6-2020

Provably robust decisions based on potentially malicious sources of information

Tim MULLER
University of Nottingham

Dongxia WANG
Singapore Management University, dxwang@smu.edu.sg

Jun SUN
Singapore Management University, junsun@smu.edu.sg

Provably Robust Decisions based on Potentially Malicious Sources of Information

Tim Muller†
School of Computer Science
University of Nottingham
Nottingham, UK
tim.muller@nottingham.ac.uk

Dongxia Wang†
School of Information Systems
Singapore Management University
Singapore
dxwang@smu.edu.sg

Jun Sun
School of Information Systems
Singapore Management University
Singapore
junsun@smu.edu.sg

***Abstract*—Sometimes a security-critical decision must be made using information provided by peers. Think of routing messages, user reports, sensor data, navigational information, blockchain updates. Attackers manifest as peers that strategically report fake information. Trust models use the provided information, and attempt to suggest the correct decision. A model that appears accurate by empirical evaluation of attacks may still be susceptible to manipulation. For a security-critical decision, it is important to take the entire attack space into account. Therefore, we define the property of robustness: the probability of deciding correctly, regardless of what information attackers provide. We introduce the notion of realisations of honesty, which allow us to bypass reasoning about specific feedback. We present two schemes that are optimally robust under the right assumptions. The "majority-rule" principle is a special case of the other scheme which is more general, named "most plausible realisations".**

***Index Terms*—Provable robustness, malicious reporting, trust-based security.**

I. Introduction

On the internet, users or agents encounter situations where they need to make decisions without sufficient direct experience or observations, e.g., deciding whether to install an app. Feedback from peers helps enrich their knowledge about the subject and make better decisions. For example, rating system of an app store enables its users to share comments about whether an app crashes, whether its user interface is friendly, and whether it respects privacy, etc. In trust-based secure routing, reports about the reliability of a node from witnesses can be referred to decide whether to choose it as the next hop [1]. Moreover, sharing security information e.g., indicators, malware reports, threat intelligence reports allows users or organisations to learn from the experience of others and seek advice, thereby improving their security posture [2], [3].

The crucial commonality between these scenarios, is the possibility of a malicious source (*attacker*) reporting fake feedback, potentially causing misguided decisions. For example, some accounts of an app store may be bribed or compromised to provide fake positive review to an app. Malicious feedback will put a system under threat, when they are used for security-critical decision making. Consider trust-based routing, wrong routing messages may cause packets to be transferred to a compromised node.

The issue of how to use potentially malicious feedback is well-studied. One way to deal with inaccurate feedback sources is to develop a trust system/model, and simulate its accuracy. Another way is to implement and use such a system, and determine the accuracy empirically. Both simulated and empirical models base their analysis on the existing types of malicious strategies employed by attackers. An alternative is to use game theory to reason about malicious behaviour before it is observed. Typically, systems are set up to punish malicious behaviour and reward honest behaviour. Finally, it is possible to apply formal reasoning to find the conditions under which decisions are good or bad. Typical examples would be: "if attackers control less than 50% of the resources", or "if the trusted third party is not compromised".

Simulated/empirical models cannot predict what happens if attackers change their behaviour; they are reactive. Although incentive-based approaches are proactive regarding attacker behaviour, they assume that the attack goals are set in stone. Given a new utility function, a certain type of malicious behaviour may now have a positive pay-off, despite there being a punishment. It is desirable to consider the entire attack space. The issue with condition-based models is that all the guarantees are merely conditional, and we do not know whether the conditions are actually met. Our approach is a generalisation of the condition-based models: although we may not know whether the conditions are met, there may be an overwhelming probability that they are. We focus on proving when there is an overwhelming probability that the conditions for making a good decision are met.

We model feedback-based decision making as a function, with the input being a set of reported data (feedback), and the output being a choice of a belief or an action. Once we know which option is true, the corresponding decision is clear. Consider an example of deciding whether to install a software based on its security property, feedback set includes two options: "is malware" and "not malware". Based on the knowledge about the reporters, if the scheme trusts the first option, then the action "not install" is selected. Honest sources

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/R01034X/1 and by the Ministry of Education, Singapore, grant MOE2016-T2-2-123.

† These authors contributed equally.

DOI 10.1109/CSF49147.2020.00036

report the true option, while malicious sources (attackers) report strategically – with an unknown strategy. Each source has a certain probability of being honest. As we want to be proactive regarding malicious behaviour, we may not assume what strategies an attacker would take, but consider the entire attack space in a given decision making context. And we allow attackers to know the decision scheme considering that to rely on secrecy is a poor practice for security (e.g. NIST recommends against this in systems security [4]).

There are several properties that we aim to achieve for a decision scheme. The most important property we define is ϵ-*robustness*, meaning that the probability that the scheme decides correctly is at least $1 - \epsilon$, no matter what feedback attackers provide. Another property is *optimality*, meaning that no other decision scheme has better robustness. The final property is *monotonicity*, meaning that robustness does not decrease if we use more feedback sources. We prove that the simple "follow majority" decision scheme is robust, monotonic and optimal, if all feedback sources are equally probable to be honest (e.g. all strangers). For the more interesting case, where sources are not equal, we propose a new decision scheme ("most plausible realisations"), and prove that it is robust, monotonic and optimal too. Trust evaluation can be used to estimate the probability of a source being malicious, but these estimates may be inaccurate or biased. We use simulations to investigate how actual attacks relate to the notion of manipulation. The simulations also show that our results are not very sensitive to the quality of the estimates. Finally, we discuss how to build upon our results to move towards applications where our assumptions need to be adapted.

II. Secure Decisions based on Feedback

The goal of this paper is to introduce a general methodology to make decisions that are almost certainly correct, in the existence of malicious feedback sources (*attackers*). The attackers try to manipulate our decisions. If the probability that the attacker successfully manipulates us is less than ϵ, then we achieved ϵ-*robustness*. Our decision, therefore, is correct with probability $1 - \epsilon$, despite manipulation attempts. We do not assume specific attacker strategy (such as rational attackers, or attackers that follow a particular template), but allow them total freedom to select their manipulative feedback.

A. Model

We aim to introduce a general methodology to approach feedback in a way that allows resistance to manipulation by design. Of course, in a situation where all sources of feedback are malicious, one can be manipulated. However, the probability of this situation occurring is typically increasingly small as more sources are used. On the other hand, it is possible that all sources of feedback are honest, in which case no manipulation can occur. The probability of this situation is also increasingly small with more sources. It gets interesting when some sources are malicious and some are honest. In those situations, whether or not manipulation can occur depends on the way decisions are made. In this section, we introduce the concepts required to reason about decisions and manipulation.

Decisions are made using a decision scheme. A decision scheme is a function that outputs a decision based on the received feedback. The feedback comes in the form of a discrete value (called an option) selected by a source[1] in a given set. Formally:

Definition 1 (Decision Scheme)**.**
- *There is a set of sources* $\mathcal{S} = \{0, \ldots, m-1\}$.
- *There is a set of feedback options* $\mathcal{O} = \{0, \ldots, n-1\}$.
- *There is a set of decisions* $\mathcal{Q} = \{0, \ldots, \nu-1\}$. *Only one decision is correct in a decision making task.*
- *Feedback* $\mathbf{f} \in \mathcal{F}$ *is an* m*-tuple:* $\mathbf{f} = (f_0, \ldots, f_{m-1})$*, where* f_s *represents the feedback option reported by source* $s : s \in \mathcal{S}$ *and* $f_s \in \mathcal{O}$.
- *A decision scheme is a function* $\mathscr{D} : \mathcal{F} \to \mathcal{Q}$.

A decision scheme works in a specific context, which is defined by $\mathcal{S}, \mathcal{O}, \mathcal{Q}$. For different contexts, a system will need to select which decision scheme is appropriate. For example, given $\mathcal{O}, \mathcal{Q}$, two schemes are required for $m = 10$ and $m = 100$. A *decision mechanism* selects an appropriate decision scheme, based on context.

Informally, manipulation is when malicious sources select their feedback to ensure that $\mathscr{D}(\mathbf{f})$ results in the incorrect decision. The malicious sources are aware of what our decision scheme is, and if we alter $\mathscr{D}$, then they change their feedback accordingly. Say $\mathscr{D}(0,1,0) = 0$ and $\mathscr{D}(0,1,1) = 1$, and we receive $(f_0, f_1, f_2) = (0,1,0)$. If the third source is malicious and 0 is the correct decision, he would provide the feedback 1, and vice versa. So if we receive $(0,1,0)$, then the correct decision is 1 and if we receive $(0,1,1)$ then it's 0. Unfortunately, if we change $\mathscr{D}$ so that $\mathscr{D}(0,1,0) = 1$ and $\mathscr{D}(0,1,1) = 0$ to reflect this, then the attacker responds by swapping his feedback around. Hence, basing the decision scheme on the feedback is problematic, as the feedback (of attackers) depends on the decision scheme. We introduce the notion of *realisations*, allowing us not to reason explicitly about feedback, bypassing this issue altogether.

When receiving feedback, some of the sources providing it will have been honest, and others malicious. Informally, this is what a realisation is. Formally:

Definition 2 (Realisation)**.** *A* realisation $\mathbf{r} \subseteq \mathcal{S}$ *is the set of sources that are honest.*
The set of all realisations $\mathcal{R}$ *is the powerset of sources* $2^{\mathcal{S}}$.
The complement of a realisation is: $\overline{\mathbf{r}} = \mathcal{S} \setminus \mathbf{r}$.

A realisation indicates which sources are honest and which are not. We use the phrase "under realisation $\mathbf{r}$" to mean "assume all $s \in \mathbf{r}$ are honest and all $s \in \overline{\mathbf{r}}$ are malicious".

Of course, when receiving feedback from sources, the recipient does not know the realisation (as the recipient does

[1]We use the abstract term "source", since it does not matter for our purposes whether or not the source is a person, an agent, a sensor, a device. As long as it provides manipulative data if it is (controlled by) an attacker, but useful information if it is not.

not know who is honest or malicious). But, depending on how the recipient makes decisions, he is open to manipulation in some realisations, but cannot be manipulated in others. Of course, some ways of making decisions are superior to others as they allow manipulation less often. Our goal is to make decisions in such a way that it is improbable to be manipulable. This is helpful for security critical decisions.

If a source s is honest under the realisation ($s \in \mathbf{r}$), then s does not try to manipulate the decision scheme. What s reports (f_s), does not depend on the decision scheme being used and is not affected by attackers' choices. We call this the *weak assumption of honesty*. If s is not honest, then he is an attacker. Attackers are aware of the decision scheme, hence able to pick f_s depending on which decision scheme is used. This means that s can provide any feedback (true or false) in $\mathcal{O}$. If there are multiple attackers under the realisation, then their feedback may contain diverse options, of which all combinations must be considered. We call the set of all the possible feedback given a realisation the *attack space* (of the realisation).

Intuitively, feedback from honest sources should lead to the correct decision. In this paper, we make a stronger simplifying assumption, namely that there is a one-to-one correspondence between which decision is correct, and which feedback honest sources provide. This means that feedback of honest sources are the same and it maps to the correct decision. We call this the *strong assumption of honesty*. To simplify notation, we can use equality to model the one-to-one correspondence, and claim that the feedback that an honest source provides *is* the correct decision. Accordingly, we have $\mathcal{O} = \mathcal{Q}$.

Under the strong assumption of honesty, given a realisation $\mathbf{r}$ and $c \in \mathcal{Q}$ as the correct decision, we can only receive feedback $\mathbf{f}$ where every $s \in \mathbf{r}$ reports c (i.e. $f_s = c$). The malicious sources, however, are not restricted and can provide any feedback in $\mathcal{O}$. We refer to this as the attack space:

Definition 3 (Attack Space)**.** *The attack space is a function $a : \mathcal{Q} \times \mathcal{R} \to \mathcal{F}$. If the correct decision is c and the realisation is $\mathbf{r}$, then $a(c, \mathbf{r}) = \{\mathbf{f} \in \mathcal{F} | \forall_{s \in \mathbf{r}}(f_s = c)\}$ is the set of all the possible feedback we could receive.*

Being non-manipulable under a realisation means that no matter what feedback the attackers report, the decision scheme always decides correctly.

Definition 4 (Non-Manipulability)**.** *A decision scheme $\mathscr{D}$ is considered* non-manipulable *under a realisation $\mathbf{r}$ when: $\forall c \in \mathcal{Q}$ and $\forall \mathbf{f} \in a(c, \mathbf{r})$, $\mathscr{D}(\mathbf{f}) = c$.*

If a scheme $\mathscr{D}$ is non-manipulable under $\mathbf{r}$, then we say $\mathbf{r}$ is non-manipulable for $\mathscr{D}$. The set $\widehat{\mathscr{D}}$ is the set of all non-manipulable realisations for $\mathscr{D}$.

Reasoning backwards, we may wonder whether for a set of realisations $R \subseteq \mathcal{R}$, it is possible to have a decision scheme which is non-manipulable under all realisations $\mathbf{r} \in R$. Unfortunately, being non-manipulable in all realisations $\mathcal{R}$ is impossible. For example, the realisation where all the sources are malicious ($\mathbf{r} = \emptyset$) is always manipulable. We define that a set of realizations R is attainable, if there exists a $\mathscr{D}$ for which all $\mathbf{r} \in R$ are non-manipulable.

Definition 5 (Attainable)**.** *A set of non-manipulable realisations $R \subseteq \mathcal{R}$ is attainable: $\mathbb{A}(R)$ if and only if there exists a decision scheme $\mathscr{D}$ such that $R \subseteq \widehat{\mathscr{D}}$.*

As it turns out, whether or not a set of realisations is attainable is characterised by a simple predicate, not involving actual decision schemes or feedback. This characterisation is the basis of our claim that we do not need to focus on actual feedback. A set of realisations is attainable, if and only if every pair of realisations shares at least one source:

Theorem 1. *$\mathbb{A}(R)$ if and only if $\forall_{\mathbf{r}_1, \mathbf{r}_2 \in R}(\mathbf{r}_1 \cap \mathbf{r}_2 \neq \emptyset)$.*

Proof. To see that $\forall_{\mathbf{r}_1 \in R, \mathbf{r}_2 \in R}(\mathbf{r}_1 \cap \mathbf{r}_2 \neq \emptyset)$ is a necessary condition for $\mathbb{A}(R)$, assume that $\mathbf{r}_1$ and $\mathbf{r}_2$ are disjoint. Select $\mathbf{f}$ s.t. $\exists s : s \in \mathbf{r}_1 \wedge f_s = c_1$. Since $s \notin \mathbf{r}_2$, $\mathbf{f} \in a(c_1, \mathbf{r}_1)$ and $\mathbf{f} \in a(c_2, \mathbf{r}_2)$. Either $\mathscr{D}(\mathbf{f}) \neq c_2$ or $\mathscr{D}(\mathbf{f}) \neq c_1$.

We show that $\forall_{\mathbf{r}_1 \in R, \mathbf{r}_2 \in R}(\mathbf{r}_1 \cap \mathbf{r}_2 \neq \emptyset)$ is a sufficient condition for $\mathbb{A}(R)$ by constructing a decision scheme $\mathscr{D}$ so that $R \subseteq \widehat{\mathscr{D}}$. If possible, pick $\mathscr{D}$ so that $\mathscr{D}(\mathbf{f}) = x$, when there exists a realisation $\mathbf{r} \in R$ such that for all sources $s \in \mathbf{r}$, $f_s = x$. If there are multiple realisations $\mathbf{r}_1$ and $\mathbf{r}_2$ where for all sources $s \in \mathbf{r}_1$, $f_s = x$ and for all $t \in \mathbf{r}_2$, $f_t = y$, then $x = y$, since $\mathbf{r}_1$ and $\mathbf{r}_2$ share at least one source. So our choice of $\mathscr{D}$ has at most one value x per $\mathbf{f}$, such that we require $\mathscr{D}(\mathbf{f}) = x$, and thus $\mathscr{D}$ exists. □

Theorem 1 implies that a realization and its complement cannot coexist in an attainable set of realizations. An important corollary follows this observation, which states that the maximum size of an attainable set of realisations is half of the total possible realisations.

Corollary 1. *If $\mathbb{A}(R)$, then $|R| \leq 1/2|\mathcal{R}|$.*

Proof. Theorem 1, that $\mathbf{r}$ and $\overline{\mathbf{r}}$ cannot both be in R. At most, $R = 1/2|\mathcal{R}|$, with $\mathcal{R} \setminus R$ being the set of its complements. □

In conclusion, our approach focuses on the possible realisations (honesty states of sources), rather than on the actual feedback. This allows us to reason more clearly about the attack spaces in different scenarios. Using the assumption that there is a bijection between honest feedback and correct decisions, we were able to further specify the model. Importantly, we show that by appropriately selecting the realisations under which we wish to be non-manipulable, we know a decision scheme exists where we indeed are non-manipulable under these realis-ations. Immediate corollaries are the fact that at most half the realisations allow us to be non-manipulable, and that only either one realisation or its complement can be non-manipulable.

B. Probability

Corollary 1 proves that given an arbitrary set of sources, at least half of the realisations are manipulable, hence the possibility that an arbitrary decision scheme gets manipulated always exists. However, it may be the case that manipulable

realisations are improbable. We can define the probabilistic notion of ϵ-robustness to capture the idea that the probability of being under a manipulable realisation, is at most ϵ. To do so, we introduce probability in this section.

Being manipulated means making incorrect decisions based on feedback. To achieve a notion of robustness, we have to consider the entire attack space of a given set of malicious sources. We cannot make assumptions about the probability distribution of the feedback that attackers provide.

Below, we introduce a way to compute the probability of deciding incorrectly for a given decision making task, named as *error rate*. First, let variable C model what the correct decision is. The outcomes of C are from the set $\mathcal{Q}$: $c \in \mathcal{Q}$. Let random variable $\mathbb{R}$ model the realisation we are under, and previously defined $\mathbf{r} \in \mathcal{R}$ is its outcome. Let random variable $\mathcal{I}$ be the decision. Based on the law of total probability, error rate: $p(\mathcal{I}{\neq}c \mid C{=}c) = \sum_{\mathbf{r}\in\mathcal{R}} p(\mathbb{R} = \mathbf{r} \mid C = c) \cdot p(\mathcal{I} \neq c \mid \mathbb{R} = \mathbf{r}, C = c)$. Whether a source is honest or not does not depend on C. Hence, $p(\mathbb{R} = \mathbf{r} \mid C = c) = p(\mathbb{R} = \mathbf{r})$. Define the distribution Δ on $\mathcal{R}$ s.t. $\Delta(\mathbf{r}) = P(\mathbb{R} = \mathbf{r})$. The distribution Δ provides a context to the sources, by defining how probable it is that certain sources are honest.

The decision $\mathcal{I}$ is defined by the decision scheme when provided with feedback. Let random variable $\mathbb{F}$ denote the received feedback, and $\mathbf{f} : \mathbf{f} \in \mathcal{F}$ is its outcome. Given realisation $\mathbf{r}$ and the correct decision c, all possible feedback is in the attack space $a(c, \mathbf{r})$, which is the support of $\mathbb{F}$. The decision $\mathcal{I}$ equals $\mathscr{D}(\mathbf{f})$, and $p(\mathcal{I} \neq c \mid \mathbb{F} = \mathbf{f}) = 1$ iff $\mathscr{D}(\mathbf{f}) \neq c$. Hence: $p(\mathcal{I} \neq c \mid \mathbb{R} = \mathbf{r}, C = c) = \sum_{\mathbf{f}\in a(c,\mathbf{r})\wedge\mathscr{D}(\mathbf{f})\neq c} p(\mathbb{F} = \mathbf{f} \mid \mathbb{R} = \mathbf{r}, C = c)$.

We use a shorthand notation to describe the probability distribution of feedback in an attack space: $\beta(\mathbf{r}, c)(\mathbf{f}) = p(\mathbb{F} = \mathbf{f}|\mathbb{R} = \mathbf{r}, C = c)$. Since honest sources only report the correct decision under the strong assumption, the distribution $\beta(\mathbf{r}, c)$ is purely determined by attackers. Different $\beta(\mathbf{r}, c)$ describes different strategy of attackers within the space $a(c, \mathbf{r})$. And $p(\mathcal{I} \neq c \mid \mathbb{R} = \mathbf{r}, C = c) = \sum_{\mathbf{f}\in a(c,\mathbf{r})\wedge\mathscr{D}(\mathbf{f})\neq c} \beta(\mathbf{r}, c)(\mathbf{f})$.

With Δ and β, we can derive a general formula of error rate $\mathtt{Err}(\mathscr{D}, \Delta, \beta) = p(\mathcal{I} \neq c \mid C = c)$:

$$\mathtt{Err}(\mathscr{D}, \Delta, \beta) = \sum_{\mathbf{r}\in\mathcal{R}} \sum_{\mathbf{f}\in a(c,\mathbf{r})\wedge\mathscr{D}(\mathbf{f})\neq c} \Delta(\mathbf{r}) \cdot \beta(\mathbf{r}, c)(\mathbf{f}) \quad (1)$$

Note, when Δ and $\mathscr{D}$ are given, the error rate is in control of attackers, specifically purely determined by β. Next, we study three properties of a decision scheme in terms of its error rate.

C. Properties

The three properties that we are interested in studying are robustness, optimality and monotonicity. A decision scheme that is robust, optimal and monotonic has the highest probability of not being manipulated and does not degrade in quality unexpectedly. These are the properties we require for secure-decision making. Note that the focus on these three properties does mean that we do not always obtain the same degree of accuracy in some scenarios.

Robustness means resistance to being manipulated. We do not want to assume any strategy for providing feedback to define robustness. Instead, we consider all possible distributions within the relevant attack spaces. Robustness in a context Δ is then determined by the maximal error rate:

$$\mathfrak{E}(\mathscr{D}) = \max_\beta \left(\mathtt{Err}(\mathscr{D}, \Delta, \beta)\right) \quad (2)$$

We can now define robustness:

Definition 6 (ϵ-robustness)**.** *Given a value ϵ, a set of sources $\mathcal{S}$ and a distribution Δ of realisations, a decision scheme is ϵ-robust when for all distributions β of feedback:*

$$\mathtt{Err}(\mathscr{D}, \Delta, \beta) \leq \epsilon.$$

Equivalently, we can say $\mathfrak{E}(\mathscr{D}) \leq \epsilon$. However, an even simpler computation of robustness – that only reasons about realisations – can be provided:

Theorem 2. *If a decision scheme $\mathscr{D}$ is ϵ-robust, then*

$$\sum_{\mathbf{r}\in\mathcal{R}\setminus\widehat{\mathscr{D}}} \Delta(\mathbf{r}) \leq \epsilon.$$

Proof. It suffices to prove that $\mathfrak{E}(\mathscr{D}) = \sum_{\mathbf{r}\in\mathcal{R}\setminus\widehat{\mathscr{D}}} \Delta(\mathbf{r})$. We first simplify the inner sum in Equation 1. If $\mathbf{r}$ is non-manipulable ($\mathbf{r} \in \widehat{\mathscr{D}}$) (Definition 4), then $\nexists \mathbf{f} : \mathbf{f}{\in}a(c, \mathbf{r}) \wedge \mathscr{D}(\mathbf{f}){\neq}c$, and $\sum_{\mathbf{f}\in a(c,\mathbf{r})\wedge\mathscr{D}(\mathbf{f})\neq c} \beta(\mathbf{r}, c)(\mathbf{f}){=}0$. Contrarily, if $\mathbf{r}$ is ($\mathbf{r} \notin \widehat{\mathscr{D}}$), then $\exists \mathbf{f} : \mathbf{f}{\in}a(c, \mathbf{r}) \wedge \mathscr{D}(\mathbf{f}){\neq}c$. Select a point distribution $\beta(\mathbf{r}, c) = 1$ for that value $\mathbf{f}$. Then, trivially, $\sum_{\mathbf{f}\in a(c,\mathbf{r})\wedge\mathscr{D}(\mathbf{f})\neq c} \beta(\mathbf{r}, c)(\mathbf{f}){=}1$. Hence, this choice of β satisfies $\mathtt{Err}(\mathscr{D}, \mathbf{r}, \beta) = \sum_{\mathbf{r}\in\mathcal{R}\setminus\widehat{\mathscr{D}}} \Delta(\mathbf{r})$. As there is no way to increase a probability beyond 1, this choice is maximal. □

By reasoning purely about the realisations, we can reach conclusions about whether or not we can be manipulated, without having to reason about the possible strategies that the attackers might employ. Given the fact that trust systems make deductions explicitly using the specific feedback, it is encouraging that we can prove that this is not necessary, simplifying the problem domain significantly.

For a sufficiently large ϵ, many decision schemes will be ϵ-robust. In general, we are interested in selecting a decision scheme that can be claimed to be robust with a minimal ϵ, or $\mathfrak{E}(\mathscr{D})$; i.e. the scheme that has maximal robustness. This idea is captured by the optimality property:

Definition 7 (Optimality)**.** *For a given distribution Δ of realisations, $\mathscr{D}$ is optimal when for all $\mathscr{D}'$, $\mathfrak{E}(\mathscr{D}) \leq \mathfrak{E}(\mathscr{D}')$.*

Or, equivalently, an ϵ-robust scheme $\mathscr{D}$ is *optimal* if there does not exist a scheme $\mathscr{D}'$ which is ϵ'-robust and $\epsilon'{<}\epsilon$.

The last property is monotonicity. Monotonicity requires that adding a source of information to the feedback does not decrease robustness. As mentioned before, a different number of sources means a different decision context. We are, therefore, comparing two different decision schemes, that arise from the same decision mechanism:

Definition 8 (Monotonicity)**.** *A decision mechanism is monotonic, when for all pairs of decision schemes $\mathscr{D}_1$ (using sources $\mathcal{S}_1$) and $\mathscr{D}_2$ (using sources $\mathcal{S}_2$), if $\mathcal{S}_1 \subseteq \mathcal{S}_2$, then $\mathfrak{E}(\mathscr{D}_1) \geq \mathfrak{E}(\mathscr{D}_2)$.*

Or, equivalently, for every ϵ, if $\mathscr{D}_1$ is ϵ-robust then $\mathscr{D}_2$ is also ϵ-robust. Using more sources does not harm robustness.

III. Majority Rule

Majority rule is a principle applied for a variety of reasons. If the decision being made affects everyone involved, then fairness is a big reason to apply majority rule. In the case that a user wants to make a decision that involves his own security, fairness towards the sources is unlikely to be a consideration. Nevertheless, majority rule occasionally pops up as a decision scheme in these scenarios too. As we demonstrate in this section, it turns out that majority rule can be the optimal way to make robust decisions under the right circumstances.

The feedback reporting scenarios in this section are simple scenarios where all sources are treated as interchangeable and independent (e.g., when it is difficult to characterise individual sources). Then, all m sources have some fixed probability p of being honest; so $\Delta(\mathbf{r}) = p^{|\mathbf{r}|} \cdot (1-p)^{m-|\mathbf{r}|} = \delta_{m,p}$. In these (simple) scenarios, majority rule is optimal.

Consider the decision scheme $\mathscr{M}_{m,p}$, which outputs the decision that more than $m/2$ sources provided as feedback. In case no such feedback exist it follows the feedback from the source $0 \in \mathcal{S}$. Formally:

Definition 9 (Majority Rule Decision Scheme)**.** *If there is a decision d s.t. $|\{s \in \mathcal{S} | f_s = d\}| > m/2$, then the* majority rule decision scheme *has $\mathscr{M}_{m,p}(\mathbf{f}) = d$. Otherwise $\mathscr{M}_{m,p}(\mathbf{f}) = f_0$.*

Our focus is typically on the decision scheme, but we may use the symbol $\mathscr{M}$ to denote the *majority rule decision mechanism*, which selects the appropriate decision scheme $\mathscr{M}_{m,p}$ based on context m and p.

Observe that if Alice's feedback is x, Bob and Charlie's is y and Dave and Elsa's is z, then we decide x, as Alice is the tie-breaker because she is the first source (source 0). This sounds counter-intuitive, but it is one of the optimal ways of deciding (as we prove later). However, if we look at this situation through the lens of realisations, we can see that whatever the truth is, at least 3 sources are malicious. These three sources could have simply provided the same fake feedback, and have obtained a majority, thus manipulating the decision scheme. Whatever the realisation is that lead to the feedback, the realisation is manipulable under the majority rule decision scheme. So, although the majority rule scheme is dependent on the actual feedback, the analysis of the decision scheme is simpler when done through the realisations.

Recall that every decision scheme has an associated set of non-manipulable realisations. The corresponding set of non-manipulable realisations for the majority rule decision scheme is straightforward: a realisation $\mathbf{r}$ is in $\widehat{\mathscr{M}_{m,p}}$, when more than half of the sources are honest, or if exactly half the sources are honest and s_0 is honest:

Lemma 1. $\widehat{\mathscr{M}_{m,p}} = \{\mathbf{r} \in \mathcal{R} \mid |\mathbf{r}| > m/2 \vee (|\mathbf{r}| = m/2 \wedge 0 \in \mathbf{r})\}$

Proof. Honest sources always report the correct decision. When $|\mathbf{r}| > m/2$, the correct decision would be the majority in the received feedback. When $|\mathbf{r}| = m/2$, we trusts the first source, and we make correct decision if he is honest. □

A. Properties of Majority Rule

The robustness of majority rule can be expressed using the cumulative binomial distribution $F_{\texttt{bin}}(k; m, p)$ (the binomial distribution is $f_{\texttt{bin}}(k; m, p)$). Via Definition 6:

Theorem 3. *If $p > 1/2$, then $\mathscr{M}_{m,p}$ is ϵ-robust for:*

$$\epsilon \geq \begin{cases} F_{\texttt{bin}}(\frac{m-1}{2}; m, p) & m \bmod 2 \neq 0 \\ F_{\texttt{bin}}(\frac{m}{2}-1; m, p) + \frac{1}{2} f_{\texttt{bin}}(\frac{m}{2}; m, p) & m \bmod 2 = 0 \end{cases}$$

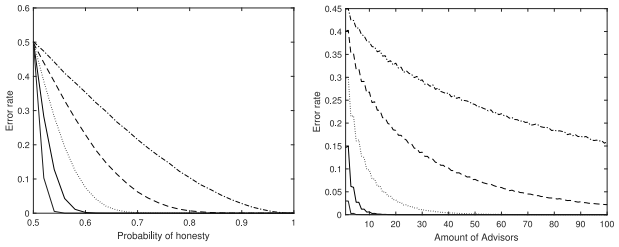
Proof. The probability of getting at most k honest sources within m sources is $F_{\texttt{bin}}(k; m, p)$. Lemma 1 states that strictly over half the sources being honest in a realisation is sufficient to be non-manipulable. If the number of sources is odd, then it is only possible to be manipulated when the honest sources are in the minority, i.e. $k < \frac{m}{2}$, with probability: $F_{\texttt{bin}}(\frac{m-1}{2}; m, p)$. If the number of sources is even, then another possibility being manipulated is in case of a tie. But even then, there is at least a 50% chance that the realisation is non-manipulable, since the first source is in the correct block in half the equiprobable permutations. This is why we must add $1/2 f_{\texttt{bin}}(\frac{m}{2}; m, p)$. □

An intuition why $\mathscr{M}$ is optimal under any possible values of m, p is that, since $p > 1/2$, a realisation r where a majority is honest is always more probable than its complement $\overline{r}$. For realisations r that have exactly the same amount of honest and malicious sources, their complements $\overline{r}$ are equiprobable. Simply selecting an arbitrary half of these realisations is therefore optimal; we select the half where s_0 is honest (leaving the half where s_0 is malicious). Formally:

Theorem 4. *Given m sources with $p \geq 1/2$, $\mathscr{M}_{m,p}$ is optimal.*

Proof. $\widehat{\mathscr{M}_{m,p}}$ includes exactly a half of all the possible realizations, which is the maximum amount of realisations, according to Corollary 1. It contains all the realisations $\mathbf{r}$ where honest sources outnumber the malicious ones, and vice versa for the complement $\overline{\mathbf{r}}$. So $\delta_{m,p}(\mathbf{r}) = p^{|\mathbf{r}|}(1-p)^{m-|\mathbf{r}|} > (1-p)^{|\mathbf{r}|}p^{m-|\mathbf{r}|} = \delta_{m,p}(\overline{\mathbf{r}})$. It may contain realisations where honest sources are equal in number to malicious ones, in which case $\delta_{m,p}(\mathbf{r}) = \delta_{m,p}(\overline{\mathbf{r}})$. Either way $\delta_{m,p}(\mathbf{r}) \geq \delta_{m,p}(\overline{\mathbf{r}})$, so $\widehat{\mathscr{M}_{m,p}}$ never contains the smaller of the pair. □

Below, we apply a Monte Carlo simulation to demonstrate how the error rate (probability of making incorrect decisions) of $\mathscr{M}_{m,p}$ changes with different values of honesty p and the number of sources m. m sources are selected from a large pool with a fraction p of honest sources. Honest sources will rate the correct decisions, but malicious sources rate according to the coordinated attack: all attackers provide the same lie. The majority rule scheme simply follows the majority, and the fraction of incorrect decisions is graphed in Fig. 1. Not



(a) Per probability of honesty. (b) Per number of sources.

Fig. 1: Effect of parameters p and m on error rate of $\mathscr{M}_{m,p}$.

surprisingly, the graphs are virtually indistinguishable from the analytical results.

In Fig. 1a, we consider scenarios with 3 (dash-dotted), 13 (dashed), 51 (dotted), 201 (solid black), or 1001 (solid gray) sources providing feedback. The x-axis depicts the probability that sources are honest. In every scenario, the error rate rapidly declines initially, and then asymptotically decreases to 0. In Fig. 1b, we consider scenarios with sources with different levels of trustworthiness: adequate (0.55, dash-dotted), moderate (0.6, dashed), somewhat (0.7, dotted), standard (0.85, solid black) and high (0.97, solid gray). Here, the x-axis depicts the amount of sources giving feedback. This graph also rapidly declines until asymptotically decreasing to 0, but with a staircase pattern. The staircase pattern is not an artifact, as proven later in Lemma 2.

Fig. 1 shows that with p values over $1/2$ and the coordinated attack, the error rate of $\mathscr{M}_{m,p}$ tends to decrease with more sources.

Monotonicity is the property that better robustness (lower maximal error rate – $\mathfrak{E}$) can be obtained, if more sources are available. It turns out that there is a disparity between even and odd numbers of sources: For even m, the $\mathfrak{E}$ for m sources and $m-1$ sources is the same, but going from m to $m+1$ the $\mathfrak{E}$ goes down.

We use the following property of the binomial distribution:

Proposition 1.
$F_{\mathtt{bin}}(k; m{+}1, p) = F_{\mathtt{bin}}(k{-}1; m, p) + (1{-}p) f_{\mathtt{bin}}(k; m, p)$

Proof. It follows from the following equality of regularised incomplete beta functions: $I_x(a, b{+}1) = I_x(a, b) + \frac{x^a(1-x)^b}{bB(a,b)}$ [5], and the fact that $F_{\mathtt{bin}}(k, m, p) = I_{1-p}(m-k, k+1)$:
$F_{\mathtt{bin}}(k; m{+}1, p) = I_{1-p}(m-k+1, k+1) = I_{1-p}(m-k+1, k) + \frac{(1-p)^{m-k+1} p^k}{k \frac{(m-k)!(k-1)!}{m!}} = I_{1-p}(m-k+1, k) + (1-p)p^k(1-p)^{m-k} \frac{m!}{(m-k)!k!} = F_{\mathtt{bin}}(k{-}1; m, p) + (1{-}p) f_{\mathtt{bin}}(k; m, p)$. □

Lemma 2. *Given m is even, $\mathfrak{E}(\mathscr{M}_{m,p}) = \mathfrak{E}(\mathscr{M}_{m-1,p})$.*

Proof. $\mathfrak{E}(\mathscr{M}_{m,p}) = F_{\mathtt{bin}}(\frac{m}{2}{-}1; m, p) + \frac{1}{2} f_{\mathtt{bin}}(\frac{m}{2}; m, p)$. $\mathfrak{E}(\mathscr{M}_{m-1,p}) = F_{\mathtt{bin}}(\frac{m}{2}{-}1; m{-}1, p)$. Let $B = \frac{1}{2} f_{\mathtt{bin}}(\frac{m}{2}; m, p)$, $A = F_{\mathtt{bin}}(\frac{m}{2} - 1; m, p)$.

Based on Proposition 1, $A = F_{\mathtt{bin}}(\frac{m}{2}{-}2; m{-}1, p) + (1{-}p) \cdot f_{\mathtt{bin}}(\frac{m}{2}{-}1; m - 1, p)$. As $\frac{1}{2}\binom{m}{\frac{m}{2}} = \binom{m-1}{\frac{m}{2}-1}$, $B = \binom{m-1}{\frac{m}{2}-1} p^{\frac{m}{2}} (1-p)^{\frac{m}{2}} = p f_{\mathtt{bin}}(\frac{m}{2}{-}1; m-1, p)$. Hence, $A + B = F_{\mathtt{bin}}(\frac{m}{2}{-}2; m{-}1, p) + (1{-}p) f_{\mathtt{bin}}(\frac{m}{2}{-}1; m - 1, p) + p f_{\mathtt{bin}}(\frac{m}{2}{-}1; m-1, p) = F_{\mathtt{bin}}(\frac{m}{2} - 1; m{-}1, p) = \mathfrak{E}(\mathscr{M}_{m-1,p})$. □

Intuitively, if one chooses a majority amongst an even number of sources, then that majority is at least 2 larger than any other option, and removing a single source would not change the decision. In case of a tie, removing an source will have a 50% chance of swaying the result in either option's favor.

Lemma 3. *Given m is even, $\mathfrak{E}(\mathscr{M}_{m,p}) > \mathfrak{E}(\mathscr{M}_{m+1,p})$.*

Proof. $\mathfrak{E}(\mathscr{M}_{m,p}) = F_{\mathtt{bin}}(\frac{m}{2}{-}1; m, p) + \frac{1}{2} f_{\mathtt{bin}}(\frac{m}{2}; m, p)$. $\mathfrak{E}(\mathscr{M}_{m+1,p}) = F_{\mathtt{bin}}(\frac{m}{2}; m{+}1, p) = F_{\mathtt{bin}}(\frac{m}{2}{-}1; m, p) + (1{-}p) f_{\mathtt{bin}}(\frac{m}{2}; m, p)$ (Proposition 1). Since $p > 1/2$, $(1{-}p) < 1/2$, proving the theorem. □

Together the two lemmas trivially prove the monotonicity of decision mechanism $\mathscr{M}$:

Theorem 5. *For sets of sources $\mathcal{S}_1 \subseteq \mathcal{S}_2$ of cardinality m and m', $\mathfrak{E}(\mathscr{M}_{m,p}) > \mathfrak{E}(\mathscr{M}_{m',p})$.*

All the properties have been proven under the assumption that $p > 1/2$. If $p \leq 1/2$, then attackers have a higher probability of achieving a majority than the honest sources. Majority rule is hardly robust in that case, as $\epsilon \geq 1/2$, no matter how many sources are used. The probability of attackers achieving a majority actually increases as the number of sources increases, so majority rule is also not monotonic for $p < 1/2$. Majority rule is also not optimal when $p < 1/2$. If the decision is binary, then the optimal decision scheme is to simply pick a decision at random, to have exactly $1/2$ probability of deciding correctly.

IV. Most Plausible Realisations

In the previous section, we studied some simplified scenarios, where all sources have a same probability of honesty. In the proposed *Majority Rule* decision scheme $\mathscr{M}_{m,p}$, all sources are treated the same in decision making. Sometimes, we may have some specific knowledge about each source and be able to evaluate their probability of honesty individually, e.g., by evaluating witness credibility [6] (see Section VII). sources with different probabilities of honesty should have different effects on decision making. Intuitively, a decision should be more inclined to feedback from a more honest source.

Say, Alice is more trustworthy than Bob and Charlie, and Alice's feedback is 0, but Bob and Charlie both say 1. How much more trustworthy does Alice need to be to make it so that the decision scheme should pick 0? Typically, the focus would be on determining the probabilities of 0 and 1 being the right decisions, given the feedback. However, these probabilities depend on the strategies of the attackers, which may change after we implement our decision scheme. As before, our approach is to reason about the realisations of honesty, rather than the actual feedback.

In this section, each source has a certain probability to be honest $p_0, \ldots, p_{m-1} = \mathbf{p}$. The assumption that the probabilities are independent remains. This means that $\Delta(\mathbf{r}) = \prod_{s\in\mathbf{r}} p_s \cdot \prod_{s\in\overline{\mathbf{r}}}(1-p_s) = \delta_{m,\mathbf{p}}$.

The *Most Plausible Realisations* decision scheme $\mathscr{R}_{m,\mathbf{p}}$, is designed to be optimal when dealing with independent sources each with their own probability of being honest. Our argument that the most plausible realisations decision scheme works is purely based on the robustness and optimality of the scheme (see Section IV-A). The motivation given below is just to provide an intuition. It uses the notion of plausibility.

Suppose we have three sources providing feedback $\mathbf{f} = (f_a, f_b, f_c)$. Alice says 0 ($f_a = 0$), whereas Bob and Charlie say 1 ($f_b = f_c = 1$). If 0 is the correct decision, then Alice reported the truth while the others lied. We use product $p_a{\cdot}(1{-}p_b){\cdot}(1{-}p_c)$ to capture the *plausibility* that 0 is the truth. The formula represents the (prior) probability of the realisation $\mathbf{r}$ where Alice is honest, and Bob and Charlie are not. This realisation $\mathbf{r}$ is the most probable realisation such that $\mathbf{f} \in a(0,\mathbf{r})$, and as such, the most plausible explanation. Similarly, let product $(1-p_a){\cdot}p_b{\cdot}p_c$ denote the plausibility that 1 is the truth. The most plausible realisation is, therefore, the realisation with the highest product. So if $p_a{\cdot}(1{-}p_b){\cdot}(1{-}p_c) > (1-p_a){\cdot}p_b{\cdot}p_c$, the scheme selects 0; and if $p_a{\cdot}(1{-}p_b){\cdot}(1{-}p_c) < (1-p_a){\cdot}p_b{\cdot}p_c$, it selects 1.

Again, sources that have a p-value below $1/2$ are not helpful. The decision scheme will simply ignore those sources. Hence, $p_i : p_i < 1/2$ will not occur in the formula of the computation for plausibility, and we assume $\exists s \in \mathcal{S} : p_s \geq 1/2$.

Definition 10 (Plausibility)**.** *The plausibility of $d \in \mathcal{Q}$ being the correct decision, given feedback $\mathbf{f}$ is defined as:*

$$g(\mathbf{f}, d) = \prod_{s\in\mathcal{S}\ :\ \mathbf{f}_s=d, p_s\geq 1/2} p_s \cdot \prod_{s\in\mathcal{S}\ :\ \mathbf{f}_s\neq d, p_s\geq 1/2} (1-p_s)$$

The decision scheme selects the most plausible decision. If no decision is the most plausible, we follow source 0. Note that, as with majority rule, in the case of a tie, any non-manipulable tie-breaker will do. There is no advantage in selecting "smart" tie-breakers, as they are equal in robustness at best, as the choice of following 0 is an optimal one:

Definition 11 (Most Plausible Realisations Decision Scheme)**.**

$$\mathscr{R}_{m,\mathbf{p}}(\mathbf{f}) = \begin{cases} \operatorname{argmax}_{d\in\mathcal{Q}} g(\mathbf{f}, d) & \text{if defined} \\ f_0 & \text{otherwise.} \end{cases}$$

Similarly as in Definition 9, $m, \mathbf{p}$ are parameters of a decision scheme that decides the context. We let $\mathscr{R}$ represent a family of such decision schemes, called the *most plausible realisation decision mechanism*.

A typical way of using sources is to aggregate them as a weighted sum, where the weight is determined by the trustworthiness of the source. Definition 11 can be restated as a weighted sum:

Lemma 4. *Let* $w(p_s) = \max(log(\frac{p_s}{1-p_s}), 0)$.

$$\mathscr{R}_{m,\mathbf{p}}(\mathbf{f}) = \begin{cases} \operatorname{argmax}_{d\in\mathcal{Q}} \sum_{s\ :\ \mathbf{f}_s=d} w(p_s) & \text{if defined} \\ f_0 & \text{otherwise.} \end{cases}$$

Proof. The argmax of a function and its logarithm are the same, as logarithm is increasing. Together with $g(\mathbf{f}, d) \propto \frac{g(\mathbf{f},d)}{\prod_{s\in\mathcal{S}}(1-p_s)}$, it suffices to show that $\log\left(\frac{g(\mathbf{f},d)}{\prod_{s\in\mathcal{S}}(1-p_s)}\right) = \log\left(\prod_{s\in\mathcal{S}\ :\ \mathbf{f}_s=d, p_s\geq 1/2}\left(\frac{p_s}{1-p_s}\right)\cdot\prod_{s\in\mathcal{S}\ :\ \mathbf{f}_s\neq d, p_s\geq 1/2}\left(\frac{1-p_s}{1-p_s}\right)\right) \propto \sum_{s\ :\ \mathbf{f}_s=d, p_s\geq 1/2} w(p(s)) = \sum_{s\ :\ \mathbf{f}_s=d} w(p(s))$. □

A. Properties of Most Plausible Realisations

In the context where sources have different probabilities of being honest, majority rule is no longer typically optimal. The decision scheme introduced in this section – most plausible realisations – is optimal, as shown in this section. It is also monotonic, and has a simple formula computing its robustness:

Theorem 6. *$\mathscr{R}_{m,\mathbf{p}}$ is ϵ-robust for:*

$$\epsilon \geq \sum_{\mathbf{r}\in\mathcal{R}|\delta_{m,\mathbf{p}}(\mathbf{r})\geq\delta_{m,\mathbf{p}}(\overline{\mathbf{r}})} \delta_{m,\mathbf{p}}(\overline{\mathbf{r}})$$

Proof. Honest sources report c. According to Definition 10, given a realization $\mathbf{r}$, if attackers want to maximize the plausibility $g(\mathbf{f}, d)$ of an option d, $d{\neq}c$, then they all need to report it. And $\operatorname{argmax}_{\mathbf{f}\in a(c,\mathbf{r})} g(\mathbf{f}, d) = \delta_{m,\mathbf{p}}(\overline{\mathbf{r}})$. The plausibility of the honest option is $g(\mathbf{f}, c){=}\delta_{m,\mathbf{p}}(\mathbf{r})$. Now if $\delta_{m,\mathbf{p}}(\overline{\mathbf{r}}) < \delta_{m,\mathbf{p}}(\mathbf{r})$, then it means no dishonest option can be more plausible than the correct decision in the entire attack space. And the correct decision will always be chosen. Hence, all realizations $\mathbf{r}$ satisfying this inequality are non-manipulable and they sum up to $1 - \mathfrak{E}(\mathscr{R}_{m,\mathbf{p}})$ (Theorem 2). □

The most plausible realisation decision scheme is optimal. No scheme is more robust, when they are given the same set of sources which are independently honest.

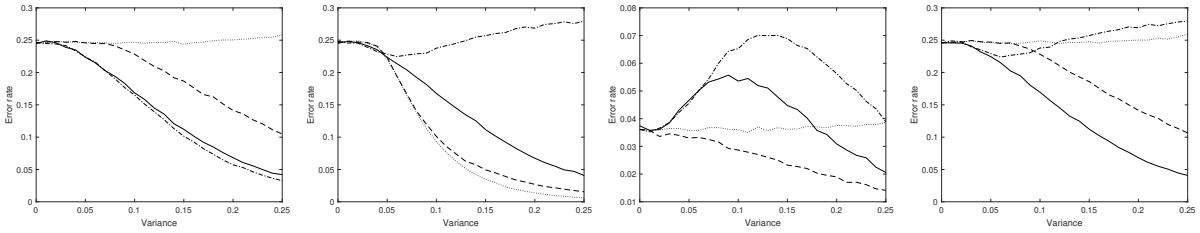
Theorem 7. *Given m sources with $\mathbf{p}$ as the probability of being honest, $\mathscr{R}_{m,\mathbf{p}}$ is optimal.*

Proof. First, $|\widehat{\mathscr{R}}_{m,\mathbf{p}}| = 1/2|\mathcal{R}|$ follows from Corollary 1. From $\forall_{R\in\mathcal{R}\ :\ \mathbb{A}(R)}\left(|R| \leq |\widehat{\mathscr{R}}_{m,\mathbf{p}}|\right)$, it follows $|\mathcal{R} \setminus \widehat{\mathscr{R}}_{m,\mathbf{p}}|$ is minimal. $|\mathcal{R} \setminus \widehat{\mathscr{R}}_{m,\mathbf{p}}|$ equals $\mathfrak{E}(\mathscr{R}_{m,\mathbf{p}})$. □

Below, we prove that decision mechanism $\mathscr{R}$ is monotonic.

Theorem 8. *Let $\mathbf{p} = p_0, \ldots, p_{m-1}$ and $\mathbf{p}' = p_0, \ldots p_m$. If $p_m > 1/2$, then $\mathfrak{E}(\mathscr{R}_{m,\mathbf{p}}) > \mathfrak{E}(\mathscr{R}_{m+1,\mathbf{p}'})$. If $p_m \leq 1/2$, then $\mathfrak{E}(\mathscr{R}_{m,\mathbf{p}}) = \mathfrak{E}(\mathscr{R}_{m+1,\mathbf{p}'})$.*

Proof. Observe $\delta_{m,\mathbf{p}}(\mathbf{r}) = p_m \cdot \delta_{m,\mathbf{p}}(\mathbf{r}) + (1-p_m)\cdot\delta_{m,\mathbf{p}}(\mathbf{r}) = \delta_{m+1,\mathbf{p}'}(\mathbf{r}\cup\{m\}) + \delta_{m+1,\mathbf{p}'}(\mathbf{r})$.
Therefore, the sum $\sum_{\mathbf{r}\in\mathcal{R}|\delta_{m,\mathbf{p}}(\mathbf{r})\geq\delta_{m,\mathbf{p}}(\overline{\mathbf{r}})} \delta_{m,\mathbf{p}}(\overline{\mathbf{r}})$ equals $\sum_{\mathbf{r}\in\mathcal{R}|\delta_{m,\mathbf{p}}(\mathbf{r})\geq\delta_{m,\mathbf{p}}(\overline{\mathbf{r}})}(\delta_{m+1,\mathbf{p}'}(\overline{\mathbf{r}}) + \delta_{m+1,\mathbf{p}'}(\overline{\mathbf{r}}\cup\{m\}))$. Every realisation or its complement is in the sum. Via Theorem 7, that sum is at least $\sum_{\mathbf{r}\in\mathcal{R}'|\delta_{m+1,\mathbf{p}'}(\mathbf{r})\geq\delta_{m+1,\mathbf{p}'}(\overline{\mathbf{r}})} \delta_{m+1,\mathbf{p}'}(\overline{\mathbf{r}})$
□



(a) Coordinated attack. (b) Devious attack. (c) Random attack. (d) Worst-case attack.

Fig. 2: Graphs depicting the sensitivity of ϵ to the quality of the estimate.

V. Analysis

We have introduced two decision mechanisms that have a form of provable robustness: Majority rule ($\mathscr{M}$) and Most Plausible Realisation ($\mathscr{R}$). In the analysis, we used a probability distribution Δ over the possible ways in which people are honest or malicious. For the first mechanism, we assumed that all sources were equally likely to be honest, and analysed the formal properties under that assumption. For the second mechanism, we assumed that all sources had individual independent probability to be honest, and analysed the formal properties under that assumption.

An interesting finding is that monotonicity by itself is insufficient to deduce that we can achieve arbitrarily low values for ϵ (see Fig. 5b). Another interesting, but not unexpected, result is that usually an attack where all malicious sources agree on the same lie is often the worst-case attack. However, in cases where sources are notoriously untrustworthy, this attack is ineffective. Finally, if we use trust-based values as estimates for the p-values involved, then we usually get a good approximation. But there seems to be a phase-shift where worse approximations suddenly become unreliable. The simulations show us interesting nuances that the aforementioned proofs does not.

A. Numerical Analysis

Fig. 2 illustrates optimality. There are four graphs, corresponding to four different attacks, and each graph contains four lines, corresponding to four decision schemes. The x-axis denotes the standard deviation of the individual honesty values p, and the y-axis the error rate. Lower lines, therefore, correspond to better decisions.

The four decision schemes that we depict contain $\mathscr{M}_{m,p}$ (dotted) and $\mathscr{R}_{m,\mathbf{p}}$ (solid), but also two example decision schemes: probability weighted sum ($\mathscr{P}$, dashed) and trust-distrusted weighted sum ($\mathscr{T}$, dash-dotted). Both decision schemes are based on straightforward approaches to aggregating information from sources with a certain trust value (the related approaches are discussed in Section VII). In the case of $\mathscr{P}$, we simply sum the p-values of the sources reporting a certain value, and pick the highest. In the case of $\mathscr{T}$, we convert the probability p, to a trust-distrust value in $[-1, 1]$, by taking $2 \cdot p - 1$. These trust-distrust values are summed, and the highest value wins. So if the result is a_1 votes A and a_2, a_3 vote B, with $p_1 = 0.9, p_2 = 0.7, p_3 = 0.4$, then $\mathscr{P}$ picks B, since $0.9 < 0.7 + 0.4$, but $\mathscr{T}$ picks A, since $0.8 > 0.4 + -0.2$. If a_i reports A and p_i is small, then in $\mathscr{P}$, $r_i = A$ makes the decision A slightly more likely, but in $\mathscr{T}$, it will actually reduce the likelihood that A is the decision.

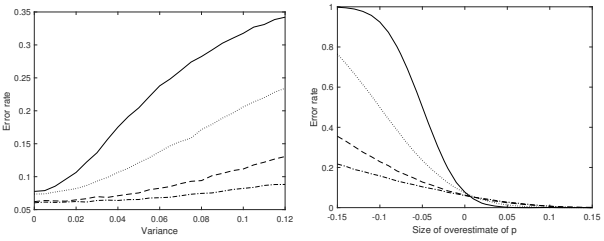
The three attacks we depict are Fig. 2a: *coordinated*, where all attackers always report the same lie; Fig. 2b: *devious*, where all trusted attackers report the same lie but untrusted attackers report the truth; Fig. 2c: *random*, where all attackers randomly lie; And Fig. 2d, we apply whichever attack that yields the highest error rate for the scheme.

Each of the simulations uses $100,000$ runs per data point, 11 sources with average p-value 0.6. The choice of a relatively small set of sources with relatively low honesty is to exaggerate the effect that specific attacks have on specific decision schemes. The p value for each individual is picked from a normal distribution with mean 0.6 and standard deviation at the x-axes. y-axis depicts the total rate of deciding incorrectly.

In Fig. 2a, we see that $\mathscr{R}_{m,\mathbf{p}}$ typically does not actually have the lowest error rate under the coordinated attack. An attacker known to be untrustworthy will lie, meaning the decision scheme could improve its decisions by doing the opposite. The $\mathscr{T}$ scheme does this, and marginally outperforms $\mathscr{R}_{m,\mathbf{p}}$. However, a smart attacker would observe the decision scheme that is being used, and alter its attack strategy in response.

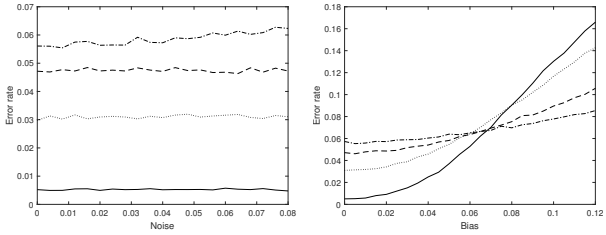
By switching to the devious attack, Fig. 2b, the performance of $\mathscr{T}$ severely degrades, since untrustworthy attackers successfully manipulate the decision scheme by reporting the truth, which $\mathscr{T}$ interprets as a lie. In $\mathscr{P}$, the report of an untrustworthy attacker is simply discounted proportionally. Since an untrustworthy attacker always tells the truth (either because he is honest, or because he is devious), a positive weight for their vote helps decrease the error rate. In majority rule, the report of an untrustworthy attacker is not even discounted, and its positive impact on the error rate is therefore even greater. Again, if either of these two decision schemes is implemented, then the attacker can simply apply the coordinated attack, where these two schemes do not perform well.

It is no coincidence, therefore, that the graph where the attacker chooses the worst-case attack based on the decision scheme (Fig. 2d), that the graph of $\mathscr{P}$ is similar to the one in Fig. 2a, and the graphs of $\mathscr{T}$ and $\mathscr{M}_{m,p}$ similar to those in Fig. 2b. The choices of $\mathscr{R}_{m,\mathbf{p}}$ are independent of the actions



(a) Per precision of estimate. (b) Per bias of estimate.

Fig. 3: Effect of parameter estimate on error rate of $\mathscr{M}$.



(a) Precision of trust opinions. (b) Bias of trust models.

Fig. 4: Effect of parameter estimate on error rate of $\mathscr{R}_{m,\mathbf{p}}$

of an untrustworthy source, so its graphs are actually the same in Figures 2a, 2b and 2d. As predicted by Theorem 7, its error rate is the lowest under the worst-case attack.

Finally, we see that the overall error rate in the random attack (Fig. 2c) is vastly lower than the error rate in any other attacks. The attacker is wasting potential by spreading lies over 4 different options, which is why it is much less likely that the decision scheme will err. The reason why $\mathscr{T}$ and $\mathscr{R}_{m,\mathbf{p}}$ are unimodal, is because they tend to favour trustworthy individuals over blocks of relatively untrustworthy sources. As the standard deviation goes up, the probability that there is a trustworthy individual goes up – and 10% of people with a p-value of 0.9 lie. But as the standard deviation continues to go up, the probability of having multiple trustworthy individuals goes up, counteracting the occasional trusted attacker. Note that the random attack, as a consequence, is not particularly appropriate to use as a basis for comparison.

B. Robustness under Estimated Honesty

To determine whether to follow the majority and how much robustness (ϵ) we can get, we need to know the probability that an source is honest (p), and also whether that $p{>}1/2$. The probability that an arbitrary unknown source is honest, is equal to the frequency of honest sources within the population of sources. It is fair to assume that by performing statistical analysis, the system can obtain a reasonable approximation of p. Obtaining such an estimate is out of the scope of this paper. However, our approach is only useful, if it is not overly sensitive to errors in the approximation of p.

In Fig. 3, we show the findings of two Monte Carlo simulations with $100,000$ runs. In Fig. 3a, we plot the change of error rate going from precise estimations of p to imprecise estimations. Whereas, in Fig. 3b, we plot the change of error rate going from underestimating p to overestimating p. The solid lines are: $m = 201, p = 0.55$; dotted are: $m = 51, p = 0.6$; dashed are: $m = 13, p = 0.7$; and dash-dotted are: $m = 3, p = 0.85$.

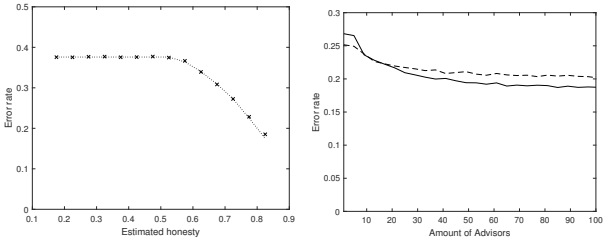
Fig. 3a models the precision of the estimations, by letting the probability p of honesty be selected from a normal distribution whose mean is the estimated p value. The standard deviation of the normal distribution increases along the x-axis. The assumption here is that the method to obtain an estimate is not biased towards overestimating or underestimating, meaning the mean of the normal distribution is equal to the estimated p value. As the SD increases, the quality of our estimate degrades, and we see that the error rate goes up. This effect is particularly pronounced when the p value is low (e.g. for the solid line), because it becomes increasing probable that the honest sources are actually outnumbered.

Fig. 3b models the effect of consistently over/underestimating p. The x-axis shows how much higher the actual probability of honesty is, compared to our estimate. On the left side, we have negative x-values, meaning the actual probability that an sources is honest is lower than our estimate. The parameters used for the different shape lines is the same as in Fig. 3a. The graphs match parts of those found in Fig. 1a, which is unsurprising, considering that the probability that the majority is right is completely determined by m and p. The most important observation here, is that underestimating p puts one in a situation where the error rate can only be lower than expected.

Trust opinions come in many formats, trust opinions may not be probabilities, or trust opinions may be defined by a network of probabilities. In Subjective Logic [7], trust opinions are an example of the former, where a quantity of uncertainty is added. SALE POMDP [8], is an example of the latter, where trust opinions are parameter estimations of partially observable Markov decision processes. In both models, and in fact most trust models in general, we can obtain a value that represents the probability that an source is honest in a specific situation. Since we are not interested in the mechanism behind the trust model, and we want to talk about trust models with generality, we simply refer to the p_i-value that represents the model's predicted probability of honesty of the source a_i as the trustworthiness of a_i.

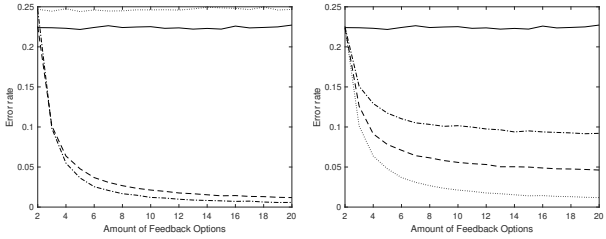
Using simulations, we can investigate what happens if the value of p that is used by the decision scheme (i.e. the trust opinion) is different from the actual probability that an source is honest. Each of the Monte Carlo simulations uses $100,000$ runs for every datapoint. The value of $n = 5$, unless stated otherwise. Its value typically is not important (see Fig. 6). Throughout the discussion, we use "trustworthiness" or "trust opinion" to refer to what the decision scheme believes the value of p is; and the actual value of p is referred to simply as the probability that the source is honest.

In Fig. 4 the trustworthiness of the individual sources



(a) Wrong perception of p. (b) A limit on error rate.

Fig. 5: Two unrelated graphs with unexpected results.



(a) Attacks on $\mathscr{M}_{m,p}$ and $\mathscr{R}_{m,\mathbf{p}}$. (b) Per size of coalition.

Fig. 6: Effect of increasing the options for feedback.

should be different, but we do not want to enforce any particular values. Therefore, the individual trustworthiness of each source is chosen from a normal distribution with mean p, and standard deviation 0.08. In both graphs, the actual probability of honesty differs from the trustworthiness in some way. The solid lines are: $m = 201, p = 0.55$; dotted are: $m = 51, p = 0.6$; dashed are: $m = 13, p = 0.7$; and dash-dotted are: $m = 3, p = 0.85$.

Fig. 4a models the precision of the individual trust opinions with a normal distribution. The actual probability that sources are honest in a simulation run, is a sample from this distribution. The assumption here is that the method to obtain an estimate is not biased towards overestimating or underestimating, meaning the mean of the normal distribution is equal to the estimated p value. The standard deviation of the normal distribution is depicted on the x-axis, with the correct estimate on the left hand side. For the first three lines, the graphs are flat. The trust opinions are allowed to be poor estimates (e.g. $p = 0.6 \pm 0.08$), and it will not affect the quality of the decision scheme. The fourth line, on the other hand, has a minute increase. The error rate goes up slightly, because there is a reasonable probability that an source has a probability of being honest larger than 0.5, but the trust opinion is below 0.5, and the source is ignored in the decision scheme.

Fig. 4b models the effect of a biased estimate, where the trust opinions may over/underestimate the average p-value. The average probability of honesty is selected from a normal distribution with the mean being the average trustworthiness, and the standard deviation is on the x-axis. This models the possibility that the whole process of obtaining trust opinions is biased, and optimistic or pessimistic. For high p values, the bias has some effect, but even with a standard deviation of 5 percent points, the error rate only goes up about a percent point, for the graphs with $p = 0.7$ and $p = 0.85$. For low p values, the bias can have a large effect. The reason is that, for $p = 0.55$ and a standard deviation of 0.06, the possibility that the majority of the raters are actually malicious (but with positive trustworthiness) less than a standard deviation away. A situation where the trustworthiness is positive, but the probability of honesty is smaller than $1/2$ is disastrous for our scheme. For security critical purposes, therefore, it is important to ensure such a situation is highly improbable.

In Fig. 5a, the trust opinion is formed based on the actual probability of honesty, and the simulation is run with that probability of honesty (but the decision scheme uses the trust opinion). The results corroborate those from Fig. 4a. The scenario is as follows: There are 4 users, a special source whose p value varies, two sources with $p = 0.6$ and one source with $p = 0.55$. The simulation performs a multitude of runs for every true p value with increments of 0.05. Each run computes the estimated p value (which comes from a normal distribution with standard deviation 0.1), and if the decision scheme decides (in)correctly, this is counted for the estimated p value. The x markers are placed at places where the decisions are counted. The dotted line is the theoretical prediction, if the trust values were always equal to the probability of honest. We can see that the decision scheme performs exceptionally close to the theoretical maximum, despite an error of 0.1 on our trust opinion being just one standard deviation.

It would seem logical that any ϵ-robustness can be reached, simply by pumping p and m. Figures 1b show that increasing m will lower the error rate asymptotically to 0. However, in both graphs the (expected) value for p is kept constant. There are realistic scenarios one can think of, where this is not the case. Often, users simply want to minimise the amount of feedback they need to gather, and an easy way is to start by requesting feedback from trustworthy sources first. But that implies that the next sources on the list to provide feedback will be less trustworthy!

In Fig. 5b, we draw the error rate, as increasingly less trustworthy sources are added to the list. As sources with a p value below $1/2$ are ignored, we ignore these values in the simulation too. So the decreasingly trustworthy sources will still have p values over $1/2$. The sequences of p values of new sources are (solid) $\frac{1}{2} + \frac{3}{m+12}$ and (dashed) $\frac{1}{2} + \frac{2}{m+7}$, which start at 0.731 and 0.750, and asymptotically decrease to 0.500. We see the error rates have asymptotes at 0.19 and 0.20, respectively.

Finally, in Fig. 6, we look at the number of options for feedback. So far, all simulations used $n = 5$. And the reason is that, under the coordinated attack, the value of n does not matter. The reason is that the attacker only ever picks 1 other value. For $\mathscr{R}_{m,\mathbf{p}}$, the coordinated attack is the attack that maximises the error rate, so this attack was used in the simulations. Fig. 6a depicts the fact that n does not matter

under the coordinated attack.

In Fig. 6a, the coordinated attack on $\mathscr{R}_{m,\mathbf{p}}$ is solid, the random attack on $\mathscr{R}_{m,\mathbf{p}}$ is dashed, the coordinated attack on $\mathscr{M}_{m,p}$ is dotted and the random attack on $\mathscr{M}_{m,p}$ is dash-dotted. The value of n has no impact on the error rate under coordinated attacks. But for random attacks, the error rate rapidly declines. As attackers spread their answers over different lies, the true value will stand out.

In Fig. 6b, we introduce a generalisation of the coordinated attack, where there is a coalition of attackers providing the same answer, and non-coalition members lie randomly. The probability of membership of the coalition is 1 (solid), 0.6 (dash-dotted), 0.4 (dashed) and 0 (dotted). The decrease of error rate is not linear w.r.t. the size of the coalition.

VI. Discussion

In this paper, we build a foundation for making decisions while resisting manipulation from malicious sources in a probabilistic way. Our results are derived based on the assumptions of an idealistic world. In the following, we discuss the practical implication of our assumptions. We start with the strong assumption of honesty and then other assumptions.

Our goal is to reason more effectively about information from potentially malicious sources. The idea is to reason about realisations and design a decision scheme; rather to decide based on concrete feedback in an ad-hoc way. The majority rule and most plausible realisations decision schemes demonstrate feasibility in an academic setting (where ad-hoc decisions are not). There are applications where our assumptions are reasonable, as we argue in this section. However, for many applications, domain-specific assumptions may be necessary. We argue, in this section, that these domain-specific assumptions do not typically form a hinder for building a robust decision mechanism based on realisations.

A. Models of Honesty

The robustness, optimality and monotonicity of $\mathscr{M}_{m,p}$ and $\mathscr{R}_{m,\mathbf{p}}$ have been formally defined based on the strong assumption of honesty. Recall that the strong assumption of honesty is that honest sources' feedback is equivalent to the correct decision. In this section, we consider the assumption of honesty in three classes of applications i.e., those where it is reasonable, those where it works as a modelling trick, and those where a weaker assumption is more appropriate. For the last case, we also look at what changes may be required.

Trusted third parties in security protocols are an example where the strong assumption of honesty is typically reasonable. Such a protocol prescribes the response (feedback) of the trusted third party, and it prescribes how the response should be used. For example, it is reasonable to assume that if a certificate authority is honest, then the link that their certificate suggests between a public key and a name is genuine. Looking through the lens of our approach, robustness of the public key infrastructure supporting the Web is obtained by having extremely high p values. Compromised certificate authorities are seldom trusted by browsers in default settings; DigiNotar [9] being a notable exception. The Web of Trust – introduced to support Pretty Good Privacy [10] – is an example of a public key infrastructure without certificate authorities. Instead, other users sign certificates linking public keys to identities; they know the person uses the public key via an offline personal connection. An honest participant is a participant who actually verifies what they sign, and is unlikely to mistakenly link a person with a wrong/different public key.

There are applications where the link between correct feedback and the right decision is straightforward, but honest sources sufficiently often fail to provide the correct feedback for the assumption of honesty to be a realistic idealisation. For example, network nodes sending routing messages may unwittingly send incorrect information (e.g., due to malfunctioning), or copyright protection using image recognition may fail to recognise (or spuriously recognise) infringement. A subtle change in semantics may be sufficient to be able to apply our results here. We can let the probability p mean "honest and accurate" and $1-p$ "malicious or mistaken". In the case of a node on a network, we can consider a node to have probability p to provide accurate information about the network, and a probability $1-p$ of being malicious, mistaken or mislead. Unlike a malicious source, a mistaken source does not have an associated attack space, but selects a specific value (or distribution) from the attack space. It follows from Definition 4 that replacing a malicious source by a mistaken one in a realisation will not make it manipulable. Hence, the results on robustness and monotonicity properties remain applicable. However, optimality may not be, as eliminating possibilities in the attack space may give rise to a better way of deciding (see example at the end of this section).

There are many domain-specific assumptions about honesty that we can make for given applications. For malware reporting, if an app is updated, older honest feedback only refers to the previous version. The more updates there are since the feedback, the more likely that the honest feedback no longer corresponds to the correct decision. But also, feedback that an app is safe is more likely to become mistaken than feedback that an app is malicious. In vehicular networks, an understanding of how traffic works needs to be built in. Some traffic information can be quickly outdated (e.g., collisions, speed traps), and some can be persistent (e.g., a new bridge or speed radar). Using the reported speeds of the GPS of road users can be a smart way to determine whether there is heavy traffic. A general way to aggregate such information, is have a stochastic relation between the correct decision and honest feedback (i.e. a probability of reporting a certain thing). Our model would need to be extended with one more probability distribution γ, determining the probabilities of honest feedback in a certain situation. Robustness is then:

$$\sum_{\mathbf{r}\in\mathcal{R}} \sum_{\mathbf{f}\in a(c,\mathbf{r})\wedge\mathscr{D}(\mathbf{f})\neq c} \Delta(\mathbf{r})\cdot\beta(\mathbf{r},c)(\mathbf{f})\cdot\gamma(\mathbf{r},c)(\mathbf{f})$$

To illustrate this further, assume we have light sensors that can distinguish red, green and blue. Assume the probability of a sensor being malicious is 25%, being mistaken is 35%,

and being correct is 40%. We can use the modelling trick of letting $p = 0.4$, and obtain a negative result: we cannot achieve meaningful robustness, since $p < 1/2$. However, we happen to know – for the sake of the example – that the sensors' mistakes are predictable: it always misreports red as green, green as blue, and blue as red. Rather than using the majority rule, the decision scheme could be to identify the least reported color, and select red/green/blue if it is blue/red/green, respectively. This rule results in the right decision for realisations where malicious sources are the smallest group – which is quite plausible, and increasingly plausible as we add more sensors. Therefore, this decision scheme clearly outperforms majority rule with $p = 0.4$, hence our earlier claim that this approach does not preserve optimality.

B. Realisation-based Decisions

Secure secret sharing [11] is an example where an assumption similar to the strong assumption of honesty does apply. In (t, n) secret sharing, there are n participants that communicate with each other, and, if at least t participants are honest, then they eventually know the shared secret. For an honest participant, $t-1$ out of the $n-1$ remaining others need to be honest. In scenarios where secret sharing is applied routinely (e.g. distributed pseudo random number generation [12]) a lower bound probability of successful secret sharing is required. Interestingly, the success of secure secret sharing protocols does not hinge on the attack space not containing misleading values, but on it being computationally difficult to find misleading values in the enormous attack space. This means that even Theorem 1 does not apply to secret sharing. Nevertheless, as demonstrated in [12], reasoning about realisations remains an effective way of accomplishing this. Reasoning based on realisations is also applied to anonymity networks like TOR [13].

An important question is whether reasoning about realisations is also a useful endeavor when the probability of honesty of sources is *not* independent. In a Sybil attack, an attacker controls multiple sources, and uses them in a coordinated way. An aspect is that malicious sources' feedback is coordinated, and this aspect is covered by our model. Another aspect of the Sybil attack, is that the attacker tries to ensure that multiple sources are malicious. If the sources are selected by the decision maker, then this may or may not be possible. But if the sources offer their information to the decision maker, then it is trivial to ensure all Sybils are included, and independence does not hold. To capture Sybil attacks in these cases, we must go beyond choices for Δ where p-values are independent.

One way to deal with dependent p values, is to use over approximation, which our simulations suggest is safe. A single source is independent by definition. Introducing a second node, we have the probabilities $P(s_2 \in \mathcal{R}|s_1 \in \mathcal{R})$ and $P(s_2 \in \mathcal{R}|s_1 \notin \mathcal{R})$, which are not necessarily equal. However, we can safely select p_2 as the minimum of these values, ensuring that if the robustness is computed with independent p_1 and p_2, then it is an overestimate. This strategy is not optimal.

None of the definitions and theorems in Section II use the notion that sources are independent, and are defined for general distributions of realisations Δ. A brute-force approach could go through all attainable sets of realisations and find the one where honesty states of sources are most probable. Unfortunately, the number of attainable sets of realisations grows exponentially. More study is required to determine whether finding the optimal decision scheme is computationally feasible, or alternatively, whether effective heuristics exist to get near optimal decision schemes.

The techniques used in this paper can be used to prove different properties and theorems, if the assumptions are changed appropriately. We hope that our approach helps develop more formal and robust ways of making trust-based decisions.

VII. Related Work

The problem of malicious feedback (a.k.a unfair/fake ratings) has been popularly studied in application and research domains such as e-commerce [14]–[20], web service [21], [22], trust and reputation systems [23]–[25], multi-agent systems [26]–[28] and recommendation systems [21], [29]. In this section, we present related works dealing with malicious feedback and decision making under it.

Trust forms the foundation for information sharing. For example, feedback from more trustworthy peers are usually considered more reliable. In the literature, filtering (or discounting) feedback based on its providers' honesty (a.k.a advisor honesty/witness trust) is a popular way of treating malicious feedback [26].

There are various factors to measure honesty of feedback sources. One of them is similarity between feedback and the self experience of a decision maker, the use of which can be seen in early works such as [14], [26], [30] and also recent ones [20], [27], [31]. In [14], [20], [31], clustering algorithms are applied to distinguish malicious sources from honest ones. Feedback identified as unfair would be filtered out. For example, Liu et al. propose to cluster feedback, and sources whose feedback belongs to the same cluster with the decision maker would be considered reliable [31]. In [27], Weng et al. propose to use the statistical correlation between the history feedback of a source and the decision maker's experience, to determine the credibility of the source. Only feedback from sources whose credibility is higher than an advisee's own confidence is aggregated, by weighted average, where the weights depend on credibility.

Besides similarity, some other criteria (e.g., feedback timestap, social relation) are also used to distinguish malicious sources. Both Yang et al. [23], Liu and Sun [32] propose to detect suspicious time intervals where attacks are more likely, and highly suspicious feedback would be abandoned. In [19], a source is considered malicious if he has reviewed two or more products targeted by crowdsourcing requests in a short while. It is assumed in [29] that similarities between malicious sources are higher than that between honest ones. In [20], the correlation among feedback criteria is considered e.g., high score for quality and low score for service time may occur

simultaneously for an honest source, while such correlation is assumed not true for an attacker.

In [28], to determine the degree of importance and reliability of feedback in multi-agent systems, Sabater and Sierra propose to exploit social relations among agents. For instance, if an advisor is found to have high level of cooperation with the agent he provides feedback, then that feedback is suspected of being biased and his social trust is assigned low. In [16], to detect fraud online reviews, Akoglu et al. build a network for users, reviews and products, and propose a network classification algorithm to label attackers and fake reviews. It is assumed that attackers would more probably provide positive (negative) review for bad (good) products compared with honest sources. This assumption is also applied in [18], where detecting fake review is also formulated as a network-based classification problem, but compared with [16], more metadata such as review texts, timestamp, relational data are considered.

Our decision scheme also relies on probability of honesty of sources. A crucial difference between our approach and the aforementioned is that there is no need to detect or filter out malicious feedback before aggregation. We have already proved that it is fallacious to discount or filtering feedback that deviate from the majority or from the first-hand evidence [33]. Filtering out deviated feedback may cause confirmation bias.

In the existing approaches, there are often assumptions about the characteristics of attacks, which make them reactive. For example, attackers are characterized as providing unfairly highly low (bad-mouthing attacks) or high (ballot-stuffing attacks) ratings are considered in [14], [16], [18], [21]. Such characterization restricts attacker behaviour to specific assumed types, ignoring other possibilities. In practice, attackers are usually adaptive: updating their strategies regarding the changes in the decision scheme or in the system. As a result, assumptions about strategies are incompatible with robustness. Therefore, to achieve robustness, we must have a proactive position on strategies, allowing them to be of arbitrary form.

In our model, we treat honesty of sources as given, as we focus on how to exploit these parameters in a way that it can lead us to provably accurate decisions. Hence, how to accurately evaluate source honesty is out of the scope of this work. In the interest of having a simple model, we assume that honest users would simply report the truth, while in practise it is more complicated. For example, as it is pointed out in [34], even if a user is honest, his feedback can be biased in multiple ways. It would be interesting to extend our decision scheme to cover situations where bias from honest sources is considered.

VIII. Conclusion

We investigated how to provably make correct decisions with high probability, using potentially malicious feedback. Our model assumed that feedback can be modeled as discrete options, one of which is reported by honest sources. The influence the feedback has on our decisions is determined by the (perceived) probability that an source is honest.

We defined two decision mechanisms: Majority Rule and Most Plausible Realisations. We defined three properties for a desired decision scheme based on unreliable feedback: robustness, monotonicity, and optimality. Regarding robustness, we proved that for both the decision mechanisms, the probability of making incorrect decisions is bound to a very small threshold, regards of what attackers report. The robustness of both the schemes monotonically increases with the number of sources whose honesty degree is over a half. Given a feedback scenario, the robustness of both the schemes is optimal, meaning there's no scheme with better robustness.

We rely on knowledge about the honesty of sources, which might be inaccurate in practise. Hence, we run simulations to test how sensitive the decision schemes are to the deviation that estimated trustworthiness has from the actual probability of honesty. We found that if our estimate is imprecise, but not biased, then it has no effect on robustness; and if it is imprecise and biased, it has limited effect. An exception is when sources that are probably malicious are considered trustworthy, in which case robustness is quickly out the window.

Using simulations, we also provided insight into how attacks and decision schemes relate. A decision scheme that attempts to exploit certain attack strategies is vulnerable. The two presented optimal decision schemes ignore untrustworthy sources completely as a result.

This work aims to improve the robustness of decision making under unreliable information sources. The robustness of such decision making is crucial especially when it is for security domains. An incorrect decision can put a system under threat. Besides the concrete schemes that we formulated, the core contribution is the demonstration and application of a novel technique to reason about manipulation. We introduced the notion of realisations, which made it possible to investigate whether one is manipulable, *without* studying the actual manipulative behaviour. Large summations of combinations of feedback that affect the actual decision could be cancelled out of the formula, by applying the idea of realisations.

Besides the obtained formal results, this work serves as a proof-of-concept for an alternative way of considering trust in the security domain. Rather than focussing on making the right decision with some given feedback, our approach takes a step back and asks under which circumstances do we want to make the right decision. Typically, we want to make the right decision under the most probable circumstances.

In this paper, we introduce two fairly simple schemes. We believe that the technique can be extended to more complex scenarios, as we address in the discussion section. In particular, the weak assumption of honesty should be sufficient to arrive at similar conclusions, and independency of sources being honest may not be necessary to obtain positive results either. An aspect that we have not yet studied is when sources provide feedback about multiple things. The next step is to apply the technique to an existing system providing us with the right parameters.

REFERENCES

[1] A. M. Pushpa, "Trust based secure routing in AODV routing protocol," in *2009 IEEE International Conference on Internet Multimedia Services Architecture and Applications (IMSAA)*. IEEE, 2009, pp. 1–6.

[2] C. Johnson, M. Badger, D. Waltermire, J. Snyder, and C. Skorupka, "Guide to cyber threat information sharing," National Institute of Standards and Technology, Tech. Rep., 2016.

[3] M. Li, X. Sun, H. Wang, Y. Zhang, and J. Zhang, "Privacy-aware access control with trust management in web service," *World Wide Web*, vol. 14, no. 4, pp. 407–430, 2011.

[4] K. Scarfone, W. Jansen, and M. Tracy, "Guide to general server security," *NIST Special Publication*, vol. 800, no. s 123, 2008.

[5] C. M. Grinstead and J. L. Snell, *Introduction to probability*. American Mathematical Soc., 2012.

[6] D. Wang, T. Muller, Y. Liu, and J. Zhang, "Towards robust and effective trust management for security: A survey," in *Trust, Security and Privacy in Computing and Communications (TrustCom), 2014 IEEE 13th International Conference on*. IEEE, 2014, pp. 511–518.

[7] A. Jøsang, R. Hayward, and S. Pope, "Trust network analysis with subjective logic," in *Proceedings of the 29th Australasian Computer Science Conference-Volume 48*. Australian Computer Society, Inc., 2006, pp. 85–94.

[8] A. A. Irissappane, F. A. Oliehoek, and J. Zhang, "A POMDP based approach to optimally select sellers in electronic marketplaces," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1329–1336.

[9] N. van der Meulen, "Diginotar: Dissecting the first dutch digital disaster," *Journal of Strategic Security*, vol. 6, no. 2, pp. 46–58, 2013.

[10] P. Zimmermann, *The Official PGP User's Guide*. MIT Press, 1995. [Online]. Available: https://books.google.co.uk/books?id=dP1SAAAAMAAJ

[11] A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, no. 11, p. 612–613, Nov. 1979. [Online]. Available: https://doi.org/10.1145/359168.359176

[12] E. Syta, P. Jovanovic, E. K. Kogias, N. Gailly, L. Gasser, I. Khoffi, M. J. Fischer, and B. Ford, "Scalable bias-resistant distributed randomness," in *2017 IEEE Symposium on Security and Privacy (SP)*. Ieee, 2017, pp. 444–460.

[13] S. J. Murdoch and G. Danezis, "Low-cost traffic analysis of tor," in *2005 IEEE Symposium on Security and Privacy (S&P'05)*. IEEE, 2005, pp. 183–195.

[14] C. Dellarocas, "Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior," in *Proceedings of the 2nd ACM conference on Electronic commerce*. Citeseer, 2000, pp. 150–157.

[15] K. Regan, P. Poupart, and R. Cohen, "Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change," in *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 2006, pp. 1206–1212.

[16] L. Akoglu, R. Chandy, and C. Faloutsos, "Opinion fraud detection in online reviews by network effects," in *Seventh international AAAI conference on weblogs and social media*, 2013.

[17] A. Post, V. Shah, and A. Mislove, "Bazaar: Strengthening user reputations in online marketplaces," in *Proceedings of the 8th USENIX conference on Networked systems design and implementation*, 2011, pp. 183–196.

[18] S. Rayana and L. Akoglu, "Collective opinion spam detection: Bridging review networks and metadata," in *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2015, pp. 985–994.

[19] P. Kaghazgaran, J. Caverlee, and M. Alfifi, "Behavioral analysis of review fraud: Linking malicious crowdsourcing to amazon and beyond," in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[20] A. A. Irissappane and J. Zhang, "Filtering unfair ratings from dishonest advisors in multi-criteria e-markets: a biclustering-based approach," *Autonomous Agents and Multi-Agent Systems*, vol. 31, no. 1, pp. 36–65, 2017.

[21] S. Wang, Z. Zheng, Z. Wu, M. R. Lyu, and F. Yang, "Reputation measurement and malicious feedback rating prevention in web service recommendation systems," *IEEE Transactions on Services Computing*, vol. 8, no. 5, pp. 755–767, 2015.

[22] Z. Malik and A. Bouguettaya, "Evaluating rater credibility for reputation assessment of web services," in *International Conference on Web Information Systems Engineering*. Springer, 2007, pp. 38–49.

[23] Y. Yang, Y. L. Sun, S. Kay, and Q. Yang, "Defending online reputation systems against collaborative unfair raters through signal modeling and trust," in *Proceedings of the 2009 ACM symposium on Applied Computing*. ACM, 2009, pp. 1308–1315.

[24] Y. Sun and Y. Liu, "Security of online reputation systems: The evolution of attacks and defenses," *IEEE Signal Processing Magazine*, vol. 29, no. 2, pp. 87–97, 2012.

[25] A. Whitby, A. Jøsang, and J. Indulska, "Filtering out unfair ratings in bayesian reputation systems," in *Proc. 7th Int. Workshop on Trust in Agent Societies*, vol. 6, 2004, pp. 106–117.

[26] W. T. L. Teacy, J. Patel, N. R. Jennings, and M. Luck, "Travos: Trust and reputation in the context of inaccurate information sources," *Autonomous Agents and Multi-Agent Systems*, vol. 12, no. 2, pp. 183–198, 2006.

[27] J. Weng, Z. Shen, C. Miao, A. Goh, and C. Leung, "Credibility: How agents can handle unfair third-party testimonies in computational trust models," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 22, no. 9, pp. 1286–1298, 2010.

[28] J. Sabater and C. Sierra, "Reputation and social network analysis in multi-agent systems," in *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*. ACM, 2002, pp. 475–482.

[29] Z. Yang, Z. Cai, and X. Guan, "Estimating user behavior toward detecting anomalous ratings in rating systems," *Knowledge-Based Systems*, vol. 111, pp. 144–158, 2016.

[30] J. Zhang and R. Cohen, "Evaluating the trustworthiness of advice about seller agents in e-marketplaces: A personalized approach," *Electronic Commerce Research and Applications*, vol. 7, no. 3, pp. 330–340, 2008.

[31] S. Liu, J. Zhang, C. Miao, Y.-L. Theng, and A. C. Kot, "iclub: an integrated clustering-based approach to improve the robustness of reputation systems," in *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*, 2011, pp. 1151–1152.

[32] Y. Liu and Y. Sun, "Anomaly detection in feedback-based reputation systems through temporal and correlation analysis," in *Social Computing (SocialCom), 2010 IEEE Second International Conference on*. IEEE, 2010, pp. 65–72.

[33] T. Muller, Y. Liu, and J. Zhang, "The fallacy of endogenous discounting of trust recommendations," in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2015, pp. 563–572.

[34] E. M. Redmiles, Z. Zhu, S. Kross, D. Kuchhal, T. Dumitras, and M. L. Mazurek, "Asking for a friend: Evaluating response biases in security user studies," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2018, pp. 1238–1255.