $S \leq_K S'$ means that source S has at least the same knowledgeability as S', where knowledgeability reflects the number of sources to which S refers. For simplicity, given the definition of $p(S)$ based on $r(S)$, knowledgeability $k(S)$ is the inverse of $p(S)$, computed as

$$k(S) = \frac{|outbound\_links| + 1}{|outbound\_links| + |missing\_links| + 2}$$

On its basis, we establish the corresponding order on $\mathcal{S}$:

Definition 3 (Knowledgeability). *For any* $S, S' \in \mathcal{S}, S \leq_K S' \leftrightarrow k(S) \geq k(S')$.

The highest value of knowledgeability corresponds to the totality of the available sources. For simplicity, we include in this count the source itself:

Definition 4 (Source Completeness). *A source* S *satisfies source completeness if* $|outbound\_links| = |\mathcal{S}|$.

The three dimensions of reputation, popularity, and knowledgeability establish a generic computable metric on the trustworthiness of a source S:

Definition 5 (Source Trustworthiness). *Source trustworthiness is computed*

$$t(S) = \Phi(\phi(r(S)), \psi(p(S)), \xi(k(S)))$$

with Φ *a given function and* ϕ, ψ, ξ *appropriate weights on the parameters.*

The choice of ϕ, ψ, ξ is essentially contextual, as it determines the role that each parameter has in the computed value of $t(s)$, e.g. to stress knowledgeability as more important than popularity, or reputation as more relevant than knowledgeability. Fixing these parameters to 1 provides the basic evaluation with all equipollent values. Φ can be interpreted e.g. as $\sum X$, $\overline{X}$, $max(X)$: again, this choice can be contextually determined.

To distinguish between different semantic strategies for information conflict resolution, we first weight the notion of source trustworthiness with respect to source order and calculate an average value.

Definition 6 (Sources with Higher Trustworthiness). *Let* $\mathcal{S}^{\sim}_{<_t S}$ *denote the set of sources with higher trustworthiness* $<_t$ *than a given source* $S \in \mathcal{S}$.

We now partition this set as follows: we denote with $\mathcal{T}$ the subset of $\mathcal{S}^{\sim}_{<_t S}$ such that $\forall S' \in \mathcal{T}$, S' trusts information ϕ; we denote with $\mathcal{T}_{\perp}$ the complement of $\mathcal{T}$.

Definition 7 (Weighted Trustworthiness). *Average trustworthiness of* $\mathcal{T}$ *is*

$$t(\mathcal{T}) = \frac{\sum_{\forall S' \in \mathcal{T}}^{|\mathcal{T}|} t(S')}{|\mathcal{T}|}$$

Let $t(\mathcal{T}_{\perp})$ *denote the average trustworthiness for the complement partition. If* $t(\mathcal{T}) > t(\mathcal{T}_{\perp})$*, then* S *trusts* ϕ*, else* S *trusts* $\neg\phi$.

In the case of weighted trustworthiness there is a possible parity outcome: either the selection of a different strategy (e.g., the simpler majority trustworthiness) or a random assignment is possible. Finally, on the basis of the trustworthiness assessment, we establish the corresponding order on $\mathcal{S}$:

Definition 8 (Trustworthiness). *For any* $S, S' \in \mathcal{S}, S \leq_t S' \leftrightarrow t(S) \geq t(S')$.

Note that the general definition allows for a partial order, as it is possible that the trustworthiness values of two distinct sources be equivalent or incomparable. The following resolution strategies assume that a strict order is being obtained.

3 Trustworthiness Selection Strategies

We define several strategies to implement negative trust based on the Trustworthiness relation defined in Section 2. Recall that distrust requires an agent to reject incoming contradictory information in favor of currently held data. In this context, we establish such a choice on the basis of higher trustworthiness.

Definition 9 (Distrust). *Assume $S <_t S'$, $S \in outbound\_links(S')$. If S' trusts ϕ and ϕ is inconsistent with the profile of S, then S distrust ϕ and trusts $\neg\phi$.*

With this protocol in place, a source with a higher trustworthiness will always reject incoming contradictory information from a lower ranked source. It is also fair to assume that where $t(S) = t(S')$, a conservative source S will not change its current information. The process of modifying currently held information to accommodate for newly incoming one (mistrust) starts therefore on the assumption that the source of incoming information has lower trustworthiness degree than the receiver. On this basis, implementing a mistrust strategy has a complex dynamic: the user can be more or less inclined to a belief change and it can require more or less evidence for it to happen. Therefore, different strategies can be designed. One strategy requires that a *majority* of agents with higher trustworthiness agree on the new incoming data. A stronger strategy requires that the *totality* of agents with higher trustworthiness agree. Reaching the desired number of agents to implement a mistrust strategy might be a dynamic process resulting from a temporally extended analysis of the set of sources. We design the different strategies assuming Definition 6 of the subset $\mathcal{S}^{\sim}_{<_t S}$ of sources with higher trustworthiness as the sources which the receiver S has to consider.

The weakest strategy is defined by an agent which allows for a mistrust operation based on the presence of *at least one* source with higher reputation that contradicts her current belief state:

Definition 10 (Weak Trustworthiness). *If $\exists S' \in \mathcal{S}^{\sim}_{<_t S}$ such that S' trusts information ϕ, then S trusts ϕ.*

To accommodate a contradicting ϕ, the source S has to modify the current set of belief, Γ, to some subset Γ' which can be consistently extended with ϕ, i.e. removing any formula implying $\neg\phi$. A stronger strategy is for the agent to accept the content on which the majority of sources with higher trustworthiness agree:

Definition 11 (Majority Trustworthiness). *Assume $\mathcal{T} \subseteq \mathcal{S}^{\sim}_{<_t S}$ such that $\forall S' \in \mathcal{T}$, S' trusts information ϕ. We denote with $\mathcal{T}_{\perp}$ the complement of $\mathcal{T}$. If $|\mathcal{T}| > |\mathcal{T}_{\perp}|$, then S trusts ϕ, else S trusts $\neg\phi$.*

In the case of a parity outcome, either the selection of a different strategy or a random assignment are possible. Note that the above strategy does not account for the order *within* the subset $\mathcal{S}^{\sim}_{<_t S}$: it only partitions it according to the truth value of a formula and then selects the partition with higher cardinality. A more refined majority strategy will weight each member $S' \in \mathcal{T}$ and $\mathcal{T}_{\perp}$ on the basis of their trustworthiness value $t(S')$. Then an average value will be assigned to the corresponding partition and the strategy will select the formula held by the partition with a higher value. If the cardinality of the partition has to be considered, the sum of the trustworthiness values of the sources can be assigned to each partition. The strongest strategy requires the agent to change her mind if all other agents with higher trustworthiness agree:

Definition 12 (Complete Trustworthiness). *If $\forall S' \in \mathcal{S}^{\sim}_{<_t S}$, S' trusts information ϕ, then S trusts ϕ.*

The Majority and Complete Trustworthiness strategies above have a strong effect on knowledge diffusion in the presence of full communication. The Consensus rule below holds even if the content from the most trustworthy source is not initially held by the majority of agents.

Proposition 1 (Consensus). *Assume $S' \in outbound\_links(S)$ holds $\forall S < S' \in \mathcal{S}^{\sim}$. Then $\mathcal{S}$ converges towards consensus on the information trusted by the most trustworthy source.*

4 Rule-based Semantics for the Strategies

The natural deduction calculus (`un`)`SecureND` [20] defines trust, mistrust and distrust protocols according to the informal semantics described in Section 1. It formalizes a derivability relation on formulas from sets of assumptions (contexts) as accessibility on resources issued by sources. In this section, we provide an extension of the calculus with a rule-based implementation of the trustworthiness selection strategies from Section 3.

Definition 13 (Syntax of (`un`)`SecureND`).

$$\begin{array}{l}
\mathcal{S}^{\sim} := \{A <_t B <_t \cdots <_t N\} \\
BF^S := a^S \mid \phi_1^S \rightarrow \phi_2^S \mid \phi_1^S \wedge \phi_2^S \mid \phi_1^S \vee \phi_2^S \mid \perp \\
mode := Read(BF^S) \mid Write(BF^S) \mid Trust(BF^S) \\
RES^S := BF^S \mid mode \mid \neg RES^S \\
\Gamma^S := \{\phi_1^S, \ldots, \phi_n^S\}
\end{array}$$

Every $S \in \mathcal{S}$ is a content producer which has a trustworthiness value based on its interactions with any other $S' \in \mathcal{S}$. Any $S \in \mathcal{S}$ is ordered with respect to the others by the trustworthiness order.[3] Formulas in the set BF^S express content produced by source S and they are closed under logical connectives. Functions on contents in the set *mode* refer to reading, writing and trusting formulas. Every source S is identified by the set of contents it produces, denoted by Γ^S called the profile of S. A formula expresses access from a source S to content issued by another source S' (metavariables S, S' are substituted by variables A, B):

[3] In other versions of this logic, the order between elements in $\mathcal{S}$ is differently defined, e.g. imposed by access policies, see e.g. [23,20,22].

Definition 14. *An* (un)`SecureND`*-formula* $\Gamma^A \vdash RES^B$ *says that under the content expressed by source A, some content from source B is validly accessed.*

The rule-based semantics of the calculus is given in Fig. 1. *Atom* establishes derivability of formulas from well-formed contexts and under consistency preserving extensions. We use the judgment $\Gamma : profile$ for a profile consistently construed by induction from the empty set. For brevity, we skip here the introduction and elimination rules for logical connectives, see [20] and focus only on the access rules. Differently from other versions of the same calculus, we drop here negation-completeness: a source without access to a content item from another source, will not assume access to its negation, i.e. uncertainty is admissible. *read* says that from any well-formed source profile A, formulas from a profile B can be read. *trust* says that if a content item is read and it preserves consistency when added to the reading profile, then it can be trusted. *write* says that a readable and trustable content can be written. By *distrust*, source A distrusts content ϕ^B if it induces contradiction when reading from Γ^A and A has higher trustworthiness than B. Its elimination uses $\rightarrow$-introduction to induce *write* from the receiver profile for any content that follows a distrust operation. This allows $Write(\neg\phi^B)$ when $\neg Trust(\phi^B)$ holds. Each of the *mistrust* rules applies one different strategy from Section 3 for a content item ϕ^B inducing contradiction when reading from Γ^A and A has lower trustworthiness than B. By *weak mistrust*, A accepts ϕ (and removes from its own profile any conflicting information) by the simple presence of B in the set of sources with a higher reputation of A: this formulation is general enough to accommodate for the substitution of B in this condition by any other source that A considers absolutely essential (appeal to authority). *majority mistrust* requires computing the partitions of the set of sources with higher trustworthiness than A and comparing their cardinality: *any* content ϕ held by the larger partition will be kept by A (even when this reduces to an application of a *distrust* rule). In *weighted majority*, the condition is expressed by the higher average reputation of the partition. By *complete mistrust* the source A requires that every element in the set of sources with higher reputation agrees on ϕ. By the rule *write*, every trusted content can be written.

5 Evaluation

5.1 Use Case Description

In 2015, a measles outbreak took place in Disneyland, California. This event received much attention online, and a quite strongly polarised discussion followed up the news regarding this event. Public authorities and pro-vaccination sources pointed out the importance of vaccination, and some of them blamed the low vaccination rate as the main reason for this outbreak. On the other hand, the anti-vaccination movement accused the government agencies and the pro-vaccination movement of misinforming the public, since the children involved in the outbreak were vaccinated. Two main factions are at work, the pro and the anti vaccinations. While sources do not always identify themselves as part of one

Fig. 1. The System `(un)SecureND`: Access Rules.

$$\frac{\Gamma^A : profile \qquad \Gamma^A;\Gamma^B : profile}{\Gamma^A;\Gamma^B \vdash \phi^B} \text{Atom, for any } \phi \in \Gamma^B$$

$$\frac{}{\Gamma^A \vdash Read(\phi^B)}\, read \qquad \frac{\Gamma^A \vdash Read(\phi^B) \qquad \Gamma^A;\phi_i^B : profile}{\Gamma^A \vdash Trust(\phi_i^B)}\, trust$$

$$\frac{\Gamma^A \vdash Read(\phi^B) \qquad \Gamma^A;\phi^B \vdash \bot \qquad A <_t B}{\Gamma^A \vdash \neg Trust(\phi^B)}\, distrust$$

$$\frac{\Gamma^A \vdash Read(\phi^B) \qquad \Gamma^A;\phi^B \vdash \bot \qquad \Delta^{B<_t A} \vdash \phi}{\Gamma'^A \vdash Trust(\phi^B)}\, weak\ mistrust, \text{ for some } \Gamma^A \supset \Gamma'^A;\phi^B \vdash wf$$

$$\frac{\Gamma^A \vdash Read(\phi^B) \qquad \Gamma^A;\phi^B \vdash \bot \qquad \Delta^{\mathcal{T}} \vdash \phi}{\Gamma'^A \vdash Trust(\phi^B)}\, majority\ mistrust, \text{ for some } \Gamma^A \supset \Gamma'^A;\phi^B \vdash wf$$

with $\mathcal{T} \subset \mathcal{S}^{\sim}_{<_t A}$ s.t. $|\mathcal{T}| > |\mathcal{T}_\bot|$.

$$\frac{\Gamma^A \vdash Read(\phi^B) \qquad \Gamma^A;\phi^B \vdash \bot \qquad \Delta^{\mathcal{T}} \vdash \phi}{\Gamma'^A \vdash Trust(\phi^B)}\, weighted\ mistrust, \text{ for some } \Gamma^A \supset \Gamma'^A;\phi^B \vdash wf$$

with $\mathcal{T} \subset \mathcal{S}^{\sim}_{<_t A}$ s.t. $t(\mathcal{T}) > t(\mathcal{T}_\bot)$.

$$\frac{\Gamma^A \vdash Read(\phi^B) \qquad \Gamma^A;\phi^B \vdash \bot \qquad \Delta^{\mathcal{S}^{\sim}_{<_t A}} \vdash \phi}{\Gamma'^A \vdash Trust(\phi^B)}\, complete\ mistrust, \text{ for some } \Gamma^A \supset \Gamma'^A;\phi^B \vdash wf$$

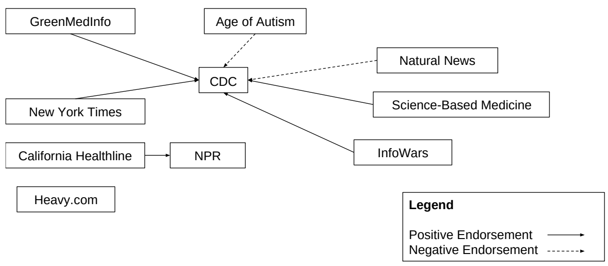
$$\frac{\Gamma^A \vdash Read(\phi^B) \qquad \Gamma^A \vdash Trust(\phi^B)}{\Gamma^A \vdash Write(\phi^B)}\, write$$

or the other, for many of them it is either clear what their stance is (e.g., when they explicitly ‘attack’ each other), or we can make safe assumptions based on our background knowledge (e.g., by assuming that authorities are pro vaccinations). We have at our disposal a set of assessments of these articles collected by means of user studies involving experts [6]. These assessments cover quality dimensions like accuracy and prediction, and present an overall quality score that is equivalent to the trustworthiness score defined here.

5.2 Data Preprocessing

We select a subset of 10 articles regarding this debate from a corpus of documents regarding the Disneyland measles outbreak[4]. The selection gives a small but diverse set of views on the topic in terms of stance (pro or anti vaccinations) and type of document (news article, official document, blog post, etc.). Provided they all discuss the specific event selected, a clear network of references emerges. However, such a network is rather sparse since a large majority of these sources do not cite each other. As we are interested in capturing their polarity to compute the three trustworthiness dimensions, we reconstruct the network as follows: (1) a source criticizing another source is considered as a negative piece of evidence regarding the reputation of the source mentioned; and (2) a source citing data from another source, even in neutral terms, is considered a piece of evidence regarding the popularity of the source cited. The resulting network of references is represented in Fig. 2 and it illustrates only the relations emerging from the corpus considered, representing a partial view on the real scenario because we derive a source's trustworthiness using one or more documents published by it as a proxy; the more documents we observe from a source, the better we can assess its trustworthiness value. For example, we estimate the source knowledgeability from the number of citations of other sources. Some sources could be cited only in some articles by the source under consideration. Also, we derive a source's trustworthiness based on the references it receives from the other sources considered, but we know that the set of sources is limited, and the scenario might change when considering other sources (e.g., the number of citations of currently poorly cited sources could rise). Given these considerations, the smoothing factor added to Definitions 1, 2, and 3, helps to cope with the resulting uncertainty.

Fig. 2. Network of references resulting from the preprocessing of our corpus. Directed arrows indicate positive (continuous line) or negative (dotted line) references.



[4] The dataset is available online at `https://goo.gl/aouDJH`.

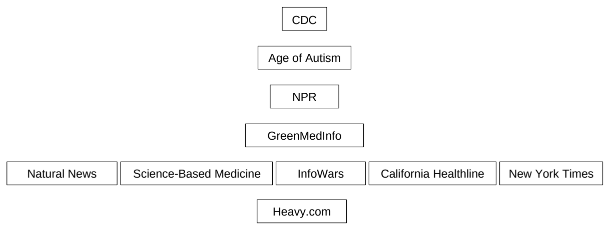
5.3 Sources Ordering

Based on the network depicted in Fig. 2, and using the formulas presented in Section 2, we compute the trustworthiness score for each of the sources in our sample. The trustworthiness score is computed by averaging the reputation, the knowledgeability, and the popularity of the sources, resulting in the scores reported in Table 1. Fig. 3 shows a graphical representation of the resulting hierarchy of sources. Since the trustworthiness thus obtained shows a weak correlation (0.2) with the overall scores provided by the users in the user study, we explore alternative ways to aggregate the scores.

Table 1. Trustworthiness scores of the sources considered for our use case. The score is computed by means of a simple average, where each component has the same weight.

Source	**Reputation**	**Knowledgeability**	**Popularity**	**Trustworthiness**
California Healthline	0.50	0.17	0.08	0.25
CDC	0.63	0.08	0.67	0.46
NYTimes	0.50	0.17	0.08	0.25
InfoWars	0.50	0.17	0.08	0.25
GreenMedInfo	0.50	0.25	0.08	0.28
Age of Autism	0.67	0.17	0.17	0.33
Science-Based Medicine	0.50	0.17	0.08	0.25
Heavy.com	0.50	0.08	0.08	0.22
Natural News	0.50	0.17	0.08	0.25
NPR	0.67	0.08	0.17	0.31

Fig. 3. Hierarchical ordering of the sources derived from the scores shown in Table 1

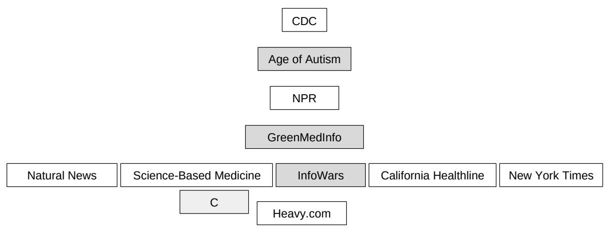


Weighted Trustworthiness Applying weights to the trustworthiness parameters can yield a different hierarchy. Instead of applying an arbitrary weighing to the scores, we apply linear regression on the parameters, targeting the overall quality scores provided by the users in the study. Once we learn the weights for the parameters, we compute the trustworthiness scores. The resulting scores show a 0.6 correlation with those provided by the users. Moreover, we also run 3-fold cross-validation (split the dataset into 3 parts and, in round, use two parts as a training set for linear regression, and one for validation). For one item only, our model is unable to make a prediction. Excluding such item, the resulting average correlation between predicted and user-provided overall quality is -0.87 (Pearson) and -0.76 (Spearman). We consider these as promising results.

5.4 Applying Trustworthiness Selection Strategies

Here we illustrate how users could apply the selection strategies described in Section 3. Fig. 4 shows the scenario where the trustworthiness selection strategies are applied. The sources analyzed in the previous step are now shown in white if they present a positive stance with respect to vaccinations, in grey otherwise. C is a new source with an unclear stance that joins the scenario. The stance of C (i.e., whether C *trusts* vaccines or not) will be determined by comparison with the other sources. Assume that the trustworthiness of C is higher than that of Heavy.com, but lower than the trustworthiness of all the other sources.

Fig. 4. Use case scenario. We adopt the same hierarchy as in Fig. 3. Sources in white trust vaccinations. Sources in grey do not. C denotes an additional source which takes part in the scenario and has not yet a clear stance.



Distrust. When C is confronted with Heavy.com and its lower trustworthiness score, following the distrust rule it will *distrust* vaccines.

Weak Trustworthiness. Let us follow up on the previous scenario. C now *distrusts* vaccines. When encountering all the other sources, if the `weak mistrust` strategy is applied, C will revise its profile: now C *trusts* vaccines because of several sources with trustworthiness higher than C *trust* ϕ. Note that `weak mistrust` requires at least one source to trust ϕ in order to follow suit.

Majority Trustworthiness. In an alternative scenario, when encountering the other sources, C can evaluate whether to trust ϕ or not based on whether the majority of the sources trusts vaccines. We partition the sources based on *vaccines* and $\neg vaccines$. With any strategy for determining the majority (partition cardinality, average trustworthiness of the sources in the two partitions, sum of the cardinalities in the two partitions), *trust* in vaccines prevails.

Complete Trustworthiness. When complete trustworthiness is applied, C needs all the sources to agree on vaccines to add it to its profile. Since three sources disagree, by applying this rule, we obtain that C *d*istrusts vaccines.

6 Discussion

The goal of our model is to provide means to mimic human thinking and provide a tool to systematically reason upon sources. The result of such reasoning is a relative reference system of sources. When oracles, fact-checkers, and other sources are available, such a reference system can be turned into an absolute one: if the user knows that a given set of statements is true or false, she can reason about the trustworthiness of the sources incorporating this additional information in the networks. When oracles are not available, the reference system can provide the user with a basis to coherently reason upon the sources she observes.

Frameworks like PageRank and its successors can be considered more evolved and successful alternatives to the present proposal. While PageRank can be applied to one or more networks to rank their sources, our system considers three distinct networks, aggregates them, and can be either extended with other networks or be used as reasoning support as it is. Hence we consider the present a viable complement to existing approaches.

While assessing the veracity of information is not the focal point of our system, the multidimensional approach we take shows promising robustness to possible attacks. Suppose that in an echo-chamber, sources cite each other positively in order to increase their own reputation and popularity. If their citations are limited to the sources in the echo chamber, their knowledgeability (and, thus, their trustworthiness) will necessarily be low. If to remedy this sources start citing others outside the echo chamber, their knowledgeability will rise, but they will also contribute to the popularity of these external sources. Still, vulnerability to the knowledgeability score is possible in sufficiently large echo chambers. Future developments will tackle this aspect more explicitly.

7 Related Work

Assessing the quality of information sources is a long-standing problem largely addressed in the fields of humanities, where specific guidelines and checklists have been proposed to address the issue of "source criticism" [3]. Such work has also been extended to Web sources in [6,7], where a combination of crowdsourcing and machine learning is adopted. Those works are complementary to the present contribution since they do not compare directly the references among sources. Counting links for a source as employed in this paper aims at mimicking the evaluation of the bibliography mentioned in the source criticism checklist. Another framework based on crowdsourcing is presented in [17].

Using fitness for purpose to assess information quality is a widely adopted strategy, see [12,13]. In the present work, we start from the assumption that where it is unclear or impossible for an agent to distinguish between contradictory data, source assessment based on trustworthiness is a valuable strategy. We show how such a protocol can be implemented through different selection strategies. A related topic is the one of fake news, tackled for instance in [25,4].

Research on trust in computational domains has been extensive in the last decades. Crucial aspects of the behavior of trust concern properties like propagation and blocking [8,10,14,16]. Solutions to these problems are various [2,9,11]. In the present work, we evaluate trust in information sources not on an absolute scale, but rather with varying degrees. A related approach is presented in [19], where a trust measure on agents is combined with the use of argumentation for reasoning about beliefs. Similarly, we propose a trust evaluation of sources to decide which information to maintain. The logic used in this work originates from a model designed to model trust in resource access control scenario, and to be able to block trust transitivity by design [23,21]. The logic has been applied to the Minimally Trusted Install Problem software management in [5], its negative counterpart [22], and tested to investigate optimal strategies to minimize false information diffusion [24]. For other accounts of negative trust, see [1,18].

8 Conclusion

In this paper, we presented an extension of (`un`)`SecureND`, a logic modeling trust on information, with strategies for assessing the trustworthiness of sources as a function (average or otherwise) of their knowledgeability, popularity, and reputation, possibly weighted. We evaluated this extension on a real-life case study on the trustworthiness of Web sources and applied the selection strategies to the resulting source hierarchy. We showed that a linear combination of these parameters presents a decent correlation with user-provided assessments.

We plan to extend this work in two main directions. First, we will work on the automation of the preprocessing phase. We expect to use natural language processing for this and, in particular, author attribution to systematically identify references among the sources, and textual entailment to capture the perspectives taken by the different sources. Second, we will improve the parameters considered for assessing the trustworthiness. For instance, knowledgeability will have to be assessed based on the estimated level of the truthfulness of the statements made by the source. We plan to run an exhaustive user study to guide the design of source trustworthiness assessment and selection. Lastly, we will experiment with network centrality measures as alternative indicators for these parameters.

References

1. Abdul-Rahman, A.: A framework for decentralised trust reasoning. Ph.D. thesis, Department of Computer Science, University College London (2005)

2. Abdul-Rahman, A., Hailes, S.: A distributed trust model. In: NSPW. pp. 48–60 (1997)
3. American Library Association: Evaluating information: A basic checklist. (1994)
4. Bessi, A., Coletto, M., Davidescu, G., Scala, A., Caldarelli, G., Quattrociocchi, W.: Science vs conspiracy: Collective narratives in the age of misinformation. PLoS ONE **2** (2015)
5. Boender, J., Primiero, G., Raimondi, F.: Minimizing transitive trust threats in software management systems. In: PST. pp. 191–198. IEEE (2015)
6. Ceolin, D., Noordegraaf, J., Aroyo, L.: Capturing the ineffable: Collecting, analysing, and automating web document quality assessments. In: EKAW. pp. 83–97 (2016)
7. Ceolin, D., Noordegraaf, J., Aroyo, L., van Son, C.: Towards web documents quality assessment for digital humanities scholars. In: WebSci 2016. pp. 315–317 (2016)
8. Chakraborty, P.S., Karform, S.: Designing Trust Propagation Algorithms based on Simple Multiplicative Strategy for Social Networks. Procedia Technology **6**(0), 534–539 (2012), iCCCS-2012
9. Chapin, P.C., Skalka, C., Wang, X.S.: Authorization in trust management: Features and foundations. ACM Comput. Surv. **40**(3) (2008)
10. Christianson, B., Harbison, W.S.: Why Isn't Trust Transitive? In: SPW. vol. 1189, pp. 171–176. Springer (1996)
11. Clarke, S., Christianson, B., Xiao, H.: Trust*: Using Local Guarantees to Extend the Reach of Trust. In: SPW. vol. 7028, pp. 171–178. Springer (2009)
12. Floridi, L., Phyllis (eds.): The Philosophy of Information Quality. Springer (2014)
13. Illari, P.: IQ: Purpose and Dimensions, pp. 281–301. Springer (2014)
14. Jamali, M., Ester, M.: A Matrix Factorization Technique with Trust Propagation for Recommendation in Social Networks. In: RecSys. pp. 135–142. ACM (2010)
15. Jøsang, A.: Subjective Logic - A Formalism for Reasoning Under Uncertainty. Springer (2016)
16. Jøsang, A., Marsh, S., Pope, S.: Exploring Different Types of Trust Propagation. In: iTrust, vol. 3986, pp. 179–192. Springer (2006)
17. Lee, Y.W., Strong, D.M., Kahn, B.K., Wang, R.Y.: Aimq: A methodology for information quality assessment. Inf. Manage. **40**(2), 133–146 (2002)
18. Marsh, S., Dibben, M.R.: Trust, Untrust, Distrust and Mistrust – An Exploration of the Dark(er) Side. In: iTrust. vol. 3477, pp. 17–33. Springer (2005)
19. Parsons, S., Tang, Y., Sklar, E., McBurney, P., Cai, K.: Argumentation-based reasoning in agents with varying degrees of trust. In: AAMAS. pp. 879–886 (2011)
20. Primiero, G.: A calculus for distrust and mistrust. In: IFIPTM. vol. 473, pp. 183–190. Springer (2016)
21. Primiero, G., Boender, J.: Managing software uninstall with negative trust. In: IFIPTM. vol. 505, pp. 79–93. Springer (2017)
22. Primiero, G., Boender, J.: Negative trust for conflict resolution in software management. Web Intelligence **16**(4), 251–271 (2018)
23. Primiero, G., Raimondi, F.: A typed natural deduction calculus to reason about secure trust. In: PST. pp. 379–382. IEEE (2014)
24. Primiero, G., Raimondi, F., Bottone, M., Tagliabue, J.: Trust and distrust in contradictory information transmission. Applied Network Science **2**, 12 (2017)
25. Zhang, A.X., Ranganathan, A., Metz, S.E., Appling, S., Sehat, C.M., Gilmore, N., Adams, N.B., Vincent, E., Lee, J., Robbins, M., Bice, E., Hawke, S., Karger, D., Xiao Mina, A.: A structured response to misinformation: Defining and annotating credibility indicators in news articles. In: WWW 18 Companion (2018)