Voting-based Methods for Evaluating Sources and Facts Reliability

Quentin Elsaesser
CRIL, CNRS, Université d'Artois
elsaesser@cril.fr

Patricia Everaere
Univ. Lille, CNRS, CRIStAL
patricia.everaere-caillier@univ-lille.fr

Sébastien Konieczny
CRIL, CNRS, Université d'Artois
konieczny@cril.fr

***Abstract*—In this work we propose a family of methods that allow to conjointly compute the reliability of a set of information sources and the confidence of the facts on a set of objects, by confronting the sources points of view. We use a (scoring-based) voting method for the evaluation of the trust of the sources, using Condorcet's Jury Theorem arguments in order to identify the truth and the reliable sources. We discuss general theoretical properties that such operators should satisfy, and we study what are the properties satisfied by our methods. We provide an experimental study that shows that we perform better than state of the art methods on the task of finding the truth among the possible facts. We show that we can also adequately evaluate the reliability of the sources of information.**

***Index Terms*—Reliability, Truth Tracking, Voting, Trust**

I. Introduction

There are numerous applications where one receives (typically conflicting) pieces of information from different sources and have to form an opinion from these pieces of information. In this situation, a standard way to solve conflicts is to believe the most reliable/trustworthy sources. We propose such an evidence-based definition of reliability (truthfulness) from available evidences.This can be useful to evaluate the reliability of an agent in a multi-agent system or social network, a source on the web, any journal/media, etc.

More precisely, we consider a set of independent sources that provide us information about different questions. Our goal is to evaluate both the reliability of the sources, and the confidence of the facts, which then allows us to find the correct answers (facts) to the different questions (objects). There are previous works that start from the same structure (sources/facts/objects), but their main objective is to find the correct answers [1], [2].

In order to find this true information, we rely on the idea of Condorcet's Jury Theorem [3], which states that it is more likely that the majority of the individuals will choose the correct solution. The intuition is as follows: suppose that among 10 sources of equal reliability, 8 tell you that the *Capital of Australia* is *Canberra*, and 2 tell you that it is *Sydney*. Following what the majority says is the safest way to find the truth. Condorcet's Jury Theorem requires a lot of hypotheses (that all the sources are equally reliable, that they are all reliable (i.e. they have more than 50% chance of finding the truth), that they are independent, and that the choice is between only 2 possibilities). However, all these hypotheses can be more or less relaxed [4]–[7]. This argument is also close to the ones that use "The wisdom of crowds" [8].

This work has benefited from the support of the AI Chair BE4musIA of the French ANR (ANR-20-CHIA-0028).

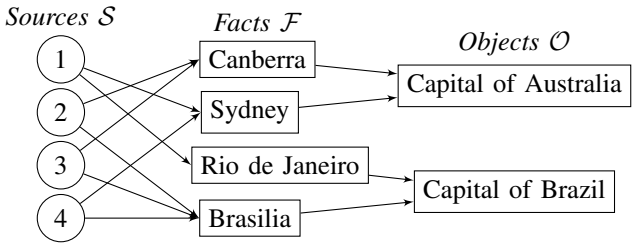

Fig. 1. Sources, Facts & Objects

In this work, we suppose that initially we have no information about the reliability of the sources, and we define an iterative procedure to determine their reliability. At the beginning, we assign the same reliability to all the sources, then we compare the answers to the different questions, and we use this "Condorcet's Jury Theorem" argument to reward the sources that provide information (facts) that are confirmed by others, and are therefore more likely to be true. Then we iterate the process with these adjusted reliabilities of the sources until convergence is reached.

To illustrate this process, consider the example of Figure 1, where four sources give information about two objects: *Capital of Brazil* and *Capital of Australia*. Note that there is initially a tie for *Capital of Australia*, with two sources giving *Canberra* and two sources giving *Sydney*. But we can use the other object. There is a majority for *Brasilia*, so *Brasilia* is considered as the good fact, and the sources giving this fact are favored over those giving *Rio de Janeiro*. And, in the next iteration, we will be able to break the tie on *Capital of Australia* since more reliable sources give us *Canberra*.

More precisely, at each iteration, sources provide some strength to the facts they claim on the different objects. This strength is the (current) reliability of the source. Thus, each object can rank the corresponding facts (possible answers) from most reliable to least reliable, just using the sum of obtained strengths. We use scoring-based voting rules in order to associate a number to each rank of facts. The simplest one is the plurality rule, where only the most reliable facts provide a score of 1 to the corresponding sources, and all the others get nothing (0). The new reliability of each source is computed

by combining all these scores. We use two normalizations for this step: one that favors sources that provide many claims, and one that favors sources that are more careful and do not fail often. Then a new iteration starts with the updated reliability of each source.

In the following, after presenting our S&F (for *Sources & Facts*) methods, we discuss logical properties for characterizing interesting methods that aim to evaluate the reliability of sources and facts. We review properties that have been proposed by [2], and discuss why some of them are not appropriate for this setting. We also propose new properties required for all methods, and some properties that characterize interesting subclasses. Then we check which properties are satisfied by our methods.

Beside this formal evaluation, we also provide experimental ones. The idea is to test if we can achieve this goal of evaluating the reliability of sources and facts in practice. There are not many real benchmarks that can be used for this task, but we test our methods on two such benchmarks. Then we also test our methods on generated benchmarks, which allows us to evaluate more parameters.

And the results are good. We show that for the tasks related to finding the true facts we are better than existing methods. But, contrariwise to the existing methods, we can also give a good evaluation of the reliability of the sources.

II. Preliminaries

We consider three sets $\mathcal{S}$, $\mathcal{F}$ and $\mathcal{O}$ respectively called *Sources*, *Facts* and *Objects*. *Sources* represent the (human or artificial) agents that provide the information. *Objects* are the questions on which we would like to have information about, and the *Facts* are the possible answers. Relatively to each object, facts are distinct and exclusive: only one fact can be claimed by each source per object.

So these objects+facts can be seen also as questions+answers or as variables+values. This is mainly a question of vocabulary here. And we stick to the one used in previous works [1], [2].

Definition 1. *Let $G = (V, E)$ be a directed graph with $V = \mathcal{S} \cup \mathcal{F} \cup \mathcal{O}$ and $E \subseteq (\mathcal{S} \times \mathcal{F}) \cup (\mathcal{F} \times \mathcal{O})$, such that:*

- *For each fact $f \in \mathcal{F}$ there is a unique object $o \in \mathcal{O}$ with $(f, o) \in E$.*
- *A source $s \in \mathcal{S}$ can claim at most one fact per object $o \in \mathcal{O}$ (i.e. $\forall s \in \mathcal{S}$ there are no $f_1 \in \mathcal{F}$, $f_2 \in \mathcal{F}$ s.t. $\{(f_1, o), (f_2, o), (s, f_1), (s, f_2)\} \subseteq E$).*

$(s, f) \in E$ means that the source s claims that the fact f is the correct answer for its corresponding object. It is possible that a fact is not claimed by any source.

For more clarity, we will use these notations: $src(f) = \{s \in \mathcal{S} : (s, f) \in E\}$, $fct(s) = \{f \in \mathcal{F} : (s, f) \in E\}$, $fct(o) = \{f \in \mathcal{F} : (f, o) \in E\}$, $obj(f) = \{o \in \mathcal{O} : (f, o) \in E\}$.

When it is not obvious (e.g. when we have more than one graph), we can specify the graph, and we write, for the graph G, $src_G(f)$, $fct_G(s)$, $fct_G(o)$ and $obj_G(f)$ instead of the above notations.

We denote $r_G(s) \in [0, 1]$ the evaluation of the reliability of a source s in the graph G and $c_G(f) \in \mathbb{R}^+$ the evaluation of the confidence of a fact f in the graph G.[1]

III. Related Work

There are some *Truth Discovery* algorithms in the literature that aim to identify the true facts.

Truth Finder [1] is an iterative algorithm, that updates the score of sources and facts at each iteration. This work focus on the confidence of the facts to find the truth. With *Truth Finder*, the reliability of a source is the average confidence in the facts provided by that source. For the confidence of a fact, the authors assume that the facts can support each other, in which case the confidence increases or decreases if the facts contradict each other. This part is beyond the scope of our basic setting (see [9] for a discussion of truth discovery methods, where there are other parameters that can be taken into account for computing the reliability).

Hubs and Authorities [10] method, defines to rank web pages, can also be used in this setting. It is also an iterative method, that defines two different scores for a page. *Hub* (which we can identify to sources) favors pages that point to many other pages, and *authority* (which we can identify to facts) favors pages that are pointed to by many different hubs.

Sums [11] is based on *Hubs and Authorities*. The main difference is the way in which the reliability of sources and facts is normalized. For *Sums*, the reliability of a source is the sum of the confidence of the facts it claims ($r(s) = \Sigma_{f \in fct(s)} c(f)$), and the confidence of a fact f is obtained as the sum of the reliability of the sources that claim it ($c(f) = \Sigma_{s \in src(f)} r(s)$). These results are then normalized by the maximal obtained value (by $max(r(s)) \forall s \in \mathcal{S}$ for $r(s)$ and by $max(c(f)) \forall f \in \mathcal{F}$ for $c(f)$). The same authors propose three other variants of *Sums*: *Average.Log*, *Investment* and *PooledInvestment* where the initial confidence of the facts is different and where the reliability of the sources is computed with different functions.

Booth and Singleton were the first to propose an axiomatic approach to the *Truth Discovery* problem in [2]. They also propose a new method, called *Unbounded-Sums*, which is based on *Sums*, but where they do not normalize the score.

We will compare the experimental results of our methods with the results of these algorithms and with *Voting*, the naive method that chooses the fact with the most claims on each object.

IV. S&F Methods

In this section, we present our methods. One method is defined by the choice of a voting method v and a normalization function n, and is denoted vn. We use an iterative method to compute the reliability of the sources and the confidence of the facts. See Algorithm 1 for a given method vn.

First, we need to initialize the reliability of the sources. Remember that we do not have information about the sources,

[1] We simply note $r(s)$ and $c(f)$, without the subscript, when the graph is clear from the context.

Algorithm 1 S&F Algorithm for vn

Input: A graph $G = (\mathcal{S} \cup \mathcal{F} \cup \mathcal{O}, E)$

Output: The reliability of the sources $r(s)$ and the confidence of the facts $c(f)$.

$r(s) = 1 \; \forall s \in \mathcal{S}$ #initial reliability of the sources
$c(f) = 0 \; \forall f \in \mathcal{F}$ #initial confidence of the facts
ts (resp. ts^{-1}) is the vector of reliability of the sources during the current (resp. last) iteration, i.e. $ts = \langle r(s) : \; \forall s \in \mathcal{S} \rangle$.
while Euclidean_distance$(ts, ts^{-1}) > 0.001$ and
number_of_iterations < 30 **do**
    # Evaluation of the confidence of the facts
    for each $f \in \mathcal{F}$ **do**
        $c(f) = \sum_{s \in src(f)} r(s)$
    end for
    # Ranking of the facts
    for each $f \in \mathcal{F}$ **do**
        $V_{\mathsf{v}}(f) = \frac{\sum_{\{y \in T^i(o) | f \in T^i(o)\}} \mathsf{v}(\succ_o, y)}{|T^i(o)|}$
    end for
    for each $s \in \mathcal{S}$ **do**
        # Evaluation of the reliability of the sources
        $r^I(s) = \sum_{f \in fct(s)} V_{\mathsf{v}}(f)$
        # Normalization of the reliability of the sources
        $r(s) = \mathsf{n}(r^I(s))$
    end for
end while

so the initial reliability is the same for all the sources: $r(s) = 1 \; \forall s \in \mathcal{S}$. An iteration runs as follows: First, the confidence of the facts is calculated (section IV-A). Then, a voting rule is used to rank the facts and assign a score to the facts depending on their ranking (section IV-B). After that, the reliability of the sources is evaluated (section IV-C) and the last step is the normalization of the reliability of the sources (section IV-D). The algorithm stops when the process converges, i.e. when the Euclidean distance between the reliability of the sources of the last iteration and the current iteration is smaller than ϵ with $\epsilon = 0.001$ or when the number of iterations is 30. It is important to note that the maximal number of iterations we obtained during our experiments is 14, and in average the convergence is obtained around 4 iterations.

A. Confidence of the facts

The confidence in a fact f is simply computed by adding the reliability of the sources that claimed it. The more reliable the sources that affirm it, the more confidence there will be in this fact.

$$c(f) = \sum_{s \in src(f)} r(s) \tag{1}$$

B. Ranking of the facts

The evaluation of the reliability of the sources is obtained by summing the rewards associated with the facts claimed. To assign a reward to a fact, we use a scoring voting rule, and for the evaluation of the reliability of the sources, each fact will transmit the reward it received to its source. The reward is given according to the rank of the fact. For each object, we rank the corresponding facts from most confident to least confident, and the scoring voting rule associates a reward (number) to each rank.

But we have to make three adjustments: The first one is that scoring voting rules are defined for linear orders, whereas we obtain total pre-orders (some facts can have the same rank). So we will use the average of the scores for the facts with the same rank. The second one is that the number of options (facts) is not the same for all objects, so we have to choose how to normalize these scores on different scales. The third one (section IV-D) is that, when the scores are received by the sources, we normalize them in order to have a result in $[0, 1]$.

Definition 2. *Let M be an integer and e be a sequence of non-decreasing integers with $e_1 \geq e_2 \geq ... \geq e_M$ and such that $e_1 > e_M \geq 0$. A scoring rule v is a function that, to each linear order $>$ on a set of at most M facts and to each fact f, associates a positive integer s.t. if fact f is ranked at the ith position in the linear order $>$, then $\mathsf{v}(>, f) = e_i$.*

When $e_1 = 1$ and $e_2 = e_M = 0$, the rule is called the plurality vote. When $e_1 = M - 1, e_2 = M - 2, ..., e_M = 0$, it is the Borda rule.

For standard voting procedures, the voters vote on a fixed set of candidates, and M is the number of candidates. In our case, the objects are related to different numbers of facts. So we state that $M = max(|fct(o)|) \forall o \in \mathcal{O}$. But, we have to do a (first) normalization by the maximal score of the facts because we want to reward fairly every winners of the vote so we state: $best\_score(\mathcal{F}) = e_1$. It means that for all objects, being the most plausible fact always provide the same score (e_1), no matter how many facts are linked to the object.

Moreover, contrary to standard hypotheses for voting rules, the ranking associated with an object is not a linear order, but a total pre-order (some facts can have the same rank). We have to adjust the scoring voting rules for possible ties. In this case, we give the averaged score, i.e. the average of the scores they were supposed to receive, as in [12]. Let us formalize this step.

A total pre-order (a reflexive, transitive, total relation[2]) $\geq$ can be seen as a set of strata. An element x belongs to a stratum $T_{\geq}$ composed of a set of equivalent elements $\{y | x \simeq y\}$. And we say that $T^i_{\geq}$ is the i^{th} stratum of the pre-order if $\exists x_1, \ldots, x_{i-1}$ such that $x_1 > \ldots > x_{i-1} > y$ with $y \in T^i_{\geq}$. If there is no $x_1 > y$ with $y \in T^i_{\geq}$ then $i = 1$.

We say that a linear order $\succ$ is compatible with a pre-order $\geq$ if $\forall x \in T^i, \forall y \in T^j, \; i < j \Rightarrow x \succ y$.

Definition 3. *For each fact f, consider its corresponding object o. Let $\mathcal{P}(o)$ be the pre-order given by the confidence of the facts (i.e. $f_1 \geq_{\mathcal{P}(o)} f_2$ iff $c(f_1) \geq c(f_2)$) and m the number*

[2]From any total pre-order $\geq$ we define the corresponding strict order $>$ as $x > y$ iff $x \geq y$ and $y \not\geq x$, and the corresponding equivalence relation $\simeq$ as $x \simeq y$ iff $x \geq y$ and $y \geq x$.

of strata in $\mathcal{P}(o)$. We have $\mathcal{P}(o) = \{T^1(o), T^2(o),...,T^m(o)\}$, where $T^k(o)$ is the k^{th} stratum in $\mathcal{P}(o)$. Then the score given to f for the pre-order $\mathcal{P}(o)$ and the scoring rule v is defined as (where $\succ_o$ is any linear order compatible with $\mathcal{P}(o)$ and $\mathsf{v}(\succ_o, f)$ is the score that the fact f gets according to the scoring rule v):

$$V_{\mathsf{v}}(f) = \frac{\sum_{\{y \in T^i(o) | f \in T^i(o)\}} \mathsf{v}(\succ_o, y)}{|T^i(o)|}$$

Example 1. *Let G be a graph with two objects. Let v be the Borda rule. The first object $o1$ and the second object $o2$ have respectively 6 and 9 facts linked to them. We have $best\_score(\mathcal{F}) = 9$. Suppose $\mathcal{P}(o1) = \{T^1(o1) = \{f_1, f_2\}, T^2(o1) = \{f_3\}, T^3(o1) = \{f_4, f_5, f_6\}\}$, i.e. $o1$ ranked 2 facts first (in the first stratum), 1 fact second and 3 facts third. So the score of f_1 and f_2 is $V_{Borda} = \frac{(9-1)+(9-2)}{2} = 7.5$, the score of f_3 is $V_{Borda} = (9 - 3) = 6$ and the score of f_4, f_5 and f_6 is $V_{Borda} = \frac{(9-4)+(9-5)+(9-6)}{3} = 4$.*

C. Reliability of the sources

The new reliability of the sources is the sum of the rewards transmitted by the facts claimed by the source:

Definition 4. *The initial reliability of a source (before normalization) is:*

$$r^I(s) = \sum_{f \in fct(s)} V_{\mathsf{v}}(f) \tag{2}$$

Let us now see the last required adjustment of the scores.

D. Normalization functions A and C

We wish to give an estimation of the reliability of a source, i.e. the probability of this source to find the true facts. So, we have to normalize the reliability of the sources to make sure that this reliability is between 0 and 1. Now we define the normalization function n. There are at least two sensible ways to normalize the reliability. The first one rewards sources that provide a lot of true information. The second focuses on quality and then on the proportion of true information.

We call the first normalization function *A* (All objects). The reliability of the sources is divided by the number of objects in the graph. If a source has a score close to 1, we know that the source is correct on almost all objects. If a source has a low reliability, it either means that the source makes a lot of mistakes and loses votes, or that the source is correct but only on few objects.

Definition 5. *The reliability of a source with the normalization function A is:*

$$r^A(s) = \frac{r^I(s)}{best\_score(\mathcal{F}) * |\mathcal{O}|} \tag{3}$$

The second normalization function is called *C* (Claimed facts). The reliability of a source is divided by the number of objects on which it claims a fact. Unlike the previous normalization, if a source has a score close to 1, we know that the source is correct but possibly on few objects.

Definition 6. *The reliability of a source with the normalization function C is:*

$$r^C(s) = \frac{r^I(s)}{best\_score(\mathcal{F}) * |obj(s)|} \tag{4}$$

With $obj(s) = \{o \in \mathcal{O} : \exists f \in \mathcal{F} : (s, f), (f, o) \in E\}$.

So this normalization favors sources that express with care, while the previous one favors the sources that express a lot (not in a silly way).

Note that the highest score a source can obtain is $best\_score(\mathcal{F})$, so we must multiply the denominator by this value. In the case of a complete graph, the two normalization functions are identical. We note $r(s)$ the normalized reliability of a source when there is no ambiguity about the normalization used.

E. Example

We give the details of the iterations for the graph of Fig. 1 with the plurality rule and the normalization function *A*. At the first iteration, we have a tie for the object *Capital of Australia* ($c(Canberra) = c(Sydney) = 2$) but *Brasilia* wins the vote on the object *Capital of Brazil* ($c(Rio\ de\ Janeiro) = 1 < c(Brasilia) = 3$). During the 2nd iteration, $r(1) = 0.25$ and $r(2) = r(3) = r(4) = 0.75$ so *Canberra* wins the vote now because we have $c(Canberra) = 1.5 > c(Sydney) = 1.0$. The algorithm ends at the 3rd iteration. The final reliabilities of the sources are $r(1) = 0$, $r(2) = r(3) = 1$ and $r(4) = 0.5$.

V. Properties

This section is twofold. First, we want to abstract the problem, to ask what properties should be satisfied by methods that aim to evaluate the reliability of sources and the confidence of facts. We recall some properties proposed in [2] and discuss them, and propose some new ones. In particular, we propose a set of properties (the basic properties) that *any* method should satisfy, as well as additional interesting properties for characterizing interesting subclasses of methods. The second aim of this section is to evaluate our methods with respect to this set of properties.

Let us first give some definitions used by the properties.

Definition 7. *We denote $B(\mathcal{F})$ the set of facts that are ranked first on their object in the whole graph : $B(\mathcal{F}) = \{f \in \mathcal{F} \mid \forall f' \in fct(obj(f)), c(f) > c(f')\}$*

We need a definition from [2] of the notion of being *"less believable"* used in their properties.

Definition 8. *For $Y, Y' \subseteq \mathcal{F}$, Y is less believable than Y' if there is a bijection $\phi : Y \longrightarrow Y'$ such that $c(f) \leq c(\phi(f))$ for each $f \in Y$ and $c(f') < c(\phi(f'))$ for some $f' \in Y$. For $X, X' \subseteq \mathcal{S}$, X is less trustworthy than X' is defined similarly.*

A. Basic Properties

Now we present the properties that have to be satisfied by any method that aims to correctly estimate the reliability of sources and to find the truth among the facts.

If a source's reliability is equal to 1 (the highest score for a source), it means that all of its facts are the most plausible i.e. they have the highest confidence on their object:

P 1 (Best). *For $s \in \mathcal{S}$, if $r(s) = 1$ then $fct(s) \subseteq B(\mathcal{F})$.*

Reliability must be at its lowest if a source has no claim:

P 2 (Null Player). *For $s \in \mathcal{S}$, if $fct(s) = \emptyset$ then $r(s) = 0$.*

We now recall four properties and one definition from [2].

If a fact is not claimed by any source, then its confidence is lower or equal than the confidence of all the other facts:

P 3 (Groundedness). *Suppose $src(f) = \emptyset$ for $f \in \mathcal{F}$. Then for any other $g \in \mathcal{F}$, $c(f) \leq c(g)$.*

If a fact is claimed by all the sources, then its confidence will be one of the greatest:

P 4 (Unanimity). *Suppose $src(f) = \mathcal{S}$ for $f \in \mathcal{F}$. Then for any other $g \in \mathcal{F}$, $c(f) \geq c(g)$.*

Definition 9. *Two graphs G and G' are equivalent if there is a graph isomorphism π between them that preserves sources, facts, and objects such that $\pi(s) \in \mathcal{S}', \pi(f) \in \mathcal{F}'$ and $\pi(o) \in \mathcal{O}'$ for all $s \in \mathcal{S}$, $f \in \mathcal{F}$ and $o \in \mathcal{O}$.*

The values computed for the reliability of the sources and the confidence of the facts depend on the graph and not on the element's name. So this property states that all sources and facts are treated in the same way:

P 5 (Symmetry). *If G and $G' = \pi(G)$ are equivalent graphs, then ($r_G(s_1) \geq r_G(s_2)$ iff $r_{G'}(\pi(s_1)) \geq r_{G'}(\pi(s_2))$) and ($c_G(f_1) \geq c_G(f_2)$ iff $c_{G'}(\pi(f_1)) \geq c_{G'}(\pi(f_2))$).*

The ranking of the elements in a connected component is not influenced by the elements outside the component:

Definition 10. *Let $G = (V, E)$ and $G' = (V', E')$ be two graphs. We say that G and G' are independent when there are no links between the elements of G and the elements in G', i.e. $V \cap V' = \emptyset$.*

P 6 (PCI). *Let $G = (V, E)$, $G1 = (V1, E1)$, $G2 = (V2, E2)$ be three graphs such that G and Gi ($i \in \{1, 2\}$) are independent graphs. Then the rank of the sources and the facts of G will be the same in $G \cup G1$ and in $G \cup G2$: $\forall s_1, s_2 \in \mathcal{S}_G$ we have $r_{G \cup G1}(s_1) \geq r_{G \cup G1}(s_2)$ iff $r_{G \cup G2}(s_1) \geq r_{G \cup G2}(s_2)$. And $\forall f_1, f_2 \in \mathcal{F}_G$ we have $c_{G \cup G1}(f_1) \geq c_{G \cup G1}(f_2)$ iff $c_{G \cup G2}(f_1) \geq c_{G \cup G2}(f_2)$.*

Facts from less reliable sources are less credible [2]:

P 7 (Fact Coherence). *If $src(f_1)$ is less trustworthy than $src(f_2)$ then $c(f_1) < c(f_2)$.*

Some additional properties seem also desirable.

Definition 11. *Let $G = (V, E)$ be a graph. We denote $dupS(G, s, n)$ the duplication of a source s and its claims n times, i.e. $dupS(G, s, n) = (V', E')$ where $V' = V \cup \{s_1, s_2, ...s_n\}$ and $E' = E \cup \{(s_i, f) | f \in fct_G(s), i = 1, ..., n\}$.*

If an opinion is popular enough, it has to be considered as the truth. So if a source is duplicated sufficiently many times, its facts should become the most plausible ones:

P 8 (Majority). *Let $G = (V, E)$ be a graph and $s \in \mathcal{S}_G$. $\exists n > 0$ such that $fct_{G'}(s) \subseteq B_{G'}(\mathcal{F})$ with $G' = dupS(G, s, n)$.*

Now, a special case, where a graph has only one object:

Definition 12. *Let $\mathcal{O}_1$ be a graph $G = (\mathcal{S} \cup \mathcal{F} \cup \mathcal{O}, E)$ with only one object i.e. such that $|\mathcal{O}| = 1$.*

When there is only one object in a graph, a fact that is claimed by more sources will have a better confidence than another fact with fewer claims. This property is important since it states that the basic strength of a fact is given by the number of claims. But with more than one object, the information gathered about other objects can be used to make better decisions. So, this property is not desirable for more than one object, since in this case we want to consider both the number of claims and the performance of the sources on other objects:

P 9 (Claims). *If $c_{\mathcal{O}_1}(f) > c_{\mathcal{O}_1}(f')$ then $|src_{\mathcal{O}_1}(f)| > |src_{\mathcal{O}_1}(f')|$*

B. Additional Properties

The properties of the last section were the ones that any method should satisfy. In this section we will give additional properties, that are not necessary for *any* method, but that characterize interesting behaviors of *some* (subclasses of) methods.

The next two properties are related to the (Best) property. A source must (correctly) claim all facts if it wants to get the highest score:

P 10 (Best *A*). *For $s \in \mathcal{S}$, $r(s) = 1$ iff $fct(s) = B(\mathcal{F})$.*

An alternative is to consider that a source is the most reliable (reliability equal to 1) if it is always correct (without having to express itself on all objects):

P 11 (Best *C*). *For $s \in \mathcal{S}$, $r(s) = 1$ iff $fct(s) \subseteq B(\mathcal{F})$.*

Note that both (Best *A*) and (Best *C*) imply property (Best).

If a source's reliability is 0 (the lowest score), this means that none of its facts (about its object) is plausible:

P 12 (Worst). *For $s \in \mathcal{S}$, $r(s) = 0$ iff $fct(s) \subseteq \mathcal{F} \setminus B(\mathcal{F})$.*

If a source claims more believable facts than another source, then the reliability of the first source will be better:

P13 (Source Dominance). *For two sources s and s', if $|B(\mathcal{F}) \cap fct(s)| > |B(\mathcal{F}) \cap fct(s')|$ then $r(s) > r(s')$.*

When a source s claims a fact with a confidence greater than another source s' on every object, then the reliability of s must be better:

P14 (Pareto). *Let $G = (V, E)$ be a complete graph and $s, s' \in \mathcal{S}$. If $c(f) > c(f')$ with $f \neq f'$, $f \in fct(s)$, $f' \in fct(s')$ and $obj(f) \cap obj(f') = \{o\}$ $\forall o \in \mathcal{O}$ then $r(s) > r(s')$.*

C. Questionable Properties

We put in this section some properties from [2] that we consider questionable for any method and discuss why we think they are not satisfactory.

The first property says that the sources that claim more believable facts have to be more reliable:

P 15 (Source Coherence)**.** *If $fct(s_1)$ is less believable than $fct(s_2)$ then $r(s_1) < r(s_2)$.*

Note that the "less believable" notion does not require facts for being on the same objects. The problem with this property is that we are comparing facts that are (potentially) about different objects, whereas the evaluation of the facts is made for each object. For example, two facts may have the same confidence, but one is the most plausible for its object, while the other one is the least plausible for another object.

The second property states that, when a fact receives a new support, then its ranking should be strictly better:

P 16 (Monotonicity)**.** *Suppose a graph G, $s \in \mathcal{S}, f \in \mathcal{F} \setminus fct(s)$. Write E for the set of edges in G, and let G' be the graph with edges $E' = \{(s,f)\} \cup E \setminus \{(s,g) : g \neq f, obj(g) = obj(f)\}$. Then for all $g \neq f, c_G(g) \leq c_G(f)$ implies $c_{G'}(g) < c_{G'}(f)$.*

This property does not take into account the rest of the graph and the changes that can occur when an edge is changed. This property seems to be associated with a local view of the problem, where an evaluation of the facts corresponding to one object is independent of the other objects. But it is important to know the performance of the sources on other objects in order to make a decision on a given object, and changing an edge on one object can change the credibility on other objects and the reliability of many sources. Finally, the evaluation of the object where the change was made will give a different result.

Another questionable property states that the confidence of the facts must depend only on the object to which it is related. Note that the authors [2] also classify this property as undesirable:

P 17 (POI)**.** *Let G, G' two graphs and $o \in \mathcal{O}$. Suppose $fct_G(o) = fct_{G'}(o)$ and $src_G(f) = src_{G'}(f)$ for each $f \in fct_G(o)$. Then $c_G(f_1) \leq c_G(f_2)$ iff $c_{G'}(f_1) \leq c_{G'}(f_2)$ for all $f_1, f_2 \in fct_G(o)$.*

This property has a similar problem than the previous one. It is important to judge the performances of the sources on other objects in order to take decision on a given object, as illustrated in the example of the introduction, where evaluating the performance on *Capital of Brazil* helps us to decide on *Capital of Australia*.

D. Properties of S&F Methods

In the tables and figures, Pl*A* and Pl*C* stand respectively for the S&F method with plurality vote and the normalization *A* or *C*. Bo*A* and Bo*C* correspond to the methods with the Borda rule and the normalization *A* and *C*.

TABLE I
PROPERTIES SATISFIED BY S&F METHODS.

	Pl*A*	Pl*C*	Borda*A*	Borda*C*	
P1 Best	✔	✔	✔	✔	Necessary
P2 Null Player	✔	✔	✔	✔	
P3 Groundedness [2]	✔	✔	✔	✔	
P4 Unanimity [2]	✔	✔	✔	✔	
P5 Symmetry [2]	✔	✔	✔	✔	
P6 PCI [2]	✔	✔	✔	✔	
P7 Fact Coherence [2]	✔	✔	✔	✔	
P8 Majority	✔	✔	✔	✔	
P9 Claims	✔	✔	✔	✔	
P10 Best A	✔	χ	✔	χ	Optional
P11 Best C	χ	✔	χ	✔	
P12 Worst	✔	✔	χ	χ	
P13 Source Dominance	✔	χ	χ	χ	
P14 Pareto	χ	χ	✔	✔	
P15 Source Coherence [2]	χ	χ	χ	χ	Undesirable
P16 Monotonicity [2]	χ	χ	χ	χ	
P17 POI [2]	χ	χ	χ	χ	

Let us check what are the properties satisfied by our methods. We focus on the two normalizations (*C* and *A*), and on the plurality rule and the Borda rule.

Proposition 1. *Pl*A *satisfies (P1-P9), (P10), (P12), (P13) and (P14). It does not satisfy (P11), (P15-P17).*

Proposition 2. *Pl*C *satisfies (P1-P9), (P11), and (P12). It does not satisfy (P10), (P13), (P14) and (P15-P17).*

Proposition 3. *Bo*A *satisfies (P1-P9), (P10) and (P14). It does not satisfy (P11-P13) and (P15-P17).*

Proposition 4. *Bo*C *satisfies (P1-P9), (P11) and (P14). It does not satisfy (P10), (P12), (P13) and (P15-P17).*

The results are summarized in Table I. First, it is important to note that our methods satisfy all the basic properties, that are the properties that are expected for all methods. It is interesting to discuss the properties that are satisfied by only some methods, in order to illustrate the difference in their behaviors. First, note that Best *C* implies to use normalization *C* for our methods, whereas Best *A* corresponds to normalization *A*. Property Worst corresponds to the behavior of the plurality rule, with other scoring rules it will not be satisfied. Conversely, the Pareto property is related to the Borda rule, and is not satisfied by plurality, that performs a more drastic evaluation of the facts. Finally, Source dominance is satisfied only by the plurality rule and the normalization *A*.

VI. EXPERIMENTAL STUDY

Beside the theoretical evaluation of our methods, we also proceeded to an experimental evaluation of their performance for identifying the true facts and for evaluating the reliability of the sources. We have carried out experiments on both real data sets and synthetic data sets.

A. Real data sets

We evaluate our methods on two data sets that come from $http://lunadong.com/fusionDataSets.htm$, namely the

Book data set and Flight data set. We abbreviate TF for Truth Finder [1], H&A for Hubs and Authorities [10], Usums for Unbounded-Sums [2] and Sums, AL for Average.Log, Inv for Investment, PInv for PooledInvestment from [11]. P stands for Precision, A for Accuracy, R for recall and C for CSI (Critical Success Index). $Precision = \frac{TP}{TP+FP}$, $Accuracy = \frac{TP+TN}{TP+TN+FN+FP}$, $Recall = \frac{TP}{TP+FN}$, $CSI = \frac{TP}{TP+FN+FP}$, see [13] for more details on these performance measures.

Book. The difficulty with this data set was to create the graph, because the data require some text processing. After data cleaning, the graph consists of 876 sources, 5685 facts and 1263 objects. The ground truth is composed of 100 objects with a known true fact. We see in Table II that the S&F method with the plurality vote and the normalization *A* is the best method with this data set.

Flight. To clean this data set, we have to put all the dates and hours in the same format. We removed the terminal of the gate because it only appears a few times. After data cleaning, the graph consists of 38 sources, 399 506 facts and 207 912 objects. The ground truth is composed of 16 089 objects with a known true fact. We see in Table III that the method with plurality vote and the normalization *A* is also the best method with this data set. This method outperforms other methods because it manages to find the truth about objects even when the majority of sources do not claim the true fact.

So we see on these two real data sets that our S&F method outperforms all existing methods from the literature for finding the true facts for all performance measures (P, A, R, C).

B. Synthetic data sets

The limited number of real data sets available does not allow us to evaluate the performance of the methods in many different situations. We have generated synthetic data sets to be able to perform this more precise evaluation.

All the generated graphs presented here are composed of 10 objects and 4 facts by object (we have carried out many other experiments with similar results, but, due to space limitation, only present this case). For each object, we randomly choose one fact to be the true value of that object. This will be our ground truth to evaluate our methods with the metrics.

For each source, we randomly choose a number of objects between 1 and $|\mathcal{O}|$ on which this source will claim a fact. To generate the links between the sources and the facts, we assign to each source an *a priori* probability p (between 0.1 and 0.9) of choosing a true fact on each object. The false facts have the probability $1-p$, uniformly distributed, of being chosen. The graphs generated may not be complete, it means that the sources may not claim a fact on every object. After the generation, we know the *a posteriori* probability of choosing a true fact for all the sources. This value represents the true reliability of these sources.

In the tests, we rank the experiments with respect to the average reliability of the sources. We can see what happen when the sources are globally more or less reliable. In the graph, an average reliability of $x\%$ means that there are $x\%$ of links between sources and true facts (and, obviously, $(100-x)\%$ of links between sources and false facts). Each point on the graphics corresponds to the mean obtained with the generation of 1000 graphs.

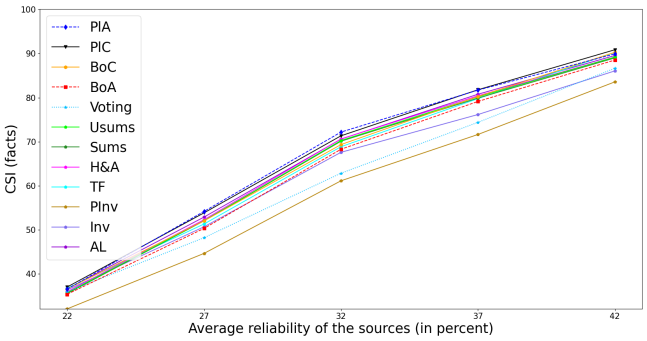

Fig. 2. CSI - 10 sources

We compare the results of our methods against related methods of the literature and Voting.

Facts Credibility - Truth Discovery. We see (Fig. 2) that the S&F methods with the plurality rule, and the two normalizations, are better for CSI than other methods from the literature when the average reliability is greater than 27%. Compared to other methods, the method with plurality finds the truth more often when there is an equal number of sources claiming the true fact and the false facts. It also finds the truth when a minority of sources claims the true fact. It is interesting to note that the methods perform very well even for low average reliability. Our methods have good results with the two normalizations. Note that all the methods find the truth when the average reliability is greater than 57%. Between 42% and 57%, we do not show the results for better readability as they are almost the same for all methods.

Reliability of the sources. We perform experiments with several metrics (number of swaps, Euclidean distance, etc.), with very convincing results, but we do not have enough space to describe all these metrics, so we will focus on averaged difference: we compute the averaged difference between the computed reliability and the (*a posteriori*) probability of choosing the true fact for every object. So this distance measures how far the estimated reliability is close to the true one (the *a posteriori* probability). For *Voting*, we define the reliability of a source as the proportion of objects on which the source claims the majority choice. We do not compare the result to Unbounded-Sums here because the score always increases for this method.

On Table IV we compare the estimated reliability obtained with the real reliability (*a posteriori* probability), for the case where the average reliability is 37% (due to space limitation, we only show the reliability of the best methods). One can check that the estimations provided by the plurality rule are very close from the true probability (recall that the results are a mean on 1000 experiments).

For a more complete picture, Fig. 3 shows the evolution of the averaged difference for different average reliability of the sources. The estimated reliability of the sources is closer to

TABLE II
RESULTS FOR *Book* DATA SET

	PlA	PlC	BoA	BoC	TF	H&A	Sums	Usums	AL	Inv	PInv
P	**78.00**	76.00	71.00	76.00	72.00	74.00	74.00	72.00	75.00	74.00	75.00
A	**90.98**	90.16	88.11	90.16	88.52	89.34	89.34	88.52	89.75	89.34	89.75
R	**78.00**	76.00	71.00	76.00	72.00	74.00	74.00	72.00	75.00	74.00	75.0
C	**63.93**	61.29	55.04	61.29	56.25	58.73	58.73	56.25	60.00	58.73	60.00

TABLE III
RESULTS FOR *Flight* DATA SET

	PlA	PlC	BoA	BoC	TF	H&A	Sums	Usums	AL	Inv	PInv
P	**91.35**	82.34	83.82	81.91	80.36	82.21	82.21	82.79	80.77	82.21	80.77
A	**91.49**	82.61	84.06	82.18	81.72	82.48	82.48	83.05	81.06	82.48	81.06
R	**91,35**	82.34	83.82	81.91	83.22	82.21	82.21	82.79	80.77	82.21	80.77
C	**84,08**	69.98	72.14	69.36	69.15	69.80	69.80	70.63	67.74	69.80	67.74

TABLE IV
SOURCE RELIABILITY - AVERAGE RELIABILITY 37%

s	Probability	PlA	BoA	Voting	Sums	TF
s1	0.11	**0.131**	0.35	0.197	0.28	0.75
s2	0.17	**0.187**	0.39	0.249	0.35	0.77
s3	0.21	**0.233**	0.42	0.288	0.41	0.78
s4	0.27	**0.296**	0.46	0.341	0.47	0.81
s5	0.33	**0.341**	0.49	0.386	0.52	0.82
s6	0.39	**0.403**	0.53	0.438	0.58	0.85
s7	0.47	**0.476**	0.58	0.503	0.66	0.86
s8	0.53	**0.528**	0.61	0.549	0.70	0.88
s9	0.57	**0.564**	0.63	0.58	0.74	0.89
s10	0.61	0.591	0.64	**0.603**	0.75	0.90

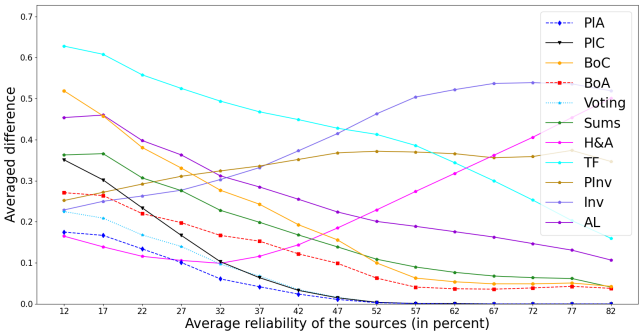

Fig. 3. Sources reliability - Averaged difference - 10 sources

the true reliability when the average reliability of the sources increases. We see that we obtain exactly the true reliability when the average reliability is better than 57% for the method using the plurality rule. With the Borda rule, we give points to all sources. This is why the reliability is not identical to the *a posteriori* probability. The sources will also get points from the false facts claimed. But, when the average reliability of sources increases, the difference tends towards 0. *Voting* has good results when the sources are reliable (average reliability greater than 57%), but before that, our method with the plurality rule is better. The iterative method helps to find the true facts even when the sources are not really reliable compared to the use of a basic voting method.

VII. CONCLUSION

In this paper, we have introduced the S&F methods for evaluating the reliability of the sources conjointly to the confidence of the facts in an information-based multi-agent system. We proposed and discuss properties that such methods should or could satisfy. And we checked which properties are satisfied by our methods. We also performed some experimental evaluations. First, we show that our methods (especially with the plurality rule) outperform methods from the literature to identify the true facts on real and generated benchmarks. But we also show that our methods allow to correctly estimate the reliability of the sources. There are numerous paths for future work. The most direct ones are to allow some similarity (or dependence) between objects, but we could also use different topics, and to take into account some a priori information about the reliability of sources.

REFERENCES

[1] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 6, pp. 796–808, 2008.

[2] J. Singleton and R. Booth, "Towards an axiomatic approach to truth discovery," *Journal of Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 2, pp. 1–49, 2022.

[3] M. de Condorcet, *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Imprimerie royale Paris, 1785.

[4] C. List and R. E. Goodin, "Epistemic democracy: Generalizing the Condorcet jury theorem," *Journal of Political Philosophy*, vol. 9, no. 3, pp. 277–306, 2001.

[5] H. P. Young and A. Levenglick, "A consistent extension of Condorcet's election principle," *SIAM Journal on Applied Mathematics*, vol. 35, no. 2, pp. 285–300, 1978.

[6] P. Everaere, S. Konieczny, and P. Marquis, "The epistemic view of belief merging: can we track the truth?" in *Nineteenth European Conference on Artificial Intelligence (ECAI'10)*, 2010, pp. 621–626.

[7] P. Hummel, "Jury theorems with multiple alternatives," *Social Choice and Welfare*, vol. 34, no. 1, pp. 65–103, 2010.

[8] J. Surowiecki, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, 2004.

[9] D. A. Waguih and L. Berti-Equille, "Truth discovery algorithms: An experimental evaluation," *arXiv preprint arXiv:1409.6428*, 2014.

[10] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[11] J. Pasternack and D. Roth, "Knowing what to believe (when you already know something)," in *International Conference on Computational Linguistics (Coling 2010)*, 2010, pp. 877–885.

[12] E. H. Rast, *Theory of Value Structure: From Values to Decisions*. Lanham: Lexington Books, 2022.

[13] R. Donaldson, R. Dyer, and M. Kraus, "An objective evaluator of techniques for predicting severe weather events," in *Preprints, Conference on Severe Local Storms, Norman, OK, American Meteoritical Society*, vol. 321326, 1975.