CHAPTER 27

Manipulation-Resistant Reputation Systems

Eric Friedman, Paul Resnick, and Rahul Sami

Abstract

This chapter is an overview of the design and analysis of reputation systems for strategic users. We consider three specific strategic threats to reputation systems: the possibility of users with poor reputations starting afresh (*whitewashing*); lack of effort or honesty in providing feedback; and *sybil attacks*, in which users create phantom feedback from fake identities to manipulate their own reputation. In each case, we present a simple analytical model that captures the essence of the strategy, and describe approaches to solving the strategic problem in the context of this model. We conclude with a discussion of open questions in this research area.

27.1 Introduction: Why Are Reputation Systems Important?

One of the major benefits of the Internet is that it enables potentially beneficial interactions, both commercial and noncommercial, between people, organizations, or computers that do not share any other common context. The actual value of an interaction, however, depends heavily on the ability and reliability of the entities involved. For example, an online shopper may obtain better or lower-cost items from remote traders, but she may also be defrauded by a low-quality product for which redress (legal or otherwise) is difficult.

If each entity's history of previous interactions is made visible to potential new interaction partners, several benefits ensue. First, a history may reveal *information about an entity's ability*, allowing others to make choices about whether to interact with that entity, and on what terms. Second, an expectation that current performance will be visible in the future may deter *moral hazard* in the present, that hazard being the temptation to cheat or exert low effort. In other words, visible histories create an *incentive* to reliably perform up to the entity's ability. Finally, because histories reveal information about abilities, entities with higher abilities will be drawn to participate, as they will be distinguishable from those of lower abilities, and respected or rewarded appropriately. In other words, visible histories *avoid problems of adverse selection.*

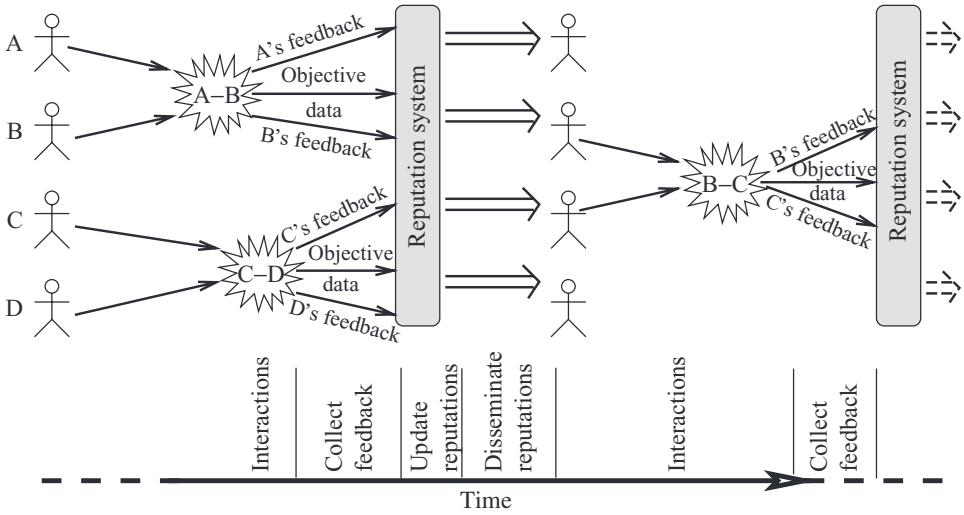

Figure 27.1. Example illustrating reputation system dynamics.

A reputation system collects, maintains, and disseminates *reputations*—aggregated records from past interactions—of each participant in a community. The rapid advance in computational power and communication capacity on the Internet has been a double-edged sword: On one hand, it has enabled the construction of reputation systems that can store, gather, and process large quantities of information. On the other hand, it has allowed more sophisticated attacks on the integrity of the reputation system to be mounted.

Reputation systems have been designed for use in many settings, including online auctions, e-storefronts, and a wide-range peer-to-peer systems. These systems naturally have differing interfaces, and track different aspects of user behavior. However, they all share certain underlying components, which are illustrated in Figure 27.1.

The core of a reputation system involves collecting records of entity A's past behavior, and then disseminating reputation information to others who may potentially interact with A in the future. (We use the term "entity" to denote the real-world entity to which we seek to attach a reputation; typically, this is an individual person, but it could also be an organized group or a firm, or a node in a computer network.) The records are based on both objective information independently collected about interactions and feedback from the entities about each other. The exact nature of both the objective information and the subjective feedback depends on the application. For an online auction, the system may record the agreed sale price and ask the buyer and seller to report their satisfaction with each other's integrity and performance after a trade. In a peer-to-peer system, we might ask each peer to monitor and report how often another peer makes its system available.

In principle, user A's reputation could simply be a concatenation of all records pertaining to A, but in practice, reputations are usually numerical summary values that permit direct comparison between users. Thus, reputation systems include an internal aggregation procedure to convert the reports to reputations. If all reports conform to a

common structure, there are two natural dimensions along which to aggregate reports: (1) Aggregating across users by computing a statistic of all other users' reports about A. (2) Aggregation across time by computing a statistic of all past reports. In addition, the aggregation function may use other structure derived from the reports, or from the reputations themselves. In particular, it often relies on some notion of *transitivity of trust*, in the sense that reports from users with high reputation are weighted more heavily than reports from users with low reputation.

Economists have studied models where entities strategically choose actions with an eye toward the histories they will generate. In these models, the link between actions and outcomes is probabilistic (bad actions sometimes lead to good outcomes and vice versa) or outcomes are observed with some error. The analysis of these models is interesting and complex, but beyond the scope of this chapter.

Rather than threats to the informativeness of a user history, we focus our attention on threats to the reputation system itself, the system that collects histories and associates them with entities. When the histories include subjective feedback, that feedback may not be reported or may not be reported honestly. Histories may even include phantom feedback from nonexistent interactions.

A second vulnerability comes from the fact that histories may not be tied directly to entities, but rather to online pseudonyms. In many systems, pseudonyms are cheap, which lead to two threats: an entity may jettison its pseudonym if it accumulates a bad reputation, and an entity may acquire many pseudonyms and have them rate each other favorably in order to inflate their reputations.

To summarize, we consider three threats to the integrity of reputation systems:

(i) *Whitewashing*. An entity may acquire a new pseudonym and start over with a clear reputation.

(ii) *Incorrectly reported feedback*. Entities may not report feedback or may not report it honestly.

(iii) *Phantom feedback*. An entity may provide feedback for interactions that never took place, perhaps using "sock puppet" identities (or sybils) created for the sole purpose of providing such phantom feedback.

We begin in Section 27.2 with a stylized model of interactions over time in a market. Initially, in Section 27.3, we assume that the available objective data about interactions are sufficient to generate informative histories, even without any reporting of subjective feedback. We consider the threat of *whitewashing*, where an entity can start over with a new pseudonym, which will not be linked to the history of actions taken under the previous pseudonym. Reputations can still create an incentive for good behavior, but only if a pseudonym with no history is forced to "pay its dues" in some fashion while it builds up a history of good actions.

Section 27.4 relaxes the assumption of objective data about actions. Feedback about interactions may not be reported correctly. Entities may not report feedback or may not report it honestly, for a variety of reasons, including fear of retaliation, or a desire to be viewed as a nice or skilled evaluator.

One approach is to treat the reporting of feedback about an action as itself an action in some other domain. A history of feedback reports made by an entity can be generated and, suitably aggregated, becomes an entity's reputation as a rater. Just as

in any reputation system, rater reputations can deter moral hazard, creating incentives for effort and honest reporting. It may, however, be difficult to assess the quality of subjectively reported feedback. We present a mechanism that does so by comparing it with other subjectively reported feedback.

Section 27.5 takes a second approach. Rather than directly assessing the quality of subjectively provided feedback, it assumes that an entity's reputation as a rater is the same as its reputation as an actor in the original domain. This leads to a notion of transitive trust: if an entity's actions in the original domain lead it to have a positive reputation, the entity is presumed to be a good rater as well, and its ratings are treated as more credible and weighted more highly in computing the reputations of other entities. For example, positive feedback from an eBay member with a good reputation would count more than positive feedback from a member with a bad reputation. This naturally leads to a graph model that represents entities and their feedback about other entities, with actions in the original domain not represented explicitly. Reputations are computed as scores for nodes in the graph, subject to the constraints imposed by the link structure of feedback among entities. We present both possibility and impossibility results on how transitive trust algorithms can handle the threats of incorrectly reported feedback and the problem of phantom feedback from sock puppet entities, the so-called *sybil attack*.

27.2 The Effect of Reputations

Economists have developed many game-theoretic models of the impact of reputations. In this section we present some of the fundamental ideas and technical tools necessary. We begin with an (over)simplified example.

Consider the "prisoners' dilemma," a classic model from the early days of game theory. There are two agents, Alice (A) and Bob (B), who interact. If both agents cooperate (C) then each gains 1 unit of utility, while if they both defect they gain 0; however if one cooperates and the other defects (D), the defector gains 2 and the cooperator loses 1. We summarize this as $\pi_A(C, C) = 1$, $\pi_A(D, D) = 0$, $\pi_A(D, C) = 2$, and $\pi_A(C, D) = -1$. π_B is similarly defined via symmetry.

Clearly the outcome of this game, when played a single time, should be (D, D) since it is a dominant strategy for both agents. In an infinitely repeated game, however, a player may choose C and accept lower payoffs in one round to increase the probability that partners will play C against her in future stage games, and thus increase her future payoffs. We denote the game played in each round as the stage game for that round.

Define the discounted payoff to player i in stage game t to be $\pi_i^t \delta^t$, where π^t is the actual payoff in round t and $0 \leq \delta < 1$ is the discount factor. The idea of a discount factor is that it is somehow preferable to get a payoff in the current round rather than in the next round. If the payoffs are monetary, the possibility of investing the payoff at some interest rate provides a good intuition for why a discount factor is needed.

We will analyze strategy alternatives that consist of decision rules about which action to play in each stage game, contingent on a player's own history and the histories of all other players. The discounted average payoff of a strategy, played infinitely into the

future, is defined as

$$\overline{\pi_i} = (1-\delta)\sum_{t=0}^{\infty}\delta^t\pi_i^t.$$

In this infinitely repeated model, consider the Grim strategy: play C unless any player has played D in a previous round. This strategy, pursued by both players, denoted (Grim, Grim), is a Subgame Perfect Nash Equilibrium (SPNE), meaning that, if all players pursue this strategy, there is no stage game at which any player would want to deviate from the strategy.

To prove this is a SPNE, we only need to consider "single deviations" in which an agent only deviates from Grim once and then returns to playing it. This follows from a generalization of the single deviation property in dynamic programming.

Consider a deviation in which Alice plays D in a round 0. Clearly this will lead to (D,D) in all future rounds for Alice (for everyone, in fact), so Alice's discounted average payoff will be $(1-\delta)(2+\delta*0+\delta^2*0+\cdots) = 2(1-\delta)$; however, if she did not deviate, then her payoff would be 1 in every period leading to $(1-\delta)(1+\delta+\delta^2+\cdots) = 1$. Thus, deviating is not advantageous when $1 \geq 2(1-\delta)$ or, equivalently, $\delta \geq 1/2$. Now, this same argument applies to any period $t > 0$ with both sides of the equations multiplied by δ^t.

Thus, when δ is small, the promise of future payoffs is not sufficient to constrain the player's current behavior. This is true in all reputation systems: if the players do not value future payoffs sufficiently, then reputations are of no value.

Other strategies that are "less grim" can also work. For example, punishing for only a small number of periods can lead to a cooperative equilibrium for higher values of δ.

Now consider a group of $N+1$ players with N odd, in which in each round players are paired up at random and play the prisoners' dilemma. In a simple reputational extension of the above analysis we consider reputational-grim, defined as follows: each agent begins with a "good" reputation and keeps it if she plays C against players with good reputations and D against those with bad ones. This reputational-grim strategy, if played by all players, is also an SPNE, for $\delta \geq 1/2$. This is because, from an defector's perspective, the punishments are the same as in the full Grim strategy.

To understand the value of shared reputations, consider an alternative system where a player remembers others' interactions with her but histories are not publicly shared. A natural strategy is to play personalized-Grim, the variant of Grim where a player views the game as being separated into N unrelated games, one with each opponent. In this case, the expected number of rounds between meeting the same opponent is N so a straightforward calculation (see exercises) yields a condition for this to be an SPNE, $\delta \geq 1 - 1/2N$, which is unreasonably close to 1, for large N.

The analysis above applies to situations in which all players have the same ability, but reputations lead them to strategies where they are reliable partners. To operationalize varying player abilities, models allow different players different action sets to choose from in the stage game. For example, a low-ability player might only have action D available (or perhaps in some percentage of stage games have only action D available). A high-ability, honest type might only have action C available. Alternatively, it might take more effort (cost more) for a low type to play C than for the high type. This could arise where C indicates the completion of a high-quality product. (Player types with

only one possible action are called "commitment" types in the economics literature.) Players with both types of action available (called "strategic" types in the economics literature) would then want to choose actions that distinguish them from low-ability players and mimic those of high-ability players.

It is also natural to extend the model to situations in which outcomes are only probabilistically linked to actions, or outcomes are reported with random error. This leads to interesting strategic opportunities, including playing C most of the time but sometimes choosing D, which would not be immediately distinguishable from the actions of high-ability honest types who also have bad outcomes only less frequently. The analysis of these models is interesting and complex, but beyond the scope of this chapter. (However, in the following section we will consider random outcomes in a limited way.)

27.3 Whitewashing

One key issue in online reputation systems is the fragility of identity. Agents with bad reputations simply reregister with a new username. This is known as *whitewashing*. It is easy to see that the ability to whitewash will disable the functioning of the reputation systems as described in Section 27.2, as agents will simply choose D and then return with a new identity in the following round.

To prevent this, there needs to be some "initiation fee" upon entry. For example, simply having an upfront cost of f to register will prevent whitewashing as long as the cost is sufficiently high. To compute this f note that the total discounted payoff for deviating once is $\overline{\pi}' = (1-\delta)(2 - f + \delta(1-f) + \delta^2 + \delta^3 \cdots)$ while following reputational grim obtains $\overline{\pi} = (1-\delta)(1 - f + \delta + \delta^2 + \cdots)$. Thus for an SPNE we need $\overline{\pi} \geq \overline{\pi}'$, which implies that $\delta f \geq 1$ or $f \geq 1/\delta$. (Note that we continue to require that $\delta \geq 1/2$ to prevent deviation without whitewashing.)

Unfortunately collecting fees is not always feasible (or politically viable); however, we can create an explicit reputational fee. The key idea is to force the new arrivals to "pay dues" upon arrival. The most efficient way to do this is to allow veterans to defect against newcomers, where newcomers are playing for the first time (apparently) and veterans have played at least once before. Thus, we can define the pay-your-dues (PYD) strategy as: play C against any veteran who has never deviated from PYD, otherwise play D against the veteran. Play D against a newcomer, unless you are a newcomer too, in which case play C.

Intuitively, this leads to the "socially most efficient" SPNE, where social efficiency measures the sums of all players' payoffs. Note, however, that the social efficiency in this equilibrium is less than the maximum social efficiency that could be attained without whitewashing. This follows because the maximum social welfare in a single pair playing the PD is 2 while choosing (D, C) yields a value of $2 - 1 = 1$. (One might consider requiring that newcomers play D against other newcomers, but this obtains a value of 0 and entails further social loss.) Thus, the possibility of whitewashing leads to an unavoidable cost being imposed on society.

Even allowing for whitewashing, PYD leads to an SPNE where every player's average discounted payoff is 1. (You should verify this as in the exercises.) However, we have left out several important details in this model that we discuss in the next section!

27.3.1 A More Dynamic Model

Stepping back, we see that the model we just analyzed has a flaw, since any newcomers in our model are clearly whitewashers. Thus, for that model, always playing D against an agent who arrived after the first period (and personalized-grim otherwise) yields a fully socially efficient SPNE, since (C, C) is played in every interaction.

Thus, it makes sense to extend our model to capture these issues; although the difficulty is retaining tractability. First, we assume αN real newcomers arrive every period and an equal number of veterans depart, where the departing veterans are chosen at random. However, once again this allows us to easily detect whitewashers—if there are more than αN newcomers in any period then the players know that there must be at least one whitewasher. Thus, there is an equilibrium in which players play PYD as long as there are exactly αN newcomers in any period and play D-always if there are ever more. However, it is clear that this equilibrium is extremely fragile, since a single deviation leads to all players defecting forever. Such fragile equilibria are artifacts of the "noiselessness" of the game and the perfect rationality assumptions inherent in game theory.

To make our model more robust, we add some "noise." We assume that in any play of the stage game a player accidently plays D with probability $\epsilon > 0$ and then returns in the following period as a whitewasher. In this model, one can show that PYD leads to the most efficient equilibrium (i.e., the highest fraction of cooperative outcomes (C, C)). Proving that PYD is an equilibrium is intuitively similar to above proofs with the addition of some ideas from dynamic programming, while proving optimality is more difficult and requires a careful stochastic analysis.

The PYD strategy in this stylized model corresponds in more practical settings to a mistrust of newcomers. Until they have proven themselves, veterans do not trust the newcomers sufficiently to allow them to undertake mutually beneficial interactions. If only the veterans could trust the newcomers, the newcomers could start right away to interact in beneficial ways with the veterans. The threat of whitewashing, however, forces a mistrust of newcomers. Because of the threat of whitewashing, in any equilibrium, newcomers must also be penalized at least the amount that a deviator would be penalized.

The only way to improve the treatment of newcomers in an equilibrium with significant cooperation is to make whitewashing difficult, by making it more difficult or expensive for existing participants to get new pseudonyms than it is for newcomers. For example, the organization running the reputation system might require entities to reveal their true names, offer them one free pseudonym, and then restrict the acquisition of addition ones or require a payment for them.

27.4 Eliciting Effort and Honest Feedback

The previous section described models in which feedback was reported automatically and objectively. Any system that actually solicits individual opinions must overcome two challenges. The first is underprovision. Forming and reporting an opinion requires time and effort, yet the information benefits others. The second challenge is honesty. A

desire to be nice, or fear of retaliation, may cause a rater to withhold negative feedback. Conflicts of interest or a desire to improve others' perception of them may lead raters to report distorted versions of their true opinions.

An explicit reward system for honest rating and effort may help overcome these challenges. When objective information will be publicly revealed at a future time, individuals' reports can be compared to that objective information. For example, weather forecasts and sports betting odds can be compared to what actually occurs. See Chapter 26 on information markets for algorithms that create incentives for honest revelation of information in such settings.

Here, we develop methods to elicit feedback effectively when independent, objective outcomes are not available. Examples include situations where no objective outcome exists (e.g., evaluations of a product's "quality"), and where the relevant information is objective but not public (e.g., a product's breakdown frequency, which is available to others only if the product's current owners reveal it).

In these situations, one solution is to compare raters' reports to their peers' reports and reward agreement.[1] However, if rewards are made part of the process, dangers arise. If a particular outcome is highly likely, such as a positive experience with a seller at eBay who has a stellar feedback history, then a rater who has a bad experience will still believe that the next rater is likely to have a good experience. If she were to be rewarded simply for agreeing with her peers, she will not report her bad experience. This phenomenon is akin to the problems of herding or information cascades.

We now describe a formal mechanism, the *peer-prediction method*, to implement the process of comparing with peers. The scheme uses one rater's report to update a probability distribution for the report of someone else, whom we refer to as the reference rater. The first rater is then scored not on agreement between the ratings, but on a comparison between the *probabilities* assigned to the reference rater's possible ratings and the reference rater's actual rating. Raters need not perform any complex computations: so long as a rater trusts that the system will update appropriately, she will prefer to report honestly.

Scores can be turned into monetary incentives, either as direct payments or as discounts on future merchandise purchases. In many online systems, however, raters seem to be quite motivated by prestige or privileges within the system. For example, at Slashdot.org, users accumulate "karma" points for various actions and higher karma entitles users to rate others' postings and to have their own postings begin with higher ratings; at ePinions.com, reviewers gain status and have their reviews highlighted if they accumulate points. Similarly, offline point systems that do not provide any tangible reward seem to motivate chess and bridge players to compete harder and more frequently.

[1] Subjective evaluations of ratings could be elicited directly instead of relying on correlations between ratings. For example, the news and commentary site Slashdot.org allows meta-moderators to rate the ratings of comments given by regular moderators. Meta-evaluation incurs an obvious inefficiency, since the effort to rate evaluations could presumably be put to better use in rating comments or other products that are a site's primary product of interest. Moreover, meta-evaluation merely pushes the problem of motivating effort and honest reporting up one level, to ratings of evaluations. Thus, scoring evaluations in comparison to other evaluations may be preferable in certain settings.

27.4.1 A Model

We now consider a model to analyze these issues. A number of raters experience a product and then rate its quality. The product's quality does not vary, but is observed with some idiosyncratic error. After experiencing the product, each rater sends a message to a common processing facility called the center. The center makes transfers to each rater, awarding or taking away points based on the raters' messages. The center has no independent information, so its scoring decisions can depend only on the information provided by other raters. As noted above, points may be convertible to money, discounts or privileges within the system, or merely to prestige. We assume that raters' utilities are linear in points. We also assume that raters are risk neutral, and hence, seek to maximize expected wealth.

We refer to a product's quality as its *type*. Assume the number of product types is finite, and the types are indexed by $t = 1, \ldots, T$. Furthermore, we assume that there is a commonly known prior probability. Let $\Pr_0(t)$ be the commonly held prior probability assigned to the product's being type t. Assume that $\Pr_0(t) > 0$ for all t and $\sum_{t=1}^{T} \Pr_0(t) = 1$.

Let I be the set of raters, where $|I| \geq 3$. I may be (countably) infinite. Each rater has a perception of a product's type, which is called her *signal*. Each rater privately observes her own signal; she does not know any other rater's signal. Let $S = \{s_1, \ldots, s_M\}$ be the set of possible signals, and let S^i denote the random signal received by rater i. Conditional on the product's type, raters' signals are independent and identically distributed; the distribution is represented by function $f(s_m|t) = \Pr(S^i = s_m|t)$, where $f(s_m|t) > 0$ for all s_m and t, and $\sum_{m=1}^{M} f(s_m|t) = 1$ for all t. We assume that this function $f(s_m|t)$ is common knowledge. Furthermore, we assume that the conditional distribution of signals is different for different values of t, so that the signals are informative about the types.

Throughout this section, we use the following simple example as an illustration. There are only two product types, H and L, with prior $\Pr_0(H) = \Pr_0(L) = 0.5$, and two possible signals, h and l. The distribution of the signals, conditioned on the true type, is as follows: $f(h|H) = .85$, $f(l|H) = 0.15$, $f(h|L) = 0.45$, $f(l|L) = 0.55$. Thus, $\Pr(h) = 0.5 * 0.85 + 0.5 * 0.45 = 0.65$.

In the mechanism we propose, the center asks each rater to announce her signal. After all signals are announced to the center, they are revealed to the other raters and the center computes transfers. We refer to this as the *simultaneous reporting game*. Let $x^i \in S$ denote one such announcement, and $x = (x^1, \ldots, x^I)$ denote a vector of announcements, one by each rater. Let $x^i_m \in S$ denote rater i's announcement when her signal is s_m, and $\bar{x}^i = (x^i_1, \ldots, x^i_M) \in S^M$ denote rater i's announcement strategy. Let $\bar{x} = (\bar{x}^1, \ldots, \bar{x}^I)$ denote a vector of announcement strategies. As is customary, let the superscript "$-i$" denote a vector without rater i's component.

Let $\tau_i(x)$ denote the transfer paid to rater i when the raters make announcements x, and let $\tau(x) = (\tau_1(x), \ldots, \tau_I(x))$ be the vector of transfers made to all agents. An announcement strategy $\bar{x}^i$ is a best response to $\bar{x}^{-i}$ for player i if for each m:

$$\forall \hat{x}^i \in S \; E_{S^{-i}}\left[\tau_i\left(\bar{x}^i_m, \bar{x}^{-i}\right)|S^i = s_m\right] \geq E_{S^{-i}}\left[\tau_i(\hat{x}^i, \bar{x}^{-i})|S^i = s_m\right]. \quad (27.1)$$

That is, a strategy is a best response if, conditional on receiving signal s_m, the announcement specified by the strategy maximizes that rater's expected transfer, where the expectation is taken with respect to the distribution of all other raters' signals conditional on $S^i = s_m$. Given transfer scheme $\tau(x)$, a vector of announcement strategies $\bar{x}$ is a Nash Equilibrium of the reporting game if (27.1) holds for $i = 1, \ldots, I$, and a strict Nash Equilibrium if the inequality in (27.1) is strict for all $i = 1, \ldots, I$.

Truthful revelation is a Nash Equilibrium of the reporting game if (27.1) holds for all i when $x^i_m = s_m$ for all i and all m, Furthermore, truthful revelation is a strict Nash Equilibrium if the inequality is strict. (In other words, if all the other players announce truthfully, truthful announcement is a strict best response.)

Continuing the two-type, two-signal example, suppose that rater i receives the signal l. Recall that $\Pr_0(H) = 0.5$, $f(h|H) = 0.85$, and $f(h|L) = 0.45$, so that $\Pr(s^i_l) = 0.35$. Given i's signal, the probability that rater j will receive a signal h is

$$\begin{aligned}\Pr(S^j = h|S^i = l) &= f(h|H)\frac{f(l|H)\Pr_0(H)}{\Pr(S^i = l)} + f(h|L)\frac{f(l|L)\Pr_0(L)}{\Pr(S^i = l)} \\ &= 0.85\frac{0.15 * 0.5}{0.35} + 0.45\frac{0.55 * 0.5}{0.35} \cong 0.54.\end{aligned}$$

If i had instead observed h, then:

$$\begin{aligned}\Pr(S^j = h|S^i = h) &= f(h|H)\frac{f(h|H)\Pr_0(H)}{\Pr(S^i = h)} + f(h|L)\frac{f(h|L)\Pr_0(L)}{\Pr(S^i = h)} \\ &= 0.85\frac{0.85 * 0.5}{0.65} + 0.45\frac{0.45 * 0.5}{0.65} \cong 0.71.\end{aligned}$$

27.4.2 Peer-Prediction Scoring

We now describe how to assign points to a rater i, based on her report and that of another player j. A *scoring rule* is a function $\mathcal{T}(s|x^i)$ that, for each possible announcement x^i of S^i, assigns a score to each possible value $s \in S$. We cannot directly access the signal s_j, but in a truthful equilibrium, we can use player j's report.

Definition 27.1 A scoring rule is *strictly proper* if the rater maximizes her expected score by announcing her true beliefs.

The literature contains a number of strictly proper scoring rules for eliciting beliefs about the probability of an event. The score can be positive or negative. For example, one proper scoring rule, the *logarithmic scoring rule*, is to penalize the player the log of the probability she assigned to the event that actually occurred. Suppose that there are only two possible events (*h,l*), and a player is asked to report her belief $\hat{p}$ of the probability of event h. The log scoring rule is defined by $\mathcal{T}(h|\hat{p}) = \ln(\hat{p})$, $\mathcal{T}(l|\hat{p}) = \ln(1 - \hat{p})$. If her true belief is that h occurs with probability p, then the expected value of announcement $\hat{p}$ is $p \ln \hat{p} + (1 - p)\ln(1 - \hat{p})$. Setting the first derivative to 0 gives the first-order condition for maximization, which requires $p = \hat{p}$.

In the peer prediction method, for each player we choose a reference rater $r(i)$. The outcome to be predicted is the reference rater's announcement $x^{r(i)}$. Player i does not directly report a probability distribution over the reference rater's report: it is inferred

from her own report and the prior probability distribution. Truthful reporting is still a best response if she believes that the reference rater will report honestly.

We write $\mathcal{T}(x^{r(i)}|x^i)$ for $\ln[\Pr_0(S^{r(i)} = x^{r(i)}|S^i = x^i)]$, i.e., the log of the inferred probability that $r(i)$ will see $x^{r(i)}$ given that S^i sees signal x^i. Then, let

$$\tau_i^*\left(x^i, x^{r(i)}\right) = \mathcal{T}\left(x^{r(i)}|x^i\right). \tag{27.2}$$

Proposition 1 *For any mapping r that that assigns to each rater i a reference rater $r(i) \neq i$, truthful reporting is a strict Nash equilibrium of the simultaneous reporting game with transfers τ_i^*.*

PROOF Assume that rater $r(i)$ reports honestly: $x^{r(i)}(s_m) = s_m$ for all m. Since S^i is stochastically informative for $S^{r(i)}$, and $r(i)$ reports honestly, S^i is stochastically informative for $r(i)$'s report as well. For any $S^i = s^*$, player i chooses $x^i \in S$ to maximize

$$\sum_{n=1}^{M} \mathcal{T}\left(s_n^{r(i)}|x^i\right) Pr\left(S_{r(i)} = s_n|S_i = s^*\right). \tag{27.3}$$

Since $\mathcal{T}(\cdot|\cdot)$ is a strictly proper scoring rule, (27.3) is uniquely maximized by announcing $x^i = s^*$. Thus, given that rater $r(i)$ is truthful, rater i's best response is to be truthful as well. □

Since $0 < \Pr(S_{r(i)} = s_n|S_i = s^*) < 1$, $\ln(\Pr(S_{r(i)} = s_n|S_i = s^*)) < 0$; we refer to τ_i^* as rater i's penalty since it is always negative in this case. (By adding a suitably large constant that depends only on the distribution f, it is in principle possible to convert this to a positive score without altering its strategic properties.)

Consider the simple example where rater i received the relatively unlikely signal l ($\Pr(S^i = l) = 0.35$). Even contingent on observing l it is unlikely that rater j will also receive an l signal ($\Pr(S^j = l|S^i = l) = 1 - 0.54 = 0.46$). Thus, if rater i were rewarded merely for matching her report to that of rater j, she would prefer to report h. With the log scoring rule, an honest report of l leads to an expected payoff

$$\begin{aligned}&\ln[\Pr(S^j = h|S^i = l)]\Pr(S^j = h|S^i = l) + \ln[\Pr(S^j = l|S^i = l)]\Pr(S^j = l|S^i = l)\\ &\quad = \ln(0.54)0.54 + \ln(0.46)0.46 = -0.69.\end{aligned}$$

If, instead, she reports h, rater i's expected score is

$$\begin{aligned}&\ln[\Pr(S^j = h|S^i = h)]\Pr(S^j = h|S^i = l) + \ln[\Pr(S^j = l|S^i = h)]\Pr(S^j = l|S^i = l)\\ &\quad = \ln(0.71)0.54 + \ln(0.29)0.46 = -0.75.\end{aligned}$$

As claimed, the expected score is maximized by honest reporting.

The key idea is that the scoring function is based on the updated beliefs about the reference rater's signal, given the rater's report, not simply matching a rater's report to the reference report. The updating takes into account both the priors and the reported signal, and thus reflects the initial rater's priors. Thus, she has no reason to shade her report toward the signal expected from the priors. Note also that she need not perform any complex Bayesian updating. She merely reports her signal. As long as she trusts the

center to correctly perform the updating and believes other raters will report honestly, she can be confident that honest reporting is her best action.

Note that while Proposition 1 establishes that there is a truthful equilibrium, it is not unique, and there may be nontruthful equilibria. To illustrate, in the example we have been considering two other equilibria are (1) report h all the time, and (2) report l all the time.[2] While such nontruthful equilibria exist, it is reasonable to think that the truthful equilibrium will be a focal point, especially when communication among raters is limited, or when some raters are known to have a strong ethical preference for honesty. In addition, the center can punish all the raters if it detects a completely uninformative equilibrium such as all h or all l.

A variety of extensions to this base scoring rule have been studied. For example, adding a constant value to the score increases the expected payoff without changing the incentives for honest revelation. Multiplying the score by a constant preserves the incentive for honest revelation but changes the amount of costly effort a rater will want to exert in order to acquire an informative signal. The points that each person earns can be debited from some other participant, so that all scores are settled through transfer payments rather than subsidies from the center. Alternative proper scoring rules to reduce the expected size of payments have also been studied.

The payments can be adapted to a sequential interaction scenario where each rater sees the previous rater's reports before reporting herself. Each rater is scored on the basis of the probability distribution inferred from the common prior beliefs, her own report, *and previous reports*. Since the center will take into account others' reports automatically, it is optimal to report just her own signal.

The most problematic aspect of the scoring mechanism is its reliance on common prior beliefs about the distribution of types and the distribution of signals contingent on types. These are needed to infer from a user's reported signal x_i the probability distribution R for the reference rater's signal, which is used to determine the user's point score. A seemingly attractive alternative is to elicit R directly, but player i may also be a reference rater for some other player, and so x_i must be truthfully elicited to score that other player.

The requirement of common priors can be relaxed somewhat if each player is asked to report her personal priors about the item's type before receiving her information signal about the item, and then to report her signal once she receives it. There still is a requirement of common beliefs about the distribution of signals contingent on types, in order to perform Bayesian updating correctly. One solution would be to define the types empirically according to the distribution of signals they elicit (e.g., type 1 yields 10% h signals; type 2 yields 20%, etc.) Then, the beliefs about distribution of signals contingent on type would, by construction, be commonly held.

Many open questions remain about the peer-prediction method. Can it be extended to situations in which raters vary in their abilities and scores are used both to assess the credibility of raters and to give them incentives for effort and honest reporting? Can the method be extended to situations in which entities choose their interactions partners

[2] To verify the "always play h equilibrium," note that if the reference rater always reports high, the rater expects $\ln(0.54)1 + \ln(0.46)0 = -0.616\,19$ if she reports l, and $\ln(0.71)1 + \ln(0.29)0 = -0.342\,49$ if she reports h. Similar reasoning verifies the "always play l equilibrium."

rather than being randomly matched? Can it be made robust to collusion among entities or sybil attacks with fake entities providing confirmatory ratings?

27.5 Reputations Based on Transitive Trust

In this section, we discuss the *transitive trust* approach to dealing with the lack of objective feedback. The foundation of this approach is the postulate that the *credibility of an agent's feedback* is tied to the *credibility of her non-feedback actions*. This assumption enables the construction of reputation systems in the absence of any external signals of interaction outcomes or feedback quality: an entity's reputation is calculated by weighting ratings of the entity according to the raters' credibilities, which are in turn calculated from those raters' reputations. Thus, if we begin with some set of credible agents, we can potentially grow this set transitively: If the currently credible agents have positive feedback about i, i can be included in the set of credible agents. This is a recursive construction; we need to carefully define how to bootstrap the credibility calculation, how to propagate the credibility through the network, and when to terminate the calculation.

One additional simplification is often employed in reputation algorithms, which is to ignore the temporal order in which feedback is received. Now, the feedback can be succinctly expressed in graphical form: At a given point of time, let $t(ij)$ denote the summary feedback (*trust*) that i reports about j, based on interactions between them thus far. We assume that the trust can be expressed as a nonnegative real value. Then, the input to the reputation system can be viewed as a "trust graph" $G = (V, E, t)$, where V is the set of agents, E the set of directed edges, and $t : E \to \Re^+ \setminus \{0\}$ the weights. (Note that typically the graph will be quite sparse, so for algorithmic considerations we explicitly include E.)

We assume that the reputations computed by our system are numeric values. Then, the reputation aggregation mechanism can be represented as a function from a trust graph to a set of reputation values, $F : G \to \Re^{|V|}$, where $F_v(G)$ is the reputation value of vertex $v \in V$. The reputation values determine an ordering or ranking of the nodes. A reputation function is trivial if the ranking induced by $F(G)$ is constant over all G; we restrict our attention to nontrivial reputation functions.

This model captures the many reputation systems that have been proposed or used in practice. One important example is PageRank, a mechanism used by Google to rank Web pages. In this case $v \in V$ is a Web page, $(v, w) \in E$ is a directed edge showing that Web page v has a hyperlink to page w and $t(v, w) = 1/Out(v)$, where $Out(v)$ is the outdegree of v. In a peer-to-peer system, $v \in V$ is a peer, $(v, w) \in E$ is a directed edge showing that peer v has interacted with w and $t(v, w)$ represents the degree of trust that v has in w, which can depend on the number, type, and outcomes of v's interactions with w.

There are numerous ways in which the reputations can be computed from the trust graph. We consider a simple version of PageRank, in which the ranking function is given by

$$F_v(G) = \epsilon + (1 - \epsilon) \sum_{v' | (v', v) \in E} F_{v'}(G) t(v', v).$$

Another interesting aggregation function, used in the Advogato system, is the max-flow algorithm, where $F_v(G)$ is the maximum flow from some start node $v_0 \in V$ to v. In the P2P setting it is natural to create personalized reputation functions where each node uses itself as the start node. In the web ranking setting one can simply choose one (or several) "trusted" nodes as the start nodes. Lastly, for comparison, we consider the Pathrank algorithm where $F_v(G)$ is the shortest path from some start node $v_0 \in V$ to v, where the length of an edge is simply the inverse of the trust value.

A reputation system is *monotonic* if adding an incoming edge to v never reduces the rank of v relative to any other node w, i.e., for $E' = E \cup \{uv\}$, $F_v(V, E) > F_w(V, E) \Rightarrow F_v(V, E') > F_w(V, E')$ and $F_v(V, E) \geq F_w(V, E) \Rightarrow F_v(V, E') \geq F_w(V, E')$. All the reputation schemes described above are monotonic. A reputation system is *symmetric* if the function F commutes with permutation of the node names, i.e., the reputations depend only on the graph structure, and not on the labels of the nodes. The simple variant of PageRank described above is symmetric, but the other reputation functions are not: the start node v_0 enjoys a privileged position.

27.5.1 Incentives for Honest Reporting

With the transitive trust model, the incentive problems are particularly acute. Entities are not rewarded or penalized directly for the quality of the ratings they provide, only for the ratings they receive from others. Thus, an entity has no incentive to provide informative feedback. Furthermore, depending on the reputation function F, she may have a strong incentive to provide *incorrect* feedback, so as to influence the credibility of other agents' feedback about herself.

Therefore, we would like a reputation function F in which an agent v cannot strategically choose feedback to boost her own standing. Define a reputation system as *rank-strategyproof* if, for every graph G and every agent $v \in V$, agent v cannot boost her rank ordering by strategic choices of how she rates other agents. This formulation allows an agent to manipulate its own or others' reputation scores as long as it is unable to improve its position in the rank ordering of reputation scores.

It turns out that rank-strategyproofness is very difficult to achieve in symmetric reputation systems: Any nontrivial, monotonic reputation system that is symmetric cannot be rank-strategyproof. For example, in the PageRank ranking system, a node v may be able to improve her rank by dropping an outgoing edge vu to a higher-ranked node u, thereby reducing u's reputation. We refer readers to the references at the end of this chapter for additional results in this vein. We note that this impossibility result does not apply to nonsymmetric reputation systems; the Pathrank function satisfies both the rank-strategyproofness and monotonicity properties.

27.5.2 Sybils and Sybilproofness

Next, we consider robustness to another attack on reputation systems: *sybil attacks*. In a sybil attack, a single agent creates many fake online identities to boost the reputation of its primary online identity. Formally, we assume that a node can create any number of sybil nodes, with any set of trust values between them. In addition, we allow the node to divide incoming trust edges among the sybils in any way that preserves the

total trust, $\sum_{v'|(v',v)\in V} t(v', v)$, and manipulate the outgoing trust links in any manner it chooses. Note that many other formulations are possible depending on the specific system being modeled. Most of the results we discuss below hold in many of the other possible formulations.

Definition 27.2 Given a graph $G = (V, E, t)$ and a user $v \in V$, we say that a graph $G' = (V', E', t')$ along with a subset $U' \subseteq V'$ is a **sybil strategy** for user v in the network $G = (V, E, t)$ if $v \in U'$ and collapsing U' into a single node with label v in G' yields G. We can refer to U' as the sybils of v, and denote a sybil strategy by (G', U').

We define two different notions of sybilproofness for reputation functions.

Definition 27.3 A reputation function F is value-sybilproof if for all graphs $G = (V, E)$, and all users $v \in V$, there is no sybil strategy for v, (G', U'), with $G' = (V', E')$ such that for some $u \in U'$, $F_u(G') > F_v(G)$.

Definition 27.4 A reputation function F is rank-sybilproof if for all graphs $G = (V, E)$, and all users $v \in V$, there is no Sybil strategy (G', U') for v (with $G' = (V', E')$) such that, for some $u \in U'$ and $w \in V \setminus \{v\}$, $F_u(G') \geq F_w(G')$ while $F_v(G) < F_w(G)$.

Theorem 27.5 *There is no (nontrivial) symmetric rank-sybilproof reputation function.*

PROOF Given a graph $G = (V, E)$ and reputation function F, let $v, w \in V$ with $F_w(G) > F_v(G)$. Now consider the graph G', which is simply 2 disjoint copies of G, where U is the second copy of G combined with v. By symmetry, there is a node $u \in U$ such that $F_u(G') = F_w(G')$. Thus F is not rank-sybilproof. □

Note that this result does not require the assumption that F is monotonic. In fact, symmetric reputation functions cannot be sybilproof even for an attack with a single sybil.

Definition 27.6 We say that a reputation function is K**-rank-sybilproof** if it is rank-sybilproof for all possible sybil strategies (G', U'), with $|U'| \leq K + 1$.

Theorem 27.7 *There is no symmetric K-rank-sybilproof nontrivial reputation function for $K > 0$.*

PROOF Consider the graphs in the previous example, where $V = \{v = v_1, v_2, \ldots, v_r = w\}$ is the original vertex set and $U = \{u_1, u_2, \ldots, u_r\}$ is the duplicate; let $V' = V \cup U$. Define G^t to be the subgraph of G' with $V^t = V \cup \{u_1, \ldots, u_t\}$ and $G^0 = G$. Then $F_w(G^0) > F_v(G^0)$, while $F_{u_r}(G^r) = F_w(G^r)$ (where u_r is the copy of node $v_r = w$), so there must exist a

t such that $\max_{i\in\{v,u_1,\ldots,u_t\}} F_i(G^t) < F_w(G^t)$, but $\max_{i\in\{v,u_1,\ldots,u_{t+1}\}} F_i(G^{t+1}) \geq F_w(G^{t+1})$. Let m be the node in $\{v, u_1, \ldots, u_t\}$ that achieves the greatest reputation in G^{t+1}. Then either $F_m(G^{t+1}) \geq F_w(G^{t+1})$ or $F_{u_{t+1}}(G^{t+1}) \geq F_w(G^{t+1})$. It follows that the addition of node u_{t+1} is a successful sybil strategy for m in G^t. Hence, F is not 1-rank-sybilproof on all graphs. □

Now, consider PageRank. It is clearly symmetric—changing the labels on the nodes does not change the reputation values. This immediately implies that it is not rank-sybilproof.

One natural approach to overcoming this result is to break the symmetry of the reputation system by using a specific trusted node (or nodes) as a seed. However, care is still needed to achieve robustness against sybil attacks. Here, we consider two simple reputation functions that are provably sybil-resistant.

We first consider the max-flow based ranking mechanism. It is easy to show that it is value-sybilproof.

Theorem 27.8 *The max-flow based ranking mechanism is value-sybilproof.*

PROOF This follows directly from max-flow equals min-cut after noticing that all sybils of $v \in V$ must be on the same side of the cut as v and thus on the other side of the cut from the source s. Thus, no sybil can have a value higher than the min-cut which is equal to $F_v(G)$. □

However, the max-flow based ranking mechanism is not rank-sybilproof, as the example in Figure 27.2 shows. This is because while $v \in V$ cannot increase its own value, it can reduce the value of nodes for which it is on a max-flow path. Nonetheless, there do exist nontrivial rank-sybilproof algorithms. The Pathrank reputation mechanism is one example:

Theorem 27.9 *The Pathrank ranking mechanism is value-sybilproof and rank-sybilproof.*

PROOF It is value sybilproof since sybils cannot decrease the length of the shortest path. Rank-sybilproofness follows from the fact that the only time a node v can affect the value of another node w is if v is on the shortest path from s to w; however, in that case, we must have $F_v(G) > F_w(G)$. □

The basic property that flow-based mechanisms are value sybilproof but not rank-sybilproof can be generalized to include a wide variety generalized flow mechanisms,

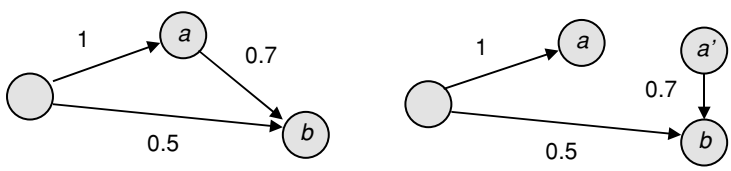

Figure 27.2. Node (a) improves its ranking by adding a sybil (a′) under max-flow.

such as those with "leaky pipes." Similarly, it can be shown that generalized path-based methods are value and rank-sybilproof and only path-based methods are rank-sybilproof in a large class of reputation mechanisms.

Lastly, we note that there are many open questions in this area. For example, while both PageRank and max-flow mechanisms are not rank sybilproof in the worst case, they are very useful reputation systems, and might be less manipulable on average. A precise formulation and analysis of this question is still open. For example, about half the pages on the Web could double their PageRank using only a single sybil.

27.6 Conclusion and Extensions

Reputations provide one of the most successful incentive mechanisms, and reputation systems are widespread on the Internet today. However, many reputation systems find themselves constantly under attack, and have to resort to fixing strategic problems after they are detected. In particular, many reputation systems are engaged in a constant arms race against attackers, where the systems change their ranking procedure and the attackers experiment until they find a weakness.

We believe that theoretical results on what can and cannot be accomplished by reputation systems, as well as provably secure system designs, would very useful. In this chapter, we have described three components of this theory; several other directions have been explored, and much research remains to be done.

27.6.1 Extensions and Open Problems

Distributed reputation systems. Up to this point, we have considered that users may strategically manipulate the feedback they provide or the identities they use, but we have implicitly assumed that they cannot directly manipulate the way in which the feedback is aggregated or the content of other users' feedback. This is a reasonable assumption as long as the users do not have any control over the communication medium or the server(s) used to compute the reputations. However, many proposed applications of reputation systems are settings, such as peer-to-peer applications or wireless ad hoc networks, in which these assumptions might be violated: there is no neutral trusted party to compute reputations, and users might be able to intercept each other's messages.

This has led many researchers to study *distributed* reputation systems in which the reputations are computed by the users themselves, but measures are adopted to minimize the risk of manipulation. One fundamental technique is to use *replication*: The same computation is performed at multiple nodes, and there are protocols to detect inconsistencies in the results. Similarly, if the users control portions of the communication network, it may be possible to send messages along multiple redundant paths so that no user can block or modify communication between two other users.

Much work remains to be done in this area. In particular, the redundancy technique is vulnerable to collusive attacks; the main design approach is to make these attacks difficult by requiring that a large number of users collude. This may be compromised by the existence of pseudonyms and sybil attacks.

Dynamic attacks. The basic model we have studied assumes that a user has full knowledge of which online identity she is interacting with. In some applications, it may be possible for users to claim credit for an interaction that another user executed, or to freeride by copying another user's actions. For example, if the contribution being measured is the number of puzzles a user solves, or the quality of ratings she gives to online articles, she may be able to garner a high reputation simply by copying another user.

On the other hand, dynamics may restrict the range of attacks in some settings. For example, in a P2P system a peer cannot divide incoming links among its sybils arbitrarily, since one needs an interaction to obtain a link and a low ranked sybil might have difficulty finding (nonsybil) partners.

Metrics and benchmarks. Strategic analysis of reputation systems often takes the form of proving robustness against attacks. While robustness against attacks is certainly desirable, we should not lose sight of the performance of the reputation system. In the extreme, a system in which everybody has zero reputation would be perfectly secure but completely useless. We need to develop metrics (or empirical benchmarks) of how well a particular aggregation method serves the users' information needs. One approach which has been taken is to formulate the performance in terms of an economic welfare measure, but a more direct formulation may be valuable.

Drawing on other social sciences. We have concentrated on economic and game theoretic approaches to reputation. Reputation has also been studied in sociology and social psychology, especially in the form of the broader, but clearly related, notion of *trust*. Insights from this literature are valuable in the design of reputation systems.

Putting it all together. The major challenge in reputation systems is to design a system that coherently puts together all the ideas that have been explored, including accurate feedback elicitation, robustness to whitewashing and sybil attacks, and distributed computation. This remains the key challenge for the reader!

27.7 Bibliographic notes

Below we provide pointers to relevant literature. Our list is meant to provide access to the literature and is certainly not comprehensive, i.e., for each topic we give one or two representative publications from which the reader can iterate the reference finding process.

Several chapters in this book extend our discussion, both providing a more detailed introduction to game theory, and discussing some examples on reputation systems. In particular, Chapter 23 on incentives in peer-to-peer systems includes a detailed discussion on the use of reputation systems in peer-to-peer environments.

There is a large literature on economic models of reputation. The following classic articles provide some foundations: Kreps and Wilson (1982), Milgrom and Roberts (1982), Fudenberg and Levine (1989), and Kandori (1992). Tadelis (1999) considers trading reputations, and shows that it is not always undesirable. Dellarocas (2001)

analyzes the economic efficiency of different feedback aggregation mechanisms. For broad overviews of this area, see Dellarocas (2003) and Resnick et al. (2000).

Our presentation of whitewashing follows Friedman and Resnick (2001). That paper includes a detailed proof that no equilibrium can yield substantially more cooperation than the Paying Your Dues equilibrium. Also see Lai et al. (2003), which introduced the term *whitewashing*.

Recently, the robustness of reputation systems to manipulation has attracted considerable research. The peer-prediction method to elicit honest feedback was originally described in an article by Miller et al. (2005). See Cooke (1991, p. 139) and Selten (1998) for a discussion of strictly proper scoring rules. Jurca and Faltings (2006) study modifications to the scoring rule to reduce the total expected payment. Bhattacharjee and Goel (2006) treat the revenues generated by a set of ratings as an objective indicator of the quality of the ratings. They provide an algorithm for dividing the revenues among raters in a way that creates incentives for entities to correct errors in the current community rating consensus.

Maintaining reputations for raters can provide signals about rater quality, in addition to incentives for good performance. Awerbuch and Kleinberg (2005) describe an algorithm that agents can use to learn who the good raters are. Their solution is robust to malicious as well as strategic attackers, provided that there are some altruistic raters who will rate accurately without incentives.

Many researchers have presented transitive-trust approaches to calculating reputations; a general framework using path algebras is described by Richardson et al. (2003). Altman and Tenneholtz (2006) study reputation systems from an axiomatic point of view, and present many possibility and impossibility results of the same flavor found in Section 27.5.1. Chien et al. (2003) prove that PageRank is monotonic. Our presentation of the sybilproofness of reputation systems follows Cheng and Friedman (2005). Many proposed solutions to the sybil attack implicitly or explicitly use the idea of a seed to break the symmetry of the reputations; for example, see Gyöngyi et al. (2004). The Advogato metric proposed by Levien (2004) also falls in this category. An alternative approach is described by Goel et al. (Zhang et al., 2004; Bhattacharjee and Goel, 2005).

Bibliography

A. Altman and M. Tennenholtz. Incentive compatible ranking systems, 2006. Available at: `http://www.technion.ac.il/ alon_a/0incentive.pdf`.

B. Awerbuch and R.D. Kleinberg. Competitive collaborative learning. In *18th Annual Conference on Learning Theory (COLT 2005)*, LNCS 3559:233–248. Springer, 2005.

R. Bhattacharjee and A. Goel. Avoiding ballot-stuffing in ebay-like reputation systems. In *P2PECON '05: Proc. 2005 ACM SIGCOMM Workshop on Economics of Peer-to-Peer Systems*, 2005.

R. Bhattacharjee and A. Goel. Incentive based ranking mechanisms. In *First Workshop on the Economics of Networked Systems (Netecon'06)*, pp. 62–68, 2006. Available at: `http://www.cs.duke.edu/nicl/netecon06/papers/proceedings.pdf`.

A. Cheng and E. Friedman. Sybilproof reputation mechanisms. In *P2PECON '05: Proc. 2005 ACM SIGCOMM Workshop on Economics of Peer-to-Peer Systems*, pp. 128–132, 2005.

S. Chien, C. Dwork, R. Kumar, D. Simon, and D. Sivakumar. Link evolution: Analysis and algorithms. *Internet Math.*, 1(3):277–304, 2003.

R. Cooke. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press, 1991.

C. Dellarocas. Analyzing the economic efficiency of ebay-like online reputation reporting mechanisms. In *Proc. 3rd ACM Conference on Electronic Commerce*, 2001.

C. Dellarocas. The digitization of word-of-mouth: Promise and challenges of online feedback mechanisms. *Management Sci.*, 49(10):1407–1424, 2003.

D. Fudenberg and D. Levine. Reputation and equilibrium selection in games with a patient player. *Econometrica*, 57:759–778, 1989.

Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *13th Intl. Conf. Very Large Data Bases*, pp. 576–587, 2004.

R. Jurca and B. Faltings. Minimum payments that reward honest feedback. In *Proc. 7th ACM Conference on Electronic Commerce*, pp. 190–199, 2006.

M. Kandori. Social norms and community enforcement. *Rev. Econ. Stud.*, 59(1):63–80, 1992.

D. Kreps and R. Wilson. Reputation and imperfect information. *J. Econ. Theory*, 27(2):253–279, 1982.

K. Lai, M. Feldman, J. Chuang, and I. Stoica. Incentives for cooperation in peer-to-peer systems. In *First Workshop on the Economics of Peer-to-Peer Systems*, 2003.

R. Levien. *Attack-Resistant Trust Metrics*. PhD Thesis, University of California, Berkeley, 2004.

P. Milgrom and J. Roberts. Predation, reputation and entry deterrence. *J. Econ. Theory*, 27(2):280–312, 1982.

N. Miller, P. Resnick, and R. Zeckhauser. Eliciting honest feedback: The peer-prediction method. *Management Sci.*, 51(9):1359–1373, 2005.

P. Resnick and E. Friedman. The social cost of cheap pseudonyms. *J. Econ. Management Strategy*, 10(2):173–199, 2001.

P. Resnick, R. Zeckhauser, E. Friedman, and K. Kuwabara. Reputation systems. *Commun. ACM*, 43(12):45–48, 2000.

M. Richardson, R. Agrawal, and P. Domingos. Trust management for the semantic Web. In *Second International Semantic Web Conference*, LNCS 2870: 351–368. Springer, 2003.

R. Selten. Axiomatic characterization of the quadratic scoring rule. *Exp. Econ.*, 1(1):43–62, 1998.

S. Tadelis. What's in a name? Reputation as a tradeable asset. *Amer. Econ. Rev.*, 89(3), 1999.

H. Zhang, A. Goel, R. Govindan, K. Mason, and B.V. Roy. Making eigenvector-based reputation systems robust to collusion. In *Workshop on Algorithms and Models for the Web Graph (WAW'04)*, 2004.

Exercises

For context, each problem is preceded by the number of the relevant section.

27.1 (27.2) Verify that if the stage game payoff is constant, the (discounted) average payoff per round equals that constant. That is, if $pi_i^t = c$ then $\overline{\pi_i} = c$.

27.2 (27.2) The well-known "tit-for-tat" (TFT) strategy can be defined as: in round i play the strategy that your opponent played in round $i-1$, starting with C. Show that TFT, played by all players, is not an SPNE for any $\delta < 1$.

27.3 (27.2) Recall our definition of the Grim strategy: play C unless some player has played D in a previous round. Explain why it should not be defined in the apparently equivalent manner: "Play C unless the other player has played D in a previous round." (Hint: SPNE strategies need to optimal even on play paths that should not arise!)