1-2024

Provably secure decisions based on potentially malicious information

Dongxia WANG

Tim MULLER

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

Citation

WANG, Dongxia; MULLER, Tim; and SUN, Jun. Provably secure decisions based on potentially malicious information. (2024). *IEEE Transactions on Dependable and Secure Computing*. 1-16.
Available at: https://ink.library.smu.edu.sg/sis_research/8650

Provably Secure Decisions based on Potentially Malicious Information

Dongxia Wang, *Member, IEEE,* Tim Muller, *Member, IEEE,* and Jun Sun, *Member, IEEE*

Abstract—There are various security-critical decisions routinely made, on the basis of information provided by peers: routing messages, user reports, sensor data, navigational information, blockchain updates, etc. Jury theorems were proposed in sociology to make decisions based on information from peers, which assume peers may be mistaken with some probability. We focus on attackers in a system, which manifest as peers that strategically report fake information to manipulate decision making. We define the property of robustness: a lower bound probability of deciding correctly, regardless of what information attackers provide. When peers are independently selected, we propose an optimal, robust decision mechanism called Most Probable Realisation (MPR). When peer collusion affects source selection, we prove that generally it is NP-hard to find an optimal decision scheme. We propose multiple heuristic decision schemes that can achieve optimality for some collusion scenarios.

Index Terms—Multi-source decision making, Provable decision making, Malicious feedback, Collusion attacks, Trust evaluation

1 Introduction

Online users or agents often experience 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 make informed decisions. For example, the 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. 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 such as indicators, malware reports and threat intelligence reports allows users or organisations to learn from the experience of peers, 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 harmful decisions. For example: compromised accounts providing fake reviews on malware in an app store, leading to more downloads; spurious routing messages causing network traffic to pass through a compromised device; or bogus reports of suspicious activities wrongfully blocking a service or an account. Malicious feedback can be a security threat.

The issue of the quality of crowd-wisdom in multi-source decision making is well-studied. Already in 1785, Marquis de Condorcet formulated Condorcet's Jury Theorem [4], which deals with a simple scenario of jurors stating a verdict which may be correct with some probability p. Current research expands upon these results for increasingly more realistic scenarios. A core assumption of Jury Theorems is that jurors are honest, but may inadvertently provide bad verdicts with some probability, e.g., due to insufficient knowledge or competence. Such an assumption is not suitable for an adversarial setting, where malicious sources may exist and how probable they report the truth is part of their strategies, which is neither deterministic, nor probabilistic and could even be adaptive depending on how decisions are made. Rather than modifying the Jury Theorems to encompass adversarial behaviour, we derive new results mathematically from the basic definitions. While our results bare some similarity to the Jury Theorems, there are fundamental differences as well, as we will discuss.

Our foundational idea is to consider malicious sources (attackers) *strategically* providing feedback to manipulate decision making, whereas honest sources report the option that corresponds to the correct decision. Rather than having a probability of feedback being correct, we use a probability of the source being honest. To make our results general, we do not assume specific attack strategies and consider the entire attack space. The attackers are assumed to know the decision scheme, since relying on secrecy is a poor practice for security (e.g. NIST recommends against this in systems security [5]). We aim to reason about the robustness of a decision scheme in a provable way, which is crucial especially for security-critical decisions.

There are two properties that are desired for a decision scheme. One 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 and with what probabilities. Another is *optimality*, meaning that no other decision scheme has better robustness.

The probability of receiving specific feedback depends partially on attack strategies, which in turn depend on how the feedback is used in decision making. Instead of reasoning about specific feedback, we investigate under what circumstances a correct decision can be ensured regardless of the attacker feedback, which is referred to as being *non-manipulable*. Typically, we want to be non-manipulable under the more probable circumstances. We propose to reason about which sources behind the feedback may be honest such that the feedback is generated, defined as *realisations*.

- *D. Wang is with the College of Control Science and Engineering, Zhejiang University, Hangzhou, 310020. E-mail: dxwang@zju.edu.cn*
- *T. Muller is with the University of Nottingham and J. Sun is with Singapore Management University.*

All possible feedback under a realisation defines its attack space. A scheme is non-manipulable under a realisation if it decides correctly in its entire attack space. The set of all the non-manipulable realisations determines the (guaranteed) accuracy level of a decision scheme. The goal is to have a decision scheme such that there are fewer and less probable realisations under which manipulation is possible.

We propose a decision scheme named *Most Probable Realisation (MPR)*. The idea is to always trust the most probable realisation behind the given feedback. Although MPR is relatively simple, it is proved to be robust and optimal, when sources are independent regarding how probable they are to be honest. Even in some settings without source independence, MPR remains to be optimal. When sources may be dependent, we prove that the general problem of finding an optimal decision scheme is NP-hard, with a representation of the probability distribution of realisations as an input. We explore several greedy heuristics for finding suitable decision schemes. For a specific distribution of realisations, we propose a polynomial-time solution for finding the optimal deterministic decision scheme. It is used to demonstrate that probabilistic schemes may outperform deterministic schemes.

Finally, we discuss how our work relates to some other domains of multi-source decision making such as the social choice theory, truth discovery and information fusion etc, specifically where they coincide and where they differ.

2 Secure Decisions based on Feedback

We aim to introduce a general methodology to approach feedback in a way that allows resistance to manipulation by design. If the probability that the attacker successfully manipulates the decision is less than ϵ, then we achieve ϵ*-robustness*. Our decision, therefore, is guaranteed to be correct with probability $1-\epsilon$, despite manipulation attempts. In this section, we introduce the concepts and the framework required to reason about decision circumstances, manipulation and decision accuracy etc.

2.1 Model

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. Consider an example of deciding whether to install a software based on its security property, the feedback set includes two options: "is malware" and "is not malware". Based on the knowledge about the sources, if the scheme outputs the second option, then the action "install" is selected. Formally:

Definition 1 (Decision Scheme)**.**

- *There is a set of sources* $\mathcal{S}=\{s_0,\ldots,s_{n-1}\}$.
- *There is a set of feedback options* $\mathcal{O}=\{0,\ldots,m-1\}$.
- *There is a set of decisions* $\mathcal{Q}=\{0,\ldots,\mu-1\}$. *Only one decision is correct in a decision making task.*
- *Feedback* $\mathbf{f}\in\mathcal{F}$ *is an* n*-tuple:* $\mathbf{f}=(f_0,\ldots,f_{n-1})$, *where* f_i *represents the feedback option reported by source* s_i : $s_i\in\mathcal{S}$ *and* $f_i\in\mathcal{O}$.
- *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 needs to select which decision scheme is appropriate. For example, given $\mathcal{O},\mathcal{Q}$, different schemes are required for $n=10$ and $n=100$. A *decision mechanism* selects an appropriate decision scheme, based on the context.

We use the following running example throughout the paper to demonstrate the relevant concepts and theorems.

Example 1. There are three sources $\mathcal{S}$=$\{s_0,s_1,s_2\}$ and three options $\mathcal{O}$=$\mathcal{Q}$=$\{A,B,C\}$. If s_0 provides feedback A, then $f_0=A$. If, furthermore $f_1=C$ and $f_2=B$, then we write $\mathbf{f}=ACB$ instead of (A,C,B).

Intuitively, the feedback of an honest source does not depend on what decision scheme is used and is not affected by attackers' choices. We call this the *weak assumption of honesty*. In this paper, we make a stronger simplified assumption, namely that there is a one-to-one correspondence between which decision is correct, and which feedback honest sources provide. We call this the *strong assumption of honesty*[2]. To simplify the notation, the strong assumption is modelled in the way that the feedback provided by an honest source *is* the correct decision.

For malicious sources, they are free to report any option in $\mathcal{O}$. Malicious sources are aware of what the decision scheme is. Their feedback depends on our choice of $\mathscr{D}$. Extending Example 1, suppose that A is the correct decision, only s_0 is honest, and that $f_0=A$ and $f_1=B$. Take $\mathscr{D}_1$ such that $\mathscr{D}_1(ABA)=A$ and $\mathscr{D}_1(ABB)=B$. Here, the malicious source s_2 will report B, because then $\mathscr{D}_1(\mathbf{f})=B$, thus forcing a wrong decision for $\mathscr{D}_1$. We may try to be clever by choosing the opposite of what the third source reports, and use $\mathscr{D}_2$ such that $\mathscr{D}_2(ABA)=B$ and $\mathscr{D}_2(ABB)=A$ to reflect this. However, s_2 will adaptively report A in this case – again forcing a wrong decision. As a result, how probable it is to receive a specific feedback is undefined, as it depends on the attacker. Hence, a decision scheme shall not be based on specific feedback or attack strategies.

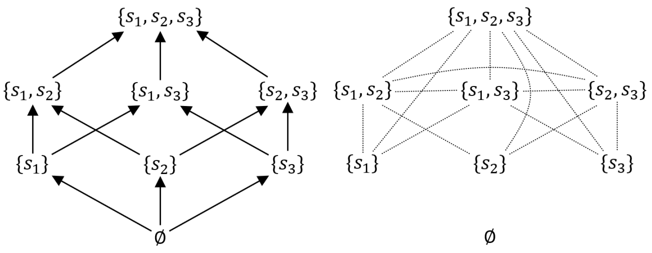
We propose to reason about the *circumstances* under which decisions are always correct, regardless of what the feedback is, to bypass investigating any specific feedback or attack strategy. We introduce the notion of realisations to capture this. Formally, a realisation defines which sources are honest and which are malicious:

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

In Example 1, we have in total of 2^3 realisations, where $\{s_1\}$ and $\{s_3\}$ are examples. Hasse diagrams can be used to represent a finite partially ordered set [7]. Figure 1(a) presents the subset relations of the realisations in Example 1.

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

[2] We've discussed in detail where the two assumptions apply in our previous work [6]



(a) Diagram of subset relations

(b) Diagram of intersections

Fig. 1. Diagrams depicting Example 1

We use the phrase "under realisation $\mathbf{r}$" to denote the decision circumstance that "all $s \in \mathbf{r}$ are honest and all $s \in \overline{\mathbf{r}}$ are malicious".

For the received feedback, a decision maker does not know the realisation behind it, i.e. who exactly are malicious. Depending on the decision scheme, manipulation is possible under some realisations, but impossible under the others. Our goal is to make decisions such that there are fewer and less probable realisations in which manipulation is possible.

The notion of realisations turns out to be a useful tool to model the entire attack space. Given a correct decision, all the feedback which can possibly be received under a realisation defines its *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* $a(c, \mathbf{r}) = \{\mathbf{f} \in \mathcal{F} \mid \forall s \in \mathbf{r}, f_s = c\}$ *is the set of all the possible feedback we could receive.*

Observe that if a realisation is a subset of another realisation, then it includes the attack space of the latter: if $\mathbf{r} \subseteq \mathbf{r}'$, then $a(c, \mathbf{r}') \subseteq a(c, \mathbf{r})$. In Example 1, if A is the correct decision, then for realisation $\{s_0, s_1\}$, $a(A, \{s_0, s_1\}){=}\{AAC, AAA, AAB\}$. And, e.g., $\{s_0, s_1\} \subseteq \mathcal{S}$, so $a(A, \mathcal{S}) = \{AAA\} \subseteq a(A, \{s_0, s_1\})$. The empty-set realisation (where all sources are attackers) is a subset of all the realisations, and accordingly, its attack space is a super set of all the possible attack spaces – in fact it is the set of all possible feedback $\mathcal{F}$. The inclusion relations of attack space can be visualised by reversing the arrows in Figure 1(a).

Also observe that $\forall\ \mathbf{r}'{\neq}\mathbf{r},\ \exists\ c, c'$ s.t. $a(c, \mathbf{r}') \cap a(c', \mathbf{r}) \neq \emptyset$. Specifically, for two disjoint realisations, there exist distinct decisions so that their attack space must be intersecting: if $\mathbf{r}{\cap}\mathbf{r}'{=}\emptyset$, then $\exists\ c{\neq}c'$ s.t. $a(c, \mathbf{r}){\cap}a(c', \mathbf{r}'){\neq}\emptyset$. In Example 1, the feedback AAB exists in both $a(A, \{s_1\})$ and $a(B, \{s_3\})$. Whatever decision is made given AAB, it will be wrong in at least one of the two cases. That is, correctness of decisions cannot be guaranteed for the two realisations at the same time. Or decision mechanisms which can achieve this do not exist. We introduce non-manipulability to capture the idea of "guaranteeing correct decisions":

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* $\forall\ \mathbf{f} \in a(c, \mathbf{r})$, $\mathscr{D}(\mathbf{f}) = c$.

The set of all non-manipulable realisations for $\mathscr{D}$ is denoted as $R_{\mathscr{D}}$.

From the observation on attack space, if $\mathbf{r}$ is non-manipulable and $\mathbf{r}{\subseteq}\mathbf{r}'$, then $\mathbf{r}'$ must also be non-manipulable (since $a(c, \mathbf{r}') \subseteq a(c, \mathbf{r})$). For any decision scheme, a realisation and its complement cannot be non-manipulable at the same time:

Lemma 1. $\forall\ \mathbf{r}{\cap}\mathbf{r}'{=}\emptyset, \nexists\ \mathscr{D}$ *s.t. both* $\mathbf{r}$ *and* $\mathbf{r}'$ *are non-manipulable.*

Proof. Refer to the observations made for attack space. □

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

An attainable set is *maximal* if there is no attainable set which is a strict superset.

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 share 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. The condition is necessary as we have already observed above that if $\mathbf{r}_1$ and $\mathbf{r}_2$ are disjoint, then they cannot both be non-manipulable.

We show that the condition is also sufficient for $\mathbb{A}(R)$ by constructing a decision scheme $\mathscr{D}$ such that $R \subseteq R_{\mathscr{D}}$. Pick $\mathscr{D}$ such that $\forall\ \mathbf{f} \in \mathcal{F}$, if there exists a realisation $\mathbf{r} \in R$ and decision $c \in \mathcal{Q}$ such that $\forall\ s \in \mathbf{r}$, $f_s = c$, then $\mathscr{D}(\mathbf{f}) = c$. If there are multiple realisations $\mathbf{r}_1$ and $\mathbf{r}_2$ where $\forall\ s \in \mathbf{r}_1$, $f_s = c$ and $\forall\ s' \in \mathbf{r}_2$, $f_{s'} = c'$, then $c = c'$, since $\mathbf{r}_1$ and $\mathbf{r}_2$ share at least one source. Hence $\mathscr{D}$ has at most one output per $\mathbf{f}$, such that we require $\mathscr{D}(\mathbf{f}) = c$, and thus $\mathscr{D}$ exists. □

An attainable set of realisations is an example of an *intersecting family*. This is a well-studied class of families in extremal set theory. And we use some of the results from the field. In Figure 1(b), every pair of intersecting realisations are linked with a dotted line. Any complete subgraph is an intersecting family, and so is an attainable set of realisations.

Realisations are a powerful tool in our framework, but are not useful for standard Jury Theorem models. Jury Theorems usually only model the probability of a source's reporting the truth, without reasoning the intention behind e.g., whether the correct feedback results from a malicious strategy. For instance, consider $\mathbf{f}{=}AAB$ and $c{=}A$ in Example 1. With our modelling of realisations, four circumstances are all possible, namely both the first two sources are honest, either of them is malicious but report the truth, or both of them are malicious. However, Jury Theorems would conclude that only the third source reports incorrectly, ignoring whether the first two are strategic. The probabilities found in Jury Theorems are, by necessity, included in the range of the probabilities we find (often as boundary probabilities). *Extremal set theory* is a branch of mathematics that is particularly useful for reasoning about realisations.

2.2 Extremal Set Theory

A *family* is a set of subsets of some fixed set. In our case, a set of realisations is a family, which are themselves subsets of the set of sources $\mathcal{S}$. Of particular interest for us are intersecting families, which are families such that each pair of members of the family is non-disjoint. This matches exactly the condition of Theorem 1. A maximal family is a family to which no set can be added without breaking a property. For instance, in Figure 1(b), the intersecting family consisting of $\{s_1, s_2, s_3\}, \{s_2, s_3\}, \{s_1, s_3\}, \{s_1, s_2\}$ is maximal.

Extremal set theory is the study of the maximising families under some restrictions. One of our contributions is that decision schemes can be seen as intersecting families, and our goal is to maximise the weight of this family. Hence, we introduce some concepts from extremal set theory.

In a *maximal* intersecting family R, if $\mathbf{r} \in R$ and $\mathbf{r} \subseteq \mathbf{r}'$, then $\mathbf{r}' \in R$, since any intersecting element between the set $\mathbf{r} \in R$ and $t \in R$ is also an intersecting element between $\mathbf{r}'$ and t. Furthermore, either the set $\mathbf{r}$ or $\overline{\mathbf{r}} = \mathcal{S} \setminus \mathbf{r}$ is an element of R, but never both (since $\mathbf{r} \cap \overline{\mathbf{r}} = \emptyset$). For the sake of contradiction, assume neither $\mathbf{r}$ nor $\overline{\mathbf{r}}$ is in R. Either $\mathbf{r}$ has an intersecting element with every $t \in R$, in which case we can add $\mathbf{r}$ to R, or there exists some $t \in R$ such that $\mathbf{r} \cap t = \emptyset$, but then $t \subseteq \overline{\mathbf{r}}$ and we can add $\overline{\mathbf{r}}$ to R

We use the notion of a rank [8]:

Definition 6. *The k^{th} rank of a family R, denoted R_k, is the subset of R consisting of members of cardinality k.*

Hence, a member (realisation) $\mathbf{r} : \mathbf{r} \in R_k$ if and only if $|\mathbf{r}| = k$. For instance, refer to Figure 1(a), $\mathcal{R}$ is the family in concern, and all the realisations with two sources (located in the second row) form a $\mathcal{R}_2$ rank. The notion of ranks allows us to formulate a non-trivial property of (maximal) intersecting families. There is a bound on the size of a rank if its members do not contain over half of the elements in the fixed set (e.g., the rank consisting of the members in the third row of Fig 1(a)). This is known as the Erdös-Ko-Rado Theorem [9]. Considering the families of realisations R and the fixed set $\mathcal{S}$, the theorem is:

Theorem 2. *For any intersecting family R, for any $k \leq {}^{n}\!/_{2}$, $|\mathcal{R}_k| \leq \binom{n-1}{k-1}$.*

We introduce the notion of a shadow (Δ) of a rank set R_k to represent the set of realisations that take away a single element. So $\Delta R_k = \{\mathbf{r} \mid |\mathbf{r}| = k-1,\ \mathbf{r} \cup \{s \in \mathcal{S}\} \in R_k\}$. Conversely, the shade (∇) of R_k is the set of realisations obtained by adding one element. The Kruskal-Katona Theorem provides a lower bound on the shadow of a rank set [10]. It provides a complicated bound on the ratio between cardinalities of neighbouring ranks. Lovász provides a simplified (but weaker) formulation [11], which is sufficient for our purpose:

Proposition 1. *For $x \in \mathbb{R}$, if $|R_k| = \binom{x}{k}$, then $|\Delta R_k| \geq \binom{x}{k-1}$.*

2.3 Probability

At least half of the realisations in $\mathcal{R}$ are manipulable (since $\mathbf{r}$ and $\overline{\mathbf{r}}$ cannot both be non-manipulable). Hence the possibility that an arbitrary decision scheme gets manipulated is always non-zero. 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.

First, let random variable $\mathbb{C}$ model what the correct decision is. The outcomes of $\mathbb{C}$ are from the set $\mathcal{Q}$: $c \in \mathcal{Q}$. Let random variable $\mathbb{R}$ model the realisation we are under, and its outcomes are from $\mathcal{R}$: $\mathbf{r} \in \mathcal{R}$. Let random variable $\mathbb{I}$ be the decision. Based on the law of total probability, the probability of making the wrong decision is: $p(\mathbb{I}{\neq}c \mid \mathbb{C}{=}c) = \sum_{\mathbf{r} \in \mathcal{R}} p(\mathbb{R} = \mathbf{r} \mid \mathbb{C} = c) \cdot p(\mathbb{I} \neq c \mid \mathbb{R} = \mathbf{r}, \mathbb{C} = c)$. Whether a source is honest or not does not depend on $\mathbb{C}$. Hence, $p(\mathbb{R} = \mathbf{r} \mid \mathbb{C} = c) = p(\mathbb{R} = \mathbf{r})$. Define the distribution Φ on $\mathcal{R}$ s.t. $\Phi(\mathbf{r}) = P(\mathbb{R} = \mathbf{r})$. The distribution Φ provides a context of the sources to a decision maker, by defining how probable it is that certain sources are honest.

Some distributions Φ may benefit decisions more than others. We define a specific well-defined class as monotonic:

Definition 7 (Monotonicity)**.** *A distribution Φ is* monotonic *when $\mathbf{r} \subseteq \mathbf{r}'$ implies $\Phi(\mathbf{r}) \leq \Phi(\mathbf{r}')$.*

Let random variable $\mathbb{F}$ denote the received feedback, and $\mathbf{f} : \mathbf{f} \in \mathcal{F}$ be 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 $\mathbb{I}$ takes the value of $\mathscr{D}(\mathbf{f})$. $p(\mathbb{I} \neq c \mid \mathbb{F} = \mathbf{f}) = 1$ iff $\mathscr{D}(\mathbf{f}) \neq c$. Hence: $p(\mathbb{I} \neq c \mid \mathbb{R} = \mathbf{r}, C = c) = \sum_{\mathbf{f} \in a(c,\mathbf{r}) \wedge \mathscr{D}(\mathbf{f}) \neq c} p(\mathbb{F} = \mathbf{f} \mid \mathbb{R} = \mathbf{r}, C = c)$.

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 strategies of the attackers within the space $a(c, \mathbf{r})$.

With Φ and β, we can derive a general formula of the probability of deciding incorrectly (or error rate) $\texttt{Err}(\mathscr{D}, \Phi, \beta) = p(\mathbb{I} \neq c \mid C = c)$:

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

Crucially, the error rate depends on the strategy of the attackers, which we cannot make assumptions about.

2.4 Properties

We are interested in two properties of a decision scheme, namely robustness and optimality. A decision scheme that is robust and optimal has a guaranteed upper-bound on manipulability, as small as possible. These are the properties desired for secure-decision making.

Robustness means resistance to manipulation. We do not want to assume any attack strategy to define robustness. Instead, we consider all possible distributions β within a relevant attack space. Robustness in a context Φ is then determined by the maximal error rate w.r.t different attack strategies:

$$\mathfrak{E}_\Phi(\mathscr{D}) = \max\nolimits_\beta \left(\texttt{Err}(\mathscr{D}, \Phi, \beta)\right) \qquad (2)$$

Definition 8 (ϵ-robustness)**.** *Given a value ϵ, a set of sources $\mathcal{S}$ and a distribution Φ of realisations, a decision scheme is ϵ-robust when for all distributions β of feedback:*

$$\forall_\beta\ \texttt{Err}(\mathscr{D}, \Phi, \beta) \leq \epsilon.$$

Equivalently, we can say $\mathfrak{E}_\Phi(\mathscr{D}) \leq \epsilon$. A simple computation exists for robustness:

Theorem 3. *If a decision scheme $\mathscr{D}$ is ϵ-robust, then*

$$\sum_{\mathbf{r}\in\mathcal{R}\setminus R_{\mathscr{D}}} \Phi(\mathbf{r}) \le \epsilon.$$

Proof. It suffices to prove that $\mathfrak{E}_{\Phi}(\mathscr{D}) = \sum_{\mathbf{r}\in\mathcal{R}\setminus R_{\mathscr{D}}} \Phi(\mathbf{r})$, meaning the maximal error rate is determined by how probable it is to be under a manipulable realisation. Refer to Definition 4, if $\mathbf{r} \in R_{\mathscr{D}}$, then $\forall_{c,\mathbf{f}\in a(c,\mathbf{r})}(\mathscr{D}(\mathbf{f}) = c)$.

Refer to the inner sum of Equation 1, if $\mathbf{r} \in R_{\mathscr{D}}$ (being non-manipulable Definition 4), then $\nexists_{\mathbf{f}\in a(c,\mathbf{r})} (\mathscr{D}(\mathbf{f}){\neq}c)$, and accordingly $\sum_{\mathbf{f}\in a(c,\mathbf{r})\wedge\mathscr{D}(\mathbf{f})\neq c} \beta(\mathbf{r},c)(\mathbf{f}){=}0$. Contrarily, if $\mathbf{r} \notin R_{\mathscr{D}}$ (being manipulable), then $\exists_{\mathbf{f}\in a(c,\mathbf{r})} (\mathscr{D}(\mathbf{f}){\neq}c)$. Let $\beta(\mathbf{r},c)(\mathbf{f}) = 1$ for that $\mathbf{f}$. Then, trivially, the inner sum $\sum_{\mathbf{f}\in a(c,\mathbf{r})\wedge\mathscr{D}(\mathbf{f})\neq c} \beta(\mathbf{r},c)(\mathbf{f}){=}1$, reaching its maximum. And accordingly, this choice of β makes $\mathtt{Err}(\mathscr{D},\mathbf{r},\beta)$ reaches its maximum, which is $\sum_{\mathbf{r}\in\mathcal{R}\setminus R_{\mathscr{D}}} \Phi(\mathbf{r})$. □

The proof suggests that the attacker can maximise the error rate by always forcing a wrong decision when the realisation is manipulable. Exactly which feedback they use to accomplish this is irrelevant. This supports our idea that one should reason about realisations rather than specific feedback or strategy.

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; i.e. the scheme that has maximal robustness. This idea is captured by the optimality property:

Definition 9 (Optimality). *For a given distribution Φ of realisations, $\mathscr{D}$ is optimal when for all $\mathscr{D}'$, $\mathfrak{E}_{\Phi}(\mathscr{D}) \le \mathfrak{E}_{\Phi}(\mathscr{D}')$.*

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

3 Most Probable Realisation

Before we propose a decision mechanism, consider Example 1: $\mathbf{f} = ACC$. Typically, with no adversaries, the focus would be on determining the probabilities of A and C being the right decisions, given some feedback e.g., $p(\mathbb{C}{=}A|\mathbb{F}{=}\mathbf{f})$. The *weighted majority voting* (WMV) mechanism[3] does exactly this – as it does not assume adversarial behaviour. However, how probable the feedback is received depends on the strategies of attackers of which we do not assume any knowledge.

Multiple realisations are possible in the example: 1) only s_0 is honest; 2) both s_1 and s_2 are honest, and 3/4) either s_1 or s_2 is honest; 5) all the three are attackers ($\emptyset$). In typical WMV, only 1) and 2) are considered. In the mechanism we propose, the decision follows the most probable realisation behind the feedback. Formally:

Definition 10 (Most Probable Realisation (MPR); $\mathscr{X}$). $\mathscr{X}(\mathbf{f}) = \operatorname{argmax}_{c\in\mathcal{O}} \max_{\mathbf{r}\in\mathcal{R}\setminus\{\emptyset\}\,:\,\mathbf{f}\in a(c,\mathbf{r})} \Phi(\mathbf{r})$.

An important technical observation is that if we make the correct decision for some feedback $\mathbf{f}$, then any feedback $\mathbf{f}'$ in which the correct feedback from $\mathbf{f}$ is present will also lead to the correct decision, even if the remaining feedback is possibly something completely different:

Lemma 2. *Given feedbacks $\mathbf{f}$ and $\mathbf{f}'$, such that $\forall_i(f_i \in \{c,d\}\wedge (f_i = c \implies f_i' = c))$, if $\mathscr{X}(\mathbf{f}) = c$ then $\mathscr{X}(\mathbf{f}') = c$.*

Proof. Since every c in $\mathbf{f}$ is also a c in $\mathbf{f}'$: $\forall\mathbf{r}(\mathbf{f}'{\in}a(c,\mathbf{r}) \implies \mathbf{f}{\in}a(c,\mathbf{r}))$; and since, for whatever $e \neq c$, every e in $\mathbf{f}'$ is a d in $\mathbf{f}$: $\forall e \neq c, \mathbf{r}(\mathbf{f}{\in}a(e,\mathbf{r}) \implies \mathbf{f}'{\in}a(d,\mathbf{r}))$. Therefore $\max_{\mathbf{r}\in\mathcal{R}:\mathbf{f}\in a(c,\mathbf{r})} \Phi(\mathbf{r}) \le \max_{\mathbf{r}\in\mathcal{R}\ :\mathbf{f}'\in a(c,\mathbf{r})} \Phi(\mathbf{r})$ and similarly $\max_{\mathbf{r}\in\mathcal{R}:\mathbf{f}\in a(d,\mathbf{r})} \Phi(\mathbf{r}) \le \max_{e\in\mathcal{O},\mathbf{r}\in\mathcal{R}\ :\mathbf{f}\in a(e,\mathbf{r})} \Phi(\mathbf{r})$. Hence if c is maximal for $\mathbf{f}$ then it is also maximal for $\mathbf{f}'$. □

This implies that the set of realisations yielded by MPR is a *maximal* intersecting family:

Theorem 4. *For every realisation $\mathbf{r}$, $\mathbf{r} \in R_{\mathscr{X}}$ or $\overline{\mathbf{r}} \in R_{\mathscr{X}}$.*

Proof. Define $\mathbf{f}$ as $f_i = c$ if $s_i \in \mathbf{r}$ and $f_i = d$ if $s_i \in \overline{\mathbf{r}}$. The most probable realisation $\mathbf{r}^{max}$ yielding $\mathbf{f}$ is non-empty, so $\exists_{s_i\in\mathbf{r}^{max}}(f_i = c \vee f_i = d)$, hence $\mathscr{X}(\mathbf{f}) = c$ or $\mathscr{X}(\mathbf{f}) = d$. Via Lemma 2, $\mathscr{X}(\mathbf{f}') = c$ for all $\mathbf{f}' \in a(c,\mathbf{r})$ – or $\mathscr{X}(\mathbf{f}') = d$ for all $\mathbf{f}' \in a(d,\overline{\mathbf{r}})$; hence $\mathbf{r} \in R_{\mathscr{X}}$ or $\overline{\mathbf{r}} \in R_{\mathscr{X}}$. □

3.1 Independent Sources

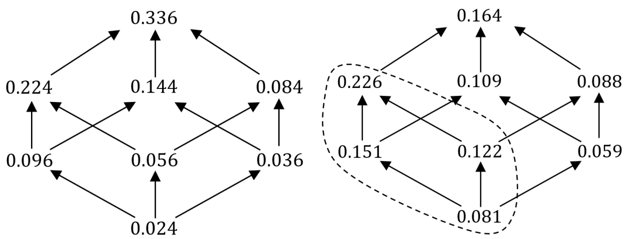
First, consider the scenario where the honesty of different sources are independent e.g., when they are selected randomly by the decision maker. Sometimes sources are treated as interchangeable (e.g., when it is difficult to characterise individual sources), the majority rule can be applied. Although simple, it has been proved in Condorcet's Jury Theorem that the decision accuracy of majority rule improves if there are more sources, and that accuracy (probability) converges to 1 (infallibility)[4], thus proving the effectiveness of relying on crowd wisdom.

In practice, we may have knowledge about each source and be able to evaluate their probability of being honest individually, e.g., by evaluating witness trustworthiness [12] (see Section 5.3). Intuitively, a decision should be more inclined to feedback from a more probably honest source.

There are n sources, and each has a probability to be honest $p_0,\ldots,p_{n-1}$, $\mathbf{p}$ is the joint probability. Assuming these n probabilities are mutually independent, $\Phi(\mathbf{r}) = \prod_{s\in\mathbf{r}} p_s \cdot \prod_{s\in\overline{\mathbf{r}}}(1-p_s)$, denoted as $\varphi_{n,\mathbf{p}}$. In Figure 2(a), we depict Example 1 with values $\mathbf{p} = (0.8, 0.7, 0.6)$; note, e.g., $\{s_0, s_1\} = p_0 \cdot p_1 \cdot (1-p_2) = 0.8 \cdot 0.7 \cdot (1-0.6) = 0.224$. A realisation $\mathbf{r}$ containing source s_i, with $p_i < 0.5$ will never be the most probable realisation, since $\varphi_{n,\mathbf{p}}(\mathbf{r}) = \varphi_{n,\mathbf{p}}(\mathbf{r}\setminus\{s_i\}){\cdot}\frac{p_i}{1-p_i} < \varphi_{n,\mathbf{p}}(\mathbf{r}\setminus\{s_i\})$, and $a(c,\mathbf{r}\setminus\{s_i\}) \supseteq a(c,\mathbf{r})$. Hence, we assume without loss of generality that $p_i \ge 0.5$ for any source s_i. In Figure 2(b), we show an example where $\mathbf{p} = (0.65, 0.6, 0.42)$ and $p_2 < 0.5$. Notice that the dashed area contains realisations that are more probable than the ones with s_2 added. Hence, we can simply focus only on the dashed area (and multiply all probabilities by $\frac{1}{0.58}$ to normalise). Let $n,\mathbf{p}$ ($\forall_{s_i}, p_i \ge 0.5$) define the context of decision making and denote MPR under independence assumption as $\mathscr{X}_{n,\mathbf{p}}$.

[3] The typical decision mechanism for extending Condorcet's Jury Theorem to arbitrary independent probabilities.

[4] Under the assumptions that sources are independent and report correctly with a same probability of over a half.



(a) Well-behaved example. (b) Example with $p_3 \leq 1/2$.

Fig. 2. Two examples of independent distributions φ.

3.1.1 Properties of $\mathscr{X}_{n,\mathbf{p}}$

First, notice $\mathscr{X}_{n,\mathbf{p}}$ is monotonic (with assumption $p_i \geq 1/2$):

Proposition 2. *Distribution $\varphi_{n,\mathbf{p}}$ is monotonic.*

Proof. If $\mathbf{r}' = \mathbf{r} \cup \{s_k\}$, then $\varphi_{n,\mathbf{p}}(\mathbf{r}') = \varphi_{n,\mathbf{p}}(\mathbf{r}) \cdot \frac{p_k}{1-p_k}$, and $p_k \geq 0.5$. General case follows inductively. □

Given a pair of complementary realisations $\mathbf{r}, \overline{\mathbf{r}}$, $\mathscr{X}_{n,\mathbf{p}}$ is non-manipulable under the more probable one.

Lemma 3. *If $\varphi_{n,\mathbf{p}}(\mathbf{r}) > \varphi_{n,\mathbf{p}}(\overline{\mathbf{r}})$, then $\mathscr{X}_{n,\mathbf{p}}$ is non-manipulable under $\mathbf{r}$.*

Proof. Via Theorem 4, $\mathbf{r} \in R_{\mathscr{X}_{n,\mathbf{p}}}$ or $\overline{\mathbf{r}} \in R_{\mathscr{X}_{n,\mathbf{p}}}$. Take $\mathbf{f}$, s.t. $f_i = c$ if $s_i \in \mathbf{r}$ and $f_i = d$ if $s_i \in \overline{\mathbf{r}}$. Since $\varphi_{n,\mathbf{p}}$ is monotonic, the maximising realisation for deciding d on $\mathbf{f}$ is $\overline{\mathbf{r}}$. As $\varphi_{n,\mathbf{p}}(\mathbf{r}) > \varphi_{n,\mathbf{p}}(\overline{\mathbf{r}})$, $\mathbf{r}$ is most probable for $\mathbf{f}$, so $\mathscr{X}_{n,\mathbf{p}}(\mathbf{f}) = c$. Finally, via Lemma 2, for all $\mathbf{f}' \in a(c, \mathbf{r})$, $\mathscr{X}_{n,\mathbf{p}}(\mathbf{f}) = c$. □

Theorem 5. *$\mathscr{X}_{n,\mathbf{p}}$ is ϵ-robust for:*

$$\epsilon \geq \sum_{\mathbf{r} \in \mathcal{R} | \varphi_{n,\mathbf{p}}(\mathbf{r}) > \varphi_{n,\mathbf{p}}(\overline{\mathbf{r}})} \varphi_{n,\mathbf{p}}(\overline{\mathbf{r}})$$

Proof. Combine Lemma 3 with Theorem 3. □

Given the same set of sources whose probabilities of being honest are independent (given $\varphi_{n,\mathbf{p}}$), there does not exist a decision scheme that is more robust than $\mathscr{X}_{n,\mathbf{p}}$, i.e., $\mathfrak{E}(\mathscr{X}_{n,\mathbf{p}})$ is minimal.

Theorem 6. *Given n sources with probability of honesty $\mathbf{p}$, $\mathscr{X}_{n,\mathbf{p}}$ is optimal.*

Proof. Lemma 1 proves for any $\mathscr{D}, \mathbf{r}$, $\mathbf{r}$ and $\overline{\mathbf{r}}$ cannot co-exist in $R_{\mathscr{D}}$. Hence, $\max |\mathbf{r}| = 1/2 |\mathcal{R}|$. $R_{\mathscr{X}_{n,\mathbf{p}}}$ consists of realisations that are at least as probable as their complementary ones. □

While MPR is optimal, it turns out that a setting with independent sources allows for lots of effective decision schemes. Any decision scheme that has an attainable set which is a maximal intersecting family does better than fifty-fifty:

Theorem 7. *For any R, such that $\mathbb{A}(R)$ and $|R| = 2^{n-1}$, $\mathfrak{E}_{n,\mathbf{p}}(R) \leq 1/2$.*

Proof. Since $\mathbb{A}(R)$: if $\mathbf{r} \in R$ then $\mathbf{r} \cup \{s \in \mathcal{S}\} \in R$. Rewrite $\sum_{\mathbf{r} \in R} \varphi_(\mathbf{r}) \geq \frac{|R|}{|\mathcal{R}|}$ to: $\sum_{\mathbf{r} \in R} \prod_{i \in \mathbf{r}} p_i \cdot \prod_{i \in \mathcal{S} \setminus \mathbf{r}} (1 - p_i) \geq \frac{|R|}{2^n}$, where $n = |\mathcal{S}|$.

Take base case $n = 1$, so w.l.o.g. $\mathcal{S} = \{s_0\}$. If $R = \emptyset$, then $\frac{|R|}{2} = 0$; if $R = \{\{s_0\}\}$, then $p_0 \geq \frac{1}{2}$, because $p_0 > 1/2$; $R = \{\emptyset\}$ is disallowed by assumption; $R = \{\emptyset, \{s_0\}\}$, then $p_0 + (1 - p_0) \geq \frac{2}{2} = 1$.

For the induction step, we have $\mathcal{S}' = \mathcal{S} \cup \{s_n\}$. Partition R into R_1^+ and R_2, where R_1^+ contains the realisations with s_n and R_2 the realisations without s_n. Let R_1 be the set of realisations from R_1^+ with the source s_n removed from each realisation. By distributing out the terms p_i and $1 - p_i$, we see: $\sum_{\mathbf{r} \in R} \prod_{i \in \mathbf{r}} p_i \prod_{i \in \mathcal{S}' \setminus \mathbf{r}} (1 - p_i) = p_n \cdot \sum_{\mathbf{r} \in R_1} \prod_{i \in \mathbf{r}} p_i \cdot \prod_{i \in \mathcal{S} \setminus \mathbf{r}} (1 - p_i) + (1 - p_i) \cdot \sum_{\mathbf{r} \in R_2} \prod_{i \in \mathbf{r}} p_i \cdot \prod_{i \in \mathcal{S} \setminus \mathbf{r}} (1 - p_i)$ which is at least $p_n \cdot \frac{|R_1|}{2^n} + (1 - p_n) \cdot \frac{|R_2|}{2^n}$, by induction hypothesis. Since $|R_1^+| = |R_1|$, and $\mathbf{r} \in R_2 \implies \mathbf{r} \in R_1^+$ by assumption, $|R_1| \geq |R_2|$. Since $p_n \geq 1 - p_n$, via Jensens inequality: $p_n \cdot \frac{|R_1|}{2^n} + (1 - p_n) \cdot \frac{|R_2|}{2^n} \geq 1/2 \frac{|R_1|}{2^n} + 1/2 \frac{|R_2|}{2^n} = \frac{|R|}{2^{n+1}}$. □

3.1.2 Relation between MPR and WMV

Speaking of multi-source decision making where sources are assumed not to be equivalent, a typical decision scheme is Weighted Majority Voting (WMV), which is popular in the domain of social choice theory [13] and also some other domains (refer to Section 5.1 for details).

WMV can be treated as an extension of a simple but commonly known decision scheme: Majority Rule (MR or $\mathscr{D}_{MR}$).

$$\mathscr{D}_{MR}(f) = \mathrm{argmax}_{c_j} \sum_{s_i} sign(f_i = c_j)$$

MR selects the option which gain majority votes, treating all the sources without distinction. This makes it not suitable for decision tasks where sources should be distinguished. WMV can be a solution sometimes by distinguishing source weights. For example, when its used for voting, weights can be determined by source importance []. And when its used to decide which feedback option corresponds to the truth (like Jury Theorems []), weights usually depend on source reliability or competence. WMV selects the option suggested by the sources of the maximal total weights, defined as below:

$$\mathscr{D}_{WMV}(f) = \mathrm{argmax}_{c_j} \sum_{s_i} w_i \cdot sign(f_i = c_j)$$

It is proved that WMV has optimal decision accuracy when $w_i = \log(q_i / 1 - q_i)$, where q_i denotes the probability of source s_i suggesting correctly [14] and q_i are independent for different sources.

WMV coincides with our scheme MPR sometimes but they are not generally equivalent, which we detail below.

Where they coincide: Given same feedback and decision setting, WMV and MPR decide the same when they both use the sources' probabilities of being honest as input, and that probabilities are independent and all above a half.

Theorem 8. *Given n sources with $p_i > 1/2, i \in \{1, \ldots, n\}$, and p_i mutually independent, $\forall f, \mathscr{X}_{n,\mathbf{p}}(f) = \mathscr{D}_{WMV}(f)$ with $w_i = \log(p_i / 1 - p_i)$.*

Proof. We need to prove that for any feedback, the most probable realisation behind it suggests the decision which receives the maximal total weights. First note that $\forall s_i, w_i > 0$. Given f, let $\mathbf{r}^* = \mathrm{argmax}_{c,r: f \in a(c,\mathbf{r})} \Phi(\mathbf{r})$, then we have $\Phi(\mathbf{r}^*) > \Phi(\overline{\mathbf{r}^*})$. This means $\prod_{s_i \in \mathbf{r}^*} p_i \cdot \prod_{s_i \in \overline{\mathbf{r}^*}} (1 - p_i) > \prod_{s_i \in \overline{\mathbf{r}^*}} p_i \cdot \prod_{s_i \in \mathbf{r}^*} (1 - p_i)$. Apply logarithm on both sides and after transposition, we can get $\sum_{s_i : s_i \in \mathbf{r}^*} w_i > \sum_{s_i : s_i \in \overline{\mathbf{r}^*}} w_i$.

Use $c^* : f{\in}a(c^*, \mathbf{r}^*)$ to denote the decision suggested by $\mathbf{r}^*$, then the left side of the inequality cannot exceed the total weights that c^* receives, while the right side cannot be smaller than the total weights that any other option receives. Hence WMV would also pick c^* as the decision. □

Where they differ: By definition, WMV is not suitable for adversarial decision setting where sources can be malicious and colluding (not independent). First, in WMV, the input probability values (used for the weights) are the sources' probability of reporting the truth. However, a malicious source has the choice of strategically reporting the truth (e.g. to camouflage itself), or not. Hence, the probability to report the truth depends on the strategy of malicious users. The problem is that malicious users may change their strategy depending on the decision scheme that is used. In our MPR, the input probability values are the sources' probability that they are honest, to avoid this problem. As a result, in MPR we find minimum probabilities (probability can only go up when malicious users tell the truth) and in WMV we find exact probabilities. Second, if WMV were used with honesty values as input, then WMV is likely non-optimal when there are sources with a $p_i < 1/2$. We use Example 1 and Figure 2(b) to illustrate the difference. Suppose $\mathbf{p} = (0.65, 0.60, 0.42)$ and $\mathbf{f}_1 = BAB, \mathbf{f}_2 = BAA$. For MPR, feedback of the third source is ignored and it trusts B for both $\mathbf{f}_1, \mathbf{f}_2$. For WMV, we get $\mathbf{w} \approx (0.27, 0.18, -0.14)$. It trusts A for $\mathbf{f}_1$ and B for $\mathbf{f}_2$. Hence source s_3 is dominating the decision of WMV. Finally, observe that if sources are colluding, then the probability that a source is malicious is not independent from the probability that another source is independent. As a result, the part of the proof of Theorem 8 that relies on the product no longer applies, as the probability of a realisation no longer equals the product of the probabilities of each source being honest. The tools used in WMV do not have ability to reason about collusion, whereas MPR continues to work (but typically no longer remains optimal).

In summary, while both dealing with multi-source decision making with source distinction, WMV is not suitable for adversarial decision setting while MPR is. When sources can be malicious, its infeasible to obtain the input WMV needs and also its non-optimal using source honesty degrees as input. MPR generalises much more nicely, as presented in the following.

3.2 General Distributions of Sources

Independent probabilities of being honest may not be realistic. For example, malicious sources may coordinate. In the previous section, our model already allows malicious sources to coordinate their feedback, by considering all possible combinations of their feedback in an attack space. But the dependence of feedback does not necessarily implies the dependence of honesty degrees. For example, Alice and Bob may be malicious with different attempts, but they can still decide to report the same sometimes.

Attackers may be able to use other ways to coordinate, like subverting the process in which they are selected. If this is the case, the probability that source s_i is honest is no longer independent from the probability that s_j is honest, indicating that their dependence is not only reflected in the feedback. Consider Sybil attacks [15] where an attack controls multiple compromised accounts. Finding that one account is actually malicious increases the probability that the others are malicious. Besides Sybil attacks, there may be various ways in which honesty degrees of malicious sources are correlated. They may want to band together and target specific decisions – in which case there is a positive correlation between sources being malicious; or they may want to spread out their efforts among many decisions – in which case the correlation is negative. In this section, we do not assume independence of sources, and accordingly, Φ is not a joint distribution of independent p-values. We assume neither the form of source dependence nor its effect to individuals. In fact, Φ can be any distribution over sets of $\mathcal{S}$.

How to obtain such a distribution Φ is outside the scope of the paper. Multiple methods may be considered. A simple example is where sources are conditionally independent under some random variable modeling the trustworthiness of the current environment. In this example, if s_1 is malicious, then it increases the conditional probability that s_2 is also malicious. More complex example could involve clustering to identify possible clusters of attackers, or game theory to model optimal attacker behaviour. We introduce the notation $\mathscr{X}_\Phi$ to represent a specific decision scheme (with Φ given) under the mechanism MPR. In this section, we study the implications to MPR of using a more general distribution Φ.

For general distributions, MPR keeps trusting the most probable realisation, regardless of how sources are colluding. For WMV, the meaning of the individual weights becomes questionable here. To reason about realisations instead of specific sources or feedback is increasingly useful.

To present our results more effectively, we introduce the function $\psi(\mathbf{r})$ which is defined as $\psi(\mathbf{r}) = \max_{\mathbf{r}' \subseteq \mathbf{r}} \Phi(\mathbf{r}')$. The function ψ is not typically a probability distribution, but can still be used as input for MPR. The crucial idea is that $\mathscr{X}_\Phi = \mathscr{X}_\psi$:

Lemma 4. $\mathscr{X}_\Phi = \mathscr{X}_\psi$.

Proof.

$$\begin{aligned}\mathscr{X}_\psi(\mathbf{f}) = \operatorname{argmax}_{c \in \mathcal{O}} \max_{\mathbf{r} \in \mathcal{R} \,:\, \mathbf{f} \in a(c,\mathbf{r})} \psi(\mathbf{r}) = \\ \operatorname{argmax}_{c \in \mathcal{O}} \max_{\mathbf{r} \in \mathcal{R} \,:\, \mathbf{f} \in a(c,\mathbf{r})} \max_{\mathbf{r}' \subseteq \mathbf{r}} \Phi(\mathbf{r}') = \\ \operatorname{argmax}_{c \in \mathcal{O}} \max_{\mathbf{r},\mathbf{r}' \in \mathcal{R} \,:\, \mathbf{r}' \subseteq \mathbf{r} \wedge \mathbf{f} \in a(c,\mathbf{r}')} \Phi(\mathbf{r}') = \\ \operatorname{argmax}_{c \in \mathcal{O}} \max_{\mathbf{r}' \in \mathcal{R} \,:\, \mathbf{f} \in a(c,\mathbf{r}')} \Phi(\mathbf{r}') = \mathscr{X}_\Phi(\mathbf{f})\end{aligned}$$

□

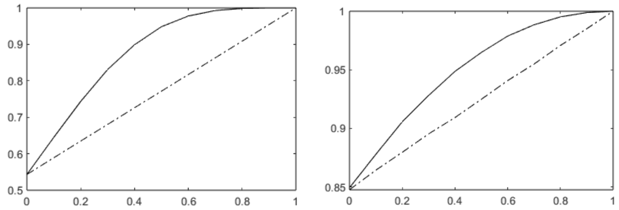
Using the fact that $\mathscr{X}_\Phi$ and $\mathscr{X}_\psi$ make the same decisions, a minor modification to Theorem 5 is sufficient to generalise the robustness of MPR mechanism to include collusion cases:

Theorem 9. *$\mathscr{X}_\Phi$ is ϵ-robust for:*

$$\epsilon \geq \sum_{\mathbf{r} \in \mathcal{R} | \psi(\mathbf{r}) > \psi(\bar{\mathbf{r}})} \Phi(\bar{\mathbf{r}})$$

Proof. Follows from Lemmas 3, 4 and Theorem 5. □

Note that Theorem 5 is a special case of Theorem 9, where $\forall \mathbf{r}, \Phi(\mathbf{r}) = \psi(\mathbf{r})$. Since if $\forall s_i, p_i{>}1/2$ and p_i are independent, then $\forall \mathbf{r}'{\subseteq}\mathbf{r}$, $\Phi(\mathbf{r}') \leq \Phi(\mathbf{r})$.



(a) With p=0.5 and n=11. (b) With p=0.7 and n=5.

Fig. 3. Decision making under collusive attacks.

MPR remains optimal even with collusion if it trusts the more probable one in any pair of a realisation and its complement:

Theorem 10. *$\mathscr{X}_\Phi$ is optimal if $\forall \mathbf{r}$ s.t. $\psi(\mathbf{r}) \geq \psi(\overline{\mathbf{r}})$, $\Phi(\mathbf{r}) \geq \Phi(\overline{\mathbf{r}})$.*

Proof. Apply Theorem 6, observing $R_{\mathscr{X}_\psi} = R_{\mathscr{X}_\Phi}$. □

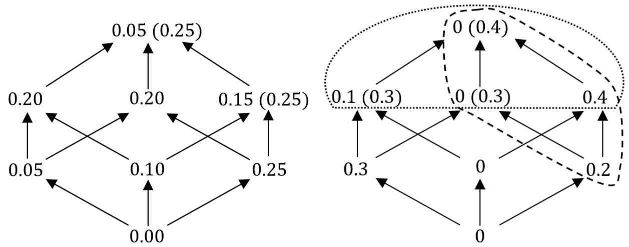
An immediate corollary is that $\mathscr{X}_\Phi$ is optimal for any *monotonic* distribution Φ. However, the condition can be true for non-monotonic distributions. For example, Figure 4(a) presents a case (extending Example 1) where sources are not independent nor monotonic, but distribution Φ satisfies the condition in Theorem 10. The function ψ only differs for two values: $\{s_2, s_3\}$ and $\{s_1, s_2, s_3\}$. It is clear that 0.15 and 0.25 are both larger than 0.05, and that 0.05 and 0.25 are both larger than 0. MPR is optimal by being non-manipulable under the family $\{\{s_3\}, \{s_1, s_3\}, \{s_2, s_3\}, \{s_1, s_2, s_3\}\}$ for a robustness of $0.25 + 0.20 + 0.15 + 0.05 = 0.65$.

The condition in Theorem 10 is not necessary for $\mathscr{X}_\Phi$ to be optimal. Recall Example 1. Let $\Phi(\{s_1\}) = 0.6, \Phi(\{s_3\}) = 0.4$. $\mathscr{X}_\Phi$ is optimal by always trusting the realisation $\{s_1\}$. But $\psi(\{s_1, s_2\})$=$\Phi(\{s_1\})$>$\psi(\{s_3\})$=$\Phi(\{s_3\})$ while $\Phi(\{s_1, s_2\})$<$\Phi(\{s_3\})$.

Proposition 3. *The optimality of* Most Probable Realisation *decision mechanism does not always hold for general distributions of sources where their honesty degrees can be dependent.*

Proof. To provide a case where MPR (here $\mathscr{X}_\Phi$ is not optimal, consider a distribution of realisations Φ as depicted in Figure 4(b). Observe that MPR is non-manipulable in the dotted area, having a lower-bound accuracy of $0.4 + 0.1 = 0.5$, but simply following s_3 provides a lower-bound accuracy of $0.4 + 0.2 = 0.6$ as shown in the dashed area. The condition of Theorem 10 is broken, since $\phi(\{s_1, s_2\}) < \phi(\{s_3\}) = \psi(\{s_3\}) < \phi(\{s_1, s_2\})$. It is MPR selecting $\{s_1, s_2\}$ that causes the suboptimality in this case. □

Even non-optimal decision schemes have the property that they are robust, even in the collusive setting. Hence, while it may be possible to make better decisions, we still have the property that the quality of our decisions is at least as good as the computed robustness value. To illustrate this, we have taken a good, but non-optimal decision scheme (Algorithm 2 as defined later), and analysed it performance under attacks that are increasingly far removed from the worst-case attack. The results are in Figure 3. On the left-hand side, at $x = 0$, the attacker always provides the feedback that maximises the probability that the decision from Algorithm 2 is wrong. The x-axis denotes the probability that a source deviates from this worst-case scenario. If $x = 1$, then the attacker's feedback is most favourable to us, and we can always decide correctly. The solid lines represent attackers that deviate with independent probability x, whereas the dash-dotted lines represent attackers that deviate in collusion, with probability x. This illustrates that our results concern the worst-case collusion attacks.



(a) Non-monotonic example satisfying the condition of Theorem 10. (b) Example where MPR is not optimal.

Fig. 4. Non-independent distributions of realisations ϕ (ψ).

4 Decision Mechanisms under Collusion

In the previous section, we presented that MPR may not be optimal for general distributions of realisations. In this section, we analyze the complexity of finding a generally optimal decision scheme, for which some heuristics methods are also explored.

4.1 Computational Complexity

As long as we find a required attainable set, we can derive a decision scheme (i.e., by referring to the set for a given feedback). For an arbitrary distribution Φ, we formally define the problems of looking for a required attainable set in the following (recall that Φ decides the number of sources n and the set of all the realisations $\mathcal{R}$), and study its complexity and various properties.

The attainable set decision problem concerns whether it is possible to find a decision scheme with ϵ-robustness, given a distribution Φ. There are various ways to define a distribution Φ (e.g. in Section 3.1 it is defined by $\mathbf{p}$). Here, we assume that the input Φ is a list of pairs of realisations and their corresponding probability values, where every non-zero value is listed. $[(\{s_0, s_1\}, 0.6), (\{s_1, s_2\}, 0.4)]$ is an example input which describes the distribution that has $\phi(\{s_0, s_1\}) = 0.6, \phi(\{s_1, s_2\}) = 0.4, \phi(\mathbf{r}) = 0$ otherwise.

Definition 11 (Attainable Set Decision Problem)**.** *Given a distribution Φ over a set of realisations, and a positive real number ϵ as input, the output is "true" if there exists a set of realisations $R \subseteq \mathcal{R}$ s.t. $\mathbb{A}(R) \wedge \sum_{\mathbf{r} \in R} \Phi(\mathbf{r}) \geq 1 - \epsilon$, and "false" otherwise.*

The optimal attainable set problem is the corresponding optimisation problem:

Definition 12 (Optimal Attainable Set Problem)**.** *Given a distribution Φ as input, the output is a set of realisations $R \subseteq \mathcal{R}$ s.t. $\mathbb{A}(R)$ and $\forall_{R' \subseteq \mathcal{R}: \mathbb{A}(R')} \left(\sum_{\mathbf{r} \in R} \Phi(\mathbf{r}) \geq \sum_{\mathbf{r}' \in R'} \Phi(\mathbf{r}')\right)$.*

Theorem 1 tells us that a set of realisations is attainable iff any two realisations are pairwise intersecting, (i.e., their intersection is nonempty). This notion of being attainable is similar to the concept of cliques and clique problems in graph theory [16], [17], which may provide hints on our study of Problems 11 and 12.

Let $G = (V, E)$ denote an undirected graph, with V, E representing the set of vertices and edges respectively. An edge is an unordered pair of vertices $\{v, u\}$. Below we prove the computational complexity of Problem 11 and Problem 12, by reducing the clique decision problem [17].

Definition 13 (Clique [16])**.** *Given an undirected graph G, a clique is a subset of vertices $C \subseteq V$, where every pair of vertices form an edge.*

A problem associated with cliques is the clique decision problem, which is one of the well-known Karp's 21 NP-complete problems [16]:

Definition 14 (Clique Decision Problem [17])**.** *Given graph G and a number k, the output is "true" if G contains a $k-$clique, and "false" otherwise.*

Definition 15 (Maximum Weighted Clique Problem)**.** *Given graph G with weights on its vertices $f : V \to W$, the output is a clique with maximum total weight.*

The *maximum clique problem* is a special case where weights of all the vertices are equal.

It has been proved that the *maximum weighted clique problem* is NP-hard [18]. If one could solve it, then the clique decision problem and the maximum clique problem can also be solved.

Theorem 11. *The* Attainable Set Decision Problem *is NP-hard.*

Proof. The input of the clique decision problem is the graph $G = (V, E)$, and an integer $k \leq |V|$. Assume without loss of generality, that the graph is connected (if the graph is not connected, we can efficiently find the connected components, and apply the reduction to each component), and that there are at least 3 vertices. The reduction to an Attainable Set Decision Problem is as follows:

Associate every edge $e \in E$ with an information source $s \in \mathcal{S}$, so $\tau(e) = s$. Associate every vertex $v \in V$ with a realisation $\mathbf{r}$ ($\tau(v) = \mathbf{r}$), so that $\tau(e) \in \mathbf{r}$ iff e is adjacent to v ($v \in e$). Note that if $v \neq u$, then $\tau(v) \neq \tau(u)$, because no vertices are adjacent to the same set of edges, since our assumptions disallow individual vertices (empty set of vertices) and pairs of vertices only connected with each other (singleton set of their shared edge).

The Attainable Set Decision Problem with input Φ, s.t. $\Phi(\tau(v))={}^{1}\!/\!_{|V|}$ and $\epsilon={}^{k}\!/\!_{|V|}$ provides output R iff $\{v | \tau(v) \in R\}$ is a clique of size at least k, and "false" otherwise.

Observe that if $\{v|\tau(v)\in R\} = C$ is a clique in G, then for every pair $\{u, v\} \in C$, $\{u, v\} \in E$ by definition. This means $\tau(\{u, v\}) \in \mathcal{S}$, and thus $\tau(u)$ and $\tau(v)$ intersect. If every pair in R intersects, then R is attainable. The number of realisations with non-zero weight in R is $|C|$; and each of those has weight ${}^{1}\!/\!_{|V|}$.

Conversely, if R is attainable and $\sum_{\mathbf{r}\in R} \Phi(\mathbf{r}) > {}^{k}\!/\!_{|V|}$, then there are at least k vertices in V such that $\tau(v) \in R$. Since every $\tau(u)$ and $\tau(v)$ intersect, there is some $\tau(e) \in \mathcal{S}$ such that $\tau(e) \in \tau(u)$ and $\tau(e) \in \tau(v)$, but then $u \in e$ and $v \in e$, so every pair of vertices is adjacent, giving a clique of size at least k. □

Corollary 1. *The* Optimal Attainable Set Problem *is NP-hard.*

Proof. It follows trivially from Theorem 11. □

An important caveat here, is that the input size is the number of realisations with non-zero probability. In the construction we provide, the number of sources involved is linear with the number of edges. The proof trivially shows, therefore, that the pseudo-complexity in terms of the number of sources n is also NP-hard. Below, we propose some practical heuristics to solve the problem efficiently.

4.2 Heuristics and Optimality

We proved that it is NP-hard to find a generally optimal decision scheme. In this section, we explore some greedy heuristics for selecting the most probable attainable set of realisations. Recall that this is equivalent to finding a decision scheme with maximal robustness.

The first heuristic is the simplest type of greedy algorithm. We start with the attainable set containing only the realisation where every source is honest. Then, we add the most probable realisation that does not break the attainability property:

```
input : A distribution $\Phi$ and the sources $\mathcal{S}$
output: A maximal attainable set of realisations $R$
1 $R := \{\mathcal{S}\}$;
2 while $R$ changed do
3     $C := \{\mathbf{r} \mid \mathbf{r} \notin R, \mathbb{A}(R \cup \{\mathbf{r}\})\}$;
4     $b := \mathrm{argmax}_{c \in C}\, \Phi(c)$;
5     $R := R \cup \{b\}$;
6 end
```

Algorithm 1: The basic greedy heuristic.

An example where Algorithm 1 performs poorly, is when $\Phi(\{s_i\})$ happens to barely be the largest value. In this case, the singleton $\{s_i\}$ is added to R first, and afterwards only realisations that contain s_i are attainable, so the result must be $R = \{\mathbf{r} \mid s_i \in \mathbf{r}\}$. The early choice for $\{s_i\}$ commits us to all other choices.

We can change Algorithm 1 into Algorithm 1b by only considering candidates C that can be added without committing to other choices. Substitute line **2** by:

$$\mathbf{2b} \quad C := \{\mathbf{r} \mid \mathbf{r} \notin R, \mathbb{A}(R \cup \{\mathbf{r}\}), \forall_{\mathbf{r}' \supset \mathbf{r}} \mathbf{r}' \in R\}$$

An example where Algorithm 1b performs badly is when $\Phi(\mathcal{S} \setminus \{s_i\})$ is small, but larger than the other realisations of rank $n - 1$ and $\Phi(\{s_i\})$ is arbitrarily large. Since the greedy algorithm does not yet consider the valuable realisation $\{s_i\}$, it picks an incompatible realisation.

Perhaps it is possible to balance the two concerns (not committing and not seeing ahead). Instead of not being able to add realisations that commit us to adding additional realisations later, we take into account the opportunity cost of doing so. The value of a realisation is the average value of all realisations that we would be committed to adding.

Instead of substituting line **2** by **2b**, we would substitute line **3** by **3c** to obtain Algorithm 1c:

$$\textbf{3c} \quad b := \mathrm{argmax}_{c \in C} \sum_{c \subseteq \mathbf{r} \subseteq \mathcal{S} | \mathbf{r} \notin R} \Phi(\mathbf{r}) / \sum_{c \subseteq \mathbf{r} \subseteq \mathcal{S} | \mathbf{r} \notin R} 1$$

None of these heuristics are generally optimal. However, interestingly, they are all optimal when Φ is monotonic:

Proposition 4. *If Φ is monotonic, then Algorithms 1, 1b and 1c are optimal.*

Proof. Algorithm 1 always picks the larger realisation from a complementary pair. Its output is therefore optimal. Since, if $\mathbf{r} \subseteq \mathbf{r}'$, then $\Phi(\mathbf{r}) < \Phi(\mathbf{r}')$, the greedy choice is the same every time in all three heuristics, hence they provide the same (optimal) output. □

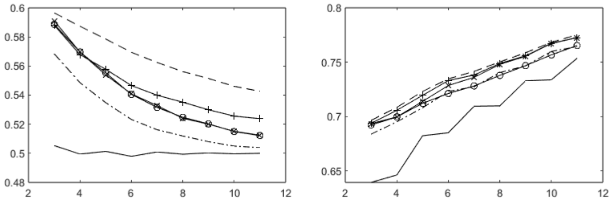
Proposition 4 means that in scenarios where more honest sources means more probable realisation, the proposed schemes can be optimal. In other words, if it is ensured that more favorable realisations are more probable, then it seems to be easier to achieve optimality, which is reasonable in practise.

Besides greedy algorithms, we can perform a local search. Given an attainable set, we can add a realisation $\mathbf{r}$, remove all realisations that do not intersect with $\mathbf{r}$, and add the complements of the realisations that were removed. The result will be an attainable set. We call this process *swapping* in $\mathbf{r}$. The local search simply looks for good candidates to swap; and does so in a greedy fashion:

```
input : A distribution $\Phi$, the sources $\mathcal{S}$ and an initial
        attainable set of realisations $R$
output: A maximal attainable set of realisations $R$
while $R$ changed do
    $C := \{\mathbf{r} \mid \mathbf{r} \notin R\}$;
    $b := \mathrm{argmax}_{c \in C} \sum_{c \subseteq \mathbf{r} \vee (c \in R \wedge c \not\subseteq \bar{\mathbf{r}})} \Phi(\mathbf{r})$;
    $R := \{\mathbf{r} \mid b \subseteq \mathbf{r} \vee (\mathbf{r} \in R \wedge b \cap \mathbf{r} \neq \emptyset)\}$;
end
```

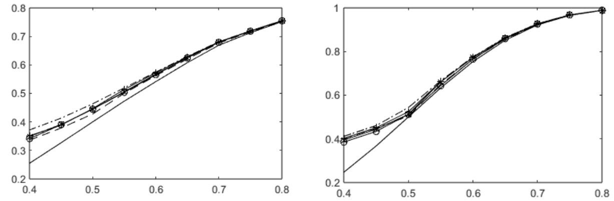
Algorithm 2: A greedy local search heuristic.

We can generalise the algorithm to consider swapping in pairs of realisations, or even k-tuples of realisations. The complexity of a single step – that is an iteration of the main while loop – is $O(N^k)$, where N is the number of realisations, and k the size of the tuples.



(a) With p=0.5 and $3 \leq n \leq 11$. (b) With p=0.6 and $3 \leq n \leq 11$.

Fig. 5. Performance of heuristics as the number of sources varies.



(a) With n=5 and $0.4 \leq p \leq 0.8$. (b) With n=11 and $0.4 \leq p \leq 0.8$.

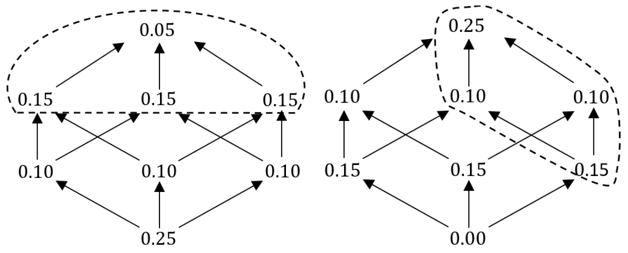
Fig. 6. Performance of heuristics as the quality of sources varies.

The various heuristics are tested on various amounts of sources of varying quality. In every graph, the solid line represents majority rule, the dash-dotted line is MPR, the +'s are Algorithm 1, the x's are Algorithm 1b, the circles are Algorithm 1c, and the dashed line is Algoithm 2. In all graphs, the probability of each coalition is randomly generated, for $p = 0.5$, these probabilities are uniformly random, but for different p-values, the probabilities of coalitions are necessarily skewed depending on their size. In Figure 5, we compare the practical strength of the heuristics, with increasingly many sources on the x-axis, from $n = 3$ to $n = 11$. Since the size of Φ is exponential in n, full analysis of heuristics for large n is infeasible. In Figure 5(a), the marginal probability of honesty is $p = 0.5$ and in Figure 5(b), it is $p = 0.65$. In Figure 6, we compare the practical strength of the heurstics, with increasing p-values on the x-axis. In Figure 6(a), there are 5 sources, and in Figure 6(b) there are 11. Majority rule is clearly inferior to the other heuristics. This is not surprising, as it does not take the probabilities of Φ into account. MPR generally performs somewhat worse than the other heuristics, and the local search algorithm (Algorithm 2) is generally the best. As p increases, Φ naturally becomes more friendly, as larger realisations become increasingly probable. As a result, the distinction between the heuristics naturally decreases, as p increases. Obviously, all heuristics increase in quality as p increases. This is generally true for n values as well, if $p \gg 1/2$. Surprisingly, if $p = 0.5$, then more sources result in worse performance. If variance between probabilities of realisations is removed, than the performance must be 0.5, but given there is variance, heuristics can smartly select positive outliers. We conjecture that this effect diminishes as n increases, so we should expect convergence to 0.5. Based on our experiments, we can conclude that MPR tends to work reasonably well, especially if malicious users are relatively rare. But if there are more malicious users, our proposed heuristics may improve accuracy.

4.3 Source-Uniform Φ-Distributions

The general problem of finding a decision scheme with a certain lower-bound robustness (or with optimal robustness) is NP-hard. However, for some specific Φ, there are polynomial-time solutions to find such decision schemes. For example, we have already shown in Theorem 10 and Theorem 6 that certain kind of Φ can make MPR be optimal.

In some contexts, there is no reason to distinguish different strangers. For instance, it does not matter whether a rating came from unknown users "Larry182" or "Bob42". Below we consider a special scenario where the probabilities of realisations are only influenced by "how many sources are honest" but not by "who are honest" – called *source-uniform* distributions. For example, consider the situation in Figure 7(a). Realisations on the same rank are equiprobable.



(a) Optimal DS is trivial here. (b) Majority rule is not optimal.

Fig. 7. Examples of source-uniform distributions ϕ.

Intuitively, since all sources are equal, there would not be a reason to trust a minority over a majority; i.e. to use the majority rule. Note that WMV defaults to the majority rule, since all sources are equally weighted. It happens to be optimal in Figure 7(a), as shown by the dashed line. However, as seen in Figure 7(b), this is not always the case. In that example, it is better to swap out a rank 2 realisation ($\mathbf{r} \in \mathcal{R}_2$) for its rank 1 complement. In the example, $\{s_1, s_2\}$ was swapped for $\{s_3\}$, going from 0.55 to 0.6 – one of the optimums. Unfortunately, in cases where the majority rule is not optimal, the useful minority opinion can only be used a small fraction of the time. We aim to quantify this below:

Lemma 5. *Let $\mathcal{R}$ be a maximal attainable set of realisations of n sources. For $k < \frac{n}{2} - 1$, the ratio $|\mathcal{R}_{k+1}| : |\mathcal{R}_k|$ is at least $\binom{n-1}{k} : \binom{n-1}{k-1}$.*

Proof. Since $\mathcal{R}$ is a maximal intersecting family, the rank $\mathcal{R}_{k+1}$ consists of $(k+1)$-element realisations which are supersets of the realisations in $\mathcal{R}_k$.

Define $A = \{\overline{\mathbf{r}} | \mathbf{r} \in \mathcal{R}_k\}$, and $B = \{\overline{\mathbf{r}} | \mathbf{r} \in \mathcal{R}_{k+1}\}$. Trivially, $|A| = |\mathcal{R}_k|$ and $|B| = |\mathcal{R}_{k+1}|$. The elements in A and B have cardinality $n-k$ and $n-k-1$, respectively. Since $\mathbf{r} \supseteq \mathbf{r}'$ iff $\overline{\mathbf{r}} \subseteq \overline{\mathbf{r}'}$, B is the set of $(n-k-1)$-element subsets of realisations in A. We can apply Proposition 1:

$$\frac{|\mathcal{R}_{k+1}|}{|\mathcal{R}_k|} = \frac{|B|}{|A|} \geq \frac{\binom{x}{n-k-1}}{\binom{x}{n-k}} = \frac{n-k}{x-n+k+1}$$

According to Theorem 2, $|\mathcal{R}_k| \leq \binom{n-1}{k-1} = \binom{n-1}{n-k}$. So $\binom{x}{n-k} = |A| = |\mathcal{R}_k| \leq \binom{n-1}{n-k}$. Therefore $x \leq n-1$, and thus $x - n + k + 1 \leq k$. Finally:

$$\frac{|\mathcal{R}_{k+1}|}{|\mathcal{R}_k|} \geq \frac{n-k}{x-n+k+1} \geq \frac{n-k}{k} = \frac{\binom{n-1}{k}}{\binom{n-1}{k-1}}$$

□

It straightforwardly follows that the maximal ratio between a rank and the ones above it, is obtained by the values dictated by the Erdös-Ko-Rado theorem (Theorem 2).

Corollary 2. *For $k < l < \frac{n}{2}$, when defined, the ratio $|\mathcal{R}_\ell| : |\mathcal{R}_k|$ is maximal when $|\mathcal{R}_\ell| = \binom{n-1}{\ell-1}$ and $|\mathcal{R}_k| = \binom{n-1}{k-1}$.*

This allows us to prove the main result on source-uniform distributions. Namely that for some $0 < k \leq n/2$, the optimal attainable set includes 0 realisations of the rank $j < k$, and maximal for $j \geq k$. The problem of finding the optimum reduces to finding k out of $n/2$ possible values:

Theorem 12. *If $|\mathbf{r}| = |\mathbf{r}'|$ implies $\Phi(\mathbf{r}) = \Phi(\mathbf{r}')$, then any optimal deterministic decision scheme $\mathscr{D}$ either has 0 or $\binom{n-1}{k-1}$ realisations of size $0 < k < n/2$.*

Proof. Define $\mathbf{v}$ such that $v_k = (\Phi(k) - \Phi(n-k)) \cdot \binom{n-1}{k-1}$, and $\mathbf{w}$ as $w_k = \frac{|\mathcal{R}_k|}{\binom{n-1}{k-1}}$. Due to Corollary 2, $0 \leq w_k \leq 1$, for all $0 \leq k < n/2$. From Lemma 5, we know $\mathbf{w}$ is monotonically non-decreasing. Every attainable set corresponds to some vector $\mathbf{w}$ on $[0, 1]$; and our goal is to maximise $S_{\mathbf{v}}(\mathbf{w}) = \sum_{0 \leq i \leq n/2} v_k \cdot w_k = \sum_{0 \leq i \leq n/2} (\Phi(k) - \Phi(n-k)) \cdot |\mathcal{R}_k|$.

Suppose $\mathbf{w}'$ maximises $S_{\mathbf{v}}$ and has w_i as the first non-zero element and w_j as the last non-unit element. Changing $w_i \ldots w_j$ by x changes $S_{\mathbf{v}}(\mathbf{w})$ linearly. Either we can increase $w_i \ldots w_j$ by $(1 - w_j)$ while not decreasing $S_{\mathbf{v}}(\mathbf{w})$ or we can decrease $w_i \ldots w_j$ by w_i while not decreasing $S_{\mathbf{v}}(\mathbf{w})$. Either w_i becomes 0 or w_j becomes 1. By repetition, we can eliminate all non-zero, non-unit values, while staying optimal. Such a $\mathbf{w}$ corresponds to a set of realisations where each rank $k < n/2$ has 0 or $\binom{n-1}{k-1}$ elements. □

Theorem 12 defines what an optimal attainable set should include, when realisations of the same size are equally probable, thus proving a clue in finding the optimal decision scheme. Interestingly, while the theorem provides an optimal set of realisations, *it does not provide a general optimal decision scheme*. A probabilistic decision scheme may outperform the deterministic one. Take an example with 11 sources. Suppose there is a 0.12 probability that everyone is honest, a 0.05 probability for each realisation of size 10, and a 0.001 probability for each realisation of size 4 (for a total of $0.12 + 11 \cdot 0.05 + \binom{11}{4} \cdot 0.001 = 1$). Following the deterministic strategy, we see that an optimal scheme is to believe any group of size 4 or larger that contains s_1 or any group of size 10 or larger, for a probability of being correct $0.12 + 11 \cdot 0.05 + \binom{10}{3} \cdot 0.001 = 0.79$. An alternative strategy is to uniformly randomly pick from any option endorsed by at least 4 sources. If 11 or 10 sources are honest, the decision is correct with probability 1, but if 4 sources are honest, there are at most 2 different options endorsed by at least 4 sources ($3 \cdot 4 > 11$), so the decision is correct with probability 0.5, for a total of $0.12 + 11 \cdot 0.5 + \binom{11}{4} \cdot 0.001 \cdot 0.5 = 0.835 > 0.79$. In the future work, we would like to explore such probabilistic decision schemes further.

5 Related Work

Multi-source decision making comes in many shapes and sizes. The purpose may be to make a collective decision or a personal decision. The feedback may be guided by knowledge, belief, preference or experience (or a mix thereof). For instance, computational social choice theory includes problems such as whether a voting system is democratic [19], or whether a majority vote is correct [20]. Ensemble methods in machine learning are proposed to improve the learning performance by aggregating the predictions of multiple learning algorithms [21]. Crowdsourcing (crowdsensing) hires a group of non-expert workers (or sensors) for a decision making task [22], [23], [24]. Multi-source information fusion like group decision making (GDM) considers how to combine diverse preferences or evidences [25]. Truth discovery aims to find the true value of a data item, based on the sources providing conflicting information [22], [26], [27], [28].

The above domains share similarities in the sense of aggregating feedback to derive a decision, but also differ regarding for example the purpose of decision making (i.e., whether the decision is to respect source preference or to figure out the truth), and whether they consider malicious sources. In this section we discuss these in detail and also how our work finds its position. Specifically, in Section 5.1, we discuss collective decision-making scenarios where voting is a representative and source preference is the main concern. In Section 5.3, we discuss personal decision-making scenarios where decision correctness faced with diverse source reliability is the main concern. And this section is further divided into two parts based on whether source unreliability results from malicious motivation.

5.1 Collective Decision Making Considering Diverse Preference

In some scenarios, the goal of decision making is to respect the preferences of individuals and derive a collective decision which is widely acceptable, e.g., voting for democracy. This type of problems also include the classical *team problem* [29] (team members sharing a same loss function work together to reach a joint decision), *linear opinion pool [30]* (individual opinions are represented by probability distributions which are aggregated using a linear function), and the more recent LSGDM (large scale group decision making [25], [31]). An underlying assumption is that the experts or individual sources are usually reliable [5], contributing diverse knowledge for decision making. In our work, however, decisions to make are personal, rather than a common one that affects the sources.

5.2 Personal Decision Making with Unreliable but Non-malicious Sources

Sometimes, the goal of decision making with multiple sources is to determine the truth or to choose the correct decision, faced with conflicting feedback and diverse reliability of sources, which is more similar to our setting. Such a decision is usually "personal" and does not need to be accepted by the sources. Different as the scenarios above, source reliability is a main concern and their preference is usually not defined or considered. Depending on the scenarios, feedback of a source can be unreliable because of insufficient experience, low competence or malice. In this section, we consider unreliable but non-malicious sources and in the next section, we consider malicious sources.

In some scenarios, unreliable feedback are assumed to be from incompetent sources who are not strategic. Typically, faced with conflicting feedback, sources competence is modelled using probabilities of reporting accurately or weights. For example, the epistemic branches of social choice theory, and specifically, Jury Theorems, model individual competence and study the relation between the correctness of decisions and the size of the crowd [4]. In ensemble learning (e.g., classifier ensembles), usually different algorithms are assigned different weights in aggregation based on their prediction accuracy [21], [32], [33]. In LSGDM, Chen et al. weight public opinion based on its distance to the expert opinion before combining them [25]. In combining belief functions, Yong et al., propose to assign less weights to the sets of evidence that are more different or highly conflicting with the others [34]. Some approaches discount the feedback before combining them e.g., *trade-off rules* [35], *discount rules* [36]. Fengrui et al. propose a crowdsensing-based framework for the collection and dissemination of the information about the urban parking spaces [37]. They assume unreliable reports are generated unintentionally. And the reliability of the parking information is determined by how knowledgeable a driver is about the parking area [37].

There also exist scenarios which does not explicitly distinguish the reason behind feedback unreliability. For example, in truth discovery, a common assumption is that a source is more reliable if its feedback is closer to the estimated truth [26], [28], regardless of whether the unreliability results from malice or other factors. Typically two steps are iteratively used: evaluating the reliability of the sources based on their distance to the truth and aggregating their feedback based on the derived reliability values as the truth.

Although we also consider source reliability, it differs intrinsically by considering an adversarial setting, where unreliable information or feedback results explicitly from *malicious* sources rather than incompetent or faulty sources. Malicious sources are a particular concern for security-critical domains. They strategically adapt what they report based on the decision mechanism, whereas honest but low-competence sources report independently of the decision mechanism. In an adversarial setting, the reliability of a source cannot simply be modelled as how probable it provides accurate feedback. Malicious sources may report accurately sometimes as part of their strategies (e.g., to camouflage), the probabilities of which are typically unknown.

5.3 Personal Decision Making with Malicious Sources

Unreliable feedback can be introduced by malicious sources who aim to manipulate decisions. In this section, we discuss related works which explicitly consider malicious sources.

5.3.1 Data Poisoning Attacks in Crowdsourcing/Crowdsensing

In crowdsourcing or crowdsensing, multiple sources (e.g. workers, sensors or smart devices) may be employed for a specific task such as providing labels for training a learning algorithm, providing location-related information (spatial crowdsourcing e.g., Uber and Waze [24], [28]), and providing sensing information in vehicular networks [24], [38] etc. However, crowd sourced (or sensed) information can be unreliable, due to insufficient knowledge, malfunction, or malice.

There are studies specifically dealing with malicious workers, sometimes named as data poisoning attackers [39], [40] in this context. Mohsen et al. propose an iterative filtering-based algorithm to defend against node collusion attacks in wireless sensor networks [41]. Francesco et al. take an approach which employs Mobile Trusted Participants (MTPs, who regularly submit reliable reports that are used to validate sensory reports of the others) to deal with malicious users in crowdsensing. [42]. Three types of

[5] Unreliable feedback may be considered, e.g., in LSGDM [25]

attacks are targeted: 1) Corruption attacks, where attackers are assigned a fixed probability of reporting reliably; 2) On-off attacks, where attackers are assumed to alternate between reporting reliably and unreliably (thus similar to Camouflage attacks [6]); and 3) Collusion attacks, where attackers are assumed to report the same unreliable information. Zonghao et al. model attackers as to maximize the decision mistakes and they propose to filter sources with high error levels before truth discovery [43]. Tahmasebian et al. propose to detect and increase the answers for boundary tasks, where the worker feedback can hardly reach a consensus [44]. Yuan et al. propose a detection mechanism that clusters workers with a similarity function and then identify attackers by introducing golden tasks or questions (pre-annotated tasks) [45]. Considering that Sybil attackers may try to evade detection by identifying such tasks or coordination, Wang et al. propose a probabilistic task assignment method to camouflage golden tasks from Sybil workers [22]. Yu proposes a deep generative model based method to identify Sybil attackers in crowdsourced navigation systems (e.g., Waze) [46]. Our work differs from these approaches in that we do not assume how malicious workers might behave or what their motivations are. we do not assume that a more unreliable worker should lie more often as in some truth discovery-based approaches [22], [26], [27]. Our modelling of source reliability (i.e., *honesty*) allows malicious sources to choose feedback from the entire attack space with any probabilities, assuming much less about malice behavior and follows a security-by-design principle.

Besides detection-based approaches to deal with malicious sources, there also exists plenty of research focusing on establishing accurate trust evaluation mechanisms to e.g., achieve more reliable source recruitment or sensing service. For exmaple, Jagabathula et al. propose two reputation algoritms to distinguish honest but unreliable workers from malicious workers, with the former taking deterministic strategies while the latter taking arbitrary strategies [47]. And they filter out malicious workers. There are approaches in crowdsensing where they evaluate trustworthiness of sensing entities like vehicles considering privacy preserving, to select or filter their sensing data, and do not consider whether to aggregate these data or to derive a correct decision from it (which is our focus). For instance, Ni et al., propose a mobile crowdsensing scheme where trust levels of recruited users are employed to filter (with thresholds) the sensing reports from their devices [48]. The selected sensing reports are delivered to customers, whose feedback is then used to update user trust levels. Whether to aggregate the selected sensing reports is not considered. Chen et al. proposes an adaptive trust management system for social IoT system [49]. Nguyen et al. propose to evaluate trust relationships between users, based on which more trustworthy users can be recruited for a sensory task [50]. Liu et al. propose a trust evaluation mechanism to select reliable data for data fusion in vehicular networks [38]. How to aggregate data is not studied. Liu et al. propose a trust management scheme for emergency message dissemination in vehicular networks [51]. A vehicular who receives an emergency message can validate its truthfulness and update the reputation of the vehicle who broadcastered it. The receiver only makes a decision if he trusts the broadcaster, and how that decision is made is not considered. Cheng et al. proposes a reputation management scheme to evaluate reliability of sensing vehicles in vehicular networks [24]. And they do not consider how to evaluate the quality of sensing data with vehicle reliability values. Considering the existence of various trust evaluation mechanisms, in this work we assume source trust values are given and focus more on secure trust-based decision making part. Moreover, while the approaches above typically filter out adversaries with trust evaluation, we allow their existence (a.l.a their honesty degrees are above a half). It has been proved to be fallacious to discount or filter feedback that deviates from the majority or from the first-hand evidence [52].

The above approaches generally take the perspective of a defender whose aim is to reduce the influence of attacks, which is the same as our perspective. There also exist approaches focusing on designing attacks against existing crowdsoucing quality control methods. Checco et al. propose a way to detect golden tasks, where malicious workers collude and use a decentralised machine learning inferential system [53] to classify tasks. Li et al. propose to attack crowdsensing systems using Truthfinder (a classical truth discovery framework) with a deep reinforcement learning-based method [54]. Miao et al. propose attacks targeting at Dawid-Skene model based crowdsoucing: malicious workers who purposely introduce wrong labels attempt to camouflage themselves by agreeing with normal workers sometimes [55]. And they formulate the mechanism as an optimization problem which aims to maximize the attacking successful rate and also attackers' reliability degrees. Fang et al. also formulate the attack as an optimization problem, where the objective is to maximize the estimation errors for data items that attackers choose [56].

For either defensive or offensive approaches, how much attackers know about the decision-making system is an important consideration. Attackers may know a system quite well s.t. they know how the aggregation mechanism works and also what honest workers report (e.g., full-knowledge attackers [56] or white-box attacks [39]), which is the same as our setting of attackers. They may also know less: e.g, only reports from part of honest workers (partial-knowledge attackers [56]). Both full-knowledge attackers and partial-knowledge attackers are considered in [39], [55], [56].

5.3.2 Fraudulent Online Reviews

There are some other scenarios which may not straightforwardly be modelled as multi-source decision making, but are very related. A typical type of examples are online review platforms like those in e-commerce (e.g., Amazon). Reviews are usually recorded experience or opinions of some users, and are supposed to provide reference to the others or be used for recommendations. Similarly as in crowdsourcing, some reviews may be fake or fraudulent, generated by malicious entities like those hired from crowdturfing sites. But there is a main difference: online reviews are generally not created for a specific decision-making task as in crowdsourcing, but can potentially influence multiple decisions, e.g., influencing multiple users and their choices like whether to make a purchase choice.

There is plenty of research aiming to detect fraudulent reviews or evaluate reviewers' trustworthiness [57], [57],

[58]. Shehnepoor et al. model reviews as a heterogeneous network and spam detection as a classification problem [57]. Kumar et al. model fairness, goodness and reliability as intrinsic properties of users, products and ratings respectively. They then define several axioms to describe the inter-dependency of these properties and further propose a formulation to satisfy the axioms [59]. Graph-based models like GNNs are very popular. Liu et al. propose a GNN-based imbalanced learning for fraud detection [60]. The problem of malicious reviews is closely related with crowdturfing, where crowdsourcing sites are used to hire workers to introduce fake reviews to a review platform [61]. Hernandez et al. propose a method of fraud de-anonymisation based on the maximum likelihood estimation method, to uncover real identities of malicious workers that control the fraud accounts in online review platforms [62].

Compared with the detection and filtering-based methods above, our work differs in several ways. First, to allow the existence of malicious sources is a core principle, and rather than attempting to detect or filter malicious feedback, we minimise the probability to be manipulated by it. We do not rely on some assumed features of malicious behaviour and are in a more proactive position faced with unknown malicious behaviour. Second, we always assume a *white-box* attack scenario where attackers are aware of the decision mechanism. This is typical for cyber-security applications and will put a decision maker in a more proactive position. Last but not the least, accurately estimating feedback or source reliability is not the focus of our work, but minimising the probability of being manipulated is.

Finally, although not a typical form of "decision making", there is also issue of malicious input in the domains of machine learning and recommender systems. For example, there exists research studying malicious behaviour of attempting to mislead a training algorithm (e.g., for image classification or recommendation) by polluting its training data e.g., data poisoning attacks in recommender systems [63] and adversarial machine learning [60], [64], [65], etc. For these approaches, manipulation is a crucial concern. For a security critical system, a proof that the effect of manipulation is minimal is – arguably – no less important than a claimed accuracy under non-strategic feedback.

6 Conclusion

In this paper, our goal is to introduce a general methodology for making decisions that are almost certainly correct in the existence of malicious sources (*attackers*), which is crucial for especially security-critical decision making. Sources provide discrete feedback options to a decision maker, one of which represents the truth and corresponds to the correct decision. The influence that the feedback has on decisions is determined by the probability that a source is honest.

We defined ϵ-robustness and optimality. Based on reasoning about the distribution of source honesty, we proposed the Most Probable Realisation decision mechanism. When sources have an independent probability of being honest, the probability of MPR deciding incorrectly is bounded to a very small threshold ϵ. MPR is also proved to be optimal, meaning that there is no mechanism with smaller bound of inaccuracy. When sources are dependent (e.g. due to collusion), MPR remains to be optimal under a class of realisation distributions (monotonic distributions), but is not generally optimal. We proved that in this case, it is NP-hard to find a decision mechanism with a given bound of inaccuracy, and thus also NP-hard to find the optimal decision mechanism. We investigated some heuristics about how to make good decisions. Finally, we looked at another class of realisation distributions (source-uniform), where we found the optimal deterministic decision scheme. With an example, we demonstrate that a stochastic decision scheme can outperform the optimal deterministic scheme (for some non-monotonic distribution).

Whereas related work considers the probability that certain feedback is truthful, we consider the probability that the feedback is honest. Honest feedback is truthful, but malicious (non-honest) feedback is not necessarily false. As opposed to existing approaches, our provably robust scheme is based on the entire attack space for malicious users, allowing them to lie or be truthful in any way. Rather than focusing on making the right decision with some given feedback, our approach takes a step back and asks under which circumstances we want to be *non-manipulable* – to always make the right decision. Typically, we want to be non-manipulable under the most probable circumstances. A core contribution is the demonstration of a novel technique, namely the use of realisations. Realisations make it possible to investigate whether a decision maker is manipulable, *without* studying the actual manipulative strategies (or studying combinations of feedback that change decisions).

Our approach can be generalised or applied to variations of the problem. In particular, it is interesting to further formalise the stochastic decision schemes. Rather than having a set of intersecting realisations, the objective is to have an assignment of weight to realisations that follows a set of rules. The deterministic model presented would be a special case of the more general formalism. Furthermore, our model can be extended to allow honest users to make mistakes, which unlike malicious users, is not strategic behaviour. Currently, we can conservatively model this as an increased probability of malice – but a more precise formalisation would be useful. Another direction for the work is to look at sequences of decisions and feedback. Intuitively, if a source has provided truthful feedback in the past, then its honesty degree p-value should increase. However, a malicious source may strategically provide truthful feedback to increase its p-value (to camouflage). A malicious source that never provides false feedback is not a threat, so there must be a balance to be struck. A realisation-based approach may lead to a formal mechanism to update p-values properly.

7 Limitation and Future Work

In the interest of having a simple model and focusing on provable decision correctness, we made the following assumptions. First, we assume that honest sources would simply report the truth, while in practise it is more complicated. For example, as it is pointed out in [66], even if a user is honest, his feedback can be biased in multiple ways. It would be interesting to extend our decision schemes to cover situations where bias from honest sources is considered. Honest sources may also make mistakes sometimes e.g., malfunctioned sensors. In [6], we made a detailed discussion about the strong 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. Second, honest sources in our setting share the same set of evidences, and there is only one ground truth which corresponds to the correct decision. And accordingly conflicting feedback results only from the existence of malice. This is appropriate for lots of security-critical decision scenarios like whether an app is a malware, but not applicable where honest sources have different observations or evidences [6]. Last but not the least, in our modeling sources do not have uncertainty with their feedback, which takes the form of discrete and finite options but not distributions or belief functions. But in reality, we may need to allow the existence of uncertainty in the modelling by using distributions or belief functions []. In future work, we would like to explore a broader range of scenarios by breaking the assumptions above to get closer to more practical solutions.

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] F. Dietrich and K. Spiekermann, "Jury Theorems," in *The Stanford Encyclopedia of Philosophy*, Summer 2022 ed., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2022.

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

[6] T. Muller, D. Wang, and J. Sun, "Provably robust decisions based on potentially malicious sources of information," in *2020 IEEE 33rd Computer Security Foundations Symposium (CSF)*. IEEE, 2020, pp. 411–424.

[7] G. Di Battista and R. Tamassia, "Algorithms for plane representations of acyclic digraphs," *Theoretical Computer Science*, vol. 61, no. 2-3, pp. 175–198, 1988.

[8] D. Gerbner and B. Patkós, *Extremal finite set theory*. Chapman and Hall/CRC, 2018.

[9] M. Deza and P. Frankl, "Erdös–ko–rado theorem—22 years later," *SIAM Journal on Algebraic Discrete Methods*, vol. 4, no. 4, pp. 419–431, 1983.

[10] J. B. Kruskal, "The number of simplices in a complex," *Mathematical optimization techniques*, vol. 10, pp. 251–278, 1963.

[11] L. Lovász, *Combinatorial Problems and Exercises*, ser. AMS/Chelsea publication. North-Holland Publishing Company, 1979.

[12] 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.

[13] S. J. Brams and P. C. Fishburn, "Voting procedures," *Handbook of social choice and welfare*, vol. 1, pp. 173–236, 2002.

[14] S. Nitzan and J. Paroush, "Optimal decision rules in uncertain dichotomous choice situations," *International Economic Review*, pp. 289–297, 1982.

[15] J. R. Douceur, "The sybil attack," in *International workshop on peer-to-peer systems*. Springer, 2002, pp. 251–260.

[16] R. D. Luce and A. D. Perry, "A method of matrix analysis of group structure," *Psychometrika*, vol. 14, no. 2, pp. 95–116, 1949.

[17] R. M. Karp, "Reducibility among combinatorial problems," in *Complexity of computer computations*. Springer, 1972, pp. 85–103.

[18] M. R. Garey, "A guide to the theory of np-completeness," *Computers and intractability*, 1979.

[19] C. List, "Social Choice Theory," in *The Stanford Encyclopedia of Philosophy*, Spring 2022 ed., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2022.

[20] P. J. Boland, "Majority systems and the condorcet jury theorem," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 38, no. 3, pp. 181–189, 1989.

[21] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[22] Y. Wang, K. Wang, and C. Miao, "Truth discovery against strategic sybil attack in crowdsourcing," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 95–104.

[23] Y. Zheng, H. Duan, and C. Wang, "Learning the truth privately and confidently: Encrypted confidence-aware truth discovery in mobile crowdsensing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 10, pp. 2475–2489, 2018.

[24] Y. Cheng, J. Ma, Z. Liu, Y. Wu, K. Wei, and C. Dong, "A lightweight privacy preservation scheme with efficient reputation management for mobile crowdsensing in vehicular networks," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[25] X. Chen, W. Zhang, X. Xu, and W. Cao, "A public and large-scale expert information fusion method and its application: Mining public opinion via sentiment analysis and measuring public dynamic reliability," *Information Fusion*, vol. 78, pp. 71–85, 2022.

[26] X. Yin, J. Han, and S. Y. Philip, "Truth discovery with multiple conflicting information providers on the web," *IEEE transactions on knowledge and data engineering*, vol. 20, no. 6, pp. 796–808, 2008.

[27] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," *ACM Sigkdd Explorations Newsletter*, vol. 17, no. 2, pp. 1–16, 2016.

[28] C. Zhang, L. Zhu, C. Xu, X. Liu, and K. Sharif, "Reliable and privacy-preserving truth discovery for mobile crowdsensing systems," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 3, pp. 1245–1260, 2019.

[29] R. Radner, "Team decision problems," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 857–881, 1962.

[30] M. H. DeGroot and J. Mortera, "Optimal linear opinion pools," *Management Science*, vol. 37, no. 5, pp. 546–558, 1991.

[31] Z.-j. Du, H.-y. Luo, X.-d. Lin, and S.-m. Yu, "A trust-similarity analysis-based clustering method for large-scale group decision-making under a social network," *Information Fusion*, vol. 63, pp. 13–29, 2020.

[32] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho, "Multiple classifiers in biometrics. part 1: Fundamentals and review," *Information Fusion*, vol. 44, pp. 57–64, 2018.

[33] J.-C. Xie and C.-M. Pun, "Deep and ordinal ensemble learning for human age estimation from facial images," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2361–2374, 2020.

[34] D. Yong, S. WenKang, Z. ZhenFu, and L. Qi, "Combining belief functions based on distance of evidence," *Decision support systems*, vol. 38, no. 3, pp. 489–493, 2004.

[35] G. Shafer, *A mathematical theory of evidence*. Princeton university press, 1976, vol. 42.

[36] R. R. Yager, "Approximate reasoning as a basis for rule-based expert systems," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 4, pp. 636–643, 1984.

[37] F. Shi, D. Wu, D. I. Arkhipov, Q. Liu, A. C. Regan, and J. A. McCann, "Parkcrowd: Reliable crowdsensing for aggregation and dissemination of parking space information," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 11, pp. 4032–4044, 2018.

[38] Z. Liu, J. Ma, J. Weng, F. Huang, Y. Wu, L. Wei, and Y. Li, "Lppte: A lightweight privacy-preserving trust evaluation scheme for facilitating distributed data fusion in cooperative vehicular safety applications," *Information Fusion*, vol. 73, pp. 144–156, 2021.

[39] F. Tahmasebian, L. Xiong, M. Sotoodeh, and V. Sunderam, "Crowdsourcing under data poisoning attacks: A comparative study," in *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 2020, pp. 310–332.

[40] C. Miao, Q. Li, H. Xiao, W. Jiang, M. Huai, and L. Su, "Towards data poisoning attacks in crowd sensing systems," in *Proceedings*

[6] For example, experts with different background may propose different suggestions in investment, or honest buyers may provide different ratings for a seller due to subjectivity

of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing, 2018, pp. 111–120.

[41] M. Rezvani, A. Ignjatovic, E. Bertino, and S. Jha, "Secure data aggregation technique for wireless sensor networks in the presence of collusion attacks," *IEEE transactions on Dependable and Secure Computing*, vol. 12, no. 1, pp. 98–110, 2014.

[42] F. Restuccia, P. Ferraro, T. S. Sanders, S. Silvestri, S. K. Das, and G. L. Re, "First: A framework for optimizing information quality in mobile crowdsensing systems," *ACM Transactions on Sensor Networks (TOSN)*, vol. 15, no. 1, pp. 1–35, 2018.

[43] Z. Huang, M. Pan, and Y. Gong, "Robust truth discovery against data poisoning in mobile crowdsensing," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[44] F. Tahmasebian, L. Xiong, M. Sotoodeh, and V. Sunderam, "Edge-infer: robust truth inference under data poisoning attack," in *2020 IEEE International Conference on Smart Data Services (SMDS)*. IEEE, 2020, pp. 45–52.

[45] D. Yuan, G. Li, Q. Li, and Y. Zheng, "Sybil defense in crowdsourcing platforms," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1529–1538.

[46] J. James, "Sybil attack identification for crowdsourced navigation: A self-supervised deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4622–4634, 2020.

[47] S. Jagabathula, L. Subramanian, and A. Venkataraman, "Identifying unreliable and adversarial workers in crowdsourced labeling tasks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 3233–3299, 2017.

[48] J. Ni, K. Zhang, Q. Xia, X. Lin, and X. S. Shen, "Enabling strong privacy preservation and accurate task allocation for mobile crowdsensing," *IEEE Transactions on Mobile Computing*, vol. 19, no. 6, pp. 1317–1331, 2019.

[49] R. Chen, F. Bao, and J. Guo, "Trust-based service management for social internet of things systems," *IEEE transactions on dependable and secure computing*, vol. 13, no. 6, pp. 684–696, 2015.

[50] N. B. Truong, G. M. Lee, T.-W. Um, and M. Mackay, "Trust evaluation mechanism for user recruitment in mobile crowd-sensing in the internet of things," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2705–2719, 2019.

[51] Z. Liu, J. Weng, J. Guo, J. Ma, F. Huang, H. Sun, and Y. Cheng, "Pptm: A privacy-preserving trust management scheme for emergency message dissemination in space–air–ground-integrated vehicular networks," *IEEE Internet of Things Journal*, vol. 9, no. 8, pp. 5943–5956, 2021.

[52] 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.

[53] A. Checco, J. Bates, and G. Demartini, "Quality control attack schemes in crowdsourcing," in *IJCAI*, 2019, pp. 6136–6140.

[54] M. Li, Y. Sun, H. Lu, S. Maharjan, and Z. Tian, "Deep reinforcement learning for partially observable data poisoning attack in crowdsensing systems," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6266–6278, 2019.

[55] C. Miao, Q. Li, L. Su, M. Huai, W. Jiang, and J. Gao, "Attack under disguise: An intelligent data poisoning attack mechanism in crowdsourcing," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 13–22.

[56] M. Fang, M. Sun, Q. Li, N. Z. Gong, J. Tian, and J. Liu, "Data poisoning attacks and defenses to crowdsourcing systems," in *Proceedings of the Web Conference 2021*, 2021, pp. 969–980.

[57] S. Shehnepoor, M. Salehi, R. Farahbakhsh, and N. Crespi, "Netspam: A network-based spam detection framework for reviews in online social media," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1585–1595, 2017.

[58] S. Sedhai and A. Sun, "Semi-supervised spam detection in twitter stream," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 1, pp. 169–175, 2017.

[59] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. Subrahmanian, "Rev2: Fraudulent user prediction in rating platforms," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018, pp. 333–341.

[60] Y. Liu, X. Ao, Z. Qin, J. Chi, J. Feng, H. Yang, and Q. He, "Pick and choose: a gnn-based imbalanced learning approach for fraud detection," in *Proceedings of the Web Conference 2021*, 2021, pp. 3168–3177.

[61] G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B. Y. Zhao, "Serf and turf: crowdturfing for fun and profit," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 679–688.

[62] N. Hernandez, M. Rahman, R. Recabarren, and B. Carbunar, "Fraud de-anonymization for fun and profit," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 115–130.

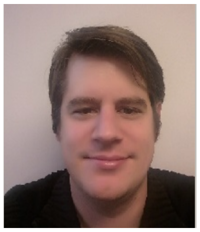
[63] X. Zhang, J. Chen, R. Zhang, C. Wang, and L. Liu, "Attacking recommender systems with plausible profile," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4788–4800, 2021.

[64] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2018, pp. 19–35.

[65] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data poisoning attacks against federated learning systems," in *European Symposium on Research in Computer Security*. Springer, 2020, pp. 480–501.

[66] 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.

Dongxia Wang is currently an assistant professor at the College of Control Science and Engineering at Zhejiang University. She obtained her PhD from School of Computer Science and Engineering in Nanyang Technological University, Singapore in 2018. Then she joined the Department of Computer Science in University of Oxford as a postdoc. She was also a research scientist in Singapore Management University. She got her Bachelor of Engineering degree in Information Engineering from Xi'an Jiaotong University, Xi'an, China in 2013. Her main research interests include multi-agent system, cyber security and trust management.



Tim Muller currently works as an assistant professor at the School of Computer Science at the University of Nottingham in the UK. He received his MSc. degree in Computer Science & Engineering at the Eindhoven University of Technology, the Netherlands, in 2009. Then he obtained his PhD degree in Computer Science at the University of Luxembourg, Luxembourg, in 2013, working on formalising computational trust. He worked on security and trust as a Research Fellow at Nanyang Technological University, Singapore, from 2013-2016. From 2016-2019, he was Departmental Lecturer in Security at the University of Oxford, United Kingdom. His research involves the security aspects surrounding computational trust, such as online ratings, reputation, attacks and robust systems. His interests are in the fields of Security, Formal Methods and AI.

Jun Sun is currently a full professor at School of Computing and Information Systems, Singapore Management University. He received Bachelor and PhD degrees in Computing Science from National University of Singapore (NUS) in 2002 and 2006. In 2007, he received the prestigious LEE KUAN YEW postdoctoral fellowship. He has been a faculty member since 2010 and was a visiting scholar at MIT from 2011-2012. Jun's research interests include software engineering, cyber-security and formal methods. He is the co-founder of the PAT model checker.