A Framework for Modeling Decision Making and Deception with Semantic Information

Christopher Griffin
Applied Research Laboratory
Penn State University
University Park, PA 16802
E-mail: griffinch@ieee.org

Kathleen Moore
College of Information Science and Technology
Penn State University
University Park, PA 16802
E-mail: kam6015@psu.edu

***Abstract*—We propose a mixed logical and game theoretic framework for modeling decision making under the potential for deception. This framework is most appropriate for online communities in which a decision maker must act upon information being provided by various sources with various different motivations. We show that in the simple three-player game we propose there are always equilibria in pure strategies. We then extend the three player game to a case where there are mixed strategy equilibria. We discuss how to approximate the truth of a given statement using a logical construct and how this can be used as a proxy in payoff functions. Finally we discuss as future directions the use of regret functions and live play.**

***Index Terms*—Deception, Game Theory, Social Networks, Formal Models**

I. Introduction

The head decision maker of a crisis response team during a regional flood, receives notice that a shelter the next town over that has taken in displaced persons and is in need of food, water, and blankets. There are conflicting reports as to whether the main road to the next town is passable, Vetted personnel are not in the area and aerial imagery is not available. Checking the message stream on Twitter offers conflicting accounts. Some posts read the road has been washed out, other read that road is passable. Who does the decision maker trust when the person tweeting is unknown? Even if the person seems trustworthy, how do you know the information offered is true?

Trust is the foundation for all human interaction. Trust allows people to feel comfortable taking risk when interacting or exchanging with others. Without trust, people would cease interacting with each other, societal functions would slow to a crawl as would business operations, countries would seal themselves off from each other if not start outright war. From buying a pack of gum, to asking for directions on a city street, trust enables people to establish expectations, weigh risks, and proceed in a course of action until events unfold that causes people to think and behave differently. Trust is a heuristic decision rule that allows decisions to be made in a matter of seconds rather than engaging in long mental processes of rational reasoning.

While trust is largely recognized as a varied concept across disciplinary study, there is little consensus among researchers as to how trust should be properly defined. Early research on the concept is grounded in interpersonal interaction on psycho-social and business relations, while more recent studies focus on a business perspective in virtual environments, privacy concerns in social media, and the use of microblogging in crisis research.

Since the Middle Ages, technology has played a part in mediating peoples perception of trust. From the invention of Gutenbergs printing press to the creation of the Internet and World Wide Web, people have always had an ambivalent relationship with communication technologies and the information gleaned from them. With the introduction of each new technological artifact, society often questions whether this new tool will benefits or harm society as we know and operate within it.

As more and more of our communication occurs online, our ability to use critical cues such as knowledge of the information source, facial or body language and common references becomes difficult or impossible. This leaves few alternatives. Dax Norman, a cryptologist in the United State National Security community, developed a trust-scale by which consumers of online information could evaluate the trustworthiness of a website [1]. Author reputation, author affiliations, expert recommendations of the site, corroborated information, identifiable sources, official branding, and professional look all serve to instill a sense of trust in the content that a consumer views [2]. While most of the attributes Norman describes were recommended by professional

DOI 10.1109/SPW.2012.25

analysts interviewed for his thesis, the overall study lacked academic rigor in the ethnographic methodology as the author also included personal perceptions as trust factors. Nonetheless, the trust scale still serves as a model within the intelligence community.

Alternate approaches include the Wikipedia approach of crowdsourcing. Enabling the public to crowdsource in online problem solving, allows the online consumer to pick and choose information they trust is relevant and useful. However, this also allows for the contribution of poor quality and even false information [OO00]. Wikipedia, the most popular interface for crowdsourcing, mitigated this problem by requiring registration in order to contribute to the wiki and by also employing wizards (editors) who manage and question content [3]. Despite a public perception of the wiki lacking veracity, a procedure of vigorously edited and reviewed articles that earn the Wikipedia bronze star for accuracy, has resulted in an populist system that has been shown these select articles to be as accurate as the Encyclopedia Britannica [3], [4], [5].

Lastly, the use of reputation systems [6] in e-commerce is common. In these systems, potential deceivers (sellers) are given a reputation which can be influenced by the users of the system. [6] attempts to deal with the problem of sybil-attack, which can bias a reputation positive. Problems of sybil-detection have also been considered [7] for cleaning reputation systems.

In this paper, we attempt to build a model for motivating deception through both game theory and semantic content. We eschew the challenging problem of automatically recognizing the semantics information in online communication and assume via *deus ex machina* that a collection of logical assertions is available that will be played by a teller. An actor will have to decide whether to act on specific information provided by the teller and will use information about the world she already has. Payoffs are received based on the actions of both players.

II. Notation

A *language* $\mathcal{L}$ is a triple $\langle \mathcal{F}, \mathcal{R}, \mathcal{C} \rangle$ where:

1) $\mathcal{F}$ is a set of function symbols f, each with positive integer arity n_f,
2) $\mathcal{R}$ is a set of relation symbols R, each with non-negative integer arity n_R, and
3) $\mathcal{C}$ is a set of constant symbols c.

For us, the constants in the logical language represent the elemental constituents of the real-world. Functions are used to describe the interactions between these basic components that produce new structures. The set of all interactions is called the set of *L-terms*; it is the smallest set $\mathcal{T}$ such that

1) $c \in \mathcal{T}$, where c is a constant of the language L,
2) $x \in \mathcal{T}$, where x is a variable in $\mathcal{V}$ and
3) if f is an n-ary function and $t_1, \ldots, t_n$ are terms, then $f(t_1, \ldots, t_n)$.

We assume a finite but arbitrarily large reserve of variables $\mathcal{V}$. Relationships between various terms are described by *formulas*. We consider only a special type of formula, namely the smallest set Φ such that:

1) $R(t_1, \ldots, t_n) \in \Phi$, where $t_1, \ldots, t_n$ are L-terms and $R \in \mathcal{R}$ is an n-ary relation symbol,
2) if $\varphi \in \Phi$, then $\neg\varphi \in \Phi$, and
3) if $\varphi, \psi \in \Phi$, then $\varphi * \psi \in \Phi$, where $* \in \{\wedge, \vee, \rightarrow, \leftrightarrow\}$.

A formula with *no free variables* is called a sentence. The reader familiar with logic will recognize that we are somewhat restricting ourselves, considering only bounded quantifier formulas. Our assumption will allow us to specify the correspondence between logical formulas and mixed integer linear programming problems. Finally it should be noted that any existential quantifier in bounded predicate calculus can be reduced to a disjunction over the universe of objects, while a universal quantifier can be reduced to a conjunction.

To complete our logical description of a system, we must describe the system itself. An *L-structure* or *model* is a tuple $\mathcal{M} = \langle M, \mathcal{F}^{\mathcal{M}}, \mathcal{R}^{\mathcal{M}}, \mathcal{C}^{\mathcal{M}} \rangle$, such that

1) M is a set called the universe,
2) $\mathcal{F}^{\mathcal{M}}$ is a set of functions $f^{\mathcal{M}} : M^{n_f} \rightarrow M$, with n_f the arity of the function symbol $f \in L$,
3) $\mathcal{R}^{\mathcal{M}}$ is a set of relations $R^{\mathcal{M}} : R^{\mathcal{M}} \subseteq M^{n_R}$, where n_R is the arity of $R \in L$,
4) and $\mathcal{C}^{M}$ is a set of constants in M.

By $|\mathcal{M}|$, we mean $|M|$. When $|\mathcal{M}|$ is finite, we call $\mathcal{M}$ a finite model. Models provide a way of assigning truth values to sentences. Let $\mathcal{M}$ be a model of a language $\mathcal{L}$. We say that an L-sentence φ is *true* in $\mathcal{M}$ if when we replace each function, relation and constant in φ with its corresponding function, relation or constant in $\mathcal{M}$, then the resulting statement holds in the grounding provided by $\mathcal{M}$. Otherwise, the sentence is false. (This is the *Tarskian* definition of truth [8].) If this is the case we write $\mathcal{M} \models \varphi$, where φ is the sentence in question. Let Φ be a set of sentences, by $\mathrm{Mod}(\Phi)$ we mean the set of all models that make each sentence in Φ true.

III. A Measurement of Truth and Falsehood

We assume the existence of a language $\mathcal{L} = \langle \mathcal{C}, \mathcal{R} \rangle$ consisting of constants $\mathcal{C}$ and predicates $\mathcal{R}$ and no functions. Any arbitary language can be converted to such a language as needed [8]. We will also assume we are working in the first order predicate calculus, though extensions to higher-order logics may be permissible.

Without loss of generality, we will impose a finite model hypothesis on sentences constructed in our language and thus given a fixed model size any sentence can be written in conjuctive normal form. A formula is in conjunctive normal form (CNF) if it may be written as:

$$\omega = \bigwedge_{i=1}^{M} \varphi_i \tag{1}$$

where for each i,

$$\varphi_i = \bigvee_{j=1}^{N_i} \psi_{ij} \tag{2}$$

and $\psi_{ij} = R_{ij}(t_1, \ldots, t_{n_{ij}})$ or $\neg R_{ij}(t_1, \ldots, t_{n_{ij}})$, with R_{ij} a predicate.

It is well known that any formula φ in propositional calculus may be rewritten as an equivalent formula in CNF [8]. Since we assume that our logic has a bounded universe, we may assume formulas are given in CNF.

Clearly given any model $\mathcal{M}$ that instantiates the language $\mathcal{L}$, the Tarskian definition of truth maybe applied. However, this definition is far to narrow to suit our purposes. To each sentence ω, we wish to associate a rational value $p_\omega|\mathcal{M}$ that is the proportion of truth within the sentence. If ω is as given in Equation 1 then:

$$I_i^{\mathcal{M}}(\omega) = \begin{cases} 1 & \text{if } \mathcal{M} \models \varphi_i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Then:

$$p_\omega|\mathcal{M} = \frac{1}{N} \sum_{i=1}^{M} I_i(\omega) \tag{4}$$

We assert that this definition gets to the very heart of deception insofar as it attempts to capture the notion of "a little true." Deception hinges on the believability of the underlying story being told. A story that is 90% true with 10% falsehood is more likely to be accepted as factual than a story that is 10% true and 90% falsehood, particularly in the presence of additional, corroborating, information. Clearly, for any sentence ω with $p_\omega|\mathcal{M} < 1$, ω is a false sentence, but the degree to which it is false is what is measured by $p_\omega|\mathcal{M}$. For the remainder of this paper, we will assume there is a special (potentially unknowable) model $\mathcal{G}$, ground truth which describes the absolutely true state of world. We will discuss $\mathcal{G}$ later.

IV. Game Theoretic Model of Deception

Consider a simplified world in which there are two players, *Actor*, Player 2 and *Teller*, Player 1. Player 2 will choose to act upon the information received from Player 1. We will assume time to be epochal and without loss of generality, we assume that at any time n both players have a common set of sentences Φ_n. These sentences may be axiomatic (e.g., "the sky is blue in the daytime") or they may be common information shared by the players over the course of the evolution of the situation. At time $n = 0$, there is an original (potentially) empty set of axioms Φ_0 that are introduced.

At time n, suppose Player 1 wishes to provide a sentence $\sigma \notin \Phi$ to Player 2. The result will be $\Phi_{n+1} = \Phi_n \cup \{\sigma\}$ if Player 2 agrees to act. There are three (gross) possibilities:

1) For every model $\mathcal{M}$ so that $\mathcal{M} \models \Phi_n$, $\mathcal{M} \models \sigma$. That is, in every possible way the world could be so that all sentences in Φ_n hold simultaneously, it is also true that σ must be true.
2) For every model $\mathcal{M}$ so that $\mathcal{M} \models \Phi_n$, $\mathcal{M} \models \neg\sigma$. That is, in every possible way the world could be so that all sentences in Φ_n hold simultaneously, it is also true that σ must be false.
3) There are two models $\mathcal{M}_1$ and $\mathcal{M}_2$ so that $\mathcal{M}_1, \mathcal{M}_2 \models \Phi_n$ but $\mathcal{M}_1 \models \sigma$ and $\mathcal{M}_2 \models \neg\sigma$. That is, σ is independent of the set Φ_n.

Since Player 1 is providing the information σ, we assume that Player 1 knows whether $\mathcal{G} \models \sigma$ even if $\mathcal{G}$ is not completely knowable. Furthermore, we assume that Player 1 knows $p_\sigma|\mathcal{G}$. In the absence of this assumption, we can assume that Player 1 can compute a proxy value:

$$p_\sigma|\Phi_n = \frac{\sum_{\mathcal{M} \in \mathrm{Mod}(\Phi_n)} p_\sigma|\mathcal{M}}{|\mathrm{Mod}(\Phi_n)|} \tag{5}$$

This can be approximated, if necessary, by sampling the space of models. An algorithm for such sampling can be obtained from the hypothesis space search algorithm in [9].

At any given time n, Player 1 will have a finite (but perhaps large) set Ψ_n of sentences that can be told to Player 2. Thus, the strategy set for Player 1 is Ψ_n while the strategy set for Player 2 is $\mathbb{B} = \{0, 1\}$, where 0 indicates no action is taken and 1 indicates an action is taken. Let Player 2's strategy be $x \in \mathbb{B}$ given $\sigma_n \in \Psi_n$. Then Player 1 receives reward $\pi^{(1)}(x, \sigma_n)$. We will assume that the marginal payoff $\pi^{(1)}(1, \sigma_n)$ is monotonically increasing with $p_{\sigma_n}|\mathcal{G}$ while $\pi^{(1)}(0, \sigma_n)$ is monotonically decreasing with $p_{\sigma_n}|\mathcal{G}$. That is, given Player 2 chooses to act, he obtains a better reward the more true σ_n is and if Player 2 chooses not to act, he obtains a higher cost (worse reward) the more true σ_n is. At any time epoch, ignoring global concerns, Player

2's problem is:

$$\max_{x\in\{0,1\}} x\cdot\pi^{(2)}(1,\sigma_n)+(1-x)\cdot\pi^{(2)}(0,\sigma_n) \tag{6}$$

However, since Player 2 does not know $p_{\sigma_n}|\mathcal{G}$, he may use $p_{\sigma_n}|\Phi_n$ as a proxy. To differentiate this case, we write: $\pi^{(2)}(x,\sigma_n|\Phi_n)$ to denote the computed payoff based on an estimate of the veracity of σ_n from Φ_n, while $\pi^{(2)}(x,\sigma_n)$ is the true payoff.

At any time epoch, we have a simple two player sequential game illustrated in Figure 1. Let

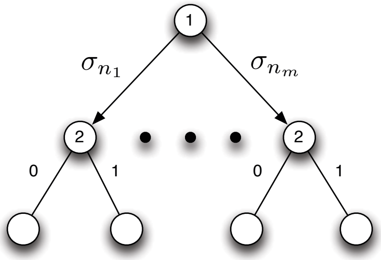


Fig. 1. A notional game tree describing the interaction between Teller and Actor.

$$x^*(\sigma_n)=\arg\max_{x\in\{0,1\}} x\cdot\pi^{(2)}(1,\sigma_n)+(1-x)\cdot\pi^{(2)}(0,\sigma_n). \tag{7}$$

Then, Player 1's problem is:

$$\max_{\sigma_n\in\Psi_n} x^*(\sigma_n)\cdot\pi^{(1)}(1,\sigma_n)+(1-x^*(\sigma_n))\cdot\pi^{(1)}(0,\sigma_n) \tag{8}$$

V. Multi-Turn Game Model

Based on the simple two player game described above, we now provide a simplified multi-turn game that describes the interaction between players over time.

1) At time 0, all players have an initial information set Φ_0
2) At each time $n\geq 0$, Player 1 chooses a strategy σ_n from Ψ_n and presents it to Player 2.
3) Player 2 chooses a strategy $x\in\mathbb{B}$.
4) Player 2 determines whether $\mathcal{G}\models\sigma_n$, Φ_n is updated to Φ_{n+1}.
5) Player 1 gains a reward of $\pi_n^{(1)}(x,\sigma_n)$ and Player 2 gains a reward of $\pi_n^{(2)}(x,\sigma_n)$.
6) Player 2 decides whether to continue the game or halt play. If play continues, then return to Step 2.

Halting play occurs when Player 2 no longer wishes to accept information from Player 1. This occurs in the real-world when an actor decides that a source is untrustworthy and will no longer act on information provided by that source.

The objective of the game is to obtain the largest net payoff possible. We can assume a time discounted payoff function to consider arbitrarily long games. That is:

$$R^{(i)}=\sum_{n=0}^{N}\beta^n\pi_n^{(i)}(x_n,\sigma_n) \tag{9}$$

where $\beta\in(0,1)$. We denote this game as $\mathfrak{G}(N)$. It is easy to see that this is just an extension of the game defined in the previous section, which can be presented by the multi-period game tree illustrated in Figure 2.

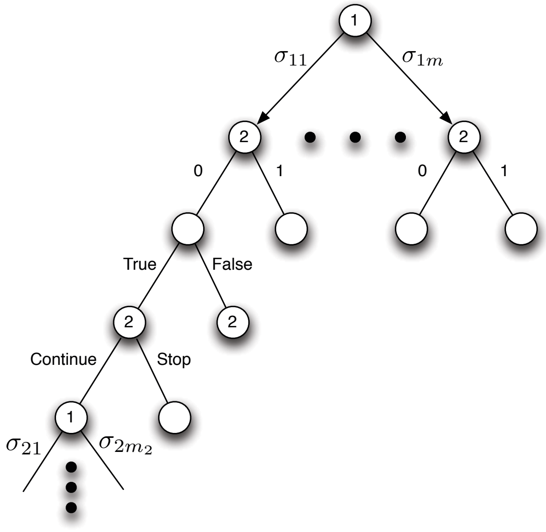


Fig. 2. A notional game tree describing the interaction between Teller and Actor in a multi-period game.

Definition 1 (Strategy)**.** Let $V^{(i)}$ represent the vertices controlled by Player i in the game tree representing the multi-period game (see Figure 2). A strategy is a rule that determines the decision made by Player i for each vertex in $V^{(i)}$.

Let $\mathbf{s}_1,\mathbf{s}_2$ be strategies for Player 1 and 2 respectively. Let $R^{(i)}(\mathbf{s}_{(1)},\mathbf{s}_{(2)})$ be the cumulative payoff for Player i when the players use these strategies.

Definition 2 (Equilibrium)**.** A strategy pair $(\mathbf{s}_1^*,\mathbf{s}_2^*)$ is an equilibrium if:

$$R^{(i)}(\mathbf{s}_i^*,\mathbf{s}_{-i}^*)\geq R^{(i)}(\mathbf{s}_i,\mathbf{s}_{-i}^*)$$

By $(\mathbf{s}_i,\mathbf{s}_{-i})$ we simply mean the strategy for Player i and a corresponding strategy for the alternate player $-i$.

Proposition 1. *For any finite N, there is at least one equilibrium strategy $(\mathbf{s}_1^*,\mathbf{s}_2^*)$ for $\mathfrak{G}(N)$.*

Proof: See the proof of Theorem 4.52 in [10]. ■

Remark 1. Note the previous proposition also holds if we allow randomness in the next Player 1 strategy set. That is, if after Step 6 in going to Step 2, we randomly choose Ψ_n from a finite set of possibilities, then there is still at least one equilibrium strategy.

Player 1 has a *motivation to deceive* Player 2, if there is at least one vertex in $v \in V^{(2)}$ so that the strategy $\mathbf{s}_1^*$ causes Player 1 to play σ_v with the property that $p_{\sigma_v}|\mathcal{G} < 1$.

We may add to the computational richness of this game in a number of ways. The easiest one is to consider a three player Markov game [11]. In this game, a player *Nature* will determine the truth or falsehood of the statement σ_n in an attempt to hurt Player 2 as much as possible. Player 2 is then playing a modified zero-sum game with Nature each time he moves; Player 2 has the four strategies: believe and continue, believe and stop, disbelieve and continue, disbelieve and stop. The resulting state of the game is then a function of choices of the players. If we denote this game by $\mathfrak{G}'(N)$, then we have:

Proposition 2. *For any finite N, there is at least one equilibrium in mixed strategies for the players in $\mathfrak{G}'(N)$.*

Proof: The result follows from [11] and the proof of Theorem 4.52 in [10] with an appropriate application of dynamic programming. ■

Remark 2. Computationally, the game $\mathfrak{G}'(N)$ is not a traditional zero-sum game. Nature's move may be constrained by the nature of Φ_n. That is, if Φ_n represents ground truth and $\Phi_n \not\models \sigma_n$ then Nature cannot force σ_n to be true. This information can be used by Player 2 to constrain his loses to Nature (and indirectly to Player 1).

Another extension motivated by the functioning of the Department of Defense or other Nongovernmental Organization (NGO's) is to break Player 2's role into two players: an analyst and an actor (generally a field officer). The analyst will still make a decision on whether to believe or disbelieve Player 1's output while the actor will decide whether to act on information provided by the analyst. In this case, the payoff to the actor will be much higher (or lower) than the payoff to the analyst. Again, this game will have at least one Nash equilibrium.

VI. Trust, Mistrust, Distrust, Misplaced Trust

Based on the above definition, we are now in a position to define several common terms used in the trust literature [12]. Fix two strategies $\mathbf{s}_1$ and $\mathbf{s}_2$ for the two players in an instance of $\mathfrak{G}(N)$. Then:

1) Player 2 *trusts* Player 1 at vertex $v \in V^{(2)}$ if $\mathbf{s}_2(v) = 1$.
2) Player 2 *trusts* Player 1 if for all $v \in V^{(2)}$, $\mathbf{s}_2(v) = 1$.
3) Player 2 *distrusts* Player 1 at vertex $v \in V^{(2)}$ if $\mathbf{s}_2(v) = 0$.
4) Player 2 *distrusts* Player 1 if for all $v \in V^{(2)}$, $\mathbf{s}_2(v) = 0$.
5) Player 2 *mistrusts* Player 1 if there is some vertex v at which Player 2 distrusts Player 1.

The concepts of trust and distrust are easy to see now that a payoff is involved. A more interesting condition is the case of *incorrect trust*. Suppose there is at least one vertex $v \in V^{(2)}$ so that $\mathbf{s}_2(v) = 1$ so that when we define a new strategy $\mathbf{s}_2'$ equal to $\mathbf{s}_2$ in every way except $\mathbf{s}_2'(v) = 0$ we have $R^{(2)}(\mathbf{s}_1, \mathbf{s}_2) < R^{(2)}(\mathbf{s}_1, \mathbf{s}_2')$, then Player 2 has incorrectly trusted Player 1.

VII. Future Directions: Regret Functions and Live Play

We would like to use this model to attempt to understand the behavior of individuals in situations where trust is required. In real life, Player 2 will never play to maximize his total payoff since computing this maybe impossible. Instead, he may attempt to minimize a regret function. Recall at vertex $v \in V^{(2)}$, Player 2's problem is:

$$\max_{x \in \{0,1\}} x \cdot \pi^{(2)}(1, \sigma_v|\Phi_v) + (1-x) \cdot \pi^{(2)}(0, \sigma_v|\Phi_v) \quad (10)$$

Player 2's *vertex regret function* [13] is defined as:

$$\rho(x) = -1\left(x\pi^{(2)}(1, \bot) + (1-x)\pi^{(2)}(0, \top)\right) \quad (11)$$

This is the cost incurred from false positive and false negative actions as a result of choosing to accept or reject the statement given by Player 1.

By way of example, suppose Player 2's payoff matrix at a given vertex is given by:

$$\begin{bmatrix} 10 & -100 \\ -10 & 5 \end{bmatrix}$$

Here Player 1 is the row player and his strategies are to tell the truth or lie. Player 2 is the column player and his strategies are to accept the statement as true and act or declare it false and act. In this case, Player 2 suffers a large loss if he fails to act on true information. Then the regret function is:

$$\rho(x) = 10x + 100(1-x)$$

which has minimum when $x = 1$, suggesting that Player 2's minimum regret strategy is to believe Player 2. We can compare this to the case when Player 1 and Player

2 are engaged in a zero-sum game (and Player 2 cannot evaluate the veracity of Player 1's assertion). In this case, Player 2's Nash equilibrium is to trust 84% of the time and distrust 16% of the time (that is $x = 0.84$). Then the regret function value is $8.4 - 16 = 24.4$. It is important to recognize that this regret function is computed as the worst possible case Player 2 can observe. Other regret functions can be defined in other ways.

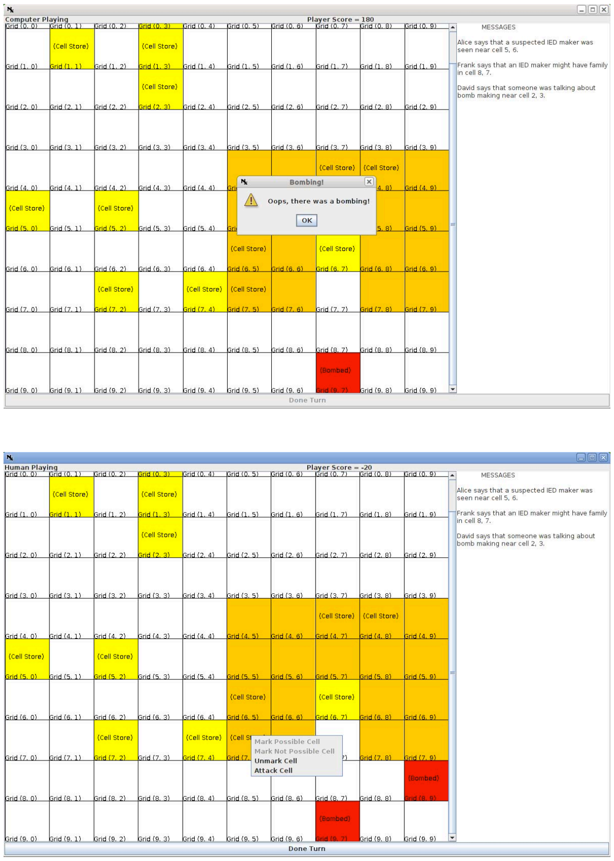
Regret can play an important role in decision making. Suppose that $p_{\sigma_n}|\Phi_n$ is very small, but the cost of making an incorrect decision is very high. The probability will bias an individual toward favoring disbelief, but regret may bias an individual toward belief, just in case. This is an important component of determining an estimate of $p_{\sigma_n}|\Phi_n$.

Computing $\pi^{(2)}(x, \sigma_v|\Phi_v)$ ($x \in \mathbb{B}$) the wall clock time must be taken into consideration. It may be that to compute the exact value of $\pi^{(2)}(x, \sigma_n|\Phi_v)$ is very computationally intensive because of the nature of Expression 5. However, sub-sampling $\mathrm{Mod}(\Phi_v)$ can lead to performance speedup. The question then becomes, how little time should be allotted to confirming the veracity of a given statement σ_n?

To answer this question, we may assign to the teller a level of trust τ_v at each vertex $v \in V^{(1)}$. When τ_v is small, exploration of the space $\mathrm{Mod}(\Phi_v)$ is extensive in an attempt to determine a good approximating value of $\pi^{(2)}(x, \sigma_v|\Phi_v)$. When τ_v is large, exploration of $\mathrm{Mod}(\Phi_v)$ is small because trust is placed in the behavior of Player 1. We can think of $1/\tau_v$ as determining the amount of time we are willing to spend "checking out Player 2's story" as opposed to getting along with the game. Naturally, we can monetize this in the game as well, incurring a cost for each time unit spent computing $\pi^{(2)}(x, \sigma_v|\Phi_v)$. The result will be an incentive to trust whenever it is unlikely that doing so will not increase the regret function (or decrease overall payoff). Techniques like those presented in [6] can be used for reputation management. Clearly this extension should be studied in greater detail.

VIII. Experimental Testing of the Model

We have designed a simple computer game for evaluating our assumptions on behavior in scenarios in which they must evaluate the truth of statements made by individuals. The game is a variation of *Minesweeper* and will be played by students who hope to become analysts for the United States Department of Defense. In the game, a terrorist cell is hiding within a 10×10 grid. A terrorist organization must be located within two units (by the Manhattan metric) of a cell phone store to efficiently obtain radio communications parts for remote detonation. A player can mark a cell as potentially containing the terrorist cell or not possibly containing the cell. The player can also initiate an attack (act). False attacks are penalized (because local sentiment turns against the aggressors). Players are given information by five computer players who provide true and false information at varying rates. The information is easy to parse, to allow post-hoc analysis of the optimal strategy for the player. The player advances the game by using a "Done Turn" button, which triggers Player 1 to provide his next statement. Occasionally, the terrorist cell will strike and the player incurs a penalty. Attacks occur at least two units from the terrorist cell but no more than five units away. A score can be optionally displayed.



MESSAGES

Alice says that a suspected IED maker was seen near cell 5, 6.

Frank says that an IED maker might have family in cell 8, 7.

David says that someone was talking about bomb making near cell 2, 3.

Fig. 3. Screenshots of the test software to be delivered to analysts in training.

We will track the player's behavior by recording their moves in the game. This will allow us to analyze strategic thinking after game play is done. We will also analyze responses to questions regarding which of the computer players seemed most trustworthy based on repeated play. The objective of these experiments is to determine whether the game-theoretic model proposed in this paper is a reasonable model of human behavior in the presence of deception provided through text input and in which the human can reason over true/false information using established rules of behavior (representing common knowledge).

IX. Conclusions

In this paper, we have presented a model of decision making for an actor and a teller in an online setting when semantic information is available about the statements being made by the teller. We show the simple results that in this game theoretic context there are equilibria in pure strategies. In a richer context in which the actor plays a constrained zero-sum game against nature in which the truth of the teller's statements is to be determined by nature, there is an equilibrium in mixed strategies. We suggest future directions that include the use of a regret function and the incorporation of a reputation management system that helps determine how much wall-clock time should be spent in evaluating the probability that a statement is false.

Acknowledgment

Christopher Griffin and Kathleen Moore were supported by the Army Research Office under Grant Number W911NF-11-1-0487.

References

[1] D. R. Norman, "How to identify credible sources on the web," Master's thesis, Joint Military Intelligence College, 2001.

[2] ——, "Web sites you can trust," *American Libraries*, vol. 36, August 2006.

[3] J. Giles, "Internet encyclopaedias go head to head," *Nature*, pp. 900–901, 2005.

[4] J. Liu and S. Ram, "Who does what: Collaboration patterns in the wikipedia and their impact on data quality," in *19th Workshop on Information Technologies and Systems*, 2009, pp. 175–180.

[5] D. M. Wilkinson and B. A. Huberman, "Cooperation and quality in wikipedia." in *Proceedings of the 2007 international symposium on Wikis - WikiSym '07*, 2007, pp. 157–164.

[6] G. Kesidis, A. Tagpong, and C. Griffin, "A sybil-proof referral system based on multiplicative reputation chains," *IEEE Comm. Lett.*, vol. 13, no. 11, pp. 862–864, 2009.

[7] A. Kurve and G. Kesidis, "Sybil detection via distributed sparse cut monitoring," in *Proceedings of the IEEE International Conference on Communications, ICC'11*, Kyoto, Japan, June 2011.

[8] J. Bell and M. Machover, *A Course in Mathematical Logic*. Amsterdam, Netherlands: North-Holland, 1977.

[9] C. Griffin, K. Testa, and S. Racunas, "An algorithm for searching an alternative hypothesis space," *IEEE Trans. Sys. Man and Cyber. B*, vol. 41, no. 3, pp. 772–782, 2011.

[10] C. Griffin, "Game theory: Penn state math 486 lecture notes (v 1.02)," http://www.personal.psu.edu/cxg286/Math486.pdf, 2010-2011.

[11] L. S. Shapley, "Stochastic games," *PNAS*, vol. 39, no. 1095-1100, 1953.

[12] S. Marsh and M. R. Dibben, "Trust, untrust, distrust and mistrust–an exploration of the dark (er) side." *Trust Management*, vol. 3477, pp. 17–33, 2005.

[13] G. Loomes and R. Sugden, "Regret theory: An alternative theory of rational choice under uncertainty," *Economic Journal*, vol. 92, no. 4, pp. 805–24, 1982.