A Reputation System for Uncertain Assertions

Mark Kramer, Arnon Rosenthal
The MITRE Corporation, 202 Burlington Road
Bedford, MA, USA
{mkramer, arnie}@mitre.org

Abstract. We investigate reputation systems that rate the performance of analysts who make uncertain assertions (claims accompanied by estimated probabilities). Accuracy metrics (based on the fraction correct) are fair only if all analysts handle identical or statistically similar cases. Furthermore, accuracy metrics discourage analysts from offering predictions on difficult-to-predict events. Because of these difficulties, we develop a class of performance scoring functions that are maximized when the analyst provides accurate probabilities, especially when these probabilities differ from the norm. Under these metrics, the disincentives to forecast low-probability events is removed and analysts are rated fairly, independent of the base event probabilities of the cases they consider. Reputation systems built around these metrics can support productivity management and increase manipulation resistance when information providers are not trustworthy. An application to citizen event reporting is presented.

Keywords: uncertainty, reputation, scoring function, citizen event reporting, counter insurgency

1 Introduction

Many analysis tasks, such as economic forecasting, criminal investigation, and medical applications, produce statements and predictions that involve significant uncertainty. In these cases, probabilistic statements are frequently more useful than simple Boolean predictions. For example, if we are planning an outdoor activity, generally we prefer predictions in the form “it will rain tomorrow with 30% probability”, instead of getting an unqualified prediction.

Now suppose there are multiple information sources (human or otherwise) providing probabilistic predictions. In planning courses of action, it would be useful to have meta-knowledge of the trustworthiness and likely novelty of the predictions. For example, a military commander who relies on knowledge of probable enemy responses would clearly benefit from knowing the skill of analysts providing the predictions.

Prediction skill differs for many reasons. Some analysts possess information unavailable to others, such as an equity analyst able to communicate with key company insiders. Others may have superior subject matter expertise or longer, richer experience. They may use different techniques, and apply different levels of skill and judgment. Moreover, information providers can be honest, manipulative, or malicious. The chance of malicious behavior can be minimized with carefully screening of information providers, but in some contexts, this is not possible. As discussed further in Section 4, citizen event reporting (CER) leverages the eyes and ears of a large population of “citizen sensors” to increase the amount of information available to decision makers. When deployed in an environment that includes a hostile subpopulation or rival clans, some of the tips gathered by CER may be aimed at deceiving decision makers, motivated by the desire to lure first responders into an ambush or induce them to attack rivals. In such a case, it is advantageous to track the history of reports obtained from the CER system, to help determine the trustworthiness of the reporters.

Providing meta-knowledge about information providers is the job of a reputation system. A reputation system can track the success of information providers over time, and provide rankings, feedback, and other information useful to the participants (information providers, consumers, or both). In this paper, we assume a centralized reputation system that possesses global information about all analysts and their prediction history. In this context, the simplest reputation system would track *accuracy*, the fraction of correct answers provided by each analyst. Unfortunately, accuracy provides a fair comparison only if all analysts consider the same or equivalent set of cases, in terms of their prior likelihood and intrinsic predictability. If analysts exercise the freedom to choose when they make predictions, then to maximize their accuracy, they will avoid making assertions about low-probability events, focusing instead on “sure things”. Obviously-correct assertions (the sun will rise tomorrow) have low value to information consumers. In general, the use of accuracy to rank analysts creates a mismatch between the needs of information consumers and information providers.

The goal of this paper is to propose a reputation system for uncertain assertions. When assertions include an estimated probability, it is possible create a system of rewards that does not skew the attention of analysts towards high-probability events. Rather than rewarding accuracy itself, our ranking system rewards both novelty, in terms of departure of a case from the norm (expressed in terms of the consumer’s prior), and accurate probability estimates. To this end, we first examine classes of scoring functions that are aware of prediction probabilities, and derive measures of producer accuracy and productivity. Next, we address how reputation scores can be used, and explore assumptions about provider motivations and work processes. For user organizations, the framework and analysis help identify questions one must ask in setting up a reputation system. For reputation researchers, the framework can help in capturing assumptions and comparing with others' results, and in identifying gaps in our knowledge. We emphasize metrics that require relatively little software and administrative labor to implement – though sometimes less accurate, they seem far more likely to be implemented.

2 Reputation Metrics

2.1 Preliminary definitions

We now define the constituents of the model and the notation to be used in the rest of the paper. For the purposes of this paper, we consider predictions consisting of a Boolean assertion about the world (A), and the provider's estimated probability ($0 \leq Q \leq 1$). An *uncertain assertion* is a pair (A, Q). Examples include:

- {it will rain in Boston tomorrow, 0.7}
- {patient Z will survive the proposed surgery, 0.9}
- {there is a roadside bomb on next route segment, 0.01}

Other types of predictions, such as quantitative, multi-valued predictions, or interval probabilities, are not considered here, but some discussion is available in [3].

An assertion may be about the past, present, or future, but reputation points will be assessed only when its truth (or falsity) becomes known. *Assertions for which we learn ground truth drive the assignment of reputation.* Ground truth is frequently available in areas such as weather, elections, and sports, and less often in fields such as medicine. We now introduce several definitions:

- *Event probability* (P): Each assertion is Boolean and may be decided at some future time. However, while the outcome remains unknown, there is uncertainty in the outcome[1]. We do not expect administrators to know or estimate P; it appears only in our mathematical analysis.
- *Prediction* (Q): Denotes the provider's estimate of P, a probability between 0 and 1.
- The *prior* (P^*) is the consumer's probability estimate for A, occurring before the provider information is taken into account. If the prior is not known, it defaults to the uninformative prior (usually 0.5).
- *Utility* (U) measures the value of learning the truth or falsity of A. When determining reputation, scores relative to assertion A can be weighted by the utility of A.
- *Scoring (payoff) functions.* We define two mathematically-related functions, $f(Q, P^*)$ and $g(Q, P^*)$, representing the analyst's reward if A does or does not occur, respectively. Since a strategy for optimizing expected score across multiple independent assertions will attempt to maximize the score separately for each assertion, we can consider scoring a single assertion without loss of generality.

We consider it essential to minimize the administrative effort, i.e., the effort to ascertain scoring functions, priors, utility, and ground truth. If a reputation system

[1] For the interpretation of probability for non-repeated event, see [1]. For our purposes, P denotes the fraction of instances like the current situation in which A holds.

requested additional information about each assertion, most providers and managers would probably refuse, or provide perfunctory estimates. For that reason, the default treatment is to use the same scoring function for all assertions, and equal weighting for utility. Still, if a large number of assertions have the same properties (e.g., weather assertions for different days), it may be feasible to elicit a few numbers (prior probability, utility, perhaps choice among loss functions) for different classes of assertions, to be applied to each instance.

2.2 Reputation Scoring Functions

Various scoring functions have already been developed with the objective of rating forecasting skill and eliciting truthful personal beliefs [2, 3, 4, 6]. Much of the work originated in the field of weather forecasting. Most commonly used is the Brier score [2], defined as:

$$\begin{aligned} f(Q) &= -(1-Q)^2 \\ g(Q) &= -Q^2 \end{aligned} \qquad (1)$$

The Brier score is the squared error based on the difference between the predicted (Q) and actual (0 or 1) probabilities (the negative sign makes higher scores better). It is independent of the prior probability, which is a drawback, because a provider's assertion is useful only if it is different than the consumer's prior. We believe a reward should be given only if the consumer learns something new (and correct) from the information provider. For example, the assertion {the sun will rise tomorrow, 1.0} has little surprise or impact, while {terrorists will attack Mumbai, 0.10} might be quite novel and useful. A higher score should be awarded for the latter assertion, if proven true, even though the asserted confidence is lower, because the difference between the prior and the prediction is larger.

We have identified several requirements for a probability-aware payoff function, as follows:

R1. **Monotonicity**. We require that $f(Q, P^*)$ increase monotonically and $g(Q, P^*)$ decrease monotonically as a function of Q. This assures that if the event occurs, larger estimated probabilities earn larger rewards. Conversely, assertions that are confident but wrong receive larger penalties than wrong answers that are explicitly declared uncertain.

R2. **No Reward for Prior**. As discussed above, a provider's assertion is useful only if it is more informative than the consumer's prior. Accordingly, we require that $f(P^*, P^*) = g(P^*, P^*) = 0$.

R3. **Expected Value**. For a Boolean assertion, the expected value of a prediction is $E(Q, P, P^*) = P\, f(Q, P^*) + (1 - P)\, g(Q, P^*)$. So each analyst is motivated provide a prediction as close to the true probability as possible, we require that $E(Q, P, P^*) < E(P, P, P^*)$ for each $Q \neq P$. A scoring rule with this property is said to be *strictly proper* [3]. This requirement assures that, over the long run, the perfect information provider who picks Q = P in each situation will earn the maximum possible score.

R4. **Hidden Knowledge**. The scoring functions cannot require knowledge of P, since this probability is assumed to be hidden from all parties.

R5. **Boundedness**. No assertion should earn an unbounded payoff (either positive or negative), because this would make accumulation of scores over multiple trials problematic.

R6. **Symmetry under Negation**. Asserting A with certainty Q is the same as asserting ~A with certainty 1-Q. Our payoffs should be indifferent to the logical "direction" of the assertion. Therefore, we require that $f(Q, P^*) = g(1–Q, 1–P^*)$ and $g(Q, P^*) = f(1–Q, 1–P^*)$.

If we assume f and g are differentiable functions, then R3 implies that $dE/dQ = 0$ at the point where $Q = P$. Therefore, strictly proper scoring functions must satisfy the differential equation:

$$P\, df/dQ + (1 – P)\, dg/dQ = 0 \text{ at } Q = P \qquad (2)$$

Furthermore, to satisfy R2, we can assume that scores depend only on the difference between the estimated probability and the prior, denoted $\Delta = Q – P^*$ (this is sufficient but not necessary). Under that assumption, $f(Q, P^*) = f(\Delta)$ and $f(\Delta=0) = 0$. In the following, we give examples of strictly proper scoring functions that satisfy these constraints.

Log/Linear Payoff. By means of example, suppose $f(\Delta) = \Delta$. This choice trivially satisfies the monotonicity condition (R1), and produces a zero reward for guessing the prior (requirement R2). To satisfy the expected value condition (R3), we solve Eq. (2). This simplifies to $P + (1 – P)\, dg/dQ = 0$ at $Q = P$, or $dg/dQ = –Q/(1 – Q)$. It follows that $g = Q + \ln(1–Q) + c$. Applying the boundary condition that $g(P^*, P^*) = 0$ (again R2) we obtain the following scoring functions:

$$\begin{aligned} f &= Q – P^* \\ g &= Q – P^* + \ln[(1 – Q)/(1 – P^*)] \end{aligned} \qquad (3)$$

These functions satisfy requirements R1 through R4. However, g is unbounded, approaching $-\infty$ as $Q \rightarrow 1$, violating requirement R5 (see Fig. 1). Additionally, the symmetry condition (R6) is not met. Two information providers who respectively assert A with certainty Q, and ~A with certainty 1 – Q, would receive different rewards. Finally, the expected value curves are essentially flat over large ranges, implying the reward function will not effectively discriminate among analysts, provided they stay away from near-certain predictions. These disadvantages eliminate log/linear payoff from further consideration.

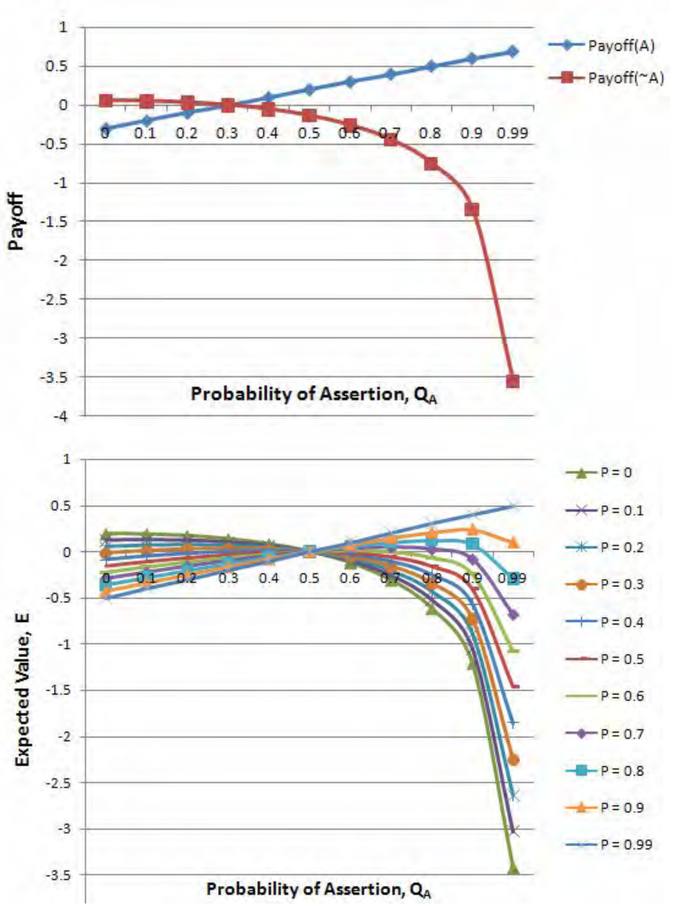

Fig. 1. The linear/log payoff function (Eq. 3), showing the payoff functions (top) and expected values (bottom). The example is for $P^* = 0.3$.

Quadratic Payoff. Assume f (Δ) has the form $a_0\Delta^2 + a_1\Delta + a_2$. We know $a_0 < 0$ because E must be concave downward for all values of Q. We can choose $a_0 = -1$ to specify the scoring function to within a multiplicative constant (which may be chosen as a function of the prior). In addition, $a_2 = 0$ (based on R2) so that f (Δ) = 0. Solving Eq.2 and applying boundary conditions (math omitted):

$$\begin{aligned} f &= 2(1-P^*)(Q-P^*) - (Q-P^*)^2 \\ g &= -2P^*(Q-P^*) - (Q-P^*)^2 \end{aligned} \qquad (4)$$

These functions are depicted in Fig. 2. Unlike the linear case, the payoff functions remain bounded, meet the symmetry condition, and do not have large flat regions. As an example of symmetry, f(0.8, 0.3) = g(0.2, 0.7) = 0.45.

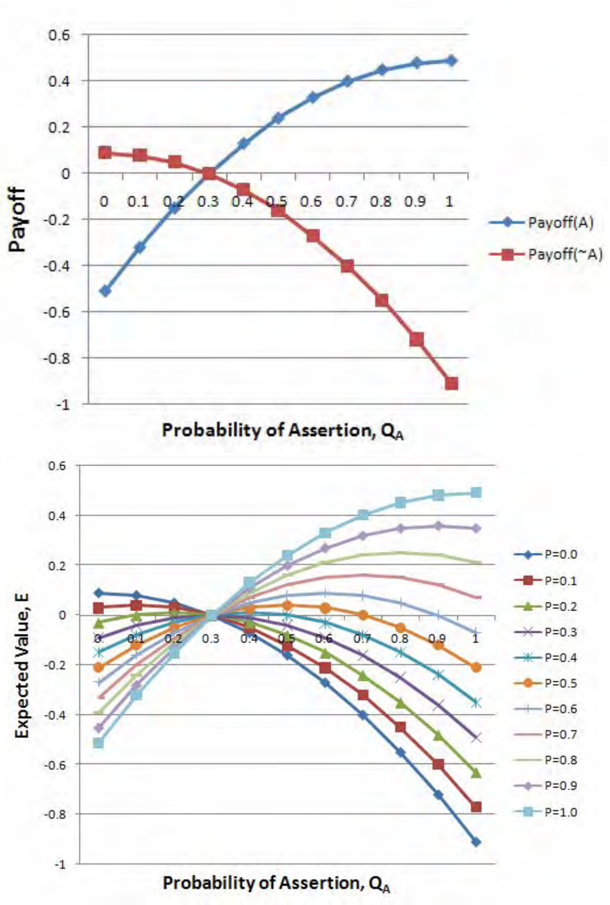

Fig. 2. The quadratic payoff function (Eq. 4), showing the payoff functions (top) and expected values (bottom). The example is for $P^* = 0.3$.

Other Payoff Functions. Among differentiable functions, we can also show that binomials in the form:

$$
\begin{aligned}
f &= -\Delta^c + (1- P^*)(c/(c-1))\Delta^{c-1} \\
g &= -\Delta^c - P^*(c/(c-1))\Delta^{c-1}
\end{aligned}
\qquad (5)
$$

generate admissible solutions for powers of $c \geq 2$ and integral. An open problem is to explore the usefulness of c fractionally greater than 1, a range that offers greater discrimination for Q near 1. The treatment will need care to avoid generating imaginary roots for negative delta. One might also explore multiplicative formulations, rewarding for the fractional change between P* and 1. However, the quadratic payoff satisfies all our conditions, and is certainly the simplest pair of functions to do so.

2.3 Analogy: Weighted Coin Tosses

In this section, we present an analogy to the current problem of predicting the probability of future events. Suppose we have a bag of coins confiscated from dishonest gamblers. Each coin may be weighted, so the chance of heads or tails is not 0.5. However, there is no reason to assume there are more coins weighted towards heads than tails. Therefore, the overall prior probability of heads is 0.5. We randomly select a coin from the bag (this represents a situation requiring analysis), and allow multiple analysts to physically examine the coin, without tossing it. Each analyst then predicts the probability of tossing heads with the coin. The coin is then tossed (producing ground truth). If the toss is heads, each analyst is paid off according to f, and if the toss is tails, each analyst is paid according to g.

Consider the following analysts:

1. *Probabilistically perfect analyst.* Always predicts the probability of A (heads) exactly. Note that "perfect" denotes accurate probability assessments, as opposed to the (impossible) clairvoyant analyst, who accurately predicts the outcome of each coin toss.
2. *Random analyst.* Produces uniform random guesses between 0 and 1.
3. *Biased analyst.* Over- or underestimates the chance of heads by a fixed amount, except where such a prediction would exceed 1 or go below 0.
4. *Noisy analyst.* Each probability estimate is off by random number drawn from a normal distribution with zero mean and given standard deviation (bounded between 0 and 1).
5. *Prior analyst.* This analyst always predicts the prior probability.

Figure 3 shows the results of 500 trials using the quadratic scoring function (Eq. 3). As expected, the perfect analyst outscores the other analysts in the coin-assessment task. The noisy analyst, shown for standard deviation 0.25, is the second best. The biased analyst, who in this case overestimates the probability of heads consistently by 0.25, is next. The analyst who picks the prior probability (0.5 in this case) earns zero, and the random analyst loses points at about the same rate as the perfect analyst earns points.

When the same experiment is carried out with the linear/log payoff function (not shown), the biased and noisy analysts both eventually make certain predictions (probability 0.0 or 1.0), which subsequently turn out to be false, causing them to earn $-\infty$ points, and fall off the chart.

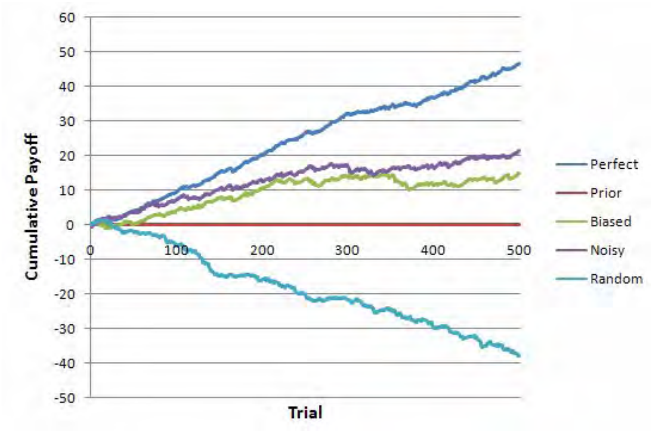

Fig. 3. Result of the weighted coin analogy for five types of analysts. In the case shown, the biased analyst consistently overestimates the probability of heads by 0.25, and the errors of the noisy analyst follows a normal distribution with standard deviation of 0.25.

Figure 4 shows the average payoff as a function of the magnitude of analyst error (bias for the biased analyst, standard deviation for the noisy analyst). In this chart, the average payoff for the perfect analyst (0.083) has been normalized to 1. Both the biased and noisy analysts perform worse than the random analyst for large biases or standard deviations because their probability predictions tend to the extremes. For example, the analyst suffering a large bias will consistently call heads (or tails) with probability 1, which is much worse than picking a random probability. The noisy analyst outperforms the biased analyst, because even in the extreme, some probability predictions will be between 0 and 1.

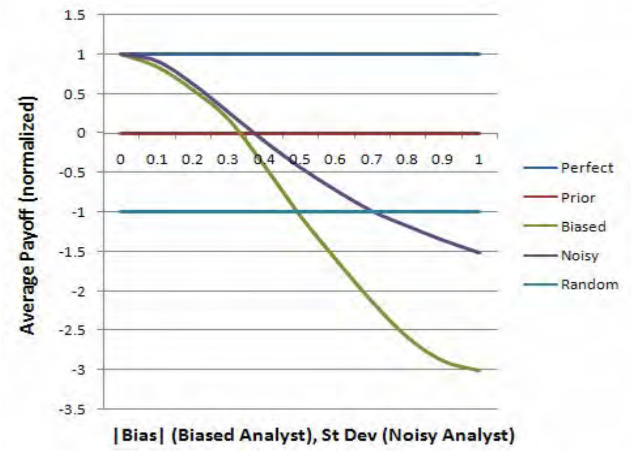

Fig. 4. Average reputation points earned per trial, for each type of analyst, as a function of average error (normalized to perfect analyst = 1).

2.4 Normalization and Analyst Comparison

Given scores for individual assertions, reputations can be developed by aggregating scores. The most obvious methods are summing and averaging. Averaging compensates for the difference in the total number of assertions scored by each analyst. However, averaging still does not ensure a fair comparison between information providers. Consider information utility. Information consumers might consider some assertions more important than others, and (formally or informally) assign a utility score to each assertion. In this case, the accumulated reputation score should derive from a weighted sum involving the product of the raw score and the utility of each assertion. The utility might be assigned by the consumer at the personal or enterprise level, the former leading to a personalized set of most trusted information providers. To reduce workload of manual assignment of utility, we suggest doing assignments to classes of assertions (e.g., probability that a patient has a certain disease). A default (utility=1) is also provided, lest additional administration discourage participation of the reputation system.

The other desirable correction involves differences in the priors between different analysts. The admissible scoring formulae are based on Δ, the difference between the asserted probability and the prior. In this way, we reward the analysts' unique contribution (novelty) as well as accuracy. To help understand the impact of priors, consider two doctors who analyze MRIs to diagnose a certain disease. The first doctor examines all MRIs. His prior reflects the incidence of the disease in the overall patient population. The second doctor is consulted for a second opinion only when the first doctor suspects the presence of the disease. Her prior reflects the incidence of disease in the subpopulation identified by the first doctor. It is easier for the first doctor to earn reputation points, because his observations change the probability of disease to a greater extent than the doctor offering a second opinion. These two different priors reflect different opportunities to earn reputation points, and should be accounted for when comparing and ranking the doctors.

To remove the impact of unequal priors, we can normalize by the size of scoring opportunity using either of the following two maximum obtainable scores:

- *Factually perfect* represents the score obtained when someone receives the maximum score at every opportunity. We obtain this standard by assuming $Q = 1$ when the event occurs and $Q = 0$ when the event does not occur.
- *Probabilistically perfect* represents the expected value of the score if the analyst chooses $Q = P$ at every opportunity. Since the scoring formulae explicitly maximize the expected value of the score for a probabilistically perfect analyst, this standard represents the best possible *obtainable* performance by any analyst.

To take the concept of probabilistic perfection slightly further, we can calculate the maximum expected value E_{max} under the assumption that $Q = P$, for the quadratic scoring function:

$$E_{max}(P, P^*) = E(P, P, P^*) = P\, f\,(P, P^*) + (1 - P)\, g(P, P^*)$$

$$= (P - P^*)^2 \qquad (5)$$

Based on Eq. 5, the *opportunity* to gain reputation points is proportional to the square of the difference between the prior and the actual event probability. In fact, if the consumer possesses an accurate prior for each event, the long-run scoring opportunity for any analyst is zero, since the analyst cannot tell the consumer anything new (although luck might prevail in the short run).

To apply Eq. (5) we must know P, which can only be estimated from relevant historical information, which may not exist. Alternatively, one may use a proxy for P, such as a group judgment. Using group judgment as a gold standard introduces a host of potential problems, such as unfairly downgrading independent-thinking analysts. Using the factually perfect score (the clairvoyant analyst) as the normalizing factor is a possibility, because it can be calculated using ground truth, but the effectiveness of approach is an open problem.

The last issue here is extrapolating to non-scored assertions. As we have discussed previously, assertions without ground truth are not scored. But given Q, the analyst will receive only one of two possible scores, f or g, the former with probability P and the latter with probability (1–P). Given P, we can determine the expected value of the score without ground truth. However, as in the case of normalizing for unequal opportunity, we are confronted with the unknown event probability. Again, one may use some proxy for P, such as group opinion, as long as one is aware of the inevitable hazards.

3 Other Considerations

3.1 Provider Behaviors

As mentioned earlier, information providers fall into three classes of behavior: honest, manipulative, and malicious.

Honest behavior has the provider doing his or her best, regardless of the reputation system. Obliviously honest behavior is likely among dedicated employees or in situations where there is no benefit to manipulation. A provider who is honest will attempt to estimate probabilities that match the true event probabilities.

Manipulative behavior aims to inflate reputation scores. The reputation system may be linked to rewards, privileges, and prestige, and thus a manipulative user may wish to accumulate undeserved rewards. As a side effect, but not as a goal, manipulators may deceive consumers about event probabilities. Manipulation can occur in two ways:

- *The choice of questions to answer.* The provider might attempt to "cherry pick" opportunities with high utility, e.g. a major crime, even if they have no special knowledge or qualification relative to the case. If evaluated by average accuracy, they may only choose cases where they have high certainty. Conversely, if scores are not weighted by utility, an analyst can manipulate the system by reporting on many obvious or uninteresting phenomena. Within an organization, this may be controlled by management oversight, for example, by penalizing time-wasting or by randomly assigning cases to analysts.

- *The probabilities.* A manipulator who seeks to optimize expected score has no incentive to mis-estimate probabilities, since in the long run, the correct probability gives the highest expected score. However, a manipulator might over- or under-estimate probabilities in the short run, seeking the “big win” or seeking to avoid a “loss”. It is thus desirable for management to reward long-term success, rather than rewarding occasional big wins or punishing mistakes.

An alternative route to combat manipulation is to keep providers unaware of the scoring system. This may be unethical within an enterprise, but it is certainly feasible for rating external providers, such in citizen event reporting or rating stock market pundits on the web.

Malicious behavior seeks to fool consumers, i.e., to convince consumers to believe and act on an incorrect probability. Methods for discouraging this behavior are well known. Casual attackers who seek instant gratification but will not invest much effort can be discouraged by requiring them prove identity, or requiring a certain number of previous postings that have been determined to be accurate before accepting their recommendations. The determined attacker is more difficult. One way to discourage sustained attack is to require that analysts provide useful information whose total utility is greater than the dis-utility of their deception. Thus, in return for possibly selling the lie, the attacker must do considerable work for the benefit of the organization. If they do not know the threshold for acceptance, this adds to their difficulty.

3.2 Applications within the Enterprise

In an enterprise, management might use scores in several ways in their relationship with employees, or with other sources that are recruited to provide information. In this section, we briefly discuss the uses for the reputation metrics by management that go beyond rating information providers.

Training. Management can teach providers how to remedy their identified weaknesses. If accuracy seems high (relatively) but productivity (quantity) is low, one might consider working more quickly. If accuracy seems low, one may need to examine the analyst’s techniques and improve subject-matter expertise. If utility weights are available, these can help the analyst focus on important cases. Bias estimates may help providers adjust for undue optimism or pessimism.

Assigning workload. Management may be responsible for giving each provider suitable tasks. Here, both accuracy and productivity is important. For example, a provider who has been accurate on relatively unimportant tasks might be given more important ones. Less critical tasks might be assigned to a provider who is moderately accurate but very fast. Finally, a provider who has shown ability to select high-payoff tasks can be given more freedom.

Judging and improving accuracy. Some providers may consistently overestimate probability of their assertions, while others may underestimate. Bias statistics attempt to estimate these tendencies, and can also be used to revise probability estimates.

4. Example: Citizen Event Reporting

In this section, we demonstrate the reputation system in the context of citizen event reporting (CER). CER involves using citizens to bolster information collection. This approach has most commonly been applied to crime-fighting efforts, but recently CER has been considered for asymmetric warfare [5]. The conflicts in Iraq and Afghanistan have highlighted the need to gather information known primarily by the local populace. Of particular concern is detection of improvised explosive devices (IEDs), bomb-making facilities, and identification of militant insurgents and terrorists. The challenge is to make CER work in hostile environments, where enemies may contribute false reports in an attempt to “game” the system, to lure first responders into ambushes, create decoys, or induce the authorities to target third parties.

In our example, we utilize a modified version of the agent-based simulation reported in [5]. In this simulation, there is one type of event to be reported, which we call a fire, but could represent sighting of a suspicious person, a weapons cache, IED, etc. Simulated citizens traverse the environment and can report an event if they come within a certain distance of it and “see” it. Reports represent an assertion of an event of interest (fire) at a given location and time. Ground truth is obtained when the authorities choose to respond to or investigate a report. For example, if a citizen reports a hidden weapons cache, upon investigation, the cache will prove either to be present or absent.

To use the reputation system, we assume the citizens are asked to attach certainty to their reports. In practice, this information could be collected on a qualitative scale (e.g. from “very unsure” to “very sure”), and mapped onto quantitative values. In the simulation, the certainty is based on the citizen’s distance from the event when they first observe it (the further away, the less certain). Unfriendly citizens (foes) can create false reports, and they may collude together to create a calling pattern that resembles a true report. The performance of the CER system is measured by the number of events responded to (fires extinguished) less non-events responded to (false alarms), divided by the total number of fires. The maximum performance is 1.

With anonymous reporting, the decision maker (DM) cannot discriminate true and false reports on the basis of reporter identity, greatly limiting the decision rules that can be implemented. There are various mechanisms for assuring calling identity using mobile devices, which will not be discussed here. For our purposes, we assume that personal reputations can be learned and factored into decision making.

The reputation system itself is implemented using the quadratic scoring rules (Eq. 4). The decision maker does not know which reports are true or false, but discovers ground truth only when the DM chooses to respond. At that point, each caller associated with the event receives reputation points. The decision to respond is taken when there is a report from any citizen with a positive reputation, with certainty exceeding the prior by 0.1. The prior is the probability of an active fire at any grid location at any time.

Unlike a black list, which was investigated earlier [5], the reputation system gives citizens the ability to “earn” their way out of a negative reputation by making accurate reports. Thus, an honest reporter who unintentionally makes an erroneous report suffers a temporary (rather than permanent) loss of reputation.

The simulation was run with 30% foes, with 50% of reports from foes involving collusion. With reputation system, the number of false alarms dropped by 95%, and the performance (defined above) increased from 0.73 to 0.95. This shows that a reputation system can greatly enhance the performance of a CER system, even in the presence of a large contingent of foes determined to undermine the system.

5. Summary and Open Problems

We have presented and illustrated a reputation scoring approach that considers prior probabilities, so that scores combine accuracy and novelty to the consumer. Extending prior work on proper scoring functions [2, 3], the approach encourages expectation-maximizing providers to give accurate probabilities, while addressing the need to keep administrative effort small. We discussed ways to normalize the scores, in order to judge providers’ accuracy, novelty, or productivity. The approach can succeed in situations where one can ascertain ground truth for a significant number of a provider’s assertions.

We demonstrated the reputation system in the context of citizen event reporting, where information is collected from many potentially unreliable sources. A reputation system is clearly needed to help decision makers identify reliable and unreliable reporters. Soliciting a degree of certainty with each report encourages citizens to provide information even if they are not 100% sure of the facts. On one hand, they have the opportunity to express when they believe something to be certain. On the other hand, they can safely transmit uncertain knowledge by declaring a low level of certainty, reducing the fear of being punished for misleading authorities, if the tip turns out to be false. Hence more tips will be gathered, allowing the decision maker additional chances to create actionable information. Over time, those that express appropriate levels of certainty will become the most reputable information sources.

For future work, we believe that normalization may be important for some purposes, to adjust for different workloads – easier and harder questions, priors’ being accurate (so agreement and zero novelty are optimal) versus inaccurate, and priors near 0.5 (no information) or .9 and .1 (unlikely to make large changes). It would also be desirable to explore scoring based on a logarithmic scale (so it is significant to move from .99 to .999 probability).

Aside from symmetry under negation, we have not considered the coherence of our scoring scheme under Boolean connectives (and, or). For example, an analyst could predict A, B, “A and B”, and “A or B” at the same time, and there should be some relationship between the scores received for the related assertions.

How reputation translates into personal trust, and how those trusts are converted to beliefs and actions, is also open to investigation. Use of trust scores to synthesize multiple sources of information has been investigated in [7] and by other authors, but we believe a stronger and more direct link to reputation is needed.

References

1. Bayarri, M. J., Berger, J.: The Interplay Between Bayesian and Frequentist Analysis. Statistical Science, 19: 58-80 (2004)
2. Brier, G. W.: Verification of forecasts expressed in terms of probability. Monthly Weather Review, 75, 1-3 (1950)
3. Gneiting, T., Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation. J. American Statistical Assn. 102, 359-378 (2007)
4. Hogarth, R.: Cognitive Processes and the Assessment of Subjective Probability Distributions. J. American Statistical Assn. 70, 271 -289 (1975)
5. Kramer, M., Costello, R., Griffith, J.: Investigating the Force Multiplier Effect of Citizen Event Reporting by Social Simulation. Conference of the European Social Simulation Association (2008)
6. Savage, L: Elicitation of Personal Probabilities and Expectations. J. American Statistical Assn. 66, 783-801 (1971)
7. Zuo, Y., Panda, B.: Information trustworthiness evaluation based on trust combination ACM symposium on Applied computing, 1880-1885 (2006)