Variable Temptations and Black Mark Reputations

Variable Temptations and Black Mark Reputations

Faculty Research Working Paper Series

Christina Aperjis

HP Labs

Yali Miao

Jane Street Capital

Richard J. Zeckhauser

Harvard Kennedy School

November 2012
RW12-055 (Revision of RWP11-020)

Variable Temptations and Black Mark Reputations[☆]

Christina Aperjis

HP Labs, 1501 Page Mill Rd, Palo Alto, CA 94304

Yali Miao

Jane Street Capital, Roppongi 6-12-4, Minato-ku, Tokyo, Japan 106-0032

Richard J. Zeckhauser

Harvard University, 79 John F. Kennedy Street, Cambridge, MA 02138

Abstract

In a world of imperfect information, reputations often guide the sequential decisions to trust and to reward trust. We consider two-player situations where the players meet but once. One player — the truster — decides whether to trust, and the other player — the temptee — has a temptation to betray when trusted. The strength of the temptation to betray may vary from encounter to encounter, and is independently distributed over time and across temptees. We refer to a recorded betrayal as a black mark. We study how trusters and temptees interact in equilibrium when past influences current play only through its effect on certain summary statistics. We first focus on the case that players only condition on the number of black marks of a temptee and study the different equilibria that emerge, depending on whether the trusters, the temptees, or a social planner has the ability to specify the equilibrium. We then show that conditioning on the number of interactions as well as on the number of black marks does not prolong trust beyond black marks alone. Finally, we consider more general summary statistics of a temptee's past and identify conditions under which there exist equilibria where trust is possibly suspended only temporarily.

Keywords: Game Theory, Trust, Reputation

1. Introduction

In a typical business transaction, one or both parties have the potential to betray. A supplier can produce low-quality goods; a debtor can default; an em-

[☆]We are grateful to John H. Lindsey II, Ramesh Johari, Paul Resnick, Ashin D. Shah, Peter Zhang, two editors and three referees for extremely helpful comments. This work was partially supported by a grant from the Alfred P. Sloan Foundation, and the NSF under award IIS-0812042.

ployee can steal; or a contractor can break the deal. Betrayals are often avoided because temptations are modest or nonexistent. But even when temptations are significant, reputations can keep untrustworthy behavior in line. Thus, betrayal is deterred, lest we lose future business with others, find ourselves without future credit or facing higher interest rates from any lender, or have great difficulty finding a job. Many economic models focus on repeat play, but often interactions between players are fleeting and knowledge of reputation comes from the broader world. Personal interactions, as between friends, present the same situation, with temptations, betrayals, and reputations all playing important roles.

Reputations are hardly sufficient statistics. They rarely tell us everything or almost everything about an individual's past performance and actions, because it may be costly or impossible to collect all the information that is potentially relevant. A typical employee reference in these litigious days is likely to be: "Joe worked here for 12 years, and there are no recorded blemishes on his record." Information on credit scores is equivalently crude. Repaying a loan counts the same whether the terms were easy or harsh. If a minimum grade-point average is necessary to keep one's scholarship, the difficulty of one's course is irrelevant.

Even when a lot of information on an individual's past is available, it may be difficult to convey, or for recipients to process all available information when making decisions. As a result, people tend to rely on summary statistics and easily accessible information. For instance, even though electronic marketplaces, such as eBay and the Amazon Marketplace, provide various summary statistics about sellers, buyers tend to rely on the information that is most prominently shown (Cabral and Hortacsu, 2010). These observations motivate us to study settings where the past influences current play only through its effect on certain summary statistics.

We focus on two-player situations, where one player — the *truster* — decides whether to trust, and the other player — the *temptee* — has the temptation to betray when trusted. (Temptee is a neologism, but one whose meaning is readily grasped.) In our model — as in real life — *the strength of the temptation to betray will vary from encounter to encounter*; formally, we assume that it is i.i.d. across time and temptees. The tempted players could be suppliers who might breach a contract that turns out to be too costly, contractors who might do a shoddy job if it saves a lot of effort, employees who might miss work often when other responsibilities are pressing, or spouses who might stray from marital vows given highly attractive opportunities.

We consider a population that consists of equal numbers of trusters and temptees. In every period, each truster is randomly matched with a temptee, learns the temptee's *reputation score*, i.e., a summary statistic of her past play, and then decides whether to trust her. We study equilibria where players condition current play on the temptee's reputation score rather than the entire history. A *reputation mechanism* specifies the rules for calculating a temptee's reputation score from the history of her past play. We allow for imperfect recording, as various studies have shown that monitoring is often imperfect in practice (Bolton et al., 2009; Dellarocas and Wood, 2008; Chwelos and Dhar,

2008), and refer to a recorded betrayal of a temptee as a *black mark*.

We start by studying the *Basic Black Mark Mechanism*, where a temptee's reputation is simply a tally of the number of black marks that she has received. In a broad range of settings, the reputation mechanism only keeps track of the number of infractions. For example, the Better Business Bureau has information on the number of complaints a particular business has received, but not the number of interactions or volume of business that might have led to complaints. On the other hand, in some instances, an infraction carries weight in and of itself, and people do not think (or recognize) that the number of trials matters. This is in the spirit of criminal justice systems, where the judge learns the number of convictions in a defendant's past before sentencing, or some systems of sexual morality which look at the number of partners someone has had. More generally, the Basic Black Mark Mechanism approximates settings where people focus on the number of negatives — even if more reputation information is provided. Such behavior is related to the Availability Heuristic (Tversky and Kahnenman, 1973), which leads individuals to judge the frequency of an event by how easily they can bring an instance to mind and, as a result, leads individuals to give significant weight to extreme bad outcomes.

We study properties of the equilibria that arise from the interactions between trusters and temptees when the Basic Black Mark Mechanism is in place. We show that in any pure equilibrium the greater the number of black marks, the less likely a temptee is to betray. Equilibria have a cutoff structure: a temptee is trusted as long as her number of black marks remains below some cutoff, but is never trusted once she reaches the cutoff. We consider the set of cutoffs that can arise in equilibrium and study which one is preferred by each side of the market, and which is socially optimal. We also present comparative static results identifying how the maximum number of black marks a temptee is allowed in equilibrium depends on the monitoring technology, on the distribution of the temptation to betray, and on how much temptees discount future payoffs.

We next study the *Enhanced Black Mark Mechanism*, where an individual's reputation consists of both the number of black marks that she has received and the total number of interactions that she has been involved in. Equilibria are again characterized by cutoffs, but now trusters may use different cutoffs depending on the total number of interactions of a temptee. Interestingly, we show that these cutoffs are upper bounded by the maximum cutoff that can arise under the Basic Black Mark Mechanism. In other words, including the number of interactions in one's reputation does not prolong trust in the sense that a temptee is not allowed to have a larger number of black marks than with the Basic Black Mark Mechanism. Moreover, we show that equilibrium behavior in the long run is identical to equilibrium behavior under the Basic Black Mark Mechanism.

With both the Basic Black Mark Mechanism and the Enhanced Black Mark Mechanism, once a temptee reaches a certain number of black marks she is never trusted again. In short, she gets permanently excluded. We then consider more general ways to aggregate the temptee's history into a reputation score and identify equilibria where trust can be suspended only temporarily.

In many reputation contexts, agents differ in types, which get revealed through their behavior through a process of adverse selection. (For a survey of such models see Mailath and Samuelson, 2006). In our model, a temptee does not have a fixed (across periods) hidden type. All agents are identical. Reputation is only used to incentivize good behavior (as in Dellarocas, 2005). On the other hand, our assumption of i.i.d. temptations means that we have repeated adverse selection, that is, adverse selection within each individual trial. This is similar to Athey and Bagwell (2001); Athey et al. (2004) and Hopenhayn and Skrzypacz (2004) who study collusion in a repeated oligopolistic game and in repeated auctions respectively.

The literature on repeated games and reputation typically assumes that players have access to the complete history of past play. Only a few recent papers consider settings where players' access to information is limited. These latter papers typically assume "finite memory", that is, players observe the last few periods of play of an individual instead of her full history (Barlo et al., 2009; Colea and Kocherlakota, 2005; Mailath and Olszewski, 2011; Liu and Skrzypacz, 2011). Doraszelski and Escobar (2012) consider a general framework where players condition on summary statistics of past play and apply a recursive characterization for the set of equilibrium payoffs. Ekmekci (2011) devises a complex rating system that entails information censoring and Liu (2011) considers a setting where players need to pay to observe past behavior of an individual. Our work relates as well to the literature on social norms and random matching (e.g., Kandori, 1992; Okuno-Fujiwara and Postlewaite, 1995), where an agent is matched with a different partner in every period and dishonest behavior against one partner leads to sanctions by other partners in the future, and on social norms in settings with fixed matchings (Bendor and Mookherjee, 1990).

In contrast to prior work, we consider more general summary statistics on which players condition. The Basic and Enhanced Black Mark Mechanisms have not been studied before even though they realistically model interactions in a number of settings. Moreover, in contrast to the existing literature on finite memory and restricted feedback, we allow the strength of one's temptation to betray to vary from encounter to encounter, as is common in real life.

The remainder of the paper is organized as follows. The problem is formulated in section 2. The Basic Black Mark Mechanism is studied in section 3. In section 4, we study the Enhanced Black Mark Mechanism, where the truster knows both the number of black marks and the number of interactions of the temptee. Then, we consider more general ways of aggregating information on past black marks in section 5. Section 6 concludes. All proofs are provided in the Appendix.

2. Model

Players are divided into two roles, trusters and temptees. For expository ease, in this analysis, those who must decide whether to trust — trusters — are males, and those who are subject to temptation — temptees — are females.

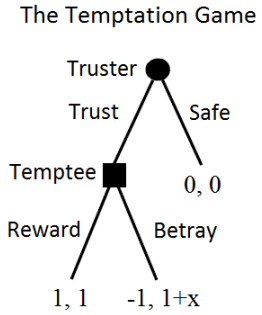

Figure 1: Extensive form representation of one-period interaction between a truster and a temptee after the temptee learns her temptation to betray x. The truster's choices are circles and the temptee's are squares, and the truster's payoff is listed first.

We model a one-period interaction between a truster and a temptee with the temptation game, shown in Figure 1. The temptee first privately observes the strength of her temptation to betray for this period, x, which is drawn from distribution F independently across periods and temptees. Then, the truster decides whether to choose "safe" or "trust." If the truster plays "safe", the temptee has no role, and both players receive zero payoff. If the truster plays "trust", then the temptee can play "reward" or "betray." If the temptee rewards, then both the truster and the temptee get a unit payoff. If the truster chooses to betray, then the temptee will get a $(1 + x)$ of payoff and the truster gets a payoff of -1. We note that the scaling of the payoffs is arbitrary. The analysis remains qualitatively the same if the truster gets a payoff of $-y$ when the temptee betrays, rather than -1, though of course the parameter values at equilibria will shift. There is no implied interpersonal comparison. For example, in dollar value a truster may gain far more than a temptee when each goes from 0 to 1.

We assume that the distribution F has a strictly positive median, which we denote by m. Then, there is a unique subgame perfect equilibrium of the one-shot temptation game where (i) the truster plays "safe" and (ii) if she were trusted, the temptee would betray whenever she had strictly positive temptation to do so. We also assume that F has a finite mean.

We now consider the repeated game. In each period, there are equal numbers of temptees and trusters, and each truster is randomly matched with a temptee. That is, one contracts with another party for just one period, and then moves on. When a truster is matched with a temptee he learns her *reputation score*, i.e., a summary statistic of her past play, and then decides whether to trust her. The strength of the temptation to betray is and remains unknown to trusters and therefore never becomes part of a temptee's reputation.

After each round, each temptee has a known probability of surviving to the next period, $s \in (0, 1)$. We leave aside discounting, except as it arises through a temptee's survival concerns. Then, the survival probability s represents how

much the temptee discounts future payoffs. In effect, as the survival probability increases, the temptee discounts future payoffs less. After each round, if a temptee dies, she will be replaced by another temptee who enters with a blank reputation record. If the reputation of a temptee ensures she will no longer be trusted, then she is not trusted until she dies (and is replaced by a new temptee with blank reputation only after she dies). We further assume that all players are risk-neutral. The temptee's goal is to maximize her expected payoff until she dies or is no longer trusted. The truster's goal is to maximize his expected payoff each period. Note that the survival probability for trusters is nonmaterial.

We refer to a recorded betrayal of a temptee as a *black mark*. We allow for imperfect recording; that is, a temptee may receive a black mark after rewarding and/or may not receive a black mark after betraying. In particular, if a temptee betrays in this period, she gets a black mark with probability $1-r$ and does not get a black mark with probability r. If a temptee rewards, then she does not receive a black mark with probability $1-q$, but does receive a black mark with probability q. Perfect monitoring is a special case with $r = q = 0$. In order to rule out settings with uninteresting equilibria for the repeated game, we assume that the imperfect monitoring probabilities are not too large, namely, $r+q < 1$.

We are interested in equilibria where a truster and a temptee that have been matched together in this period condition current play on the temptee's reputation score (i.e., the statistic on the temptee's past history that is shown to the truster when he encounters her) rather than the entire history. We thus restrict attention to *Markov Perfect Equilibria* (MPE) where the state is the temptee's reputation score. In other words, the past influences current play only through its effect on reputation scores. Note, however, that in our setting payoffs are not state (i.e., reputation score) dependent. This is along the lines of the state-strategy equilibrium framework of Doraszelski and Escobar (2012).

A *reputation mechanism* specifies the rules for calculating a temptee's reputation score from the history of her past play. Consider a specific temptee and let ρ^t represent her reputation score in period t. The reputation score could be a scalar or a vector. Let τ^t be the indicator variable of whether the temptee was trusted by the truster she was matched with in period t, that is, $\tau^t = 1$ if she was trusted and $\tau^t = 0$ otherwise. Similarly, denote by β^t the indicator variable of whether the temptee received a black mark in period t; $\beta^t = 1$ if yes, $\beta^t = 0$ otherwise. A reputation mechanism is a function that determines the temptee's reputation score in period $t+1$ from the tuple $(\rho^t, \tau^t, \beta^t)$.

Formally, if the reputation score takes values from some set $\mathcal{P}$, then the reputation mechanism is a function $h : \mathcal{P}\times\{0,1\}\times\{0,1\} \to \mathcal{P}$ and the reputation score at time $t+1$ is $\rho^{t+1} = h(\rho^t, \tau^t, \beta^t)$. Note that even though the reputation mechanism is a deterministic function, the reputation score at time $t+1$ may not be deterministically determined by the reputation score and the action profile at time t because β_t records imperfectly whether the temptee betrayed at time t.

The first reputation mechanism we study is the Basic Black Mark Mechanism, where a temptee's reputation is simply the number of black marks that she has received in the past; that is $\rho^t \in \mathbb{N}$ and $h(\rho^t, \tau^t, \beta^t) = \rho^t + \beta^t$. We then

study the Enhanced Black Mark Mechanism, where in additional to the number of black marks, a temptee's reputation also reveals the total number of past interactions of a temptee, that is, the number of times that the temptee has been trusted in the past. In this case, $\rho^t \in \mathbb{N}^2$ and $h(\rho^t, \tau^t, \beta^t) = \rho^t + (\beta^t, \tau^t)$. Other reputation mechanisms are considered in section 5. A temptee's reputation could also consist of her whole feedback history; however, we do not consider this extreme reputation mechanism in this paper.

The MPE that arise depend on which reputation mechanism is in place. We observe that there always exists a degenerate MPE where trusters never trust and temptees never reward when the temptation to betray is positive, that is, in every period players play the unique subgame perfect Nash equilibrium of the one-shot temptation game. Throughout the paper, we focus on pure equilibria, because mixed equilibria provide no additional insights. For completeness, mixed-strategy equilibria are discussed in Appendix C.

3. Basic Black Mark Mechanism

In this section, we consider the *Basic Black Mark Mechanism*, where a temptee's reputation is simply the number of black marks she has received in the past. We denote the number of black marks by b. When monitoring is perfect, b equals the actual number of betrayals. In general, however, its value may differ.

3.1. Characterization of Equilibria

We start by characterizing the (pure) MPE that arise under the Basic Black Mechanism. A truster's strategy consists of determining whether he trusts a temptee as a function of her reputation. A temptee's strategy is to choose whether she rewards as a function of her reputation b and her temptation to betray x in that period.

For a fixed strategy of trusters, let b^* be the minimum number of black marks at which a truster no longer trusts a temptee. That is, a truster trusts when $b < b^*$ and does not trust when $b = b^*$. Since a truster does not trust a temptee at b^*, a temptee will never have more than b^* black marks. We thus refer to b^* as the *cutoff* at which trusters stop trusting a temptee.

We first consider the best response of a temptee when trusters use the cutoff b^*. Let $v(b)$ be the maximum expected infinite horizon payoff to the temptee when her reputation score is b black marks. Since the cutoff is b^*, the trusters will never trust the temptee once her reputation becomes b^*, and thus

$$v(b^*) = 0. \tag{1}$$

For $b \in \{0, 1, ..., b^* - 1\}$, $v(b)$ is given by the following dynamic program[1]

$$v(b) = \int \max\{1+x+s((1-r)\cdot v(b+1)+r\cdot v(b)), 1+s((1-q)\cdot v(b)+q\cdot v(b+1))\}dF(x)$$

In particular, given that her temptation to betray is x, the temptee chooses the action that maximizes her expected payoff. Should she choose to betray, her expected payoff is $1 + x + s((1 - r) \cdot v(b + 1) + r \cdot v(b))$, since she receives $1 + x$ now and her reputation deteriorates to $b + 1$ black marks with probability $1 - r$ and remains the same (i.e., stays at b black marks) with probability r. On the other hand, if the temptee chooses to reward, her expected payoff is $1 + s((1 - q) \cdot v(b) + q \cdot v(b + 1))$, since she receives 1 now and her reputation remains the same with probability $1 - q$ and deteriorates to $b + 1$ black marks with probability q. Note that the continuation value is either $v(b)$ or $v(b + 1)$, since the temptee's total number of black marks either remains the same or increases by one.

Let

$$x_b^* \equiv s(1 - r - q) \cdot (v(b) - v(b + 1)). \tag{2}$$

Straightforward calculations show that $v(b)$ satisfies the following recursion:

$$(1 - s(1 - q))v(b) = sq \cdot v(b + 1) + 1 + \int_{x_b^*}^{\infty} (y - x_b^*)dF(y). \tag{3}$$

It is optimal for the temptee to reward if her temptation to betray is $x \leq x_b^*$ and betray if $x > x_b^*$. The temptee is indifferent between rewarding and betraying when $x = x_b^*$. For simplicity, we will assume that she chooses to reward if and only if $x \leq x_b^*$.[2] This simplifies the presentation because now the set of thresholds $\{x_b^*, b = 0, 1, ..., b^*\}$ characterizes the best response of the temptee. However, this assumption is not essential for our results. Since the temptee gets strictly positive immediate payment whenever she is trusted, the value $v(b)$ is strictly decreasing for $b \leq b^*$. Moreover, the assumption $r + q < 1$ implies that x_b^* is strictly positive for $b \in \{0, 1, ..., b^* - 1\}$.

We next consider the strategy of trusters. Consider a truster who is matched with a temptee who has b black marks in this period. Given x_b^*, his expected payoff is $2F(x_b^*) - 1$ if he trusts; and 0 otherwise. We conclude that the truster trusts if $F(x_b^*) > 1/2$; does not trust if $F(x_b^*) < 1/2$; and is indifferent between trusting and not trusting if $F(x_b^*) = 1/2$.[3] Note that the condition $F(x_b^*) \geq 1/2$

[1] In our setting, it is easier to study directly the dynamic program that represents the temptee's problem than it is to apply a generalization of the methods of Abreu et al. (1990) or Doraszelski and Escobar (2012), partly because we do not assume that players can coordinate using a randomization device.

[2] In most cases, it is not essential to specify what the temptee does when her temptation to betray is exactly x_b^*, because this occurs with probability zero. This is clearly true for a continuous distribution. On the other hand, when the distribution is discrete, then x_b^* is usually at a point of zero mass.

[3] The number 1/2 arises because we are assuming that the truster's payoff is equal to -1

is equivalent to $x_b^* \geq m$, where m is the median of the distribution F. It is important to emphasize here that because each truster is randomly matched with a temptee in every period, each truster is essentially myopic in the sense that in every period his strategy only depends on the temptee that he is matched with.

We conclude that the cutoff $b^* \geq 0$ and the thresholds $\{x_b^*, b = 0, 1, ..., b^*\}$ constitute an MPE under the Basic Black Mark Mechanism if (i) there exists a function $v : \mathbb{N} \to \mathbb{R}$ such that (1), (2) and (3) hold, and (ii) $F(x_b^*) \geq 1/2$ for $b < b^*$; $F(x_{b^*}^*) \leq 1/2$.

An MPE can be computed by recursively solving Equations (2) and (3) to obtain $x_{b^*-i}^*$ and $v(b^* - i)$ starting from the initial condition given by (1). Then, the cutoff b^* and the computed set of thresholds $\{x_b^*, b = 0, 1, ..., b^* - 1\}$ constitute an MPE if $\mathbb{P}[X < x_b^*] \geq 1/2$ for $b < b^*$. On the other hand, if $\mathbb{P}[X < x_b^*] < 1/2$ for some $b < b^*$, then no (pure) MPE exists with cutoff greater or equal to b^*. In general, if there exists an MPE with cutoff $b^* = k$, there also exists an MPE with cutoff $b^* = k'$, where $k' < k$ (assuming that both k and k' are positive integers). We use B^* to denote the maximum cutoff that can arise in equilibrium. Then, the set of equilibrium cutoffs is $\{0, 1, ..., B^*\}$.

3.2. Betrayal as a Function of Reputation

This section considers how reputations work when the Basic Black Mark Mechanism is in place. We find that temptees are less likely to betray when they have more black marks, and that (for a plausible class of distribution functions) the likelihood of betraying decreases faster when the temptee's reputation consists of a larger number of black marks. The following proposition states this result formally.

Proposition 1. *For every MPE* $(b^*, \{x_b^*, b = 1, 2, ..., b^*\})$ *under the Basic Black Mark Mechanism,* x_b^* *is strictly increasing and convex in* b *for* $b \in \{0, ..., b^* - 1\}$.

Proposition 1 shows that the threshold x_b^* is increasing and convex in the number of black marks. The following corollary of Proposition 1 characterizes the probability of rewarding $F(x_b^*)$ as a function of the number of black marks.

Corollary 1. *For every MPE* $(b^*, \{x_b^*, b = 1, 2, ..., b^*\})$ *under the Basic Black Mark Mechanism:*

(i) the probability of rewarding $F(x_b^*)$ *is increasing in* b *for* $b \in \{0, ..., b^* - 1\}$

(ii) if F *is linear or convex, then* $F(x_b^*)$ *is convex in* b *for* $b \in \{0, ..., b^* - 1\}$

In words, when a temptee's reputation depends solely on the number of black marks, the more black marks to date, the less likely the temptee is to

---

when the temptee betrays. More generally, if the truster got a payoff of $-y$ (instead of -1) when the temptee betrayed, our subsequent analysis would go through with $1/2$ replaced by $y/(y + 1)$.

betray. This follows from the fact that x_b^* is increasing in b. It may seem counterintuitive at first glance that those with worse reputations would behave better. However, since the truster is using a cutoff strategy and the temptee survives after every period with probability $s < 1$, when the temptee has more black marks she is more likely to use up all her black marks up to the cutoff before she dies. Thus, it is optimal for her to be more thrifty with black marks, to betray with a smaller probability (that is, only for very large temptations) when her reputation becomes worse. Equivalently, when the temptee is far from the cutoff, she can "afford" to spend black marks more freely, to succumb to temptation to a greater extent.

This insight is relevant for the design of reputation mechanisms in electronic marketplaces. For instance, EachNet, a Chinese auction site, implemented a warning system where a seller found guilty upon buyers' complaints received a warning and a seller with three warnings had to leave EachNet (Cai et al., 2011). eBay is implementing a similar warning system to complement its reputation mechanism. Corollary 1(i) suggests that a given seller would be less likely to betray for each warning she received.[4]

More generally, the structure of the temptation game resembles settings where players have a choice between playing safe at some cost, or taking a risk of adding a "black mark." Examples include the California criminal justice system, driver's license suspension, and several sports. For instance, California has a three-strikes-and-you-are-out rule for criminals: one who gets convicted of three felonies gets jailed for life. In each period, a person can decide whether to commit a crime or not. If she commits a crime, there is a chance of being caught. Following our model, as she comes closer to getting put away for life, she is less likely to commit a crime. Consistent with our model, she could have a payoff from the crime if she does not get caught, her temptation, which might be the expected amount of money she would steal. Corollary 1(i) suggests that recidivism rates in California should reveal a lesser propensity to criminal activity after two felony convictions. Indeed, recent literature has found reduced participation in criminal activity among second and third time offenders (e.g., see Iyengar, 2008, and the references therein). However, we note that our model does not capture certain aspects of this setting, such as the possibility to migrate to other states or multiple levels of crime severity.

In addition to showing that the likelihood of betraying decreases as the number of black marks increases, Corollary 1 shows that if the distribution function F is convex or linear (as is the case for the uniform distribution),

[4]We note that this does not necessarily imply that real world sellers with more warnings are less likely to betray than a seller with fewer warnings, because adverse selection effects could be affecting these probabilities. That is, sellers may differ in terms of payoff structure, self-control, or the distribution of the temptation to betray. Therefore, Corollary 1 does not contradict the finding of Cabral and Hortacsu (2010) that on eBay the interarrival time between the first and second negative is shorter than the arrival time of the first negative. As discussed in Appendix D, it is possible to include multiple types of temptees in our model and study adverse selection effects.

then the more black marks to date, the larger the marginal decrease in the likelihood of betraying. This follows from the convexity of x_b^*. That is, under a convex distribution function, the likelihood of betraying decreases faster when the temptee has a worse reputation.

3.3. Maximum Equilibrium Cutoff

In this section we study the properties of the maximum cutoff B^* that can arise in a (pure) MPE. The value of B^* depends on the distribution F, the survival probability s, and the imperfect monitoring probabilities r and q.

We first observe that B^* is finite for any fixed survival probability $s < 1$; that is, the temptee is not allowed an infinite number of black marks in equilibrium. In particular, if a temptee knew that she would be trusted even after an infinite number of black marks, then her best response would be to always betray whenever her temptation is positive. However, the truster's best response would then be to never trust, because we are assuming that the distribution F has a positive median. Intuitively, if a temptee was trusted irrespectively of her number of black marks, then she would not be incentivized to reward trust sufficiently often.

We next show that B^* increases without bound as s approaches 1. For the purposes of this result, we write $B^*(s)$ to denote that the maximum equilibrium cutoff B^* depends on the survival probability s. We also show that when there is imperfect monitoring with $q > 0$, $B^*(s)$ scales asymptotically like $1/(1-s)$.

Proposition 2. *Suppose r, q, F are fixed and $B^*(s) \geq 1$ for some $s < 1$. Then:*

(i) $B^*(s)$ is non-decreasing in s and $B^*(s) \to \infty$ as $s \uparrow 1$.

(ii) If $q > 0$, there exist constants $c_1, c_2 > 0$ and a threshold $\tilde{s} < 1$ such that $B^*(s) \in [c_1/(1-s), c_2/(1-s)]$ for $s \in [\tilde{s}, 1)$.

Proposition 2(i) says that the maximum number of black marks that a temptee is allowed in equilibrium increases without bound as the temptee's survival probability approaches 1. This implies that the number of periods for which cooperation is sustained (in the sense that the temptee is trusted) also increases without bound. Recall that the survival probability essentially represents how much the temptee discounts future payoffs. Thus even though the cutoff $B^*(s)$ is finite for any fixed $s < 1$ and is often reached with probability 1 (e.g., when $q > 0$ or $F(x^*_{B^*-1}) < 1$), the time it takes to reach the maximum cutoff increases without bound as the temptee discounts future periods less. This limit applies for any distribution F and regardless of whether monitoring is imperfect.

Proposition 2(ii) considers settings of imperfect monitoring where a temptee may get a black mark after rewarding, i.e., $q > 0$. For this case, we can characterize how the maximum equilibrium cutoff scales with the survival probability s when s is close to 1. The fact that $B^*(s)$ scales asymptotically like $1/(1-s)$

implies that there exist MPE where the temptee's *normalized* discounted payoff[5] is bounded away from 0. This result contrasts with a grim trigger equilibrium for the prisoner's dilemma with imperfect monitoring, where the timing of the switch to a punishment phase is independent of how patient the players are and, as a result, a player's normalized expected payoff approaches 0 as the discount factor approaches 1 (e.g., see Mailath and Samuelson, 2006, p. 235). In the case of the Basic Black Mark Mechanism, this inefficiency does not intrude — despite having restricted the information to simply the number of black marks.

We note that when $q = 0$, the temptee's maximum normalized discounted payoff is always bounded away from 0 as $s \uparrow 1$; specifically, it is lower bounded by 1. In particular, if the temptee always rewards she gets a payoff of 1 in every period, never receives a black mark (because $q = 0$) and is therefore trusted until she dies. The optimal strategy will give a normalized discounted payoff that is at least as large. Finally, using similar arguments to those used in the proof of Proposition 2, we can show that if $q = 0$ and r is sufficiently small,[6] then $B^*(s)$ scales asymptotically like $1/(1-s)$.

We next study the dependence of the maximum equilibrium cutoff on the imperfect monitoring probabilities, for a fixed survival probability $s < 1$. The following proposition considers r, that is, the probability that a temptee escapes a black mark despite betraying.

Proposition 3. *Suppose s, q, F are fixed. Then, B^* is non-increasing in r.*

Intuitively, if a temptee is less likely to receive a black mark when she betrays, she will find it advantageous to betray more often. Knowing this, a truster needs to decrease the maximum number of black marks he will allow in equilibrium if he is to avoid a negative expected payoff whenever he trusts a temptee. The result is that the maximum equilibrium cutoff is non-increasing in r.

We next consider q, that is, the probability that a temptee receives a black mark after rewarding. Interestingly, the dependence of B^* on q may be non-monotonic. In particular, the maximum equilibrium cutoff may increase for small values of q and then decrease for large values of q. That is, it is possible that a temptee is allowed more black marks in equilibrium for some $q > 0$ than when $q = 0$, as the following example illustrates.

Example 1. *Assume that $s = 0.98$, $r = 0$ and the temptation to betray is uniformly distributed on $[0, 30]$. If $q = 0$, then $B^* = 10$. On the other hand, for $q = 0.01$ we have that $B^* = 11$, that is, the maximum equilibrium cutoff increases even though the noise increased. For larger values of q, B^* is non-increasing and becomes 0 for $q \geq 0.23$.*

[5] The normalized discounted payoff is equal to the infinite horizon discounted payoff when the temptee has no black marks (i.e., $v(0)$) times $1 - s$. The value $v(0)$ depends on the attributes of the temptees (i.e., s, r, q, and F) and the cutoff b^*; it is maximized when $b^* = B^*(s)$.

[6] The condition for this is $r/(1-r) < \int_m^\infty (y - m) dF(y)$.

A larger q might increase the number of black marks allowed in equilibrium because when a temptee is more likely to get a black mark despite rewarding, in marginal cases she may be more careful and reward rather than betray. However, in most cases, a larger q is associated with a smaller B^*.

3.4. Optimal Cutoffs

This section identifies the (pure) MPE that are most favorable for the trusters, most favorable for the temptees, and that optimize social welfare. We first discuss how those three might be chosen among all possible equilibria in practice. We might think of these three situations as ones where the two different parties, or some uninvolved but benevolent coordinator can select among the various possible equilibria. They may have this capability because they can establish customs or laws that apply in a particular community, or simply because they have the ability to communicate. Such communication can establish what Schelling (1980) labels a focal point. Myerson (2009, p. 1111) comments on Schelling's insight: "anything in a game's environment or history that focuses the players' attention on one equilibrium may lead them to expect it, and so rationally to play it. This focal-point effect opens the door for cultural and environmental factors to influence rational economic behavior."

In the temptation game, if only the trusters can communicate, they can say to the temptees: "We will allow you one black mark, but once you get to two, no one will trust you." If that message is transmitted, we would expect the temptees to behave as if $b^* = 2$. Cheap talk in such circumstances can determine which equilibrium is chosen: an equilibrium becomes focal because it is "agreed on" through cheap talk, and it is then followed (Farrell, 1987). See Crawford and Sobel (1982) and Farrell and Rabin (1996) for detailed discussions on cheap talk for coordination games and games of incomplete information. On the other hand, humans often use information on the opponent's past actions as coordinating devices, see Dalea et al. (2002) for an experimental study. In the context of the temptation game, if the trusters have not been trusting temptees with one black mark in the past, this may lead the temptees to expect that this behavior may continue in the future. In the temptation game, players belong to identifiable groups, which may give them an additional incentive to adhere to the rules set out in pre-game communications. If the players are to deviate from the announced group behavior, they will possibly suffer another type of reputation loss, that with their peers.

Posit that the trusters, and only the trusters have the ability to communicate verbally. They will tell the temptees that they are employing the cutoff in number of black marks that maximizes the expected welfare of trusters assuming that temptees respond optimally to that cutoff. On the other hand, if the temptees have the power to choose the equilibrium — whether through communication or by setting the rules or laws in some group — they will commit to their thresholds x_b^* for $b \in \{0, 1, ..., b^*\}$, thus letting a truster pick his b^* in response. This produces the first-best condition for the temptees. Because a truster responds to the temptees' strategy in a predictable way, the temptees are effectively choosing the trusters' cutoff b^*. The *socially optimal* equilibrium

emerges when a third party, perhaps a government agency or an e-commerce site, proposes a set of strategies and associated equilibrium to optimize a weighted sum of the payoffs going to trusters and temptees.[7]

A truster's payoff in a given period depends heavily on the number of black marks of the temptee that he interacts with. In particular, when matched with a temptee with b black marks, the truster gets a positive expected payoff of $2F(x_b^*)-1$ if $b < b^*$ and a zero payoff if $b = b^*$. As far as a truster is concerned, the number of black marks of any temptee evolves according to a Markov chain. The state is the temptee's number of black marks, which either increases by 1 (if the temptee is trusted and a betrayal is recorded), or remains the same (if either the temptee is trusted and a reward is recorded or the temptee is not trusted), or becomes 0 (if the temptee dies and thus is replaced with a new player with a blank reputation).[8] Let π denote the stationary distribution of this Markov chain and assume that the Markov chain has reached stationarity.[9] Then, a truster's expected payoff from every temptee that he may interact with in a given period is equal to $\sum_{b=0}^{b^*-1} \pi_b(2F(x_b^*)-1)$. We do not include a term for $b = b^*$ in the sum, because a truster gets zero payoff when matched with a temptee who has b^* black marks.

We define b_C^* and b_D^* to be the optimal cutoffs for trusters and temptees respectively, that is, the cutoffs of the equilibria that maximize the corresponding payoffs. We let $b_S^*(\alpha)$ denote the cutoff that maximizes the sum of the trusters' payoff and α times the temptees' payoff, where $\alpha \geq 0$. Recall that the set of equilibrium cutoffs is $\{0, 1, 2..., B^*\}$. The following proposition shows that the optimal cutoff will be greatest for the temptee, least for the truster, and in between for the social optimum.

Proposition 4. *If $B^* \geq 1$, then $1 \leq b_C^* \leq b_S^*(\alpha) \leq b_D^* = B^*$ for any $\alpha \geq 0$.*

Proposition 4 tells us that the first-best equilibrium for temptees has the maximum possible cutoff. Intuitively, a temptee prefers to be trusted longer.

Proposition 4 also says that the first-best equilibrium cutoff for trusters is in $\{1, 2, ..., B^*\}$. There are two effects that influence a truster's expected payoff. On the one hand, if he is matched with a temptee whom he decides to trust, he is better off if the cutoff is small because then that temptee is more likely

[7] We note that the third party could also propose an equilibrium that achieves some other goal, e.g., if Amazon could specify the equilibrium that buyers and sellers play in the Amazon Marketplace, it would perhaps choose the equilibrium that maximizes Amazon's revenue. We do not consider that situation in this paper.

[8] A truster does not care how long a specific temptee lives, because he is guaranteed to meet a new temptee each period.

[9] Because trusters are essentially myopic maximizers, the set of equilibria we derive in section 3.1 does not depend on the reputation distribution of the population of temptees. We have the same set of equilibria irrespectively of whether the reputation distribution has reached stationarity. This result contrasts with the norm equilibrium of Okuno-Fujiwara and Postlewaite (1995), where players effectively play a best response to the stationary distribution. We only use the stationary distribution in this section because it is natural to define a truster's expected payoff and optimal cutoff with respect to this distribution.

to reward trust. On the other hand, when the cutoff is smaller, the truster is less likely to be matched with a temptee who is below the cutoff. If the former (resp., latter) effect dominates, then the first-best equilibrium for the truster involves a smaller (resp., larger) cutoff.

The following example demonstrates that even for the extremely simple case where temptations are uniformly distributed, the first-best equilibrium cutoff for trusters b_C^* may take any value in $\{1, 2, ..., B^*\}$. In other words, it can involve the minimum non-trivial equilibrium cutoff, the maximum equilibrium cutoff, or any value in between.

Example 2. *Assume that* $s = 0.95$, $r = 0.1$, $q = 0.01$ *and the temptation to betray is uniformly distributed on* $[0, A]$. *Figure 2 shows the optimal cutoffs of trusters and temptees for various values of* A. *We observe that when* $A = 10$, *a truster's payoff is maximized at* $b_C^* = 1$, *that is, the one-betrayal-and-you-are-out strategy is best for trusters. This is the strategy that many societies have employed to deal with marital infidelities, particularly those of women. A temptee's payoff on the other hand is maximized at* $b_D^* = 4$. *Thus, in this case,* $1 = b_C^* < b_D^* = 4$. *We next observe that when* $A = 20$ *we have that* $1 < b_C^* = 2 < b_D^* = B^* = 4$. *Finally, for* $A \in \{49, 50, ..., 82\}$, *we have that for any* $\alpha \geq 0$, $b_C^* = b_S^*(\alpha) = b_D^* = B^* = 3$, *that is, both trusters and temptees prefer the same equilibrium cutoff.*

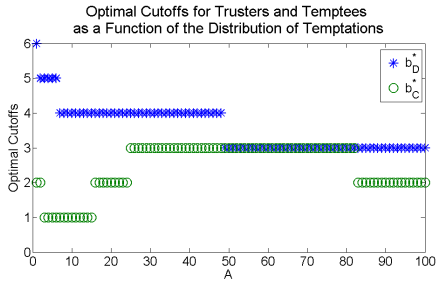

Figure 2: Optimal cutoffs for trusters and temptees when $s = 0.95$, $r = 0.1$, $q = 0.01$ and the temptation to betray is uniformly distributed in $[0, A]$.

In Example 2, the optimal cutoff for the truster is non-monotonic, whereas the optimal cutoff for the temptee — and therefore also the maximum equilibrium cutoff — decreases as the support of the distribution increases. We next show that this is always the case for the uniform distribution.

Proposition 5. *If the temptation to betray is uniformly distributed on* $[0, A]$, *then the maximum equilibrium cutoff* B^* *is non-increasing in* A.

Note that when $A' > A$, the uniform distribution on $[0, A']$ stochastically dominates the uniform distribution on $[0, A]$. Intuitively, when A increases, temptations are stronger overall, and a temptee will give in to temptation more frequently. This result does not generalize to non-uniform distributions. Thus,

it is possible to have two distributions, F and G, such that G stochastically dominates F, yet the temptee is less likely to betray at b black marks when the temptation to betray is drawn from G. That is because she would be giving up more in terms of opportunity cost in the future. As a result, the maximum equilibrium cutoff with G may be larger than the maximum equilibrium cutoff with F, even though G stochastically dominates F. Example 3 illustrates.

Example 3. *Suppose that $s = 0.6$ and $r = q = 0$. With distribution F, the temptation to betray equals 1 with probability 0.4 and equals 2 with probability 0.6. With distribution G, the temptation to betray equals 2 with probability 0.9 and equals 10 with probability 0.1. Clearly, G stochastically dominates F. If the temptation to betray is drawn from G, then the maximum equilibrium cutoff is 1. (If $b^* = 1$ and the temptee currently has no black marks, it is optimal for her to reward if $x = 2$ and betray if $x = 10$. That is, she betrays with probability 0.1.) On the other hand, if the temptation to betray is drawn from F, then the maximum equilibrium cutoff is 0. If the temptee were trusted, it would be optimal for her to reward if $x = 2$; that is, she would betray with probability 0.6 or higher. But then the truster would be better off not trusting her, so $B^* = 0$.*

Example 3 shows that the maximum equilibrium cutoff does not necessarily decrease when the temptation distribution "increases" in the sense of (first-order) stochastic dominance. We next show that (under some conditions) second-order stochastic dominance implies that the maximum equilibrium cutoff decreases. Equivalently, when the temptation distribution is more likely to take on "extreme" values, then the maximum equilibrium cutoff increases. The reason is that higher variability in the temptation to betray is associated with higher opportunity costs which incentivize a temptee to betray less frequently.

Proposition 6. *Consider two distributions F_1, F_2 with the same median m and the same mean. Let B_i^* be the maximum equilibrium cutoff when F_i is the distribution of the temptation to betray. If F_1 second-order stochastically dominates F_2 and $F_1(x) \geq F_2(x)$ for all $x \geq m$, then $B_2^* \geq B_1^*$.*

For instance, Proposition 6 implies that the maximum equilibrium cutoff is at least as large when the temptation is uniformly distributed on $\{1, 2, 3, 4\}$ as when it is uniformly distributed on $\{2, 3\}$.

In addition to the distribution of the temptation to betray, the optimal cutoffs for trusters and temptees also depend on the parameters s, r, and q. From Proposition 4 we know that $b_D^* = B^*$. Thus, Propositions 2 and 3 imply that the optimal cutoff for temptees is increasing in the survival probability s and decreasing in r. Furthermore, it follows from Example 1 that b_D^* may be non-monotonic in q. On the other hand, the optimal cutoff for trusters b_C^* is generally non-monotonic in each of the parameters s, r, and q, because of its dependence on the distribution of black marks in the population of temptees.

We conclude this section by considering the effect of the imperfect monitoring probabilities on the players' expected payoffs. This is important not merely to understand comparative statics, but to know what choices both classes of

players might want to make to improve monitoring capabilities. Thus, if we had a tradeoff between accuracy on q and r (as would seem quite reasonable, as there are often tradeoffs between type 1 and type 2 errors), then we provide the ingredients to determine how a temptee would tune r and q within the technological constraint and what it would be worth to temptees to have a more accurate system.

We first observe that for a fixed cutoff, a temptee is better off when r is larger, that is, when her betrayals are less likely to be recorded, and worse off when q is larger, that is, when she is more likely to get a black mark despite rewarding. However, r and q also affect the maximum equilibrium cutoff, which is the preferred equilibrium cutoff for temptees.

There are two effects as r increases: the temptee gets a higher expected payoff for any fixed cutoff b^*, but at the same time the maximum cutoff B^* may decrease (by Proposition 3). As a result, a temptee's maximum equilibrium payoff, i.e., her payoff at her preferred equilibrium, increases in an interval over which the maximum equilibrium cutoff remains the same, then drops whenever the maximum equilibrium cutoff decreases. That is, the temptee's maximum payoff is non-monotonic in r. (See Figure 3 in Appendix B.) Therefore, if the maximum equilibrium cutoff is played, in some cases the temptee may prefer a larger r at which her betrayals are less accurately recorded. But it is also possible that the temptee is better off when betrayals are more accurately recorded.

Interestingly, a temptee's maximum equilibrium payoff may increase for small values of q. This may occur when the maximum equilibrium cutoff increases for small values of q, as in Example 1. However, in most cases a temptee is worse off when q increases. In general, there are jumps in the temptees' maximum equilibrium payoff when B^* changes, but there are downward drifts for any given B^* with increases in q and upward drifts within any B^* for increases in r. We provide some examples in Appendix B. We conjecture that trusters are worse off whenever monitoring becomes less accurate, i.e., when either r or q increases. We thus expect that trusters would prefer to improve the monitoring technology as long as it is not too costly to do so.

4. Enhanced Black Mark Mechanism

Thus far, we have considered the Basic Black Mark Mechanism, which tracks only the number of black marks. In this section, we study the *Enhanced Black Mark Mechanism*, which in addition to the number of black marks also reveals the *number of interactions*, that is, the number of times that the temptee has been trusted in the past. Our main result in this section is that the Enhanced Black Mark Mechanism has the same maximum equilibrium cutoff as the Basic Black Mark Mechanism. That is, including the number of interactions in the temptee's reputation does not prolong trust.

We denote a temptee's reputation score by (b, n), where n is the number of interactions that she has completed. A strategy of the trusters in this more general model consists of a cutoff for each number of interactions. With a slight abuse of notation, let $b^*(n)$ be the cutoff for n interactions, that is, a truster

trusts a temptee with reputation (b, n) if and only if $b < b^*(n)$. We refer to $b^*(n)$ as the cutoff function or the trusters' strategy. On the other hand, a temptee's strategy will consist of a threshold $x^*_{b,n}$ for every possible reputation (b, n). That is, when a temptee has b black marks in n interactions, then she betrays if the strength of her temptation to betray exceeds the threshold $x^*_{b,n}$.

The cutoff function $b^*(\cdot)$ and the thresholds $\{x^*_{b,n}, b = 0, 1, ..., b^*(n), n \in \mathbb{N}\}$ constitute a (pure) MPE under the Enhanced Black Mark Mechanism if:

(a) There exists a function $v : \mathbb{N}^2 \to \mathbb{R}$ such that (i) $v(b^*(n), n) = 0$ for all $n \in \mathbb{N}$ and (ii) for $b < b^*(n)$ and $n \in \mathbb{N}$:

$$v(b,n) = 1 + s \cdot (1-q) \cdot v(b, n+1) + s \cdot q \cdot v(b+1, n+1) + \int_{x^*_{b,n}}^{\infty} (y - x^*_{b,n}) dF(y),$$

where $x^*_{b,n} \equiv s \cdot (1 - r - q) \cdot (v(b, n+1) - v(b+1, n+1))$.

(b) For all $n \in \mathbb{N}$: (i) $F(x^*_{b,n}) \geq 1/2$ when $b < b^*(n)$ and (ii) $F(x^*_{b^*(n),n}) \leq 1/2$.

Condition (a) guarantees that the set of thresholds $\{x^*_{b,n}, b = 0, 1, ..., b^*(n), n \in \mathbb{N}\}$ is a best response of temptees to the cutoff function, and condition (b) guarantees that the trusters are playing a best response.

The derivation of these equilibrium conditions parallels that of the derivation for the Basic Black Mark Mechanism in section 3.1. Moreover, if we set $b^*(n) = b^*$ for some cutoff $b^* \leq B^*$, then the Enhanced Black Mark Mechanism essentially reduces to the Basic Black Mark Mechanism. In this case, $b^*(n)$ is a constant that does not vary with the number of interactions n. However, under the Enhanced Black Mark Mechanism there also exist equilibria where $b^*(n)$ varies with n. In that case, trusters are effectively using a moving quota of permitted betrayals and stop trusting if there are $b^*(n)$ betrayals in n interactions. As a result, we get a larger set of equilibria with the Enhanced Black Mark Mechanism, because there is more available information on which players can condition their strategies.

We next show two fundamental properties of the cutoff function; the first intuitive, the second less so. First, $b^*(n)$ is non-decreasing in n. The larger the number of interactions of a temptee, the larger the number of black marks that a truster will tolerate. Second, $b^*(n)$ is bounded above by B^*, i.e., the maximum cutoff for which there exists a (pure) equilibrium when reputation only consists of the number of black marks. That is, including the number of interactions in the reputation information does not increase the maximum number of black marks that a temptee will be allowed in equilibrium.

Proposition 7. *At any MPE under the Enhanced Black Mark Mechanism:*

(i) $b^*(n)$ *is non-decreasing*
(ii) $b^*(n) \leq B^*$ *for all* n

The fact that $b^*(n)$ is upper-bounded by B^* may at first seem counterintuitive. One could expect that a truster would allow a temptee more black marks when he knows that she has completed a very large number of interactions than when he has no information on the number of interactions. However,

if the trusters tolerated a larger number of black marks, a temptee would not be properly incentivized in the sense that her probability of betraying would be greater than 1/2; thus, a truster would be better off not trusting her. This result critically depends on the fact that, because of the random matching, trusters are essentially myopic in our model.

Intuitively, for a temptee's incentives, the only thing that matters is how far she is (in terms of black marks) from no longer being trusted. This distance depends on the cutoff function $b^*(n)$ and the temptee's current reputation. If the temptee knows that she can get an additional B^* black marks and still be trusted, then the punishment of no longer being trusted will arrive too far into the future, and the temptee is not properly incentivized. This means that in equilibrium the temptee cannot be further than B^* black marks from no longer being trusted. But if a temptee has no black marks, then the distance in terms of black marks from no longer being trusted is lower bounded by $b^*(n)$. This implies that $b^*(n)$ cannot exceed B^*, no matter how large is the number of past interactions n.

Observe that there always exists an equilibrium with $b^*(n) = B^*$ for all n. Thus, Proposition 7(ii) implies that the maximum equilibrium cutoff under the Enhanced Black Mark Mechanism is equal to B^*, i.e., the same as for the Basic Black Mark Mechanism. Then, Propositions 2, 3, 5 and 6 also characterize how the maximum equilibrium cutoff of the Enhanced Black Mark Mechanism depends on s, r and F. Moreover, similarly to Proposition 4, we can say which equilibrium cutoff functions each side of the market prefers when the Enhanced Black Mark Mechanism is in place. The best equilibrium cutoff function for the temptees is $b^*(n) = B^*$. Trusters also prefer $b^*(n) = B^*$ in some cases, but in other cases their best cutoff function takes lower values.

Given that $b^*(n)$ is increasing but upper bounded (by Proposition 7), we conclude that $b^*(n)$ is constant for all large n, which implies that the number of interactions plays no role after some point. Then, the thresholds $x^*_{b,n}$ correspond to thresholds that arise with the Basic Black Mark Mechanism and Proposition 1 applies. Thus, after that point a temptee is less likely to betray when she has more black marks; in other words, $x^*_{b,n}$ is increasing in b when n is sufficiently large. However, $x^*_{b,n}$ may not be increasing in b for small values of n when there is a high probability of misrecording a betrayal and the cutoff function $b^*(n)$ is not constant.[10]

5. Temporary Exclusion

With both the Basic Black Mark Mechanism and the Enhanced Black Mark Mechanism, once a temptee reaches a certain number of black marks she is never trusted again. That is, she is permanently excluded once she reaches a cutoff. Another possibility is *temporary exclusion*, that a temptee is temporarily

[10] We thank John H. Lindsey II for constructing an example where $x^*_{b,n} > x^*_{b+1,n}$ and $b + 1 < b^*(n)$ at an equilibrium.

not trusted but later she is trusted once again. In other words, with temporary exclusion the temptee is essentially "punished" by not being trusted for a number of periods, and is trusted again once this punishment phase is over. In this section, we consider what reputation mechanisms give rise to MPE with temporary exclusion.

We next show that given the attributes of a temptee (i.e., s, r, q, and F), we can compute the minimum length of punishment that can arise in an MPE with temporary exclusion.

Proposition 8. *An MPE with temporary exclusion exists if and only if the reputation information allows the trusters to punish the temptee (by not trusting her) for at least*

$$T^*(s,r,q,F) \equiv \left\lceil \log\left(1 - \frac{1/s - 1}{(1-r-q)\left(1+\int_m^\infty (x-m)dF(x)\right)/m - q}\right) / \log s \right\rceil$$

periods after every black mark.

In other words, we can have equilibria where trust is not lost forever if and only if the reputation mechanism provides sufficient information to allow trusters to punish a temptee after a black mark for at least T^* periods. If the reputation mechanism does not provide enough information to allow trusters to punish a temptee for T^* periods, then we either have a unique MPE where trusters never trust temptees, or there also exist MPE with permanent exclusion, as with the Basic and Enhanced Black Mark Mechanisms.

We note that the minimum punishment length T^* is increasing in both r and q, that is, a longer punishment is required when recording is less accurate. On the other hand, T^* decreases when the term $g_F \equiv \left(1+\int_m^\infty (x-m)dF(x)\right)/m$ increases. To obtain the underlying intuition, consider two distributions F_1 and F_2 with the same median and the same mean. If F_1 second-order stochastically dominates F_2, then $g_{F_2} \geq g_{F_1}$ and, as a result, when the temptation to betray is given by F_2 it is possible to have an equilibrium with shorter punishments after every black mark. In other words, *greater variability in the temptation to betray allows for equilibria with shorter punishments.* This is along the lines of Proposition 6, where greater variability in the temptation to betray allows for equilibrium with larger cutoffs. Intuitively, when the distribution is more variable, by betraying now the temptee would be giving up more in terms of opportunity cost in the future.

Observe that trusters cannot coordinate a punishment of T^* periods with the Basic or the Enhanced Black Mark Mechanism, since these mechanisms provide no information about when black marks occurred. For instance, consider the Basic Black Mark Mechanism and suppose a truster is matched with a temptee who has one black mark. The truster has no way of knowing whether the temptee has already been punished for that black mark and whether he should trust her in this period. Similar issues arise with the Enhanced Black Mark Mechanism.

We next provide examples of reputation mechanisms for which temporary exclusion may arise in equilibrium. Recall that ρ^t denotes the reputation score in period t, τ^t is the indicator variable of whether the temptee was trusted in period t, and β^t is the indicator variable of whether the temptee received a black mark in period t. A reputation mechanism is a function h such that $\rho^{t+1} = h(\rho^t, \tau^t, \beta^t)$; see section 2 for details.

Example 4. *Consider a finite memory mechanism where a temptee's reputation score consists of her history of play in the last K periods, that is, $\rho^t = (\tau^{t-1}, \beta^{t-1}, \tau^{t-2}, \beta^{t-2}, ..., \tau^{t-K}, \beta^{t-K})$. Proposition 8 tells us for which values of K there exist MPE with temporary exclusion. In particular, if $K < T^*(s, r, q, F)$, there exists a single MPE where players play the equilibrium of the one-shot temptation game in every period and thus trusters never trust. On the other hand, if $K \geq T^*(s, r, q, F)$, there also exist MPE with temporary exclusion where a truster trusts the temptee only if she has not received a black mark in the last T periods. There may also exist other MPE, e.g., where a temptee is allowed $b > 1$ consecutive black marks (with no punishment inbetween) and a punishment of $T' > T^*(s, r, q, F)$ periods afterwards.*

Example 5. *Consider a reputation mechanism where the reputation score in period $t + 1$ is a weighted average of the reputation score in period t and the indicator β^t, that is, $h(\rho^t, \tau^t, \beta^t) = (1-\alpha)\rho^t + \alpha\beta^t$ for some parameter $\alpha \in (0, 1)$. Thus, the reputation score is a scalar taking values in $[0, 1]$ and α measures how strongly recent black marks affect the reputation score. This updating rule is a good model of how people update their impressions without a reputation mechanism in place (Anderson, 1981; Hogarth and Einhorn, 1992; Kashima and Kerekes, 1994). Note that with this mechanism a larger value of ρ^t is worse, as with the Basic Black Mark Mechanism. Suppose that a truster trusts a temptee at time t if and only if her reputation score is $\rho^t < \alpha(1-\alpha)^T$. Then, a temptee is not trusted for at least T periods after she gets a black mark. Thus, we can have MPE with temporary exclusion for any $\alpha \in (0, 1)$. Moreover, if $\alpha > 1/2$, we can guarantee that a temptee is not trusted for exactly T periods after she receives a black mark.*

Example 6. *Assume that a temptee's reputation score reveals (i) the number of black marks she has received, and (ii) the number of times she has not been trusted. Formally, the reputation mechanism is $h(\rho^t, \tau^t, \beta^t) = \rho^t + (\beta^t, 1 - \tau^t)$; we can denote by ρ_1^t the number of black marks and by ρ_2^t the number of periods that the temptee has not been trusted. In contrast to Examples 4 and 5, this mechanism does not weight information from the temptee's recent past more heavily. However, it is still possible for trusters to coordinate a punishment of T periods per black mark by trusting a temptee if and only if $\rho_1^t \cdot T \leq \rho_2^t$.*

6. Conclusion

This paper studies how trusters and temptees interact in equilibrium when past influences current play only through its effect on certain summary statistics. The Basic Black Mark Mechanism establishes the equilibria that emerge

when players condition solely on the number of recorded betrayals of a temptee. The Enhanced Black Mark Mechanism allows players to condition on both the number of recorded betrayals and the number of interactions of a temptee. The same qualitative results apply and the maximum number of black marks a temptee can get in equilibrium does not increase when the number of interactions is recorded. In closing, the paper considers more general summary statistics and identifies conditions under which there exist equilibria where trust is only suspended temporarily.

Throughout, the paper considers a setting with multiple trusters and multiple temptees, where in every period each truster is randomly matched with a temptee. That is, one engages with another party for just one period, and then moves on. However, our results also apply to long-term interactions between one truster and a large number of temptees. In this setting, the truster interacts with multiple temptees simultaneously (in each period). For instance, the truster can be a big employer interacting with multiple employees, a university interacting with many students, or a state interacting with a large number of citizens.

Two further extensions immediately suggest themselves. First, some relationships have a natural termination or sunset date quite apart from black marks. Thus, for a college and a student, rule infractions, e.g., plagiarism or disorderly behavior, would be the equivalent of betrayals. But once graduation occurs, the relationship is ended no matter what and past black marks become irrelevant. Second, many long-term relationships — and some one-time-only relationships — have both parties trusting and both parties tempted. Thus, a business and its supplier or a husband and wife may both rely on each other; each has a reputation, each can trust, and each can betray.

Across a wide swath of societal concerns, we live with the notion that a single betrayal does not end a relationship. Thus, there are second chances (and possibly more). Religions routinely allow for forgiveness. "The God I believe in is a God of second chances," Bill Clinton once said (Clinton, 1994). And George W. Bush, not known for being soft on crime, observed: "America is the land of second chance — and when the gates of the prison open, the path ahead should lead to a better life" (Bush, 2004). That is the way two successive Presidents outlined the theme that motivated this analysis: The game of life accommodates betrayals, but not without putting betrayers on warning.

From the time of the snake in the Garden of Eden, temptation has always been with us. Betrayals must be expected from all of us, and reputations are required to keep them within bounds. And should betrayals exceed some critical value, expulsion will be our fate. Such is the life of the temptee.

Appendix A: Proofs

Proof of Proposition 1: Let

$$g(y) \equiv 1 + \int_y^{\infty} (x - y) dF(x).$$

We observe that $g'(y) = -(1 - F(y))$. This implies that $g'(y)$ is negative and increasing in y, and thus g is decreasing and convex.

We first show that x_b^* is strictly increasing in b for $b \in \{0, ..., b^* - 1\}$. From (2) and (3) we have

$$\frac{1 - s(1-q)}{s(1-r-q)} x_b^* + (1-s)v(b+1) = g(x_b^*)]. \tag{4}$$

Let $b_1 < b_2$, and let $x_1 = x_{b_1}^*$ and $x_2 = x_{b_2}^*$ be the corresponding solutions of (4). Then,

$$\frac{1 - s(1-q)}{s(1-r-q)} x_1 + (1-s)v(b_1+1) = g(x_1). \tag{5}$$

$$\frac{1 - s(1-q)}{s(1-r-q)} x_2 + (1-s)v(b_2+1) = g(x_2). \tag{6}$$

Suppose that $x_1 \geq x_2$. Then,

$$\begin{aligned} &\frac{1 - s(1-q)}{s(1-r-q)} x_2 + (1-s)v(b_2+1) < \\ &\frac{1 - s(1-q)}{s(1-r-q)} x_1 + (1-s)v(b_1+1) = \\ &g(x_1) \leq \\ &g(x_2), \end{aligned}$$

which contradicts (6). We note that the first inequality follows because v is decreasing in b and $s < 1$; the equality follows from (5), and the second inequality holds because $x_1 \geq x_2$. We conclude that $x_1 < x_2$, and thus x_b^* is strictly increasing in b for $b \in \{0, 1, ..., b^* - 1\}$.

We next show that x_b^* is convex in b for $b \in \{0, ..., b^* - 1\}$. From (4) we find that

$$\frac{1 - s(1-q)}{s(1-r-q)} (x_b^* - x_{b-1}^*) + (g(x_{b-1}^*) - g(x_b^*)) = (1-s)(v(b) - v(b+1))$$

Moreover, by (2) we have that $v(b) - v(b+1) = x_b^*/(s(1-r-q))$. Thus,

$$\frac{1 - s(1-q)}{s(1-r-q)} (x_b^* - x_{b-1}^*) + (g(x_{b-1}^*) - g(x_b^*)) = \frac{1-s}{s(1-r-q)} x_b^*.$$

Let $b_1 < b_2$. Since x_b^* is increasing in b (by the first part of this proof), we have that

$$\frac{1-s(1-q)}{s(1-r-q)}(x_{b_1}^* - x_{b_1-1}^*) + (g(x_{b_1-1}^*) - g(x_{b_1}^*)) <$$
$$\frac{1-s(1-q)}{s(1-r-q)}(x_{b_2}^* - x_{b_2-1}^*) + (g(x_{b_2-1}^*) - g(x_{b_2}^*)) \tag{7}$$

Suppose that $x_{b_1}^* - x_{b_1-1}^* > x_{b_2}^* - x_{b_2-1}^*$. Then, by the convexity of g we have that

$$\begin{aligned}
&g(x_{b_1-1}^*) - g(x_{b_1}^*) \geq \\
&g(x_{b_2-1}^*) - g(x_{b_2-1}^* + (x_{b_1}^* - x_{b_1-1}^*)) \geq \\
&g(x_{b_2-1}^*) - g(x_{b_2-1}^* + (x_{b_2}^* - x_{b_2-1}^*)) \geq \\
&g(x_{b_2-1}^*) - g(x_{b_2}^*)
\end{aligned}$$

which contradicts (7). We note that the first inequality holds because g is convex, the second inequality is a consequence of $x_{b_1}^* - x_{b_1-1}^* > x_{b_2}^* - x_{b_2-1}^*$, and the third inequality holds because g is decreasing. Thus, $x_b^* - x_{b-1}^*$ is nondecreasing in b and x_b^* is convex for $b \in \{0, 1, ..., b^* - 1\}$. □

Proof of Proposition 2: Throughout the proof, we use the notation $B^*(s)$ and $x_i^*(s)$ to denote the dependence on the survival probability s. We assume that r, q and F are fixed.

We first show that $B^*(s)$ is increasing in s. Consider some fixed cutoff b^* and suppose that $s_1 < s_2$. We will show that $x_i^*(s_1) \leq x_i^*(s_2)$ for $i = 0, 1, ..., b^* - 1$, which implies that $B^*(s_1) \leq B^*(s_2)$. (2) and (3) imply that for any $i \leq b^* - 1$

$$\frac{1-s(1-q)}{s(1-r-q)}x_i^*(s) + \frac{1-s}{s(1-r-q)}\sum_{j=i+1}^{b^*-1} x_j^*(s) = 1 + \int_{x_i^*(s)}^{\infty}(y - x_i^*(s))dF(y). \tag{8}$$

Let

$$g_i(s;x) \equiv \frac{s(1-r-q)}{1-s(1-q)}\left(1 + \int_x^{\infty}(y-x)dF(y)\right) - \frac{1-s}{1-s(1-q)}\sum_{j=i+1}^{b^*-1} x_j^*(s).$$

Then $x_i^*(s)$ is the unique fixed point of $g_i(s;x)$, that is, $x_i^*(s) = g_i(s; x_i^*(s))$. First observe that $g_{b^*-1}(s_2;x) > g_{b^*-1}(s_1;x)$ for all x, which implies that $x_{b^*-1}^*(s_1) \leq x_{b^*-1}^*(s_2)$. We use this as the induction basis and show that $x_i^*(s_1) \leq x_i^*(s_2)$ for $i = b^* - 2, b^* - 3, ..., 0$ using backward induction. Suppose that $x_i^*(s_1) \leq x_i^*(s_2)$. Then, because both $g_i(s;x)$ and $g_i(s_2;x) - g_i(s_1;x)$

are decreasing in x, we have that:

$$\begin{aligned}
&\frac{1-s_2}{1-s_2(1-q)}x_i^*(s_2) - \frac{1-s_1}{1-s_1(1-q)}x_i^*(s_1) \\
&< \frac{1-s_1}{1-s_1(1-q)}(x_i^*(s_2) - x_i^*(s_1)) \\
&\leq x_i^*(s_2) - x_i^*(s_1) \\
&= g_i(s_2; x_i^*(s_2)) - g_i(s_1; x_i^*(s_1)) \\
&\leq g_i(s_2; x_i^*(s_2)) - g_i(s_1; x_i^*(s_2)) \\
&\leq g_i(s_2; x) - g_i(s_1; x) \text{ for all } x \in [0, x_i^*(s_2)]
\end{aligned}$$

This implies that $g_{i-1}(s_1; x) \leq g_{i-1}(s_2; x)$ for all $x \in [0, x_i^*(s_2)]$. From Proposition 1 we know that $x_{i-1}^*(s_1), x_{i-1}^*(s_2) \in [0, x_i^*(s_2)]$. We therefore conclude that $x_{i-1}^*(s_1) \leq x_{i-1}^*(s_2)$, which shows the induction step. Thus, we have shown that $B^*(s)$ is increasing in s.

We now show that $B^*(s) \to \infty$ as $s \uparrow 1$. First consider the case that $q = 0$. We will show that for every finite N there exists an $s_N < 1$ such that $B^*(s_N) \geq N$. We fix the cutoff to be $b^* = N$ and consider the thresholds $x_i^*(s)$ for $i \in \{0, 1, 2, ..., b^* - 1\}$ that represent the temptee's best response. It suffices to show that there exists $s < 1$ such that $F(x_i^*(s)) \geq 1/2$ for $i \in \{0, 1, 2, ..., b^* - 1\}$. Since $q = 0$, (8) can be written as

$$\frac{1-s}{s(1-r)} \sum_{j=i}^{b^*-1} x_j^*(s) = 1 + \int_{x_i^*(s)}^{\infty} (y - x_i^*(s)) dF(y). \tag{9}$$

We will use backward induction to show that for $i = b^* - 1, b^* - 2, ...$ we have that (i) $F(x_i^*(s)) \geq 1/2$ for sufficiently large s and (ii) $\frac{1-s}{s(1-r)} \sum_{j=i}^{b^*-1} x_j^*(s) \to 1$ as $s \uparrow 1$.

Basis: We start with $i = b^* - 1$. From (9) we have that

$$x_{b^*-1}^*(s) = \frac{s(1-r)}{1-s} \left(1 + \int_{x_i^*(s)}^{\infty} (y - x_i^*(s)) dF(y) \right) \geq \frac{s(1-r)}{1-s} \to \infty \text{ as } s \uparrow 1,$$

which implies that (i) holds for $i = b^* - 1$ and that

$$\int_{x_{b^*-1}^*(s)}^{\infty} (y - x_{b^*-1}^*(s)) dF(y) \to 0 \text{ as } s \uparrow 1.$$

The previous limit together with (9) imply that (ii) holds for $i = b^* - 1$.

Inductive Step: Suppose that (ii) holds for $i = b + 1$. If the distribution F has unbounded support then (9) and the inductive hypothesis imply that $x_b^*(s) \to \infty$ as $s \uparrow 1$; it follows that both (i) and (ii) hold for $i = b$. We now consider the case that the distribution F is supported on a bounded interval and let $M \equiv \min\{x : F(x) = 1\}$ be the maximum point in the support. Then it follows from (9) and the inductive hypothesis that $x_b^*(s) \uparrow M$ as $s \uparrow 1$, so (ii)

trivially holds for $i = b$. Then, if F is continuous at M we have that $F(x_b^*(s)) \uparrow 1$ as $s \uparrow 1$, i.e., (i) holds for $i = b$. On the other hand, if F has a jump at M then $F(x_b^*(s)) \to \lim_{x \to M^-} F(x)$ as $s \uparrow 1$. We observe that if $B^*(s) \geq 1$ for some s, it must be that $\lim_{x \to M^-} F(x) \geq 1/2$ and therefore (i) holds for $i = b$.

Thus, we have shown that $\lim_{s \uparrow 1} B^*(s) = \infty$ when $q = 0$. To conclude the proof, it suffices to show (ii), since (ii) implies that $\lim_{s \uparrow 1} B^*(s) = \infty$ when $q > 0$.

We now consider the general case where q is not restricted to be equal to 0. Define

$$c(s) \equiv 1 + \int_m^\infty (y - m) dF(y) - \frac{1 - s(1-q)}{s(1-r-q)} m.$$

Then, (8) implies that $F(x_i^*(s) \geq 1/2$, or equivalently $x_i^*(s) \geq m$, if and only if

$$c(s) \geq \frac{1-s}{s(1-r-q)} \sum_{j=i+1}^{b^*-1} x_j^*(s).$$

If the maximum equilibrium cutoff is $B^*(s)$, then it must that $x_i^*(s) \geq m$ for $i = 0, 1, ..., B^*(s) - 1$ and

$$\frac{1-s}{s(1-r-q)} \sum_{j=1}^{B^*(s)-1} x_j^*(s) \leq c(s) < \frac{1-s}{s(1-r-q)} \sum_{j=0}^{B^*(s)-1} x_j^*(s).$$

(If the second inequality did not hold, then the maximum equilibrium cutoff would be strictly greater than $B^*(s)$.) We then have that:

$$\frac{1}{1-s} \frac{(1-r-q)sc(s)}{\max_j x_j^*(s)} < B^*(s) < \frac{1}{1-s} \frac{(1-r-q)sc(s)}{\min_j x_j^*(s)} \leq \frac{1}{1-s} \frac{(1-r-q)sc(s)}{m} \tag{10}$$

We observe that $c(s)$ is increasing in s and upper-bounded by $1 + \int_m^\infty (y - m) dF(y) < \infty$. Moreover, if $B^*(s) \geq 1$ then $c(s) > 0$. Choose some $\tilde{s} < 1$ such that $B^*(\tilde{s}) \geq 1$ (such an $\tilde{s}$ exists by the assumption of the proposition). Observe that if $q > 0$, for every $i \in \{0, 1, 2, ..., B^*(s) - 1\}$

$$\begin{aligned} 0 \leq x_i^*(s) \leq x_{B^*(s)-1}^*(s) &= \frac{s(1-r-q)}{1-s(1-q)} \left(1 + \int_{x_{B^*(s)-1}^*(s)}^\infty (y - x_{B^*(s)-1}^*(s)) dF(y) \right) \\ &< \frac{1-r-q}{q} \left(1 + \int_0^\infty y dF(y) \right), \end{aligned}$$

where the second inequality follows from Proposition 1 and the equality from (8). This implies that each $x_j^*(s)$ is upper bounded by some constant (independent of s) when $q > 0$. Set

$$c_1 = \frac{q\tilde{s}c(\tilde{s})}{1 + \int_0^\infty y dF(y)}; \; c_2 = \frac{(1-r-q)c(1)}{m}$$

Note that $0 < c_1 < c_2$, because $c(\tilde{s}) > 0$. (10) implies that

$$\frac{c_1}{1-s} \leq B^*(s) \leq \frac{c_2}{1-s}$$

for $s > \tilde{s}$, which concludes the proof.

□

Proof of Proposition 3: Throughout the proof, we use the notation $B^*(r)$ and $x_i^*(r)$ to denote the dependence on r. We assume that s, q and F are fixed. Consider some fixed cutoff b^* and suppose that $r_1 < r_2$. We will show that $x_i^*(r_1) \geq x_i^*(r_2)$ for $i = 0, 1, ..., b^* - 1$, which implies that $B^*(r_1) \geq B^*(r_2)$.

From (2) and (3) we have that that for any $i \leq b^* - 1$, $x_i^*(r)$ is given by the solution to the following equation

$$(1 - s + sq)x = g_i(r; x), \tag{11}$$

where we define

$$g_i(r; x) \equiv s(1 - r - q)\left(1 + \int_x^\infty (y - x)dF(y)\right) - (1 - s)\sum_{j=i+1}^{b^*-1} x_j^*(r).$$

(This is equivalent to (8) from the proof of Proposition 2.) Note that $g_i(r; x)$ is decreasing in x.

We observe from (11) that if $g_i(r_1; x_i^*(r_1)) \geq g_i(r_2; x_i^*(r_2))$ then $x_i^*(r_1) \geq x_i^*(r_2)$. We will show that for each $i \leq b^*-1$ there exists $A_i \geq \max\{x_i^*(r_1), x_i^*(r_2)\}$ such that $g_i(r_1; x) \geq g_i(r_2; x)$ for all $x \in [0, A_i]$. The proof uses backward induction starting with $i = b^* - 1$.

For $i = b^* - 1$, the induction hypothesis trivially holds, since $g_{b^*-1}(r_1; x) - g_{b^*-1}(r_2; x) \geq 0$ for all x.

Now suppose that the induction hypothesis holds for some $i \leq b^* - 1$. Then we have that $x_i^*(r_1) \geq x_i^*(r_2)$. We will show that the induction hypothesis holds for $i - 1$ with $A_{i-1} = x_i^*(r_1)$. This will conclude the proof since we know from Proposition 1 that $x_{i-1}^*(r) \leq x_i^*(r)$. We have that

$$\begin{aligned}(1 - s)(x_i^*(r_1) - x_i^*(r_2)) &\leq (1 - s + sq)(x_i^*(r_1) - x_i^*(r_2)) \\ &= g_i(r_1; x_i^*(r_1)) - g_i(r_2; x_2^*(r_2)) \\ &\leq g_i(r_1; x_i^*(r_1)) - g_i(r_2; x_1^*(r_2)) \\ &\leq g_i(r_1; x) - g_i(r_2; x) \text{ for all } x \leq x_i^*(r_1) \end{aligned} \tag{12}$$

The first inequality holds because $sq \geq 0$ and $x_i^*(r_1) \geq x_i^*(r_2)$. The equality follows from (11). The last two inequalities hold because both $g_i(r; x)$ and $g_i(r_1; x) - g_i(r_2; x)$ are decreasing in x. We therefore have that for $x \leq x_i^*(r_1)$,

$$g_{i-1}(r_1; x) = g_i(r_1; x) - (1 - s)x_i^*(r_1) \geq g_i(r_2; x) - (1 - s)x_i^*(r_2) = g_{i-1}(r_2; x),$$

where the equalities follow from the definition of g_i and the inequality from (12). This concludes the proof. □

Proof of Proposition 4: We first show that $b_D^* = B^*$. Let $u(b, b^*)$ be equal to $v(b)$ when the cutoff b^* is used. We observe that $u(b, b^*)$ only depends on the difference $b^* - b$ (given the same F, s, r and q), and is increasing in $b^* - b$. Thus, $u(0, b^*)$ is maximized when b^* is maximized.

Now that we have shown that $b_C^* \leq b_D^*$, how about the socially optimal equilibrium $b_S^*(\alpha)$ which optimizes the weighted return of trusters and temptees? Because the return for temptees decreases when b^* decreases, it is not possible that $b_S^*(\alpha) < b_C^*$, because both players would be better off with $b_S^*(\alpha) = b_C^*$. It is also not possible that $b_S^*(\alpha) > b_D^*$, because $b^*(D)$ is the highest cutoff possible in the equilibrium set. Then we have $b_C^* \leq b_S^*(\alpha) \leq b_D^*$ for all $\alpha \geq 0$. To conclude the proof we note that a truster gets zero payoff when $b^* = 0$; thus if $B^* \geq 1$, the truster strictly prefers $b^* = 1$ to $b^* = 0$. Therefore, $1 \leq b_C^* \leq b_S^*(\alpha) \leq b_D^* = B^*$ for all $\alpha \geq 0$. □

Proof of Proposition 5: Throughout the proof, we assume that s, r, q are fixed and use the notation $B^*(A)$ and $x_i^*(A)$ to denote the dependence on A. Define $y_i^*(A) \equiv x_i^*(A)/A$, which represents the probability of rewarding at i black marks. Consider some fixed cutoff b^* and suppose that $A_1 < A_2$. We will show that $y_i^*(A_1) \geq y_i^*(A_2)$ for $i = 0, 1, ..., b^* - 1$, which implies that $B^*(A_1) \geq B^*(A_2)$.

Since the temptation to betray is uniformly distributed on $[0, A]$, we have that $F(x) = x/A$ for $x \in [0, A]$. Then, from (2) and (3) we have that that for any $i \leq b^* - 1$

$$\frac{(1 - s + sq)}{s(1 - r - q)} y_i^*(A) = k_i(A) + \frac{1}{2}(1 - y_i^*(A))^2, \tag{13}$$

where

$$k_i(A) \equiv \frac{1}{A} - \frac{1 - s}{s(1 - r - q)} \sum_{j=i+1}^{b^*-1} y_j^*(A).$$

We observe from 13 that if $k_i(A_1) \geq k_i(A_2)$ then $y_i^*(A_1) \geq y_i^*(A_2)$. To conclude the proof we will show that $k_i(A_1) \geq k_i(A_2)$ for $i = b^* - 1, b^* - 2, ..., 0$ using backward induction. The induction basis trivially holds, because $k_{b^*-1}(A) = 1/A$.

Now suppose that $k_i(A_1) \geq k_i(A_2)$ holds for some $i \leq b^* - 1$. Then, $y_i^*(A_1) \geq y_i^*(A_2)$. Moreover, by the definition of $k_i(A)$, we have that

$$k_{i-1}(A_1) - k_{i-1}(A_2) = k_i(A_1) - k_i(A_2) - \frac{1 - s}{s(1 - r - q)}(y_i^*(A_1) - y_i^*(A_2)).$$

We will show that $k_i(A_1) - k_i(A_2) \geq \frac{1-s}{s(1-r-q)}(y_i^*(A_1) - y_i^*(A_2))$ which implies

that $k_{i-1}(A_1) \geq k_{i-1}(A_2)$. Indeed, from (13),

$$\begin{aligned} k_i(A_1) - k_i(A_2) &= \frac{1-s+sq}{s(1-r-q)}(y_i^*(A_1) - y_i^*(A_2)) + \frac{1}{2}((1-y_i^*(A_2))^2 - (1-y_i^*(A_1))^2 \\ &\geq \frac{1-s+sq}{s(1-r-q)}(y_i^*(A_1) - y_i^*(A_2)) \\ &\geq \frac{1-s}{s(1-r-q)}(y_i^*(A_1) - y_i^*(A_2)) \end{aligned}$$

where the first inequality holds because $y_i^*(A_1) \geq y_i^*(A_2)$ and the second because $q \geq 0$. This concludes the proof. □

Proof of Proposition 6: By the definition of second-order stochastic dominance, we have that $\int h(y)dF_1(y) \geq \int h(y)dF_2(y)$ for every concave function h (Mas-Colell et al., 1995). Setting $h(y) = -\max\{y - x, 0\}$, which is concave in y, we conclude that for every x,

$$\int_x^\infty (y-x)dF_2(y) \geq \int_x^\infty (y-x)dF_1(y). \tag{14}$$

Throughout the proof, we assume that s, r, q are fixed and use the notation $x_i^*(F_j)$ to denote the dependence on the distribution. Given some cutoff b^*, it suffices to show that $x_i^*(F_2) \geq x_i^*(F_1)$ for $i = 0, 1, ..., b^* - 1$. Similarly to the other proofs on comparative statics (e.g., the one of Proposition 2), let

$$g_i(F; x) \equiv \frac{s(1-r-q)}{1-s(1-q)}\left(1 + \int_x^\infty (y-x)dF(y)\right) - \frac{1-s}{1-s(1-q)}\sum_{j=i+1}^{b^*-1} x_j^*(s)$$

so that $x_i^*(F)$ is the unique fixed point of $g_i(F; x)$. Observe that g_i is decreasing in x. From (14) it follows that $g_{b^*-1}(F_2; x) \geq g_{b^*-1}(F_1; x)$ for all x and therefore $x_{b^*-1}^*(F_2) \geq x_{b^*-1}^*(F_1)$.

Using backward induction, we will now show for $i \leq b^* - 1$ that if $x_i^*(F_2) \geq x_i^*(F_1)$ then $g_{i-1}(F_2; x) \geq g_{i-1}(F_1; x)$ for all $x \leq x_i^*(F_2)$, which in turn implies that $x_{i-1}^*(F_2) \geq x_{i-1}^*(F_1)$. By the Leibniz integral rule we have that the derivative of $g_i(F_j; x)$ with respect to x is equal to $-(1 - F_j(x))$. Thus, if $z > x > m$, we have that

$$g_i(F_1; x) - g_i(F_1; z) = \int_x^z (1 - F_1(y))dy \leq \int_x^z (1 - F_2(y))dy = g_i(F_2; x) - g_i(F_2; z), \tag{15}$$

where the inequality follows from the assumption that $F_1(y) \geq F_2(y)$ for $y \geq m$. Therefore,

$$\begin{aligned} \frac{1-s}{1-s+sq}(x_i^*(F_2) - x_i^*(F_1)) &\leq x_i^*(F_2) - x_i^*(F_1) \\ &= g(F_2; x_i^*(F_2)) - g(F_1; x_i^*(F_1)) \\ &\leq g(F_2; x_i^*(F_2)) - g(F_1; x_i^*(F_2)) \\ &\leq g_i(F_2; x) - g_i(F_1; x) \end{aligned}$$

for $x \in [m, x_i^*(F_2)]$. The first inequality holds because $q \geq 0$, the equality by the definition of g, the second inequality because g is decreasing and $x_i^*(F_2) \geq x_i^*(F_1)$, and the last inequality follows from (15). The previous inequality together with the fact that $g_{i-1}(F_j; x) = g_i(F_j; x) - (1-s)/(1-s+sq)x_i^*(F_j)$ implies that $g_{i-1}(F_2; x) \geq g_{i-1}(F_1; x)$ for all $x \leq x_i^*(F_2)$. This concludes the proof. □

Proof of Proposition 7: Suppose that $b^*(n) > b^*(n+1)$ and consider a temptee who has $b^*(n+1)$ black marks and n interactions. Trusters will trust the temptee at this reputation, because $b^*(n) > b^*(n+1)$. However, the temptee knows that whatever she does in this period, she will not trusted in the next period. Thus, it is optimal for her to betray when her temptation to do so is positive. But then trusters have no reason to trust. So it must be that $b^*(n) \leq b^*(n+1)$, which shows (i).

To show (ii), we first show that at a pure equilibrium,

$$b^*(n) \leq \frac{\log((1+\mathbb{E}X)/m) + \log(1/(1-s))}{\log(1/s)} \equiv K,$$

where $\mathbb{E}X \equiv \int x dF(x)$ is the expected value of the temptation to betray.

Consider an MPE where the trusters' strategy is given by the cutoff function $b^*(n)$ and the temptees' strategy is given by the set of thresholds $\{x_{b,n}^*\}$ and suppose that there exists some n with $b^*(n) > K$. A necessary condition for this to be an equilibrium is that $x_{0,n}^* \geq m$. We show that a temptee is better off using some strategy $\{x_{b,n}\}$ with $x_{0,n} < m$. In particular, consider a strategy with $x_{b,n'} = x_{b,n'}^*$ for $n' > n$. Suppose that a temptee's current reputation is $(0, n)$ and let x be her current temptation to betray. If the temptee betrays now, her current payoff will increase by x. We next upper bound the amount that the temptee will lose by betraying now. First observe that the earlier time that the temptee may be expelled is K periods later. This is a very conservative estimate, because the temptee will probably not get a black mark in every period and $b^*(n)$ may be increasing. We next observe that the temptee misses at most $1 + \mathbb{E}X$ in expectation for each period after she is expelled. Considering that the temptee only survives with probability s in every period, in total she misses at most $(1 + \mathbb{E}X)/(1-s)$. But this payment is at least K periods away, so she discounts it by at most s^K. Thus, if the temptee betrays now, her future payment will decrease by at most $(1 + \mathbb{E}X)s^K/(1-s)$. We conclude that the temptee will be better off betraying if

$$x > \frac{s^K}{1-s}(1 + \mathbb{E}X).$$

Because of the way we defined K, note that $(1 + \mathbb{E}X)s^K/(1-s) < m$. So the temptee is better off using $x(0, n) < m$.

Thus, for any given problem there exists some constant (independent of the number of interactions n) upper bound on $b^*(n)$. We already know that $b^*(n)$ is non-decreasing, so there must exist a $\bar{n}$ and a $\bar{b}$ such that $b^*(n) = \bar{b}$ for $n \geq \bar{n}$.

Then, after $\bar{n}$ interactions the exact number of interactions does not affect the cutoff (which is constant). Without loss of generality, we can restrict the set of possible reputation scores to $\{(b,n) : b \leq b^*(n), n < \bar{n}\} \cup \{(b,\bar{n}) : b \leq \bar{b}\}$, which is a finite set. For $n = \bar{n}$, the equilibrium conditions are $v(\bar{b},\bar{n}) = 0$ and,

$$v(b,\bar{n}) = 1 + s\cdot(1-q)\cdot v(b,\bar{n}) + s\cdot q\cdot v(b+1,\bar{n}) + \int_{x^*_{b,\bar{n}}}^{\infty}(y - x^*_{b,\bar{n}})dF(y) \text{ for } b < \bar{b};$$

$$x^*_{b,\bar{n}} = s(1-r-q)\cdot(v(b,\bar{n}) - v(b+1,\bar{n})) \text{ for } b < \bar{b}.$$

Observe that the variable $\bar{n}$ does not affect the recursion in the previous equations, since it appears in all the terms. More importantly, if we ignore $\bar{n}$ these are exactly equations (1), (2), and (3), that is, the equations we had under the Basic Black Mark Mechanism, where a temptee's reputation consisted only of the number of black marks. This observation implies that $\bar{b} \leq B^*$, and also $b^*(n) \leq B^*$, which concludes the proof for (ii). □

Proof of Proposition 8: Suppose that every time the temptee receives a black, she is not trusted for T consecutive periods. Let V be the maximum infinite horizon discounted payoff to the temptee if she will be trusted in the current period. Then,

$$V = 1 + \int \max\{x + srV + s^{T+1}(1-r)V, s(1-q)V + s^{T+1}qV\}dF(x),$$

since is the temptee gets a black mark now she will not be trusted until $T+1$ periods later. Equivalently, $(1 - s(1-q) - s^{T+1}q)V = 1 + \int_{x^*}^{\infty}(y - x^*)dF(x)$, where $x^* \equiv s(1-r-q)(1-s^T)V$ is the threshold above which it is optimal for the temptee to betray. It then follows that

$$x^*\frac{1 - s + sq(1-s^T)}{s(1-s^T)(1-r-q)} = 1 + \int_{x^*}^{\infty}(y - x^*)dF(x).$$

Observe that x^* is increasing in T. In order to have an equilibrium, we need that $x^* \geq m$, so that the temptee rewards with probability greater or equal to $1/2$ whenever she is trusted. Observe that the LHS is increasing and continuous in x^*, whereas the RHS is decreasing and continuous in x^*. Thus, $x^* \geq m$ if and only if T is large enough so that

$$m\frac{1 - s + sq(1-s^T)}{s(1-s^T)(1-r-q)} \leq 1 + \int_{m}^{\infty}(y - m)dF(x).$$

The latter is equivalent to $T \geq T^*(s,r,q,F)$, where T^* is given in the statement of the proposition. This concludes the proof. □

Appendix B: Monitoring and the Temptee's Maximum Payoff

Figure 3 shows the temptee's maximum payoff, i.e., her payoff when the maximum equilibrium cutoff is used, as a function of r for two examples. Each jump corresponds to an decrease in the maximum equilibrium cutoff (per Proposition 3). In each interval between two consecutive jumps, the maximum equilibrium cutoff is constant. Thus, between each two consecutive jumps, the temptee's payoff increases monotonically in r.

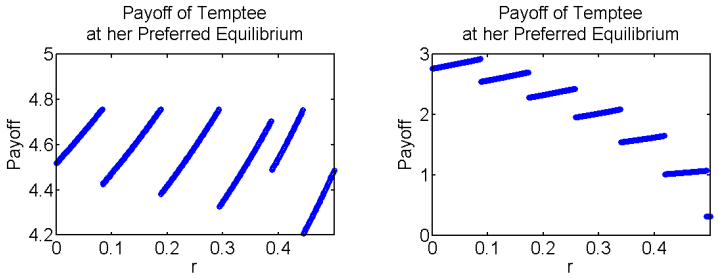

Figure 3: Temptee's normalized expected payoff at her preferred equilibrium as a function of r when $s = 0.98$ and the temptation to betray is uniformly distributed on $[0, 30]$. In the left plot, $q = 0$; in the right $q = 0.1$.

Figure 4 shows the temptee's maximum payoff as a function of q for two examples. Each jump corresponds to a change in the maximum equilibrium cutoff. In most cases, a jump corresponds to a decrease of the maximum equilibrium cutoff and thus the temptee's expected payoff decreases. However, it is possible that an increase in q leads to an increase in the maximum equilibrium cutoff, implying that the temptee's payoff increases. This is the case for the first jump of the right plot of Figure 3, which corresponds to Example 1. In each interval between two consecutive jumps the maximum equilibrium cutoff is constant and therefore the temptee's payoff decreases monotonically in q.

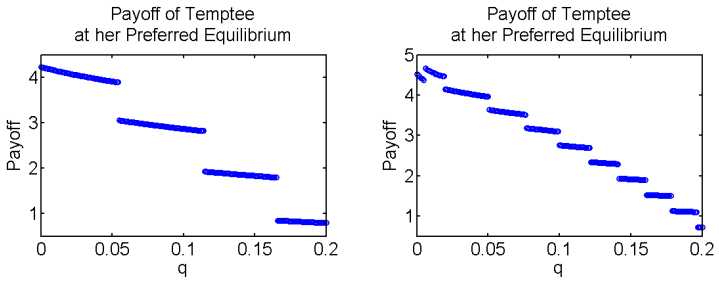

Figure 4: Temptee's normalized expected payoff at her preferred equilibrium as a function of q when $r = 0$ and the temptation to betray is uniformly distributed on $[0, 30]$. In the left plot, $s = 0.95$; in the right $s = 0.98$.

Appendix C: Mixed-Strategy Equilibria

Our analysis in this paper considers pure-strategy equilibria. In this appendix, we characterize mixed-strategy equilibria and discuss why many of the insights of the paper still hold for mixed-strategy equilibria. We focus on mixed strategies for trusters, where a truster trusts a temptee with some probability that depends on her reputation. We consider the same model as in section 2 and allow for imperfect monitoring, assuming that a temptee's reputation does not change in periods that she did not interact with a truster because she was not trusted by the truster she was matched to.

We start with the Basic Black Mark Mechanism. We do not consider mixed strategies for temptees here.[11] Thus, a temptee's strategy shows (as in the pure equilibrium case) whether she rewards as a function of her reputation and her temptation to betray. Then, a temptee's strategy can be represented by a set of thresholds $\{x_b^*, b = 0, 1, ...\}$.

A truster's mixed strategy represents the probability he trusts a temptee as a function of her reputation. Thus, a truster's strategy can be summarized by $\{p_b^*, b = 0, 1, ...\}$, where p_b^* is the probability that the truster trusts a temptee when her reputation consists of b black marks.

The probabilities $\{p_b^*, b = 0, 1, ...\}$ and the thresholds $\{x_b^*, b = 0, 1, ...\}$ constitute an MPE if the following conditions are satisfied:

(a) There exists a function $v : \mathbb{N} \to \mathbb{R}$ such that

$$(1 - s(1 - q \cdot p_b^*))v(b) = p_b^* \left(1 + s \cdot q \cdot v(b+1) + \int_{x_b^*}^{\infty} (y - x_b^*) dF(y) \right),$$

where $x_b^* \equiv s \cdot (1 - r - q) \cdot (v(b) - v(b+1))$

(b) $p_b^* = 1$ if $F(x_b^*) > 1/2$; $p_b^* \in [0, 1]$ if $F(x_b^*) = 1/2$; $p_b^* = 0$ if $F(x_b^*) < 1/2$.

Condition (a) guarantees that the set of thresholds $\{x_b^*, b = 0, 1, ...\}$ is a best reponse of temptees to the probabilities $\{p_b^*, b = 0, 1, ...\}$, and condition (b) guarantees that the set of probabilities $\{p_b^*, b = 0, 1, ...\}$ is a best reponse of trusters to the thresholds $\{x_b^*, b = 0, 1, ...\}$. The derivation of these equilibrium conditions parallels that of the derivation for pure equilibria in section 3.1. Pure equilibria are a special case of the mixed equilibria discussed here with $p_b^* \in \{0, 1\}$.

Given an equilibrium $\{(x_b^*, p_b^*), b = 0, 1, ...\}$, we define $b^* \equiv \min\{b : p_b^* = 0\}$. In words, b^* is the cutoff (in terms of the number of black marks) at which trusters no longer trust a temptee. We already know that a finite cutoff is

[11] The results can be easily extended to consider mixed strategies for temptees; such strategies would only be relevant if the temptation to betray is drawn from a discrete distribution. In particular, a temptee with b black marks would only randomize if her temptation to betray is exactly equal to x_b^*. If F is continuous, the temptation to betray will be equal to x_b^* with zero probability, and thus what a temptee does at x_b^* does not affect the players' payoffs. If F is discrete, a temptee could mix optimally at most at one point (i.e., x_b^*).

associated with every pure equilibrium. The following lemma shows that this is also the case for mixed equilibria.

Lemma 1. *For every MPE* $\{(x_b^*, p_b^*), b = 0, 1, ...\}$ *there exists a finite cutoff* b^* *such that* $p_b^* > 0$ *for* $b < b^*$ *and* $p^*(b^*) = 0$.

Proof. Suppose that $p_b^* > 0$ for all b. Then,

$$v(0) = \frac{1}{s(1-r-q)} \sum_{b=0}^{\infty} x_b^* \geq \frac{1}{s(1-r-q)} \sum_{b=0}^{\infty} m = \infty,$$

because $m > 0$.

On the other hand, the equilibrium conditions imply that

$$(1-s)v(0) = p^*(0) \left(1 + \int_{x_b^*}^{\infty} (y - x_b^*) dF(y) - \frac{q}{1-r-q} x_b^* \right),$$

which cannot hold if $v(0) = \infty$, since F has a finite mean and $p^*(0) \leq 1$. We conclude that there always exists a finite cutoff. □

Proposition 1 shows that for pure equilibria, x_b^* is strictly increasing and convex in b for b in $\{0, 1, ..., b^* - 1\}$. This is not the case for mixed equilibria in general, since at a mixed equilibrium we may have $x_{b^*-1}^* = m$ and $x_b^* > m$ for $b < b^* - 1$. However, the insights of Proposition 1 still hold if we do not consider the b's for which $x_b^* = m$.

Note that the maximum equilibrium cutoff among all mixed-strategy MPE is at least as large as B^*. Thus, the insight of Proposition 2 still applies when we consider mixed-strategy equilibria in the sense that the maximum equilibrium cutoff among all mixed-strategy MPE approaches $+\infty$ as $s \uparrow 1$. Moreover, if $q > 0$, the maximum equilibrium cutoff among all mixed-strategy MPE scales at least as fast as $1/(1-s)$ when s is close to 1.

Similarly to section 3.4, we can define a Markov chain that tracks how the number of black marks of any temptee evolves. The only difference is that the probabilities p_b^* will also influence the transitions; in particular, the number of black marks of a temptee remains the same if she is not trusted now and survives to the next period. Then, in stationarity, a truster's expected payoff from every temptee that he may interact with in a given period is equal to $\sum_{b:p_b^*=1} \pi_b (2F(x_b^*) - 1)$, where π is the stationary distribution of the Markov chain. A truster's expected payoff is 0 if $p_b^* \in (0, 1)$; otherwise he wouldn't be indifferent between trusting and not trusting.

Having defined a truster's expected payoff at a given equilibrium, we can then consider which equilibrium is preferred by each side of the market. As in Proposition 4, a temptee prefers equilibria where she is trusted more. In particular, given two MPE $\{(x_b^*, p_b^*), b = 0, 1, ...\}$ and $\{(\tilde{x}_b^*, \tilde{p}_b^*), b = 0, 1, ...\}$ such that $p_b^* \geq \tilde{p}_b^*$ for all b, a temptee prefers the former MPE. Of course this property does not produce an ordering among all MPE. The following example introduces a specific MPE which, when it exists, prolongs trust and is preferred by the temptee compared to any pure MPE.

Example 7. *A Dominant Extend Equilibrium is an MPE such that* $p_b^* = 1$ *for* $b = 0, 1, ..., B^*-1$*;* $p^*(B^*) \in (0, 1)$*; and* $p^*(B^*+1) = 0$*. A temptee always prefers a dominant extend equilibrium to any pure equilibrium, because her expected payoff is maximized the longer she can expect to be trusted. Thus, rather than being expelled for sure at* $b = B^*$*, she would prefer to have a probabilistic chance there, with expulsion at* $b = B^* + 1$*.*

As far as more general reputation mechanisms are concerned, we note that even within the class of mixed-strategy MPE, revealing the number of interactions in addition to the number of black marks does not prolong trust compared to the Basic Black Mark Mechanism (as in Proposition 7). Finally, Proposition 8 still holds for mixed-strategy MPE, since not trusting at all is a more severe punishment than trusting with some probability.

Appendix D: Multiple Temptee Types

This paper has focused on a pure moral hazard setting, where all temptees are the same in terms of payoff structure and self-control. Attention has thus been strictly on inducing good behavior despite temptation. This appendix allows for multiple types. Hence adverse selection rears its ugly head alongside moral hazard. Moreover, a truster will appropriately update his belief that a temptee is of a particular type depending on her reputation score. For simplicity, we focus on the Basic Black Mark Mechanism and pure MPE.

There are k types of temptees. Each type is defined by its distribution of temptations. Let F_i denote the distribution of the temptation to betray for type i. For a given cutoff b^*, we can compute the best response for type i (as in section 3.1), which consists of a threshold $x_{i,b}^*$ for each $b < b^*$. Each $x_{i,b}^*$ is increasing and convex in b for $b \in \{0, ..., b^* - 1\}$ (this can be shown with the arguments used in the proof of Proposition 1). Moreover, the probability that a player of type i betrays is decreasing in b for $b \in \{0, ..., b^* - 1\}$. In words, the more black marks to date, the less likely a temptee of a particular type is to betray.

Which cutoffs b^* can arise in equilibrium in a world that allows for multiple types, hence adverse selection as well as moral hazard? A cutoff b^* can be an equilibrium if the truster's expected payoff from trusting is nonnegative for all $b \in \{0, 1, ..., b^* - 1\}$. Utilizing the framework of section 3.1, we first compute the maximum cutoff that can arise at a (pure) equilibrium when the population is comprised solely of temptees of type i. Denote this maximum cutoff as B_i^*. Then, for any $b^* \leq \min_i B_i^*$, we have an equilibrium.[12]

In a world of multiple types, some will benefit from the presence of others, while others will lose. If the proportion of temptees of type j with $B_j^* = \min_i B_i^*$

[12]Since $b^* \leq \min_i B_i^*$, we have that at each $b < b^*$ each type rewards with probability greater or equal to 0.5. This implies that a randomly chosen temptee with b black marks rewards with probability greater or equal to 0.5 when $b < b^*$. This in turn implies that the truster is better off trusting a temptee with $b < b^*$ black marks.

is relatively small, there can be equilibria with $b^* > B_j^*$. Conversely, temptees of type h, where $B_h^* = \max_i B_i^*$, if they are not numerous, may find some of their forgiving (high b^*) equilibria disappear. The comparative statics results of section 3.3 apply for each B_i^*.

The insights of section 3.4 still hold in a world with multiple types. Temptees still prefer the equilibrium with the maximum cutoff (this follows from the proof of Proposition 4), whereas trusters may prefer a smaller cutoff. However, the truster's expected payoff is now given by a more complex formula than the one of section 3.4, a formula that attends to the proportions of different types in the system.

Propositions 5 and 6 give conditions on two distributions F_i and F_j under which $B_i^* \geq B_j^*$, that is, type i has a larger maximum equilibrium cutoff than type j. We also note that it is possible that type i is more likely (than type j) to betray with b black marks, but less likely with b' black marks.

Multiple types introduce great richness to a world with black mark reputations. In particular, the distribution of types may shift over time with the number of black marks, a classic outcome with adverse selection. The shifting happens as the distribution of black marks converges to the steady state for every type of temptees. If initially all temptees have zero black marks and the number of temptees of each type is equal, initially a randomly selected temptee with zero black marks will belong to each type with probability 0.5. But as time goes by, a randomly selected temptee with zero black marks is more likely to be of the type that betrays with a smaller probability.

In real world play, there are sure to be multiple types with temptations that vary over time. Temptees will exhibit both adverse selection and moral hazard. Trusters will seek to deter their bad behavior with an ultimate refusal to play once black marks reach a certain level.

References

Abreu, D., Pearce, D., Stacchetti, E., 1990. Toward a theory of discounted repeated games with imperfect monitoring. Econometrica 58, 1041–1063.

Anderson, N.H., 1981. Foundations of information integration theory. New York: Academic Press.

Athey, S., Bagwell, K., 2001. Optimal collusion with private information. The RAND Journal of Economics 32, 428–465.

Athey, S., Bagwell, K., Sanchirico, C., 2004. Collusion and price rigidity. The Review of Economic Studies 71, 317–349.

Barlo, M., Carmona, G., Sabourian, H., 2009. Repeated games with one-memory. Journal of Economic Theory 144, 312–336.

Bendor, J., Mookherjee, D., 1990. Norms, third-party sanctions, and cooperation. Journal of Law, Economics, and Organization 6, 33–63.

Bolton, G., Greiner, B., Ockenfels, A., 2009. Engineering trust - Reciprocity in the production of reputation information. Working Paper Series in Economics 42. University of Cologne, Department of Economics.

Bush, G.W., 2004. Address before a joint session of the congress on the state of the union. Public Papers of the Presidents of the United States.

Cabral, L., Hortacsu, A., 2010. Dynamics of seller reputation: Theory and evidence from eBay. J. of Industr. Econom. 58, 54–78.

Cai, H., Jin, G., Liu, C., Zhou, L.A., 2011. How to Promote Trust: Theory and Evidence from China. Working Paper.

Chwelos, P., Dhar, T., 2008. Differences in "truthiness" across online reputation mechanisms. Working Paper, Sauder School of Business.

Clinton, B., 1994. Interview with Peggy Wehmeyer. World News Tonight. ABC.

Colea, H.L., Kocherlakota, N.R., 2005. Finite memory and imperfect monitoring. Games and Economic Behavior 53, 59–72.

Crawford, V.P., Sobel, J., 1982. Strategic information transmission. Econometrica 50, 1431–1451.

Dalea, D.J., Morganb, J., Rosenthal, R.W., 2002. Coordination through reputations: A laboratory experiment. Games and Economic Behavior 38, 52–88.

Dellarocas, C., 2005. Reputation mechanism design in online trading environments with pure moral hazard. Inform. Systems Res. 16, 209–230.

Dellarocas, C., Wood, C., 2008. The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. Management Sci. 54, 460–476.

Doraszelski, U., Escobar, J.F., 2012. Restricted feedback in long term relationships. Journal of Economic Theory 147, 142–161.

Ekmekci, M., 2011. Sustainable reputations with rating systems. Journal of Economic Theory 146, 479–503.

Farrell, J., 1987. Cheap talk, coordination, and entry. The RAND Journal of Economics 18, 34–39.

Farrell, J., Rabin, M., 1996. Cheap talk. The Journal of Economic Perspectives 10, 103–118.

Hogarth, R.M., Einhorn, H.J., 1992. Order effects in belief updating: The belief-adjustment model. Cognitive Psychology 24, 1–55.

Hopenhayn, H.A., Skrzypacz, A., 2004. Tacit collusion in repeated auctions. Journal of Economic Theory 114, 153–169.

Iyengar, R., 2008. I'd rather be hanged for a sheep than a lamb: The unintended consequences of 'Three-Strikes' Laws. Working Paper 13784. National Bureau of Economic Research.

Kandori, M., 1992. Social norms and community enforcement. The Review of Economic Studies 59, 63–80.

Kashima, Y., Kerekes, A.R.Z., 1994. A distributed memory model of averaging phenomena in person impression formation. Journal of Experimental Social Psychology 30, 407 – 455.

Liu, Q., 2011. Information acquisition and reputation dynamics. Review of Economic Studies 78, 1400–1425.

Liu, Q., Skrzypacz, A., 2011. Limited Records and Reputation. Working Paper.

Mailath, G.J., Olszewski, W., 2011. Folk theorems with bounded recall under (almost) perfect monitoring. Games and Economic Behavior 71, 174–192.

Mailath, G.J., Samuelson, L., 2006. Repeated Games and Reputations. Oxford University Press.

Mas-Colell, A., Whinston, M.D., Green, J.R., 1995. Microeconomic Theory. Oxford University Press.

Myerson, R.B., 2009. Learning from Schelling's *Strategy of Conflict*. Journal of Economic Literature 47, 1109–1125.

Okuno-Fujiwara, M., Postlewaite, A., 1995. Social norms and random matching games. Games and Economic Behavior 9, 79–109.

Schelling, T.C., 1980. The Strategy of Conflict. Harvard University Press, Cambridge, MA.

Tversky, A., Kahnenman, D., 1973. Availability: A heuristic for judging frequency and probability. Cognitive Psychology , 207–232.