Autonomous Agents and Multi-Agent Systems manuscript No.
(will be inserted by the editor)

How to trust a few among many

Anthony Etuk, Timothy J. Norman, Murat Şensoy, and Mudhakar Srivatsa

Received: date / Accepted: date

Abstract The presence of numerous and disparate information sources available to support decision-making calls for efficient methods of harnessing their potential. Information sources may be unreliable, and misleading reports can affect decisions. Existing trust and reputation mechanisms typically rely on reports from as many sources as possible to mitigate the influence of misleading reports on decisions. In the real world, however, it is often the case that querying information sources can be costly in terms of energy, bandwidth, delay overheads, and other constraints. We present a model of source selection and fusion in resource-constrained environments, where there is uncertainty regarding the trustworthiness of sources. We exploit diversity among sources to stratify them into homogeneous subgroups to both minimise redundant sampling and mitigate the effect of certain biases. Through controlled experiments, we demonstrate that a diversity-based approach is robust to biases introduced due to dependencies among source reports, performs significantly better than existing approaches when sampling budget is limited and equally as good with an unlimited budget.

Keywords Trust, reputation, diversity, sampling

1 Introduction

The trustworthiness of information sources is an important factor in making informed and reliable decisions about what is true in the world. Decisions about what to do often depend on accurate assessments of environmental states: river water level, pressure in a pipeline, numbers and locations of casualties following a disaster. Typical approaches to estimating the value of some environmental state is to rely on reports from as many sources as possible; the underlying assumption being that exploiting the 'wisdom of the crowd' effect [38]

Anthony Etuk and Timothy J. Norman
University of Aberdeen, UK
Murat Şensoy
Department of Computing Science, Ozyegin
Mudhakar Srivatsa
IBM T. J. Watson Research Center, Hawthorne, NY, USA.
E-mail: murat.sensoy@ozyegin.edu.tr

minimises the influence of biased opinions. In many real-world contexts, however, capturing and distributing evidence can be costly. In distributed environments such as peer-to-peer and sensor networks, and pervasive computing applications, each participant is responsible for collecting and combining evidence from others due to the lack of a central authority or repository. In emergency response, for example, a decision-maker at some node in the network must make decisions in real time on the basis of high volumes of streaming information received from a variety of sources through different routes. Major constraints in these systems are bandwidth, delay overheads, and energy, motivating the need to minimise the number of messages exchanged. Furthermore, there is often no guarantee that evidence obtained from different sources are based on direct, independent observations. A notable example is in social sensing, where information shared by individuals (e.g., via a social network such as Facebook or Twitter) can be accessed by a wide audience [6, 26]. They may, in turn, report the same information later, possibly without any acknowledgement[1]. Accounting for dependencies in the source population is advantageous, not only as a means for minimising the cost of information acquisition, but also for making better assessments.

We argue that agents operating in complex, dynamic and constrained environments can adopt a model of *diversity* in order to minimise redundant sampling and mitigate the effect of certain biases. Exploiting source diversity may, for example, provide evidence from groups of sources with different perspectives on a problem. This has the potential to mitigate the risk of double-counting evidence due to correlated biases among group members. Broadly speaking, our view of diversity is a stratification of the source population, such that sources likely to provide similar reports are grouped together. Our requirement for diversity, however, goes beyond simply accounting for dependencies among information sources. We seek to capture richer information contexts, such as differences in expertise and perspectives, which may be exploited by a decision-maker for making more informed decisions. For example, the cost and risk analysis of interacting with certain groups of sources may serve to inform *whom* to approach under what circumstances [23]. Such groups may, for example, be communities in a geographic region, divisions in a corporation, or sensors owned by a specific organisation.

Thinking about diversity in populations of information/opinion providers is not a new idea; this is a common tactic used in the social sciences and by polling organisations. Shiller suggests that people who interact with each other regularly tend to think and behave similarly, and describes how, for example, political beliefs or opinions on policy issues tend to show geographical and social patterns [37]. This is often referred to as *herd mentality* (or herding) [32, 38]: the alignment of thoughts or behaviours of individuals in a group through local interactions. For example, individuals from the same organisation tend to behave in a similar manner based on certain codes of conduct or policies. In general, entities in different populations may have diverse beliefs about the state of the world. These populations, or subgroups, are often defined by a range of features (e.g., age, nationality, geography, religion) that may influence their behaviour. Exploiting correlations between behaviour and observable features of agents has also been explored in computational models of trust, where the problem addressed is which agent should a task be delegated to. Liu *et al.* use clustering techniques to learn stereotypes on the basis of past transactions and assess agents according to those stereotypes [27], and Burnett *et al.* use model tree learning to form stereotypes that are used as a prior to a Beta trust model such that direct evidence, when acquired, gradually overrides the effect of the stereotype [5]. More recently, Sensoy *et al.* demonstrate the use of

[1] See: http://www.bbc.co.uk/news/uk-14490693 for how Twitter was used to spread false rumours during the England riots of 2011.

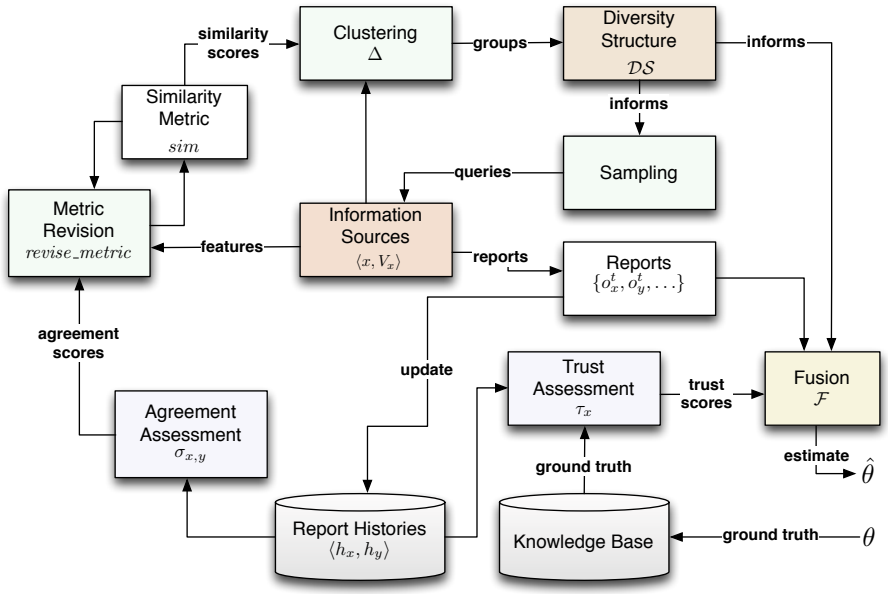

Fig. 1 The TIDY framework

graph mining techniques to formulate stereotypes from structured features, such as patterns in a social network, that may be used to inform trust assessments [36].

In this research, we are interested in the question of what information sources should be asked for an opinion regarding the state of the environment given that there are costs associated with acquiring information and we need to operate within resource constraints. The contributions that we claim in this research are: (i) that the TIDY framework offers a general approach for resource-constrained information fusion from variously trusted sources; and (ii) that an instantiation of this framework may be used to select sources such that the accuracy of the assessment of an environmental state is significantly better than existing approaches under resource constraints, and robust to dependencies among sources. The instantiation we present employs model tree learning to create a metric to assess the similarity of sources given histories of reports from those sources, and trust-based heuristics for sampling.

2 The TIDY Framework

The Trust in Information through DiversitY (TIDY) framework for source selection and fusion is centred around the idea of a Diversity Structure. The framework, illustrated in Figure 1, uses histories of reports from information sources that exhibit certain features to learn a similarity metric. This metric is then used to cluster sources on the basis of their features to form a diversity structure. A sampling strategy is then employed that is informed by this diversity structure, and reports acquired from the sampling process are fused to provide an estimate of the environmental state.

In formalising the TIDY framework, we assume a decision-maker (or *agent*) that has the task of monitoring an environmental state (e.g., the weather condition at a location, or the number of casualties following a disaster). Also, to aid in our presentation of the TIDY framework, we consider as a running example, a weather station or agent, the task of whom is to provide weather information.

Definition 1 (Task) *A task is the activity of deriving an estimate of some environmental state θ^t, at each time $t \in T$ within an interval $[t_1, t_2]$, such that $t_2 \geq t \geq t_1$.*

The domain of the variable θ^t may be different for different query types, such as "is it snowing?" and "what is the temperature?". For a particular query, Θ represents the set of possible values of θ^t. The value of $\theta \in \Theta$ can change over time, and the agent must, therefore, repeatedly update its estimate at time t, $\hat{\theta}^t$, of the environmental state, θ^t.

For instance, the task of the weather agent, in our hypothetical scenario, is to make periodic assessment of the weather situation (e.g., hourly or daily weather updates).

To acquire an estimate of θ^t, $\hat{\theta}^t \in \Theta$, sources of varying trustworthiness may be queried, the result of which will be a set of reports from the selected sources.

Definition 2 (Information Source) *An information source is a tuple $\langle x, V_x \rangle$, where $x \in \mathcal{N}$ such that $\mathcal{N} = \{1, \ldots, n\}$ is a unique identifier and V_x is a vector containing values for x's features.*

We consider, in our running example, that the weather agent has access to a group of sensors, the reports from whom can be used to make weather predictions. Each sensor is furnished with a unique identifier, and has a set of features.

Definition 3 (Report) *A report received from source $x \in \mathcal{N}$ at time t regarding θ^t is denoted o_x^t. The set of all reports from source x, at each time $t \in T$, within an interval $[t_1, t_2]$ is denoted O_x^T.*

For example, each sensor can provide reports with continuous values such that, $o, \theta \in \mathbb{R}$. In particular, we assume that at some time point, t, a sensor, x, can provide a temperature reading, o, of say, 0.6°C, relating to $\theta \in \mathbb{R}$ being ground truth (or actual temperature in this instance).

The decision-maker maintains histories of reports received from each information source.

Definition 4 (Report History) *A history of reports from a source is a sequence, defined as a function $h_x : T \to O_{x,\perp}^T$ where $O_{x,\perp}^T = O_x^T \cup \{\perp\}$. If, for some t, $h_x(t) = \perp$, then no report was received from source x at time t. For convenience, we refer to $h_x(t)$ as h_x^t, and we define the reports received at time t as $O_X^t = \bigcup_{x \in X} h_x^t$, $X \subseteq \mathcal{N}$.*

	t_1	t_2	t_3	t_4
x_1	-0.9°C	0.9°C	$\perp$	4.6°C
x_2	-1.0°C	3.0°C	0.7°C	4.7°C

Table 1 Example representation of report history

Table 1 provides an example representation of histories of report for sensors, x_1 and x_2, as maintained by the weather station from our scenario. For instance, no temperature reading was received from sensor x_1 at time t_3. Also, reports received from all the sensors sampled, (i.e., $x_1, x_2 \in X$), at time t_1, denoted as $O_X^{t_1} = \{0.6, -1.0\}^\circ$C.

We assume that sources have observable features; for example, the number and types of followers it has in a social network.

Definition 5 (Feature) *Let $F = \{f_1, \ldots, f_d\}$ be the set of all features. A feature $f_i \in F$ is an observable attribute of an information source.*

Each feature $f_i \in F$, as observed by the decision-maker, has some domain D_i, and for each source $x \in \mathcal{N}$, there exists a feature value $v_i \in V_x$, also, as observed by the decision-maker, such that $v_i \in D_i \cup \{null\}$. If a feature is unobserved or not relevant, its value is $null$ for that source.

Specific details on how features and their corresponding values may be obtained is application-specific, and beyond the scope of this work. The following illustration shows an example feature representation for three sensors, $x, y, z \in \mathcal{N}$, in our scenario:

$F = \langle \texttt{ownership}, \texttt{cost}, \texttt{battery-life} \rangle$

$V_x = \langle null,\ 0.11,\ 80 \rangle;\ V_y = \langle \text{UOA},\ 0.12,\ 80.5 \rangle;\ V_z = \langle \text{UOE},\ 0.6,\ 12 \rangle$

In order to group sources according to their features, we need a good similarity metric that allows the decision-maker to estimate the degree of similarity between sources.

Definition 6 (Similarity Metric) *A similarity metric is a function $sim : \mathcal{N} \times \mathcal{N} \to \mathbb{R}$.*

The idea behind this definition of a similarity metric for the TIDY framework is that similarity in reporting patterns may correlate with similarity in some of the sources' features. If this is the case, then given two sources $\langle x, V_x \rangle$ and $\langle y, V_y \rangle$, the function $sim(x, y)$ will give a score representing the degree of similarity that is expected in reports from these sources. This allows us to overcome the challenge of insufficient or sparse evidence regarding sources' reporting patterns, and to easily generalise to unseen cases [5].

For instance, the weather agent, in our running example, may over time "learn" the importance of the composite feature $\texttt{ownership} \wedge \texttt{battery-life}$ in defining similarity among the sensors. That is, sensors owned by the same (or a similar) organisation and with a similar battery life tend to provide similar reports when queried.

We can then use this similarity metric to stratify, or cluster sources to form a diversity structure.

Definition 7 (Diversity Structure) *A diversity structure, $\mathcal{DS}$, is a stratification of the set of all sources in the system into exhaustive and disjoint groups. $\mathcal{DS} = \{G_1, \ldots, G_K\}$, such that $\bigcup_{k=1}^{K} G_k = \mathcal{N}$ and $G_k \cap G_l = \emptyset$ for any $k, l \in \{1, \ldots, K\}$ with $k \neq l$.*

Within the context of our weather predication scenario, a diversity structure may, for example, capture sensors across different geographical locations, or sensors belonging to different organisations.

In forming a diversity structure, we assume there is some function Δ, such that, given some set of sources and a similarity metric, will compute a diversity structure. That is, $\mathcal{DS} = \Delta(sim, \mathcal{N})$. This function may be realised through an off-the-shelf clustering algorithm such as hierarchical or k-means clustering [19].

2.1 Source Agreement

Our aim is to generalise from similarity in sequences of reports from different sources to similarity of sources on the basis of their observable features. In general, we require a function that, given histories of reports from two sources, provides an assessment of the level of agreement between the reports received from those sources. In the TIDY framework, we

make the assumption that agreement between histories of reports from two sources can be derived from assessments of agreement between individual reports when reports are received from the two sources at the same time. The rationale for this is that we may then define a mechanism for assessing agreement between sources that operates efficiently on streams of reports received.

Definition 8 (Report Agreement) *An assessment of the extent to which reports from two sources, $x, y \in \mathcal{N}$, agree is determined by the function $\nu_{\mathbf{agr}} : T \times O_x^T \times O_y^T \to \Pi_{\mathbf{agr},\bot}$, where $\Pi_{\mathbf{agr},\bot} = \Pi_{\mathbf{agr}} \cup \bot$. We require that $\nu_{\mathbf{agr}}(t, h_x^t, h_y^t) = \bot$ if either $h_x^t = \bot$ or $h_y^t = \bot$; i.e. agreement can only be assessed if we receive reports from two sources at the same time.*

To illustrate this definition, we adapt the illustration provided in Table 1. We assign a value of 1, to denote an agreement. if the difference between two reports is minimal or negligible (e.g., $<= 0.1$). We assign a value of 0, otherwise, to denote a disagreement.

	t_1	t_2	t_3	t_4
x_1	$-0.9°$C	$0.9°$C	$\bot$	$4.6°$C
x_2	$-1.0°$C	$3.0°$C	$0.7°$C	$4.7°$C
$\Pi_{\mathbf{agr}}$	1	0	$\bot$	1

Table 2 Example representation of report agreement

The report agreement function will depend on the underlying model of report and source agreement. If, for example, a Beta distribution [20] is used to model agreement, $\Pi_{\mathbf{agr}} = \{0, 1\}$, where 0 indicates that reports do not agree and 1 that they do agree. Now, given an assessment of the extent to which two reports agree, we may define a means to compute the agreement between sources.

Definition 9 (Source Agreement) *Agreement between sources is some aggregation of a sequence of agreements between reports that have been received from the two sources at the same time. $\sigma : (\mathcal{N} \times \mathcal{N}) \times \left(T \times O_x^T \times O_y^T \to \Pi_{\mathbf{agr}}\right) \to \mathbb{R}$*

As illustrated in Table 2, reports from the two sensors, x_1 and x_2, are in agreement at times t_1 and t_4, and in disagreement at time t_2. However, agreement couldn't be assessed at time t_3, since no report was received from sensor x_1 at that time point. This information can then be aggregated to determine the degree of agreement between the two sensors.

With assessments of how sources agree, we may revise the function used to assess similarity between sources (Definition 6).

Definition 10 (Similarity Metric Revision) *The similarity assessment function is revised on the basis of (dis)agreements between reports received from sources: $revise\_metric$: $\left((\mathcal{N} \times \mathcal{N}) \times \left(T \times O_x^T \times O_y^T \to \Pi_{\mathbf{agr}}\right) \to \mathbb{R}\right) \to (\mathcal{N} \times \mathcal{N} \to \mathbb{R})$*

The source agreement function, σ provides a means to compute an agreement score for each source pair. Through the identifiers of the source pairs we have the values of the observable features of each source. The $revise\_metric$ function represents the problem of computing a classifier that assigns a similarity score for two sources on the basis of the values of their features.

For instance the weather agent may, over time and with fresh evidence (from source agreement), learn that the feature `ownership` is not informative, and therefore no longer a useful metric for stratification.

2.2 Trust Assessment

In addition to source similarity, an important factor in making source querying decisions is the extent to which we trust a source to provide an accurate report. We assume that the agent is able to observe ground truth, θ^t, but this is only available at a time after which it is useful for decision making. This observation of ground truth may, however, be used to revise our assessments of information sources, given we have a history of reports from those sources.

Definition 11 (Report Assessment) *The assessment of a report against ground truth is determined by the function* $\nu_{\mathbf{tru}} : T \times O_x^T \times \Theta^T \to \Pi_{\mathbf{tru},\perp}$*, where* $\Pi_{\mathbf{tru},\perp} = \Pi_{\mathbf{tru}} \cup \perp$*. We require that* $\nu_{\mathbf{tru}}(t, h_x^t, \theta^t) = \perp$ *if* $h_x^t = \perp$*.*

Again, this function will depend on the underlying model, and as with source agreement we can define a source trust assessment function.

Definition 12 (Source Trustworthiness) *The trustworthiness of a source is determined by assessments of the sequence of reports received from that source over time.* $\tau : \mathcal{N} \times \left(T \times O_x^T \times \Theta^T \to \Pi_{\mathbf{tru}}\right) \to \mathbb{R}$*.*

Information about the trustworthiness of sources is recorded for each source using an appropriate instantiation of this function (e.g., a Beta probability density function).

2.3 Sampling

While monitoring the environmental state, the decision-maker will acquire reports from various sources over time. The decision-maker must make a decision on how to sample for evidence. In particular, the agent must decide which sources to sample. The objective of a sampling strategy is to select a subset, $N \subseteq \mathcal{N}$, in order to maximise its utility. The utility of a sampling decision is a function of the accuracy of the estimate $\hat{\theta^t}$ (or information quality), and the cost of sampling (or information cost).

The quality of information measures the degree of accuracy of an estimate of the environmental state with respect to ground truth. In the context of our hypothetical scenario, this would be the extent to which a temperature reading received from a sensor reflects the actual temperature or weather outcome.

Definition 13 (Information Quality) *The information quality obtained from sampling a set of sources is a function:* $qual : \Theta \times \Theta \to \mathbb{R}$*. For example, if* θ^t *is the environmental state and* $\hat{\theta^t}$ *is the estimate of that state obtained, the information quality is* $qual(\theta^t, \hat{\theta^t})$*.*

In sampling sources, the decision-maker incurs cost. We assume that the cost of sampling a specific source remains fairly stable over time, or changes at a very slow and predictable rate.[2] Nevertheless, costs may vary across sources; e.g., the cost of asking an expert may be different from polling a group of friends.

Definition 14 (Sampling Cost) *Sampling cost is a function:* $cost : 2^{\mathcal{N}} \to \mathbb{R}$*. In many settings, sampling costs are strictly additive:* $cost(N) = \sum_{x \in N} cost\left(\{x\}\right)$

The decision-maker's task is to select a subset of sources in order to maximise its utility, or, more generally, maximise its expected utility over a sequence of sampling decisions given that the act of sampling provides information about the characteristics of the information sources sampled.

[2] This is consistent with most real-world economic settings [11].

2.4 Fusion

Fusion provides the decision-maker with the necessary tools for combining reports from various sources to arrive at an estimate.

Definition 15 (Fusion) *Information fusion is a function, $\mathcal{F} : 2^O \rightarrow \Theta$, that computes an estimate of the environmental state, $\hat{\theta}^t$, given a set of reports, O^t.*

For example, the weather agent may take the average temperature reading, from all the sensors sampled, as an estimate of the actual temperature.

In performing fusion, the decision-maker may take into account estimates of the trustworthiness of sources, such as in Subjective Logic [20] where these assessments are used to discount reports received from sources. Now, given this framework for resource-constrained information fusion, we present a realisation of this framework that we evaluate in Section 4.

3 A Realisation: TIDY$_0$

Generally speaking, a task refers to the process of determining what is true in the world such as the current river level. A task may be repetitive, requiring a periodic assessment of the ground situation (e.g., hourly or daily weather updates). Repetitive task models arise naturally in areas such as time series analysis [29]. In the course of our discussion, we assume a weather station tasked with the provision of weather information. In order to achieve this objective, the decision-maker needs to constantly sample available sources at the location of interest for weather reports. The rationale for this is that the decision-maker is then able to form an opinion over time regarding the behaviour of the sources. This is as opposed to a one-shot task, where an agent may need only carry out a single transaction (e.g., buying a life insurance from a broker) without the need for long-term monitoring. In our experiments, we assume that interactions are ordered, and refer to each time period that an interaction occurs (i.e., querying a group of sources and deriving an estimate) as a *sampling round*.

Information sources can be soft (e.g., human) or hard (e.g., wireless sensors). They can be structured (e.g., databases) or unstructured (e.g., open-source data on the Internet). Sources may have different reporting capabilities (or expertise) depending on the context. For example, a UAV (unmanned aerial vehicle) may do a better job than a human in providing surveillance coverage of a disaster region. On the other hand, human sources may be better in differentiating between different kinds of wildlife affected in the aftermath of the disaster. Sources may also exhibit different behaviours based on a variety of reasons. For example, a sensor whose battery-life is low may provide imprecise measurements, or drop packets. Other sources may obfuscate their reports before sharing in order to avoid revealing sensitive information, or may maliciously report misleading information in order to bias the decision-making process.

Reports obtained from information sources are used to derive an estimate of the environmental state. A report being an opinion about the state of the world can assume values from a wide range of domains such as binary, continuous, etc. For example, the report $\mathrm{o_x} = 0$, is interpreted differently for the queries "is it snowing?" and "how many casualties?". The first query belongs to a binary domain, s.t. $\theta^t \in \{1, 0\}$, and the reported value is taken to represent a negative opinion from source x about the event snow. The domain of the second is the set of natural numbers, i.e., $\theta^t \in \mathbb{N}$, and the report is interpreted as no casualties. The confidence measure, $\delta \in [0, 1]$, represents a degree of confidence in the measured value, such that a 0 would indicate an absolute lack of confidence or uncertainty, and 1 indicates

an absolute confidence attached to the measured value. In our evaluation, we assume that reports are continuous s.t. $\theta^t \in \mathbb{R}$ (e.g., temperature readings).

Information sources are described by a set of observable features. These features represent attributes such as organisational affiliation, location, age, etc. Similar to reports, features can assume a wide range of values in both quantitative (i.e. continuous, discrete, and interval) or qualitative (i.e. nominal, ordinal) domains. For convenience, we assume numeric values for features in our experiments.

3.1 Computing Source Agreement and Trust

A decision-maker can form opinions based on evidence obtained by interacting, and subsequently evaluating the behaviour of sources in the system. Evidence for computing both the agreement and trustworthiness of sources is gathered from different interaction contexts.

3.1.1 ***Evidence of Source Agreement***

Evidence of agreement between pairs of sources can be obtained following a sampling activity. That is, after obtaining reports from sampled sources, the decision-maker is able to evaluate these reports, and thus update evidence parameters $\langle r_{x,y}, s_{x,y} \rangle$ of the agreement for each source pair, x, y, where $r_{x,y}$ denote the number of positive experiences regarding the agreement between sources, x and y, and $s_{x,y}$ denote the number of negative experiences. Using Equation 1, both the positive ($r_{x,y}$) and negative ($s_{x,y}$) evidence parameters can be updated in light of new evidence obtained at time step t. The parameter, δ_{agr} represents an application-specific threshold for the agreement between two reports.

$$(r^t_{x,y}, s^t_{x,y}) \;=\; \nu_{\mathbf{agr}}(t, h^t_x, h^t_y) = \begin{cases} (1,0), & \text{if } |h^t_x - h^t_y| \leq \delta_{agr} \\ (0,1), & \text{if } |h^t_x - h^t_y| > \delta_{agr} \\ (0,0), & \text{otherwise} \end{cases} \tag{1}$$

3.1.2 ***Evidence of Source Trustworthiness***

After an estimate of the environmental state has been made, we assume that the decision-maker is able to observe ground truth, θ^t. The reliability of a source can be assessed on the basis of the conformity of its report to fact. Evidence used in computing the trustworthiness of a source, x is accumulated over time as a $\langle r_x, s_x \rangle$ pair, where r_x denote the number of positive experiences regarding the conformity of x's report to fact, and s_x denote the number of negative experiences. Each experience is obtained using Equation 2, where δ_{tru} is an application-specific threshold value for report reliability.

$$(r^t_x, s^t_x) \;=\; \nu_{\mathbf{tru}}(t, h^t_x, \theta^t) = \begin{cases} (1,0), & \text{if } |h^t_x - \theta^t| \leq \delta_{tru} \\ (0,1), & \text{if } |h^t_x - \theta^t| > \delta_{tru} \\ (0,0), & \text{otherwise} \end{cases} \tag{2}$$

Having described how evidence may be aggregated, we now describe how these experiences may be used by an agent to compute both source agreement and trust.

We adopt Beta distribution [20] as a representative model for both source agreement and trust. The Beta distribution provides a means of forming opinions based on available evidence. For instance, opinions about the degree of agreement of two sources, x and y can be formed on the basis of positive ($r_{x,y}$) and negative ($s_{x,y}$) evidence. These opinions may be updated in light of new evidence. The pair $\langle r_{x,y}, s_{x,y} \rangle$, provides a source of $\alpha_{x,y}$ and $\beta_{x,y}$ parameters of the Beta distribution such that: $\alpha_{x,y} = r_{x,y} + 1$ and $\beta_{x,y} = s_{x,y} + 1$. The expected value of Beta($\sigma_{x,y} \mid \alpha_{x,y}, \beta_{x,y}$) can be derived using these parameters:

$$E(\sigma_{x,y}) = \frac{\alpha_{x,y}}{(\alpha_{x,y} + \beta_{x,y})} \tag{3}$$

Similarly, opinions about the trustworthiness of a source, x can be formed on the basis of positive (r_x) and negative (s_x) evidence, which may also be updated as new evidence becomes available. The expected value of Beta($\tau_x \mid \alpha_x, \beta_x$) can be derived:

$$E(\tau_x) = \frac{\alpha_x}{(\alpha_x + \beta_x)} \tag{4}$$

If considering the trustworthiness of a group of sources, G_i, then group trust τ_i can be calculated as the average trust score of group members:

$$\tau_i = \frac{\sum\limits_{x \in G_i} \tau_x}{|G_i|} \tag{5}$$

While the agreement assessment we have described above provides evidence of similarity among known sources, we still require mechanisms for generalising from this evidence to a structure that can be used to stratify sources based on their observable features.

3.2 Learning a Similarity Metric

As mentioned, a similarity metric allows a decision-maker to generalise from similarity in the reports of sources, to similarity on the basis of their observable features. A possible realisation of this generalisation can be obtained by employing techniques from machine learning.

Decision trees [3] provide an appropriate representational abstraction for modelling a similarity metric. They are classification tools, which allow a label to be found for a given input by selecting paths through a tree based on conditions specified at branching nodes. Each node of a decision tree represents a particular feature, and branches from nodes are followed depending on the value of the feature represented by that node. In our own case, each input feature value holds the *distance* between the values of that feature for a source pair. This intuitively captures the notion of similarity in features. Each leaf of the tree represents a similarity score (or a function producing a similarity score), which is assigned to every source pair or classification examples reaching that leaf.

Classical decision tree induction techniques are not suitable for problems where the class value to be predicted is real-valued [12,31]. As our aim is to estimate the degree of similarity between sources represented by a real-valued similarity score, we require a decision tree induction technique which accommodates real-valued class labels. One possible technique that can be employed for this is model tree learning, which allows us to learn a classifier capable of predicting similarity scores from a real-valued domain [31, 47].

In model tree learning the leaves of a tree are linear regression models, which can be used to estimate a target value (similarity score in our own case). Using this technique, a similarity metric can be induced by using training examples from features of sources as well as available evidence from their report histories. In particular, we make use of the M5 model tree algorithm [31][3].

Each training instance to the M5 algorithm is of the form: $\langle dis(v_{1,x}, v_{1,y}), \ldots, dis(v_{d,x}, v_{d,y}), \sigma_{x,y} \rangle$, comprising of the feature value distances of a source pair, and their degree of agreement, $\sigma_{x,y}$ representing the class label. For each feature $f_i \in F$, we obtain a value for a source pair x, y as $dis(v_{i,x}, v_{i,y})$, where $v_{i,x}$ is the value of feature f_i for source x. Any suitable distance function (e.g., Euclidean distance) can be employed for this task. Our specific instantiation computes the distance for each feature value $dis(v_{i,x}, v_{i,y})$ for a source pair, x, y, as the absolute difference of their features; $dis(v_{i,x}, v_{i,y}) = |v_{i,x} - v_{i,y}|$[4]. Where a feature has a *null* value for either or both sources, no reasonable comparison can be made. In such instance, the entry for the feature value for the source pair is assigned a *null* value. The agreement between a pair, $\sigma_{x,y}$, is derived using Equation 3. This value is 0.5 when the system is initially instantiated, portraying an equal likelihood of both outcomes (i.e., (dis)agreement) before any positive (agreement) or negative (disagreement) evidence is observed (i.e., $r_{x,y} = s_{x,y} = 0$). Lack of evidence may impact the efficacy of the learned metric in identifying the desired correlations. Therefore, it is necessary that a revision of the model be carried out periodically as evidence is accumulated through repeated interactions.

As well as the ability to handle features of different kinds (e.g., nominal), the model tree algorithm is robust to missing values that pose a risk of overfitting a learned model [47, p. 86]. This is particularly important to us, given that some of the sources might not have values for certain features in F. For instance, a human source might not have a value for the feature `battery-life`, which may be relevant to other kinds of sources (e.g., wireless sensors). One way of handling this challenge is to use the class value as a *surrogate* attribute in a training set. In a test set, one possible solution is to replace the unknown attribute value with the average value of that attribute for the training examples that reach the node.

An example representation of a similarity metric (model tree) is shown in Figure 2. The linear models at the leaf nodes are linear combinations of the attributes with assigned weights, and are of the form: $w_0 + w_1 f_1 + w_2 f_2 + w_3 f_3$. This metric can be used to classify pairs of sources by tracing a path through the tree, in order to determine an appropriate (linear) model to employ. The output is a real value that represents the similarity score for the source pair. Using this structure, an agent can easily generalise to a notion of similarity on the basis of sources' features, and thus being equipped with a useful tool to form a diversity structure.

3.3 Creating a Diversity Structure

To form a diversity structure, $\mathcal{DS}$, we employ hierarchical clustering [19]. This is a well-known technique that can be employed for group formation. In contrast with other clustering techniques such as *k*-means clustering, hierarchical clustering allows us to cluster into a set of groups the cardinality of which we do not know in advance. The stratification uses

[3] We use the M5 implementation of Weka [14], a popular open-source machine learning toolkit written in Java.

[4] The M5 algorithm can accommodate other feature types including qualitative. As well as this, different metrics exist for computing the distance between features of other kinds [17].

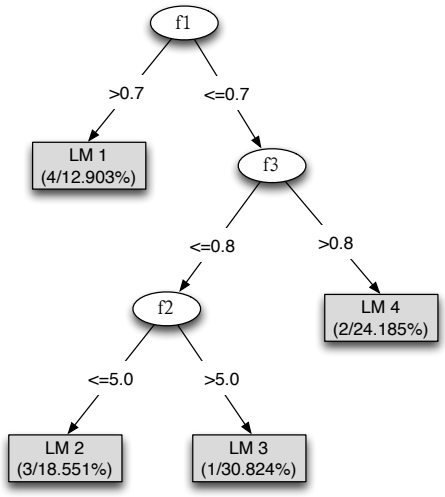

Fig. 2 An example induced similarity metric

an agglomerative method, where each source starts in a singleton group. The proximity between each source pair, x, y is then computed using the similarity metric, $sim(x, y)$. That is, given their feature vectors, an appropriate regression model can be selected for deriving a similarity score. The two most similar groups are then merged. Merging of groups continues until a stopping criterion is satisfied. One can, for instance, decide to stop either when the groups are too far apart to be merged (distance criterion) or when there is a sufficiently small number of clusters (number criterion). We model the stoppage criterion using a (diversity) threshold parameter, ψ. This parameter value lies in the interval $[0, 1]$, and specifies the maximum level of diversity required in the system. For instance, if $\psi = 1$, all the sources will be assigned to singleton groups; a condition of extreme diversity. On the other hand, if $\psi = 0$, all the sources are assigned to one group; a condition of no diversity.

3.4 Model Validity

A diversity structure once constructed, provides a static estimate of appropriate groupings of source in the population. Sources and their availability may, however, change. New sources may appear and sources may become unavailable. Although we can assign new sources to groups on the basis of their features rather than waiting for behavioural evidence, this does require us to consider which cluster is the best fit for any new source. Unavailable sources can be simply removed from their clusters. The behaviour of sources may also change over time, which may warrant a revision to the model of their relative similarity. New evidence from the behaviour of sources in previously unseen situations may also provide evidence that could lead to a more refined similarity metric.

One way of incorporating fresh evidence would be to revise the model periodically by defining a learning interval L. This interval may be determined by the number of interactions the decision-maker carries out with the environment before invoking the $revise\_metric$ function.

In revising the model, new examples, $\langle dis(v_{1,x}, v_{1,y}), \ldots, dis(v_{d,x}, v_{d,y}), \sigma_{x,y}\rangle$ are added to the training set for each source pair in the population, and the model tree is then reconstructed. It is not necessarily the case that features of sources would change over time. For example, while it is possible that features such as `battery-life` (of say wireless sensors) may change over time, other features such as `ownership` may remain fairly stable over a period of time. Evidence of agreement between sources is accumulated over time as a $\langle r_{x,y}, s_{x,y}\rangle$ pair, which is then used in Equation 3 to obtain an updated agreement score, $\sigma_{x,y}$ (see Section 3.1).

Although quite straightforward, employing a learning interval for model revision is insensitive to the dynamics in the population, and therefore may lead to unnecessary overheads. It is preferrable that the diversity model be revised based only on evidence. For instance, available evidence may suggest merging groups previously thought to be different, given the high rate of agreement in the reports of sources belonging to those groups. There may also be evidence suggesting the need to split certain groups in which members are observed to disagree a lot. An agent could, for example, employ a threshold level of error that a current model should operate within before being revised. These sort of evidence-based revisions are necessarily heuristics, and have the advantage of being much quicker than rebuilding the model from scratch. However, a limitation with this approach is the chance of anomalous revisions which may not adapt well to the global population.

We have discussed one possible reaction of an agent to changes in the source population. However, that does not preclude other forms of responses from a decision-maker. For instance, instead of always resorting to model revision, the decision-maker may adapt its sampling strategies in line with evidence pointing to the current state of the model. A decision-maker may, for example, decide to sample more from groups where available evidence suggests higher rates of disagreement in the reports and vice versa. This is a non-trivial decision problem [9, 51], the discussion of which is left to our future work.

3.5 Sampling

The primary objective of a diversity structure is to aid the process of source selection. The use of a diversity structure as a basis for sampling has some similarities to stratified sampling [7], a technique that has been shown to perform well in many survey applications including social media analytics [28]. It involves partitioning a population into disjoint subgroups according to some stratification variable (e.g., geography, culture, age). Samples are then taken from each subgroup independently to estimate the variable of interest. While similar in some aspects, the sampling strategy we propose in this research is significantly different from stratified sampling. It is also important to emphasise here that group membership is not simply based on similar level of trustworthiness, rather it is a measure of the consistency of sources in giving similar reports in response to different queries. Therefore, it is possible for sources in different groups to have similar level of trustworthiness (e.g., sources from different, but equally reputable organisations). Furthermore, the cost of sampling individual sources might differ from source to source. Therefore, the decision-maker should sample intelligently accordingly budgetary constraints. This sort of budget-constrained problem solving model is becoming quite popular in many application settings including crowdsourcing [24, 30, 42]

The number of groups in a diversity structure is $|\mathcal{DS}|$. For each $G_i \in \mathcal{DS}$, the subset of sources sampled is $g_i \subseteq G_i$. The set $\mathcal{G}$ contains all the groups sampled. We consider two sampling strategies, contingent on a sampling budget, Φ (defined in Section 4.1):

– Strategy I ($\Phi \geq |\mathcal{DS}|$): The number of candidates to be sampled, or the budget assigned to a group G_i is determined by the size of the group:

$$budget(G_i) = |G_i| \times (\Phi/|\mathcal{N}|) \tag{6}$$

Individual sources are then randomly selected from G_i according to this budget. Applying this technique may, however, lead to information exclusion in much smaller groups (e.g., not selecting from singleton groups). For this reason, we select at least one candidate from each group, and correspondingly reducing the number of candidates to be sampled from much larger groups.

– Strategy II ($\Phi < |\mathcal{DS}|$): This strategy is applied only if the sampling budget is insufficient to cover all groups. Groups are ranked in order of members' trustworthiness. Then a single source is selected from the most trustworthy group, then the second most trustworthy group and so on until the budget is exhausted. The intuition here is that, although information is lost from some of the groups, it is more beneficial for a decision-maker to prioritise available resources to more trustworthy groups. A similar intuition is adopted by most trust models in making source selection decisions [13]. Also, to some extent, available budget is distributed across diverse group of sources, rather than individual "trustworthy" sources, who may, for example, be relying on others for their reports. In our evaluation, we demonstrate in a concrete manner the effect of correlated bias to the fusion results.

We do not suggest these to be the only methods for sampling. We have selected these heuristic methods because they exploit our source diversification mechanism, and allow us to assess (through experiments) the merits of learning a diversity structure. In addition, these sampling methods are reasonable heuristics that are related to practical survey methods.

3.6 Fusion

Reports from sampled sources are combined in order to derive an estimate, $\hat{\theta^t}$. The reports from sources within a group, G_i, are aggregated to form a group estimate $\hat{\theta^t}_i$ according to Equation 7.

$$\hat{\theta^t}_i = \frac{\sum_{x \in g_i} o_x}{|g_i|} \tag{7}$$

The resulting estimates from each of the groups sampled are then discounted by their corresponding trust scores, τ_i (see Equation 5). Finally, the normalised opinions from all groups are combined to obtain the estimate $\hat{\theta^t}$ using Equation 8.

$$\hat{\theta^t} = \frac{\sum_{i \in \mathcal{G}} \hat{\theta^t}_i \times \tau_i}{\sum_{i \in \mathcal{G}} \tau_i} \tag{8}$$

One advantage of this fusion approach as will be demonstrated in our evaluation, is the potential of minimising the adverse effect of large groups of unreliable sources attempting to undermine the credibility of the fusion process.

4 Evaluation

We are interested in understanding the effectiveness of a diversity-based sampling approach to estimating environmental states in resource-constrained environments where information sources vary in trustworthiness. To explore this, we conducted two sets of experiments: in the first set, the independent variables are sampling budget and the proportion of malicious sources (i.e. sources that are more likely to provide misleading but independent reports); in the second, the independent variables are sampling budget and the proportion of colluding sources (i.e. sources that are more likely to copy each other's reports). In each case the dependent variable is the mean absolute error in the resulting estimate of the environmental state.

Our research question is: as budgetary constraints and source trustworthiness (independent misleading reports and source collusion) vary, how effective is a diversity-based sampling strategy? To answer this question, we compare the following models of source selection and environmental state estimation:

Diversity-Based Sampling (DBS) Diversity-based sampling uses the realisation of the TIDY framework, TIDY_0, defined above.

Observation-Based Sampling (OBS) Observation-based sampling uses assessments of the trustworthiness of individual sources to guide sampling. This is a common approach in trust-based service selection models [13, 22, 41]. Various algorithms have been proposed, but we model the trustworthiness of each source using a Beta probability density function. In addition to driving source selection, trust assessments are used to discount reports received during fusion; an approach referred to as *exogenous* discounting [22]. When constrained by budget, OBS selects the most trusted sources according to the budget allowance. In particular, we compare our model, DBS, to the sensor framework, RFSN, proposed in [13]. RFSN consists of two modules: (1) an outlier detection scheme for sensor readings, and (2) a trust representation and update scheme. Similar to DBS, RFSN uses a sensor node's trust score, obtained using a Beta probability density function, as a weight for data reading reported by the node. Also, when faced with the decision of selecting only a subset of sensor nodes, the trust score is used as a decision making criteria, resulting in the most trusted sensor nodes being selected.

Majority-Based Sampling (MBS) Majority-based sampling is based on *endogenous* filtering [22]. This technique uses the statistical properties of the reports themselves as a basis for assessment [46, 50]. In particular, we compare our model, DBS, to that proposed in [50]. In fusion, reports deviating from mainstream opinion are filtered out. In particular, [50] filters out reports that deviate more than one standard deviation from the mean report. Therefore, estimation of the environmental state is based on the mean report of the selected sources. In source selection, sources that are closer to the majority (mean) opinion are selected preferentially.

Random Sampling (RBS) Random-based sampling is a popular method in conducting surveys [44]. In this approach, each source has an equal probability of being sampled irrespective of previous performance. Similar to MBS, RBS estimates the environmental state using the average report of the sampled sources. The difference being that it does not perform filtering (as in MBS) or weighting (as in OBS).

The representative models under each class of existing approaches (OBS, MBS, and RBS) have been selected because of their ability to accommodate continuous reports from sources.

Table 3 Experimental Parameter

Parameter	Value	Description
$\mathcal{N}$	100	No. of sources in popl.
P_l	0.1	Popl. change probability
ψ	0.4	Diversity threshold
L	30	Learning interval
δ_{agr}	0.1	Report agreement threshold
δ_{tru}	0.1	Report reliability threshold

Table 4 Source profiles

ID	f_1	f_2	f_3
p_1	x		x
p_2		x	x
p_3	x	x	x
p_4			x
p_5	x	x	

4.1 **Experimental Environment**

A summary of the experimental parameters used is provided in Table 3. Each information source in our experiments is assigned a profile, which determines its reporting pattern in relation to other sources in the system. Each profile has three features, and for each feature, a distribution is defined from which feature values may be drawn for individual sources in the profile. Each feature value is drawn from a Gaussian distribution, with informative profile features having a small standard deviation $N(\mu, 0.01)$, and uninformative profile features following a uniform distribution $N(\mu, 1.0)$. In addition, each profile has a conformity parameter, P_c, that specifies the degree to which reports of sources in a profile tend to be correlated. Therefore, with probability P_c a source will provide a similar report to other profile members, and with probability $1 - P_c$, it provides an independent report. Specifically, a source that does not conform, deviates from mainstream opinion held by its profile. A low P_c value means that more sources in a profile will report independently, according to their individual reliability model. A conforming source when reporting, first finds out about opinions maintained by its profile members. If any exist, it randomly selects one of these opinions to report, discarding its own private opinion. In this way, we define the target relationships among groups of sources we wish our model to identify. The P_c parameter adds an extra challenge to the learning algorithm, and allows us to evaluate the ability of our model to cope with *noise* due to uncorrelated feature-behaviour similarity. Unless otherwise stated, the P_c parameter is set at 0.8 for all profiles. A summary of the profiles is provided in Table 4. In the figure, informative feature for defining similarity are marked with an "x", while unmarked ones are *noise* features.

Since our work is also placed in the context of large and open environments, sources may freely join and leave the system at any time. We model this condition in the system using the population change probability, P_l. Specifically, P_l is used to specify in each interaction round, the probability that a source would leave the system. When a source leaves, it is replaced with a new source of the same profile in order to keep the number of sources fixed throughout the simulation. This property impacts on the ability of the different approaches to accurately model the behaviour of sources, and emphasises the need for a good exploration of the population. However, dynamic activity is relaxed in all cases for the first 30 sampling

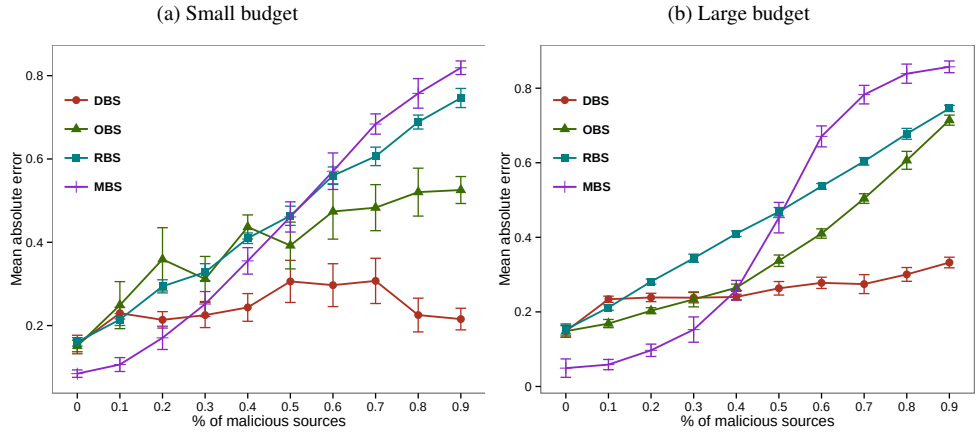

Fig. 3 Increasing proportion of malicious sources with different budget (Φ) constraints

rounds of each experiment to enable the different approaches gather information to build their individual models.

Each source has a reliability parameter, P_r that determines the type of reports it provides (i.e., honest, malicious). We define the following report types:

- Reliable report: This type of report is closer to the ground truth, θ^t, and is drawn from the distribution $N(\theta^t + 0, 0.01)$. Sources with high reliability ratio P_r are more likely to provide this type of report when queried.
- Malicious report: Reports of this kind are significantly deviated from the ground truth, and follow the distribution $N(\theta^t + 1, 0.01)$. Sources with low P_r are more likely to provide this type of report, which, if left unmanaged, could potentially undermine the fusion result.

The report reliability threshold, δ_{tru} is set to 0.1, which reflects the intuition that information is still useful if it has a small amount of noise or is slightly discounted [35].

To permit a clear discussion and evaluation of our source diversification model, we assume a setting with a fixed budget, Φ, such that: $cost_N \leq \Phi$. In particular, we define budget, Φ, in terms of the number of sources that may be sampled for evidence, such that the subset $N \leq \Phi$. Consequently, we define a small budget as $\Phi = 5$, a medium budget as $\Phi = 25$, and a large budget as $\Phi = 75$. This allows us to evaluate the performance of the different approaches under different sampling constraints.

4.2 Results

Each instance of our simulation was repeated 10 times, with each run having 100 sampling rounds. Statistical significance of differences between strategies was computed using ANOVA at 95% confidence interval. Analyses of significant differences between pairs of strategies was performed using Tukey's honest significant difference (HSD) test. We present and analyse the mean absolute error (information quality) averaged over multiple runs for the different strategies considered.

Budget	small			medium			large		
Proportion of malicious sources	low [0 – 0.3] p-value: 0.167	medium [0.3 – 0.6] p-value: 0.092	high [0.6 – 0.9] p-value: 7.55×10^{-6}	low [0 – 0.3] p-value: 0.028	medium [0.3 – 0.6] p-value: 0.13	high [0.6 – 0.9] p-value: 4.38×10^{-6}	low [0 – 0.3] p-value: 0.01	medium [0.3 – 0.6] p-value: 0.24	high [0.6 – 0.9] p-value: 5.16×10^{-5}
DBS vs. OBS	p-adjusted: 0.628 No sig. diff.	p-adjusted: 0.22 No sig. diff.	p-adjusted: 0.003 Diff: 24%	p-adjusted: 0.973 No sig. diff.	p-adjusted: 0.378 No sig. diff.	p-adjusted: 0.002 Diff: 26%	p-adjusted: 0.909 No sig. diff.	p-adjusted: 0.929 No sig. diff.	p-adjusted: 0.007 Diff: 26%
DBS vs. RBS	p-adjusted: 0.826 No sig. diff.	p-adjusted: 0.091 No sig. diff.	p-adjusted: 0.0001 Diff: 39%	p-adjusted: 0.800 No sig. diff.	p-adjusted: 0.107 No sig. diff.	p-adjusted: 0.0001 Diff: 38%	p-adjusted: 0.840 No sig. diff.	p-adjusted: 0.226 No sig. diff.	p-adjusted: 0.001 Diff: 34%
DBS vs. MBS	p-adjusted: 0.735 No sig. diff.	p-adjusted: 0.192 No sig. diff.	p-adjusted: 0.0001 Diff: 44%	p-adjusted: 0.135 No sig. diff.	p-adjusted: 0.304 No sig. diff.	p-adjusted: 0.0001 Diff: 50%	p-adjusted: 0.035 (-) Diff: 13%	p-adjusted: 0.510 No sig. diff.	p-adjusted: 0.0001 Diff: 49%

Fig. 4 Analysis result with different budget constraints and different proportions of malicious sources

4.2.1 Error vs. Malicious sources

Figure 3 shows how our model compares to other approaches under different budget settings. In Table 4 we present the results of the statistical analysis on the experimental data under different conditions. In particular, we analyse the results within the contexts of small and large budgets, and with increasing proportions of malicious sources (i.e., low [0 - 0.3], medium [0.3 - 0.6], and high [0.6 - 0.9]).

The result of the analysis (Table 4) suggests that there is no significant difference in the performance of the different approaches ($p > 0.05$), when considering the case with small budget and low proportion of malicious sources. In ther words, this result suggests that a model of diversity does not necessarily lead to a better (or worse) performance under this condition. The graph in Figure 3(a) however indicates a slight edge in the performance of Majority-Based Sampling (MBS). This is because MBS benefits from the high proportion of honest sources, who are likely to provide reliable reports, to filter bogus reports. In addition, the approach is not affected by the dynamic nature of sources in the system; its filtering is based only on statistics on the reports, and not on any knowledge of sources' behaviour. The performance lag in the case of Diversity-Based Sampling (DBS) and Observation-Based Sampling (OBS) can be attributed to the discounting of opinions. Both approaches use the trust scores of sources as discounting weights. When the correct weights are not known, reports from sources could be misrepresented. This problem is amplified by the dynamic nature of sources in the system, thus making it even more challenging for the decision-maker to determine the true reliability of sources, hence appropriate weights for their reports. This observation in itself suggests that, in environments with low proportion of malicious sources, discounting may lead to non competitive results, especially when appropriate discounting weights are not applied.

In the context of small budget, we consider the case with medium proportion of malicious sources. The statistical analysis suggests that there is no significant difference in the performance of the different approaches ($p > 0.05$). We do not reject the null hypothesis, and cannot conclude that employing a model of diversity leads to a better (or a worse) performance. While not statistically significant, the graph in Figure 3(a) shows that DBS performs better than the other approaches under this condition. In this instance, the performance of the MBS approach is observed to degrade comparatively. This result is expected since majority-based sampling approaches are not robust in the presence of increasing num-

ber of malicious reports. Though not by huge margins, we also begin to see the merits of discounting of reports as the number of malicious sources increases.

We examine the condition when there is a high proportion of malicious sources under a small budget scenario. The statistical analysis suggests that there is a highly significant difference between the performance of the different approaches ($p = 7.55 \times 10^{-6}$). This leads to the rejection of the null hypothesis. We conclude that there is a significant difference between the performance of the different approaches. A *post hoc* analysis (using Tukey's HSD) allows us to examine specifically how our model compares to the other approaches in terms of performance difference. The test: DBS vs. OBS records an adjusted *p-value* of 0.003. This suggests a highly significant difference between the performance of both the DBS and MBS approaches, with DBS having on average an estimation accuracy of about 24% higher than OBS. The Random-Based Sampling (RBS) is equally outperformed by DBS. The adjusted *p-value* for the test DBS vs. RBS is 0.000, with DBS having on average 39% higher estimation accuracy. The test: DBS vs. MBS suggests that there is also a highly significant performance difference between DBS and MBS. The adjusted *p-value* is 0.000, with DBS having on average about 44% higher estimation accuracy. These results demonstrate that in contexts of limited budget, a model of diversity leads to better assessments in the presence of a high proportion of malicious sources. As observed from the graph in Figure 3(a), both the DBS and OBS approaches, tend to perform better than the other approaches. This observation points to the merits of discounting when the proportion of malicious sources is high in the system. The MBS technique on the other hand, completely falls over when the majority of sources are unreliable. This technique is observed to be out-performed even by the random selection strategy (RBS), which may, at certain times select a reliable source by chance. Performance of DBS remains relatively stable under this condition. In comparison to OBS, DBS is much more robust to a dynamic population: it exploits knowledge about the groups of unknown sources to appropriately evaluate their reports.

We consider the case with medium budget (graph not shown due to page constraint). The result of the statistical analysis suggests a significant difference in the performance of the different approaches when the proportion of malicious sources is low. However, a *post hoc* test indicates that our model does not perform better (or worse) than the other approaches. The only significant variation in performance is between MBS and RBS (not shown in the result). The result suggests that there is no significant difference in the performance of the different approaches with medium proportion of malicious sources. However, DBS once again shows a consistent high performance when there is a high proportion of malicious sources. The statistical analysis suggests that the performance of the different approaches have a highly significant difference ($p = 4.38 \times 10^{-6}$). A *post hoc* analysis indicates that DBS performs significantly better than OBS, with a 26% higher accuracy level. DBS also performs significantly better than RBS and MBS with 38% and 50% higher accuracy levels respectively.

There is an interesting observation in the large budget case, with a low proportion of malicious sources. The analysis result suggests that there is a significant difference in the performance of the different approaches. A pairwise comparison (DBS vs. MBS) reveals that MBS actually performs significantly better than DBS. The p-adjusted value in this instance is 0.035, with MBS having on average a 13% higher performance than DBS. The intuition behind this result is that, with larger budgets, trust-based approaches are able to gather more evidence with which to make their assessments. In the case of MBS techniques, they can sample more (and possibly) honest sources, the reports from whom can be used to filter out the few unreliable ones in the set. When compared to the other approaches (OBS and RBS),

DBS shows no significant difference in performance, even though the graph in Figure 3(b) indicates a slight improvement of OBS over DBS.

The result of the statistical analysis suggests that there is no significant difference between the different approaches under the condition of large budget with medium proportion of malicious sources. However, the graph in Figure 3(b) under this condition shows a slight performance improvement of DBS over the other approaches.

DBS performs significantly better than the other approaches when the proportion of malicious sources is high under the large budget scenario. Although more budget is assigned, these other techniques cannot make better assessments given the high proportion of malicious sources. DBS is more robust in this context, given that it deals with groups rather than individual sources. Even though evidence about a sampled source might not be available, DBS uses the trust of the group to which a source belongs to weight its reports. The statistical analysis suggests that the performance of the different approaches have a highly significant difference ($p = 5.16 \times 10^{-5}$). A *post hoc* analysis shows that DBS performs significantly higher than the other techniques, with DBS performing on average 26%, 34%, and 49% better than OBS, RBS, and MBS respectively.

In summary, we can make the following conclusions regarding our first set of experiments:

- The diversity model leads to significantly better assessments under different budget settings when there is a high proportion of malicious sources in the population. The performance of the model remains fairly stable with increasing number of malicious sources.
- Although not statistically significant, the performance of the diversity model is observed to be relatively better under different budget constraints, with a moderate proportion of malicious sources.
- The performance of the diversity model tend to diminish with a low proportion of malicious sources, under all budget constraints. However, the model performs no worse than any of the benchmark approaches within this context.

4.2.2 Error vs. Dependent sources

The degree of corroboration of evidence is often used as an indication of trustworthiness, especially in systems where there are no clear experts. In such scenarios, for example, one would be more likely to believe an event reported by numerous sources than conflicting evidence supplied by only a few sources. This is the case in applications such as crowdsourcing and citizen sensing [4,25], where information is often sought from numerous and mostly unreliable sources. If those sources are simply relaying what they heard from others, then this may lead to correlated biases and misinformation.

In this set of experiments, we demonstrate the robustness of our source diversification model to varying degrees of source dependence. There are no clear experts, and the decision-maker relies on the degree of corroboration of reports to estimate environmental states. We vary the proportion of sources depending on others for reports from 0 to 0.9, where 0 represents a lack of dependence, and 0.9 represents high correlations.

We present analysis of our results, which demonstrate the significance of our source diversification model under this setting. Figure 5 shows the performance of our model in comparison to the baseline approaches. In Table 6, we present the results of the statistical analysis on the experimental data under different conditions.

Figure 5(a) shows the condition with small sampling budget. The result of the statistical analysis (Table 6) suggests a highly significant difference in the performance of the different

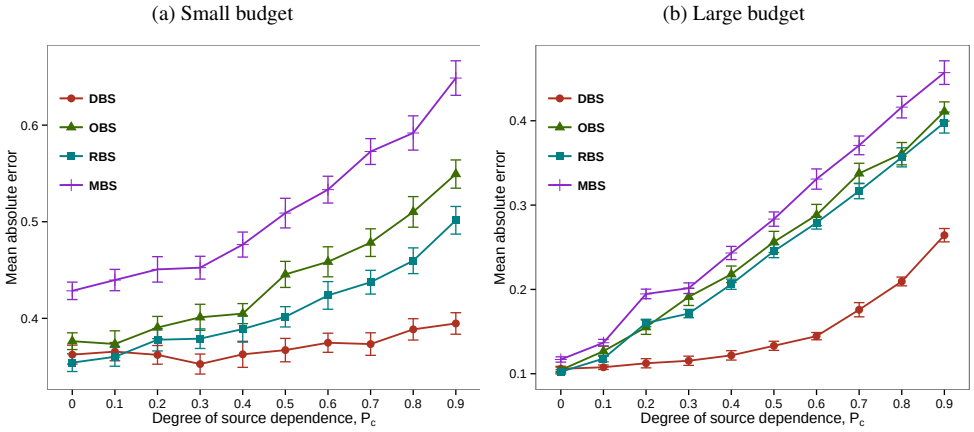

Fig. 5 Increasing proportion of dependent sources with different budget (Φ) constraints

Budget	small			medium			large		
Proportion of dependent sources	low [0 – 0.3] p-value: $6.92\text{x}10^{-7}$	medium [0.3 – 0.6] p-value: $8.27\text{x}10^{-5}$	high [0.6 – 0.9] p-value: $3.21\text{x}10^{-5}$	low [0 – 0.3] p-value: 0.043	medium [0.3 – 0.6] p-value: 0.004	high [0.6 – 0.9] p-value: $6.94\text{x}10^{-4}$	low [0 – 0.3] p-value: 0.211	medium [0.3 – 0.6] p-value: 0.004	high [0.6 – 0.9] p-value: 0.001
DBS vs. OBS	p-adjusted: 0.034 Diff: 2.4%	p-adjusted: 0.019 Diff: 6.3%	p-adjusted: 0.003 Diff: 12%	p-adjusted: 0. 491 No sig. diff.	p-adjusted: 0.025 Diff: 9%	p-adjusted: 0.006 Diff: 16%	p-adjusted: 0.483 No sig. diff.	p-adjusted: 0.015 Diff: 11%	p-adjusted: 0.007 Diff: 15%
DBS vs. RBS	p-adjusted: 0.797 No sig. diff.	p-adjusted: 0.278 No sig. diff.	p-adjusted: 0.059 No sig. diff.	p-adjusted: 0.722 No sig. diff.	p-adjusted: 0.069 No sig. diff.	p-adjusted: 0.018 Diff: 13%	p-adjusted: 0.641 No sig. diff.	p-adjusted: 0.033 Diff: 10%	p-adjusted: 0.012 Diff: 14%
DBS vs. MBS	p-adjusted: 0.0001 Diff: 8.2%	p-adjusted: 0.0001 Diff: 12.8%	p-adjusted: 0.0001 Diff: 20%	p-adjusted: 0.03 Diff: 5%	p-adjusted: 0.003 Diff: 13%	p-adjusted: 0.001 Diff: 22%	p-adjusted: 0.164 No sig. diff.	p-adjusted: 0.003 Diff: 14%	p-adjusted: 0.001 Diff: 20%

Fig. 6 Analysis result with different budget constraints and different proportions of dependent sources

approaches ($p = 6.92 \times 10^{-7}$), with a low proportion of dependent sources. We reject the null hypothesis and conclude that there is a significant difference in the performance of the different approaches in contexts of dependent sources. A pairwise comparison between DBS and OBS indicates that DBS performs significantly better than OBS. The adjusted *p-value* as captured in the result is 0.034, with DBS having on average 2.4% better performance level than OBS. The performance of OBS is also observed to be increasingly worse in comparison to that of DBS under the conditions of medium and high proportions of dependent sources. Two factors may explain the performance of OBS. First, in settings where sources are not necessarily highly reliable or unreliable, OBS cannot easily exploit models of sources to gain a competitive performance. Second, since OBS approaches usually assume independence, they are not robust to correlated biases present in the source population. As observed, the performance of this approach becomes increasingly worse with an increase in the proportion of dependent sources. Although the trust component of DBS also does little in learning source reliability, by modelling the diversity among sources it can better select candidate sources for fusion in a way that is sensitive to the correlations among the sources. Also,

by using *local fusion* based on identified groups, the effect of correlated biases in the final fusion result is minimised.

The diversity-based approach also dominates when compared to majority-based approach, MBS. Under the small budget setting, DBS is observed to consistently outperform MBS in all cases of dependencies. On the other hand, MBS shows a much worse performance than all the other approaches. Since there are no clear experts, mainstream opinion becomes inadequate for filtering out outliers. As the proportion of dependent sources increases in the system, MBS is more inclined towards opinions held by larger groups of sources. This can be problematic under the considered setting. First, sampling only these groups may not necessarily lead to reliable assessments, given the lack of experts. Secondly, by not aiming at diverse reports, MBS, and in fact OBS cannot effectively compensate for the errors in individual reports obtained. Interestingly, the RBS approach copes much better under this setting than the OBS and RBS approaches. Although the graph in Figure 5(a) shows that DBS performs better than RBS under the small budget condition, the result of the statistical analysis suggests that there is no significant difference between the performance of the two approaches (DBS vs. RBS). This is true in all the different cases of dependencies under the small budget setting. By randomly sampling the population, RBS may, by chance select from diverse groups, hence being better able to compensate for the errors in individual reports.

Performance of all the approaches is affected by budget. The graphs (Figure 5) show that performance tends to improve with increase in budget. This observation confirms our intuition: more evidence tends to approximate to better assessments, especially in environments where the sources have equal likelihood of being accurate. A more flexible budget provides an agent with more information with which to deal more robustly with biases. While our model does not necessarily experience a performance lag under these conditions, the other approaches are better equipped, with an increased budget, to mitigate the effect of correlated biases. Worthy of note is the comparison of our model to OBS (DBS vs. OBS) under the medium budget setting. With a low proportion of dependent sources and an increased budget, OBS picks up in performance. The statistical analysis suggests that under this condition, DBS performs no better (or worse) than OBS. However, the dominance of our model over OBS is again emphasised as the proportion of dependent sources becomes high. Similarly, DBS shows evidence of a superior performance to the random sampling approach under the condition of medium budget and high dependency. This is because given the opportunity to sample more sources, RBS becomes equally as vulnerable to the effect of correlated biases as are OBS and MBS. The adjusted *p-value*, p-adjusted in this instance is 0.018, with DBS having on average 13% better performance than RBS.

Performance of all the baseline approaches appear to be marginally similar to that of DBS under the large budget condition. The result in Table 6 suggests that DBS does no better (or worse) than the other approaches when budget is large and the proportion of dependent sources is low. However, DBS dominates these other approaches as the proportion of dependency increases from low to medium and from low to high. The specific performance gains under these conditions are recorded in Table 6.

We summarise our conclusions in our second set of experiments:

- With an increasing proportion of dependent sources, a model of diversity can lead to a significantly higher performance than non-diversified approaches.
- DBS tends to cope much better than the baseline approaches when sampling budget is low.

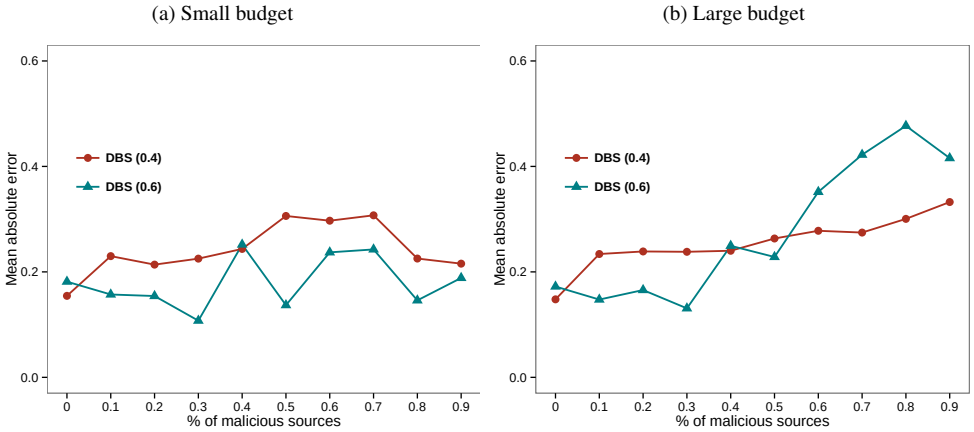

Fig. 7 Comparing different ψ (diversity threshold) parameter values: 0.4 and 0.6

– The merits of the diversity model tend to diminish with a low proportion of dependent sources. However, the model does not perform any worse than non-diversified approaches.

4.2.3 Sensitivity to parameter settings

As earlier mentioned, the diversity threshold, ψ, allows us to control the process of group formation. In particular, $\psi \in [0, 1]$ models a stoppage criterion during the merging of groups to form a diversity structure. In our earlier experiments, we set this parameter at 0.4 (see Table 3), which, we believe, provides a reasonable cut-off mark for the identification of groups capable of accommodating different degrees of source dependence, P_c. If for instance, ψ is set too high, we face the risk of not identifying (or forming) groups, even when evidence of group effect exists in the system. On the other hand, setting this parameter too low may lead to assigning all sources to a single group. This may not be the desirable outcome, especially if it does little to reflect the underlying group effect in the system. To test the sensitivity of our model to different parameter values, we conduct a separate set of experiments, with a different ψ value of 0.6. We compare the performance of our model, DBS when $\psi = 0.4$, as used in previous experiments, and when ψ is set at 0.6. We label these instances DBS (0.4) for $\psi = 0.4$, and DBS (0.6) for $\psi = 0.6$ for ease of reference.

Figure 7 shows how the two settings DBS (0.4) and DBS (0.6)) compare under varying budgetary constraints. Performance of both instances appears similar under the small budget condition (Figure 7(a)), with DBS (0.6) having a slight edge over DBS (0.4). However, the performance of DBS (0.4) appears to be slightly more stable with increasing proportion of malicious sources. The performance of DBS (0.4) is also observed to be more stable than that of DBS (0.6) under the large budget condition (Figure 7(b)). With large sampling budget, the performance of DBS (0.6) is observed to degrade significantly from that of DBS (0.4), when the proportion of malicious sources is high (> 0.6).

One possible explanation for the instability in the performance of DBS (0.6) as opposed to that of DBS (0.4), lies in the nature of groups identified under this setting. While DBS (0.4) is able to produce groups that more closely reflect the underlying source profiles, this isn't the case when ψ is set at 0.6, as captured by DBS (0.6). For instance, DBS (0.6) may

not be able to group similar sources together given the high diversity threshold of 0.6. In such an instance, sources that would otherwise have belonged to the same group are not so identified. This may, for example, lead to inaccurate discounting weights assigned to reports from certain groups, thereby leading to unstable performance.

5 Related Work

While we have referred to a few trust and reputation mechanisms in this article, we will highlight here some related approaches relevant in our problem space.

Hang *et al.* proposed to use mixture of beta distributions to model trustworthiness of service providers [15, 16]. Wang *et al.* proposed a probabilistic approach for maintaining trust based on evidence. The evidence in this approach is used to update the parameters of trust models based on beta distributions. Similar to these approaches, in this work, we also used beta distributions to model trust. In the literature, beta distributions are used to model subjective degree of belief for binomial propositions. The Beta Reputation System (BRS) [21] uses beta probability density functions to estimate the likelihood of the probability that a binary proposition, such as 'agent x is trustworthy' or 'agent y provides good services', is correct . Aggregated evidence from information sources are used as parameters of beta distribution. BRS is extended in [46] to handle misleading reports from malicious sources using a majority-based algorithm. By assuming mainstream opinion to be reliable, minority opinions statistiscally deviated from the majority opinion are filtered out. This approach, as demonstrated in our work, performs badly in environments where there are correlated biases in the reports of sources: sources might simply be copying from others that are not necessarily reliable.

Teacy *et al.* attempt to address the shortcoming of BRS [41]. In their TRAVOS model, the reports of individual sources are weighted or discounted based on their perceived reputation before fusion. While we adopt a similar discounting measure, we do not seek to model the trustworthiness of each individual source as in TRAVOS. Interaction with all possible sources may not be feasible, for instance, in large and dynamic systems. Rather, similar to [5], an estimate of the trustworthiness of an unknown source may be based on that of the group members with which it is identified. Regan *et al.* propose a Bayesian trust model (BLADE) to deal with the problem of biases in sources' reports [34]. The evaluation function used by a source for reporting may be learned by exploiting features of the source. This approach is very effective in the sense that an agent can make use of all available reports, with a limited requirement to discount or discard reports considered to be misleading. One type of bias not covered by BLADE is correlations or dependencies among sources, which might impact on their reporting patterns. Similarly, Reece *et al.* [33] propose a model that enables agents to evaluate the outcome of a transaction based on multiple, and possibly, correlated features (e.g., time, cost). By decomposing the outcome of a transaction into multiple dimensions, an evaluating agent can more directly form opinions on different aspects of a transaction, especially as these might affect estimates of its expected utility. However, in contrast with their work, our focus is to disambiguate the features defining similar reporting patterns among a group of agents, and not necessarily how those features may affect or contribute to the success or failure of a transaction. Furthermore, the authors in [33] propose an approach for addressing double-counting of evidence in decentralised systems. In particular, they propose a mechanism whereby a reporting agent labels its reports as either "private" (not previously communicated to any other agent) or "shared" (communicated to, or received from, another agent). By doing this, the authors assume that such metadata about re-

ports will always be present, let alone provided honestly. As discussed, information sources, especially in social platforms (e.g., Facebook or Twitters), may provide reports without any acknowledgement, thus limiting the use of this approach in many settings. In contrast, our approach does not require agents to provide any extra information. We learn a similarity function, given report histories and observable features of agents, and show that this can effectively mitigate the problem of rumour propagation in reputation systems. Teacy *et al.* proposed HABIT, which implicitly estimates trustworthiness of information sources using a hierarchical Bayesian modeling [39]. HABIT exploits similarities between source reports and direct evidence. Therefore, it may not work when direct evidence is not available.

Yin *et al.* proposed a truth discovery system, TruthFinder, that works with multiple conflicting information sources on the Web [49]. It is based on the assumption that an information source is trustworthy if it provides many pieces of true information, and a piece of information is likely to be true if it is provided by many trustworthy sources. Unlike our approach, this system assume that sources are independent, so it is vulnerable when the sources collude to provide misleading reports, e.g., copy from one another. Several authors have identified the role of source independence when integrating opinions from multiple sources. A misleading report, for instance, can be spread through copying (e.g., on the Web) thereby making the discovery of the true state of the world challenging. In order to deal with this problem, the truth-finder system [43] presents an algorithm for diversifying the selection of information sources. Although their intuition on diversity is similar to ours, the authors assume a static and prior knowledge of a stratification metric. For instance, the authors form a dependence graph on Twitter using the connectedness of sources, such that individuals "following" others are assumed to be dependent on those sources. While the proposed metric is quite relevant in the specific system, such knowledge may not always be available, necessitating the need for an agent to learn a general metric for stratification. Dong *et al.* use an iterative heuristics to estimate the probability of dependence and conflicts between sources [8, 45]. Their approach relies on knowledge of ground truth, and works on the assumption that sources providing same false reports are likely to be dependent. This can be problematic in environments where ground truth is delayed in being observed or is not altogether available. In contrast, our model does not require knowledge of ground truth to function: similarity assessments are made solely on reports provided by the sources. More recently, [10] use a two-layered clustering approach to categorise sources into different groups based on their subjectivity or reporting patterns. Reports from sources in different (subjectivity) groups are then "aligned" (i.e., adjusted or re-interpreted based on observed pattern) according to those of the evaluating agent. This approach assumes that the decision-maker is in a position to make personal observations of the variable of interest, which might not necessarily be the case.

Deciding what sources to ask for opinions regarding the state of the world is a familiar problem. The underlying motives are also varied depending on the application domain and individual system requirements and/or constraints. In some domains, the selection of sources might be constrained in certain ways (e.g., cost, energy, timely decisions) [48]. Approaches that operate under resource constraints often resort to the use of a subset of the source population. Even quality-driven approaches may (e.g., [21], [41], without any constraints on resources), at some point, need to draw a line on the number of sources to sample for information, even if they may not engage any active source selection strategy.

Whereas trust and reputation mechanisms focus on maximising the quality of estimates of ground truth, cost minimisation or the need to balance the trade-off between information quality and cost has been the focus of other approaches. We adopt a similar view in this article. Notable work within this context include [18, 24, 40], which formulate the problem

of source selection as a Markov Decision Process (MDP). Although MDPs and related techniques (e.g., multi-armed bandits [1]) provide a sound mathematical framework for modelling the problem of choice under uncertainty, they suffer from complexity issues and tend to be overwhelmed by the explosion of the problem space. As a result, these approaches can only be applied to problems with limited scale. Still within this context, active learning [2] presents an interesting body of work, the goal of which is to sample a distribution (or group of sources) proportionate to the variance of members' reports. Active learning has been applied to stratified sampling and multi-armed bandits for optimum allocation [1,2,9]. Given that these approaches do not take the reliability of groups into account, more sampling effort may be appropriated to unreliable groups, which may in turn impact estimations of ground truth. Recently, the problem of interdependent tasks allocation under budgetary constraints has been studied within the context of a crowdsourcing application [42]. In particular, given a budget, the aim of the approach is to determine the number of micro-tasks to be performed and the price to pay for each task. The authors use a quality control procedure, known as *AccurateAlloc*, to efficiently allocate micro-tasks to minimise the probability of error. The procedure employs a tuning or error bound parameter, such that only those candidate solutions that are within the error bound of the most popular candidate are progressed to a subsequent phase. While the proposed approach provides performance guarantees under fixed budget, its applicability might be limited in environments characterised by biased sources. This is because the authors assume that sources are not malicious or dependent, and therefore depend on basic consensus to predict correct answers. Thus, the number of sources required to attain a certain level of accuracy might increase dramatically depending on the number of 'lazy' workers in the system.

6 Discussion

We have demonstrated that a model of diversity can lead to more accurate estimates of environmental states under varying budget constraints. This is particularly true if there are correlations in the reports provided by the sources. It is encouraging, that even when generalising on the behaviour of sources, our approach still performs as well as classical trust approaches, the focus of which is on modelling the behaviour of individual sources.

The overarching motivation of our work is to mitigate the double-counting of evidence. We achieve this goal by ensuring agents in the population belong to distinct groups, hence reducing the risk of double-counting individual agents, which might not be known apriori. Our requirement that agents belong to one group, is to enable us calculate the value of querying individuals from different groups without any form of ambiguity. An agent selected from a group takes on the known properties of that group (e.g., trust score).

One limitation of our diversity model in terms of generality, which constitutes a key future direction, is the assumption that sources will exhibit the same kind of correlations. Sources may, for instance, behave differently or show different kinds of *affinity* based on the type of query issued. In other words, a decision-maker may need to strategically stratify the population in different ways according to different information needs. For example, a source may respond differently when a question bordering on national interest is posed, as opposed to one on organisational interest. A possible approach for addressing this concern is by defining the similarity metric in a way that is sensitive to the query type. In this way, evidence received may be better utilised given the specific goal of the query.

The use of a learning interval, L, in order to revise the learned model could lead to computation overheads, especially if there wasn't such need for model revision. Model revisions

should rather be based on evidence obtained from interaction with the system. Where learning is not expected to lead to a significant change in the diversity structure, a decision-maker may instead adapt its sampling strategy in line with the current state of the model.

Another key future direction involves designing a sampling strategy that takes the dynamics in the diverse groups into account when sampling from them. Our current strategy, which is based solely on heuristics doesn't adapt well under such circumstances. In particular, by sampling proportionally according to the size of groups, a decision-maker may potentially waste scarce resources, especially in environments where a decision-maker also has to make decisions concerning the number of sources to sample. For example, a decision-maker may allocate unnecessary budget to large but unreliable groups. Also, samples drawn from groups with trustworthy sources may be superfluous, and not necessarily lead to a further improvement in the estimate. One way of addressing this problem is to find a way to incorporate the optimum allocation strategy adopted in active learning approaches [1, 2, 9], without trading-off our model's sensitivity to source trustworthiness.

7 Conclusions

In this article we presented a source diversification model, that allows a decision-maker to group sources based on their perceived similarity. The model is aimed at supporting a decision-maker to intelligently sample for evidence, such that reliable assessments can be made within different budget constraints. The results of our experiments show the efficacy of our model in guiding reliable assessments, especially in situations with high proportions of biased sources in the population. Where hidden networks or patterns defining correlated behaviour exist in the population, our source diversification model is able to identify and exploit such structures in order to cope with constraints in budget, while maintaing the quality of information. We have demonstrated that our model outperforms classical trust and survey approaches in making reliable assessments of ground truth.

We have identified the need to incorporate a more principled and rigorous decision-theoretic mechanism to handle complex source selection strategies. This will enable us to meet different information needs. For example, the cost and risk analysis of interacting with certain groups of sources in a diversity structure may serve to inform how sampling decisions are made. The source diversification model presented in this article provides a good basis for driving such intelligent source selection strategies.

We believe that the sort of problems addressed by our model have significant impacts in environments such as sensor networks with real resource constraints, and in social networks where dependencies among sources increase risks of double-counting of evidence and other biases.

References

1. A. Antos, V. Grover, and C. Szepesvári. Active learning in multi-armed bandits. In *Proceedings of the 19th International Conference on Algorithmic Learning Theory*, pages 287–302, 2008.
2. A. Antos, V. Grover, and C. Szepesvári. Active learning in heteroscedastic noise. *Theoretical Computer Science*, 411(29):2712–2728, 2010.
3. L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. Wadsworth, 1984.
4. J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava. Participatory sensing. *ACM Sensys World Sensor Web Workshop*, 2006.

5. C. Burnett, T. J. Norman, and K. Sycara. Bootstrapping trust evaluations through stereotypes. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, pages 241–248, 2010.
6. A. T. Campbell, N. D. Lane, E. Miluzzo, R. A. Peterson, H. Lu, X. Zheng, M. Musolesi, K. Fodor, S. B. Eisenman, and G.-S. Ahn. The rise of people-centric sensing. *Internet Computing*, 12(4):12–21, 2008.
7. W. G. Cochran. *Sampling Techniques*. John Wiley & Sons, 1977.
8. X. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment (PVLDB)*, 2, 2009.
9. P. Etoré and B. Jourdain. Adaptive optimal allocation in stratified sampling methods. *Methodology and Computing in Applied Probability*, 12:335–360, 2010.
10. H. Fang, J. Zhang, and N. Magnenat Thalmann. Subjectivity grouping: Learning from users' rating behavior. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1241–1248, 2014.
11. M. Feldstein. The costs and benefits of going from low inflation to price stability. In *Reducing Inflation: Motivation and Strategy*, pages 123–166. University of Chicago Press, 1997.
12. E. Frank, Y. Wang, S. Inglis, G. Holmes, and I. H. Witten. Using model trees for classification. *Machine Learning*, 32(1):63–76, 1998.
13. S. Ganeriwal, L. Balzano, and M. Srivastava. Reputation-based framework for high integrity sensor networks. *ACM Transactions on Sensor Networks*, 4(3):15, 2008.
14. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11:10–18, 2009.
15. C.-W. Hang and M. P. Singh. Selecting trustworthy service in service-oriented environments. In *The 12th AAMAS Workshop on Trust in Agent Societies*, pages 1–12, 2009.
16. C.-W. Hang and M. P. Singh. Trustworthy service selection and composition. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 6(1):5, 2011.
17. M. Ichino and H. Yaguchi. Generalized minkowski metrics for mixed feature-type data analysis. *IEEE Transactions Systems on Man and Cybernetics*, 24:698–708, 1994.
18. A. Irissappane, F. Oliehoek, and J. Zhang. A pomdp based approach to optimally select sellers in electronic marketplaces. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1329–1336. International Foundation for Autonomous Agents and Multiagent Systems, 2014.
19. A. Jain, M. Murty, and P. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31:264–323, 1999.
20. A. Jøsang. *Subjective Logic*. Book Draft, 2013.
21. A. Jøsang and R. Ismail. The beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference*, pages 41–55, New York, USA, 2002.
22. A. Jøsang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007.
23. A. Jøsang and S. Presti. Analysing the relationship between risk and trust. In *Trust Management*, pages 135–145, 2004.
24. E. Kamar, S. Hacker, and E. Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, pages 467–474, 2012.
25. A. Kittur, E. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 453–456. ACM, 2008.
26. H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600. ACM, 2010.
27. X. Liu, A. Datta, K. Rzadca, and E.-P. Lim. StereoTrust: A group based personalized trust model. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 7–16, 2009.
28. B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 122–129, 2010.
29. M. A. Osborne, S. J. Roberts, A. Rogers, S. D. Ramchurn, and N. R. Jennings. Towards real-time information processing of sensor network data using computationally efficient multi-output gaussian processes. In *Proceedings of the 7th International Conference on Information Processing in Sensor Networks*, pages 109–120, 2008.
30. R. Ouyang, M. Srivastava, A. Toniolo, and T. J. Norman. Truth discovery in crowdsourced detection of spatial events. Technical report, International Technology Alliance (ITA), https://www.usukitacs.com/node/2637., 2014.
31. J. Quinlan. Learning with continuous classes. In *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, pages 343–348, 1992.

32. R. Raafat, N. Chater, and C. Frith. Herding in humans. *Trends in Cognitive Sciences*, 13:420–428, 2009.
33. S. Reece, A. Rogers, S. Roberts, and N. R. Jennings. Rumours and reputation: Evaluating multi-dimensional trust within a decentralised reputation system. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*, page 1063?1070, 2007.
34. K. Regan, P. Poupart, and R. Cohen. Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 206–212, 2006.
35. M. Şensoy, A. Fokoue, J. Pan, T. J. Norman, Y. Tang, N. Oren, and K. Sycara. Reasoning about uncertain information and conflict resolution through trust revision. In *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems*, pages 837–844, 2013.
36. M. Şensoy, B. Yilmaz, and T. J. Norman. STAGE: Stereotypical trust assessment through graph extraction. *Computational Intelligence*, DOI: 10.1111/coin.12046, 2014.
37. J. R. Shiller. Conversation, information, and herd behaviour. *American Economic Review*, 85:181–185, 1995.
38. J. Surowiecki. *The wisdom of crowds*. Anchor, 2005.
39. W. L. Teacy, M. Luck, A. Rogers, and N. R. Jennings. An efficient and versatile approach to trust and reputation using hierarchical bayesian modelling. *Artif. Intell.*, 193:149–185, Dec. 2012.
40. W. T. L. Teacy, , G. Chalkiadakis, A. Rogers, and N. Jennings. Sequential decision making with untrustworthy service providers. In *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems*, pages 755–762, 2008.
41. W. T. L. Teacy, J. Patel, N. R. Jennings, and M. Luck. Travos: Trust and reputation in the context of inaccurate information sources. *Autonomous Agents and Multi-Agent Systems*, 12:183–198, 2006.
42. L. Tran-Thanh, T. D. Huynh, A. Rosenfeld, S. D. Ramchurn, and N. R. Jennings. Budgetfix: Budget limited crowdsourcing for interdependent task allocation with quality guarantees. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems*, pages 477–484, 2014.
43. M. Uddin, M. Amin, H. Le, T. Abdelzaher, B. Szymanski, and T. Nguyen. On diversifying source selection in social sensing. In *Proceedings of the 9th International Conference on Networked Sensing Systems (INSS)*, pages 1–8, 2012.
44. J. Waksberg. Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73(361):40–46, 1978.
45. D. Wang, L. Kaplan, and T. F. Abdelzaher. Maximum likelihood analysis of conflicting observations in social sensing. *ACM Trans. Sen. Netw.*, 10(2):30:1–30:27, Jan. 2014.
46. A. Whitby, A. Jøsang, and J. Indulska. Filtering out unfair ratings in bayesian reputation systems. In *Proceedings of the 7th International Workshop on Trust in Agent Societies*, 2004.
47. I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.
48. O. Wolfson, B. Xu, and R. Tanner. Mobile peer-to-peer data dissemination with resource constraints. In *Proceedings of the 8th International Conference on Mobile Data Management*, pages 16–23, 2007.
49. X. Yin, J. Han, and P. Yu. Truth discovery with multiple conflicting information providers on the web. *Knowledge and Data Engineering, IEEE Transactions on*, 20(6):796–808, June 2008.
50. W. Zhang, S. Das, and Y. Liu. A trust based framework for secure data aggregation in wireless sensor networks. In *Proceedings of the 3rd Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, volume 1, pages 60–69. IEEE, IEEE, 2006.
51. Z. Zheng and B. Padmanabhan. On active learning for data acquisition. In *Proceedings of IEEE Conference on Data Mining*, pages 562–569, 2002.