NBER WORKING PAPER SERIES

VARIABLE TEMPTATIONS AND BLACK MARK REPUTATIONS

Christina Aperjis
Yali Miao
Richard J. Zeckhauser

Working Paper 16423
http://www.nber.org/papers/w16423

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2010

This work was partially supported by a grant from the Alfred P. Sloan Foundation, and the NSF under award IIS-0812042. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Variable Temptations and Black Mark Reputations
Christina Aperjis, Yali Miao, and Richard J. Zeckhauser
NBER Working Paper No. 16423
September 2010
JEL No. C71,C73,D83,K12

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 trusted player, called the temptee, has a temptation to betray. The strength of the temptation to betray varies from encounter to encounter. We set aside any information about types and examine how reputations work when the temptees are the same in terms of morals and self control. We refer to a recorded betrayal as a black mark and focus on mechanisms that only reveal the number of black marks of a temptee. We show that the greater the number of black marks, the less likely the temptee is to betray. We then study the different equilibria that emerge, depending on which side of the market has the ability to specify the equilibrium. In closing, we generalize to cases where the number of encounters is also recorded.

Christina Aperjis
HP Labs
1501 Page Mill Rd.
Palo Alto, CA 94304
christina.aperjis@hp.com

Yali Miao
Jane Street Capital
Roppongi 6-12-4, Minato-ku
Tokyo, Japan 106-0032
yalimiao@post.harvard.edu

Richard J. Zeckhauser
John F. Kennedy School of Government
Harvard University
79 John F. Kennedy Street
Cambridge, MA 02138
and NBER
richard_zeckhauser@harvard.edu

Variable Temptations and Black Mark Reputations

CHRISTINA APERJIS
Social Computing Lab, HP Labs

YALI MIAO
Harvard University

RICHARD ZECKHAUSER
Kennedy School, Harvard University

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 trusted player, called the temptee, has a temptation to betray. The strength of the temptation to betray varies from encounter to encounter. We set aside any information about types and examine how reputations work when the temptees are the same in terms of morals and self control. We refer to a recorded betrayal as a black mark and focus on mechanisms that only reveal the number of black marks of a temptee. We show that the greater the number of black marks, the less likely the temptee is to betray. We then study the different equilibria that emerge, depending on which side of the market has the ability to specify the equilibrium. In closing, we generalize to cases where the number of encounters is also recorded.

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 employee 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 the concern with 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 prior credit dealings or employment history. 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, it is irrelevant if one's courses are easy or hard. The Better Business Bureau[1] only has information on the number of complaints for a business, but does not provide details on each complaint. On the well-known and highly successful eBay reputation system, the summary score tells us how many positive, neutral, and negative feedbacks a seller has received, but not the highly relevant information of the prices of all the items that received negative feedback.[2]

In this paper, we study the *black mark reputation mechanism*, a mechanism where an individual's reputation is simply a tally of the number of bad ratings or complaints 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 for a particular business, but not the number of transactions 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.

Black mark reputations approximate settings where buyers focus on the number of negatives — even if more reputation information is provided. A strongly negative rating is a black mark. The Internet is now bristling with websites where users rate firms. They include Tripadvisor (for hotels and restaurants), Angie's List (for service providers), and Yelp (for restaurants, attractions, etc.). Participants give individual feedback scores after an encounter. Most rated entities, even those with good average reputations, have some very low scores, usually from some disastrous encounters. Some potential buyers focus on the extreme negatives: "The restaurant lost our reservation and could not seat us;" "The plumber showed up 6 hours late, and left the place a mess." Businesses presumably know that further highly negative encounters could be extremely damaging, and will strive to improve their reservation system or their promptness.

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. We set aside any information about types and examine how reputations work when all players are essentially the

[1] www.bbb.org

[2] The concern, of course, is that the seller would be generally trustworthy, but dishonest on rare occasions when describing a very high-priced item. Information regarding sold items remains available for 90 days on eBay. It is thus possible to scroll through to see the sale prices of recently sold items. This just complicates the strategy of the dishonest seller, who must take a break after doing an untrustworthy transaction at a high price. In any case, information that is prominently shown to buyers on eBay has a larger effect than information that is available but harder to find (Cabral and Hortacsu 2010).

same in terms of morals and self-control. Behaviors differ in our model — as in real life — because *the strength of the temptation to betray varies from encounter to encounter*. 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. The strength of the temptation to betray is then assumed to vary according to some probability distribution.

This paper addresses two major questions.

(1) Are people with worse reputations more likely to betray?

(2) Do different equilibria for the treatment of reputations emerge depending on which side of the market has the ability to specify the equilibrium?

The answer to even the first question is hardly clear. Suppose we have a "two strikes and you're out" system. Will a player with one strike be more likely to betray than a player with no strikes? We show that in any pure equilibrium of a game where a temptee's reputation depends solely on the number of black marks, *the greater the number of black marks, the less likely the temptee is to betray*. Moreover, under certain probability distributions of the temptation's strength, the likelihood of betraying decreases faster when the temptee's reputation consists of a larger number of black marks.

It may seem counterintuitive that those with worse reputations would behave better. Our intuition is led astray because we are used to reputational models that posit a difference among types. In such models, more betrayals indicate that one is a worse type, and can be expected to perform worse. However, the model developed here focuses on moral hazard, which leads to a completely different outcome.

The qualitative equivalent of the situation we address is seen in the Goldman Sachs (GS) situation after its revenue was tarnished by allegedly proffering portfolios designed by famed short seller John Paulson, who wished to sell them short. GS allegedly did not reveal this highly relevant information and is currently subject to both regulatory impositions and significant lawsuits. Quite apart from any legal action, its reputation has suffered gravely. Holding GS's type fixed, it now seems much less likely that it would allow itself to engage in another ploy of this sort. Another round of such allegations could be the death knell for GS as a leading investment house.

We address the second question by considering which pure equilibria are best for each of the two players, that is, the truster and the temptee. We demonstrate that the preferences of the two

players may in general be dramatically different. However, there are cases where both the truster and the temptee prefer the same pure equilibrium, making a socially optimum available.

We then consider mixed strategies for the truster. That is, for at least one reputation value he trusts the temptee with a probability that is strictly between 0 and 1. We focus on a special class of mixed equilibria, at which trust is prolonged, and show that the temptee strictly prefers such an equilibrium to a pure equilibrium.

In closing, we consider mechanisms that track both the number of black marks and the number of encounters of a temptee, and show that the equilibrium behavior in the long run is identical to equilibrium behavior under a black mark reputation mechanism.

The remainder of the paper is organized as follows. Section 2 discusses related literature on reputation. The problem formulation is given in section 3, and a general characterization of pure equilibria is presented in section 4. Section 5 shows that in any pure equilibrium the temptee is less likely to betray when she has more black marks. Section 6 studies how the equilibrium is determined by the one who specifies it. Section 7 considers mixed equilibria. Section 8 considers a setting where the truster knows both the number of black marks and the number of transactions of the temptee. Section 9 discusses extensions to long-term relationships and section 10 concludes. All proofs are provided in the Appendix.

2. Related Literature

To the best of our knowledge, in all prior equilibrium approaches to reputation, agents with high reputations perform better.

Reputation is often studied in settings with both adverse selection and moral hazard. The agent is assumed to have a type that her counterparts try to infer from her past behavior (for a survey of such models see Mailath and Samuelson 2006). In this setting, an agent's reputation represents the belief that other players have about her type, and the analyses focus on sorting among types. An agent with a worse reputation is thought less likely to be of a good type and is considered more likely to betray. For instance, Sobel (1985) considers an adverse selection model, where one player (the sender) is either a friend or an enemy and the other player (the receiver) has a prior belief on the sender's type; it turns out that an enemy is less likely to lie when she has a better reputation. Imperfect monitoring (Fudenberg and Levine 1992), learning (Bar-Isaac 2003), social norms (Kandori 1992) and settings where reputations can be bought and sold (Mailath and Samuelson 2001) have been considered in models with both adverse selection and moral hazard.

By contrast, in the setting we study, the temptee does not have a hidden type. Reputation is only used to incentivize good behavior. This approach has also been taken by Dellarocas (2005)

in the context of an electronic marketplace, where the seller might betray the buyer; his paper studies mixed equilibria in which the seller cheats (betrays) the buyers with some probability that decreases with her reputation. Thus, a seller with a better reputation is less likely to betray.

Alternatively, agents can build reputations. Shapiro (1983) derives an equilibrium price-quality schedule for markets in which buyers cannot observe product quality prior to purchase. At the equilibrium, sellers initially invest in reputation in order to be able to sell high-quality items, which can then be sold at a premium above their cost. The premium compensates sellers for their investments in reputation. At this equilibrium, reputable sellers sell high-quality items; in other words, high reputation sellers behave better.

Reputation has been extensively studied in the context of e-commerce. A number of papers have empirically studied the effect of the seller's reputation on the average payment she receives in electronic marketplaces. Some sample studies include data about eBay auctions for coins (Lucking-Reiley, Bryan, Prasad, and Reeves 2007); Palm Pilots (Kalyanam and McIntyre 2001); Pentium III processors (Houser and Wooders 2006); collectible coins, Thinkpads, and Beanie babies (Cabral and Hortacsu 2010); and postcards (Resnick, Zeckhauser, Swanson, and Lockwood 2006). Finally, the effect of different dimensions of a seller's reputation — assessed by considering text comments — on pricing power has been studied on both Amazon and eBay (Ghose, Ipeirotis, and Sundararajan 2005, Pavlou and Dimoka 2006). On the other hand, a number of papers have considered the question of designing a reputation mechanism that optimally incentivizes the seller (Fan, Tan, and Whinston 2005, Aperjis and Johari 2010b, Ekmekci 2010).

Although we do not consider incentives for trusters to leave honest feedback in this paper, another extensive line of research considers how truthful feedback can be elicited. In online markets, players might undertake fake transactions in order to enhance their reputations. This stratagem is unattractive, however, if a specific relation between the reputation premium and the transaction cost holds (Bhattacharjee and Goel 2005). On the other hand, even if fake transactions cannot be undertaken, buyers might not leave honest feedback after a transaction. Nevertheless, it is still possible to devise a scoring system that induces honest reporting of feedback (Miller, Resnick, and Zeckhauser 2005). In this paper, we posit that trusters leave honest feedback, since they have no reason not to do so. We focus on the temptee's decision and its influence on the decision to trust. We do, however, allow for imperfect monitoring in our model, as various studies have shown that monitoring is often imperfect in practice (Bolton, Greiner, and Ockenfels 2009, Dellarocas and Wood 2008, Chwelos and Dhar 2008).

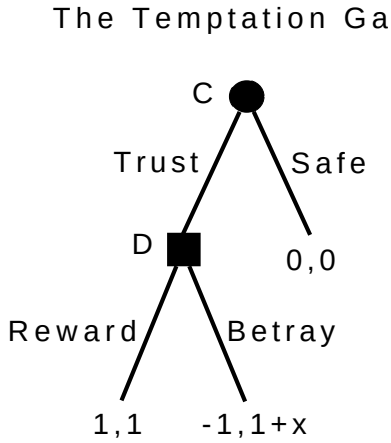

Figure 1 **Extensive form representation of one-period interaction between C and D. (C's payoff listed first.)**

3. Problem formulation

Players in the temptation game are divided into two roles, Cs and Ds, trusters and temptees, with equal numbers of each in each period. For expository ease, those who must decide whether to trust — trusters — are males, and those who are subject to temptation — temptees — are females in our analysis.

In every period, each C is randomly matched up with a D. Then a C decides whether to choose "trust" or "safe." If C plays trust, then D can play "reward" or "betray." If D rewards, then D and C both get a unit payoff. If D chooses to betray, then D will get a $(1 + x)$ of payoff, where x is the strength of the temptation to betray. Its magnitude is the realization of a random variable X drawn from a continuous probability distribution with density $f(x)$.[3] In each round, prior to choosing whether she will reward or betray, D learns her x for that round, namely her level of temptation. The value of x is and remains unknown to C. Thus, no information on the strength of the incentive to betray is ever part of a D's reputation. If D chooses to betray, then C will get a payoff of -1. Figure 1 shows the extensive form representation of a one-period interaction between C and D, where C's choices are circles and D's are squares, and C's payoff is listed first. The analysis remains qualitatively the same if C gets a payoff of $-y$ when D betrays, rather than -1, though of course the parameter values at equilibria will shift.

We note that the scaling of these payoffs is arbitrary. There is no implied interpersonal comparison. For example, in dollar value C may gain far more than D when each goes from 0 to 1.

We refer to a recorded betrayal of D as a black mark. The number of black marks a D has received is known to a C when he encounters her. We allow for imperfect recording; that is, recording

[3] It is possible to extend the results of the paper for the case where X has a discrete distribution in whole or in part.

rewards as betrayals and betrayals as rewards. In particular, if D betrays, the number of black marks increases by 1 with probability $1-r$ and remains the same with probability r. If D rewards, then the number of black marks remains the same with probability $1-q$ and increases by 1 with probability q. Perfect monitoring is a special case with $r=q=0$. We refer to the number of recorded betrayals (or black marks), b, as the *reputation* of D. When monitoring is perfect, b equals the actual number of betrayals. In general, however, it may differ. In the discussion below, when we refer to betrayals as part of a player D's reputation we mean recorded betrayals or black marks.

At each round, each D has a certain probability of surviving to the next period, s. We leave aside discounting, except as it arises through D's survival concerns. Absent discounting, the survival rate for Cs turns out to be nonmaterial. After each round, if a D dies, she will be replaced by another D who enters with a blank reputation record. Ds also leave the game (in our language get expelled) if their reputation ensures they will no longer be trusted. We further assume that all players are risk-neutral. The goal of a D is to maximize her expected payoff until she dies or is expelled. The goal of a C is to maximize his expected payoff each period, since his actions have no effect on his survival.

Key notation introduced in this and subsequent sections is summarized in Table 1. Table 1 also shows some general assumptions we make on the model's parameters; as discussed in Section 4, these assumptions are only introduced to rule out settings with uninteresting equilibria. In particular, we are assuming that the random variable X has a finite mean and a strictly positive median, that the imperfect monitoring probabilities are not too large, and that the survival probability of the temptee is strictly between 0 and 1.

4. Characterization of Pure Equilibria

In this section we characterize pure equilibrium. In the following sections we use this characterization to show properties of these equilibria. Mixed equilibria are analyzed in section 7.

For the game to be in equilibrium, at each round, both C and D must have no incentive to deviate from the strategies defining the equilibrium. C's strategy consists of whether he trusts D as a function of D's reputation. D's strategy determines whether she rewards as a function of her reputation b and her realization of X in that period.

For a fixed strategy of C, let b^* be the minimum number of betrayals at which C does not trust D. That is, C trusts when $b<b^*$ and does not trust when $b=b^*$. Since C does not trust D at b^*, D will never have more than b^* black marks. We thus refer to b^* as the *cutoff* at which C stops trusting D.

Notation	Definition	Assumptions (if any)
X	strength of temptation to betray	$\mathbb{E}X < \infty$
$f(x)$	probability density function of X	
m	median of X	$m > 0$
r, q	imperfect monitoring probabilities	$r + q < 1$
s	survival probability of the temptee (player D)	$0 < s < 1$
b	number of (imperfectly) recorded betrayals, or black marks	
b^*	number of black marks at which C stops trusting D	
$v(b)$	D's maximum expected infinite horizon payoff at b black marks	
$x^*(b)$	threshold on X below which D rewards, and above which she betrays	
$w(b)$	probability that D betrays when she has b black marks	

Table 1 Notation used in the paper.

We first consider the best response of player D when C uses a cutoff b^*. Let $v(b)$ be the maximum expected infinite horizon payoff to D when her reputation consists of b black marks. We note that if monitoring is imperfect (i.e., $r + q > 0$), then b may be different than the actual number of D's betrayals. Since the cutoff is b^*, C will never trust D once her reputation becomes b^*, and thus

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

For $b \in \{1, ..., b^* - 1\}$, $v(b)$ is described by the following dynamic program:

$$v(b) = \mathbb{E}[\max\{1 + X + s((1-r) \cdot v(b+1) + r \cdot v(b)), 1 + s((1-q) \cdot v(b) + q \cdot v(b+1))\}]$$

In particular, given that the realization of the random variable X is x, D 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., equal to b black marks) with probability r. On the other hand, if D 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 (equal to b) with probability $1 - q$ and deteriorates to $b+1$ black marks with probability q. The temptee selects to reward or betray depending on which action gives her the largest expected payoff.

Straightforward calculations show that

$$v(b) = 1 + s(1-q) \cdot v(b) + sq \cdot v(b+1) + \mathbb{E}[(X - s(1-r-q) \cdot (v(b) - v(b+1)))^+], \tag{2}$$

where $y^+ \equiv \max(y,0)$ is the positive part of y. Let

$$x^*(b) = s(1-r-q)\cdot(v(b)-v(b+1)). \tag{3}$$

Equation (2) implies that it is optimal for D to reward if $X < x^*(b)$ and betray if $X > x^*(b)$. Thus, the set $\{x^*(b), b = 0,1,...,b^*\}$ characterizes the best response of D.[4]

Substituting (3) into (2) we conclude that

$$(1-s(1-q))v(b) = sq\cdot v(b+1) + 1 + \mathbb{E}[(X-x^*(b))^+]. \tag{4}$$

(We assume that X has a finite mean, so that $\mathbb{E}[(X-x^*(b))^+]$ is well defined.)

Since D gets strictly positive immediate payment whenever she is trusted, the value $v(b)$ is strictly decreasing for $b \le b^*$. This is shown formally in the following lemma.

LEMMA 1. *For a given cutoff $b^* < \infty$, suppose that $v(b)$ and $x^*(b)$ satisfy* (1), (3), *and* (4). *Then $v(b)$ is strictly decreasing in $\{0,1,...,b^*\}$.*

If the level of imperfect monitoring is sufficiently high, namely if $r+q \ge 1$, then it is impossible to incentivize D to reward when the strength of the temptation to betray is positive (since then Lemma 1 and Equation (3) imply that $x^*(b) \le 0$). We are not interested in studying such cases and assume that $r+q<1$ throughout the paper. Lemma 1 then implies that $x^*(b)$ is strictly positive for $b \in \{0,1,...,b^*-1\}$.

We next consider player C. Given $x^*(b)$, the expected payoff of C is

$$\mathbb{P}[X < x^*(b)] - \mathbb{P}[X > x^*(b)] = 2\mathbb{P}[X < x^*(b)] - 1$$

if he trusts D; and 0 otherwise. We conclude that C trusts D if $\mathbb{P}[X < x^*(b)] > 1/2$; C does not trust D if $\mathbb{P}[X < x^*(b)] < 1/2$; and C is indifferent between trusting and not trusting if $\mathbb{P}[X < x^*(b)] = 1/2$.

Let m be the median of X, i.e., m is such that $\mathbb{P}[X < m] = 1/2$. If $m \le 0$, then C always trusts D at the equilibrium. We are interested in settings where this is not the case and thus assume that $m > 0$. We first observe that, if $m > 0$, then there cannot exist an equilibrium where C always trusts D. In particular, if C always trusts D, then D's best response is to always betray whenever there is a positive temptation to betray, i.e., $x^*(b) = 0$ for all b; however, C's best response to $x^*(b) = 0$ for all b is to never trust, since $m > 0$. We conclude that $b^* < \infty$.

At an equilibrium, both C and D play a best response to the other player's strategy. The following lemma gives necessary and sufficient conditions for an equilibrium.

[4] D is indifferent between rewarding and betraying when $X = x^*(b)$. For expository simplicity, we assume that X is a continuous random variable, so $X = x^*(b)$ with zero probability and it does not matter what D does in that case. The results can be extended for the case that X has a discrete distribution in whole or in part.

LEMMA 2. *A cutoff $b^* \geq 0$ and the set $\{x^*(b), b = 0, 1, ..., b^*\}$ constitute a pure equilibrium of the temptation game if the following conditions are satisfied:*

1. *$x^*(b)$ is a best response of D to C's strategy, i.e., there exists a function $v(b)$ such that $x^*(b)$ and $v(b)$ satisfy* (1), (3), *and* (4)

2. *b^* is a best response of C to D's strategy, i.e.,*

- *$x^*(b) \geq m$ for $b < b^*$*
- *$x^*(b^*) \leq m$*

A pure equilibrium can be computed by recursively solving Equations (3) and (4) 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 $\{x^*(b), b = 0, 1, ..., b^* - 1\}$ constitute an equilibrium if $x^*(b) \geq m$ for $b < b^*$. On the other hand, if $x^*(b) < m$ for some $b < b^*$, then there does not exist a pure equilibrium with cutoff greater or equal to b^*.

We observe that there always exists a degenerate equilibrium where C never trusts D and D never rewards, that is, $b^* = 0$. The temptation game has a non-degenerate equilibrium (i.e., with $b^* \geq 1$) if the solution y of the following equation

$$\frac{1 - s(1-q)}{s(1-r-q)} y = 1 + \mathbb{E}[(X - y)^+]$$

is greater or equal to m. This follows by considering Equations (1), (3), and (4) for $b = b^* - 1$.

Even though $b^* < \infty$ and, therefore, C does not trust D after a finite number of black marks, there may exist equilibria at which cooperative behavior is sustained for the duration of D's lifetime. Alternatively, if we interpret D's survival probability as a discount factor and the monitoring is perfect, then cooperation can be sustained forever, along the lines of a folk theorem (Myerson 1997). In particular, suppose that $b^* = 1$ and D always rewards when she has no black marks (i.e., $P[X < x^*(b^* - 1)] = 1$). Straightforward calculations show that this is the case if the maximum possible value of X is less than $s(1-r-q)/(1-s(1-q))$. On the other hand, there may also exist equilibria with $b^* > 1$, where D may betray when she has strictly less than $b^* - 1$ black marks, but always rewards when she has $b^* - 1$ black marks. For instance, if $b^* = 2$ represents an equilibrium and the maximum possible value of X is less than $s(1-r-q)/(1-s(1-q))$, then D will betray for X greater than some amount when she has no black marks, but will never betray after this. Thus, in some sense, cooperation is sustained, with one betrayal. The outcome is similar for other equilibria where the permissible number of black marks is greater.

In general, if there exists an equilibrium with cutoff $b^* = k$, there also exists a pure equilibrium with cutoff $b^* = k'$, where $k' < k$ (assuming that both k and k' are positive integers). Let B^*_{pure} be

the maximum cutoff b^* for which there exists a pure equilibrium. The value of B^*_{pure} depends on the density f, the survival probability s, and the imperfect monitoring probabilities r and q.

5. Betrayal as a Function of Reputation

In this section we consider how reputations work. We find that temptees are less likely to betray when they have bad reputations, 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 pure equilibrium* $(b^*, \{x^*(b), b = 1, 2, ..., b^*\})$ *of the temptation game,* $x^*(b)$ *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. What are the implications of this result on how likely player D is to betray? Let

$$w(b) \equiv \mathbb{P}[X > x^*(b)]$$

be the probability of betraying with b black marks. Moreover, let

$$F(x) \equiv \int_{-\infty}^{x} f(y)dy$$

be the cumulative distribution function (CDF) of the random variable X.

The following corollary of Proposition 1 characterizes $w(b)$.

COROLLARY 1. *For every pure equilibrium* $(b^*, \{x^*(b), b = 1, 2, ..., b^*\})$ *of the temptation game:*
(i) $w(b)$ *is decreasing in* b *for* $b \in \{0, ..., b^* - 1\}$
(ii) if F *is linear or convex, then* $w(b)$ *is concave in* b *for* $b \in \{0, ..., b^* - 1\}$

In words, in any game where 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 betray. This follows from the fact that $x^*(b)$ is increasing in b. It may seem counterintuitive that those with worse reputations would behave better. Our intuition is led astray because we are used to reputational models based on differences among types; more black marks are seen as an indicator that one is of a worse type and can be expected to perform more dishonestly. However, here we are considering moral hazard. To be sure, in any one trial in our model, players D do have a type, namely the draw they receive for X, their level of temptation. But this type does not carry over from trial to trial. Thus, there is no learning about types.

In addition to showing that there is a decrease in the likelihood of betraying as the number of black marks increases, Corollary 1 shows that if the random variable X is drawn from a distribution

with a convex or linear CDF (such as the uniform distribution), 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 CDF, the likelihood of betraying decreases faster when the temptee's reputation consists of a larger number of betrayals. In other words, in this case the likelihood of rewarding increases faster when the temptee has a bad reputation.

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." We briefly discuss the California criminal justice, driver's license suspension, tennis, baseball, and basketball below.

California criminal justice. 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 the chance of being caught. Following our model, as she comes closer to getting put away for life (three convictions, hence three strikes), 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. In theory, recidivism rates in California should reveal a lesser propensity to criminal activity after two felony convictions. Unfortunately, two significant complications make this evidence almost impossible to assess. First, criminals are heterogeneous as to type. Those with two felony convictions presumably commit more crimes and are more likely to be caught, on average, than those with only one. Second, there is evidence that some two-strike criminals have migrated to other states.

Driver's licence suspension. In some states, there are penalties for getting certain numbers of traffic infractions, or a certain number of accidents. If a driver gets a certain number of traffic violations in a period, then her license is suspended. Note that the motor vehicle bureau does not know how far a driver has traveled during that period, nor the level of temptation. (It may be more tempting to speed when one is late for an important meeting.) However, this is only part of the damage, since traffic infractions also cause a boost in insurance rates. This suggests that if we could fine people for a black mark in the temptation game, the additional instrument might afford a superior outcome.

Tennis. A tennis-player gets two serves. This is equivalent to being allowed two black marks before she is expelled (loses the point). On the first serve, it is optimal to be more aggressive. Indeed, the first serve is typically aggressive. It is struck with power and placement to have a high chance of winning the point outright or soon thereafter, assuming that it goes in. The second serve is usually much more conservative — slower speed, less risky placement — to make it exceedingly

likely to go in and thereby avoid a double fault. This strategy has important, albeit not exact, parallels to being less willing to betray when one's reputation consists of more black marks in the temptation game. One key difference, of course, is that tennis is a game of strictly opposed interests, in contrast to the temptation game. A second key difference is that a betrayal is a certain move, whereas whether a serve goes in is a probabilistic phenomenon.

Baseball. A related situation occurs in baseball. Holding the number of balls constant, batters can afford to be more picky on pitches when they have no strikes than with one strike, and with one strike than with two strikes.[5]

Basketball. In the NBA, if a player commits six personal fouls over the course of a game, he fouls out and is disqualified from participation for the remainder of the game. This is a setting, like our model, with imperfect monitoring. A player may get called for a foul that he did not commit, so it is dangerous to get up to five fouls. One of the features of basketball, but not the temptation game, is that an infraction is also punished by giving out foul shots. This situation is similar to the cumulative violations system in driving that was discussed above.

6. Optimal Cutoffs

This section identifies the pure equilibria (as characterized by Lemma 2) 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)[6] 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." Myerson goes on to observe that: "Schelling's focal-point effect should be counted as one of the most important ideas in social theory. Recognizing the fundamental problem of selecting among multiple equilibria can help us to better understand the economic impact of culture on basic social phenomena such as social relationships, property and justice, political authority and legitimacy, foundations of social

[5] An opposite situation applies to the pitcher, of course. He must make more of an effort to get the ball over the plate when there are more balls thus far, holding strikes constant, lest he walk the batter.

[6] pages 1111-1112

institutions, reputations and commitment." Rosenthal and Landau (1979) study possible decision rules or "customs" which players might use to determine their moves in a game as a function of reputation. Our attention here is on ways of choosing among multiple equilibria in the temptation game.

		Player B	
		I	II
Player A	i	5, 2	0, 0
	ii	0, 0	3, 6

Figure 2 A coordination game

Figure 2 shows a game matrix taken in spirit from Schelling. The two parties have a joint interest in arriving at one of the two equilbria. One is superior for A and the other for B. Note also, as in the game between trusters and temptees, that the equilibria are not symmetric. If one party can verbally communicate, say A, he can simply say: "I will play i". If B takes this as a commitment, he will play I, and A gets the payoff at his superior equilibrium. Similarly, in our game, if only the trusters can communicate, they can say to the temptees: "We will allow you one betrayal, but once you get to two, no one will trust you." (That could be the equivalent of saying: "We will play i.") If that message is transmitted, we would expect the temptees to behave as if $b^* = 1$. 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.

Often the players on one side actually have the power to establish customs or laws that can establish which equilibrium will prevail. For example, in most traditional societies men make the rules relating to marital behavior, such as the punishment for infidelity or the bases for divorce. Similarly, banks have historically made the rules for lending practices. But we could imagine a feminist movement changing the mores relating to marital behavior in a society[7] or, in the wake of the recent financial meltdown, we could easily envision Congress remaking the rules on lending to be more favorable to the consumer so as to maximize social welfare.

The temptation game differs from the normal multi-play, two-person game-theoretic situation, in that players only meet any partner once. This probably makes it easier for an equilibrium, once

[7] For instance, in Lysistrata by Aristophanes, the women of Greece try to change the rules regarding who makes decisions about war. In particular, they withhold sexual privileges from their husbands as a means of forcing the men to negotiate peace.

identified, to stick. Posit that a particular pair of players A and B above were tied together for many rounds, and that A had announced that he would play i. We could easily imagine that B might try to break away from the upper-left equilibrium as the exclusive outcome. Thus, he might alternate I and II, hoping that A would perceive the pattern and cooperate on sharing the big payoff.

Such a strategy would not make sense, however, when a specified A and B only meet once. In the temptation game, players belong to identifiable groups, which may give them 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, with their peers.

6.1. Equilibrium Selection by a Third Party

Consider a third party (such as an electronic marketplace) that wants to have the Cs and Ds play some equilibrium b^*_M. This may either be an equilibrium that maximizes a weighted sum of payoffs to the Cs and Ds, or an equilibrium that the third party prefers for some other reason (e.g., because it maximizes the expected revenue of the marketplace). In the context of an electronic marketplace, the Cs are the buyers who decide whether to trust sellers with certain reputation, and the Ds are the sellers who decide whether to reward or betray. The electronic marketplace cannot force the Cs and Ds to play a certain equilibrium, but can expel a D with a large number of black marks from the marketplace.

If the marketplace sets the maximum number of allowed black marks to be equal to b^*_M, this restricts the set of feasible equilibria to those with $b^* \in \{0, 1, ..., b^*_M\}$. Moreover, the equilibrium with cutoff b^*_M becomes a focal point, and as a result the Cs and the Ds may be more likely to play it. On the other hand, if the marketplace sets the maximum number of allowed black marks to be equal to some number that does not correspond to an equilibrium of the temptation game between the Cs and the Ds, then effectively the marketplace is not intervening and lets the players choose which equilibrium will be played.

6.2. An Inequality Between Optimal Cutoffs

In the temptation game, posit that Cs, and only Cs have the ability to communicate verbally. They will tell the Ds that they are employing the cutoff in number of black marks that maximizes the expected welfare of Cs assuming that Ds respond optimally to that cutoff. On the other hand, if the Ds 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 x*(b) for $b \in \{0, 1, ..., \}$, thus letting C pick

his b^* in response; this produces the first best condition for D. By Lemma 2, because C responds to D's strategy in a predictable way, D is effectively choosing C's cutoff b^* for him. 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 C and D.

We define $b^*(C)$ and $b^*(D)$ to be the optimal cutoffs for C and D respectively, that is, the cutoffs of the pure equilibria that maximize the corresponding payoffs. We let $b^*_\alpha(S)$ denote the cutoff that maximizes the sum of C's payoff and α times D's payoff, where $\alpha \geq 0$. If α is equal to 1, then $b^*_\alpha(S)$ maximizes the total payoff going to C and D. The following proposition shows that the optimal numbers of betrayals will be greatest for the temptee, moderate for the social optimum, and least for the truster.

PROPOSITION 2. *For any* $\alpha \geq 0$,

$$b^*(C) \leq b^*_\alpha(S) \leq b^*(D) = B^*_{pure}.$$

Recall that B^*_{pure} is the maximum cutoff that can arise at a pure equilibrium (for given f, s, r and q). Proposition 2 tells us that the first best equilibrium for D has the maximum possible cutoff. Intuitively, D prefers to be trusted longer. Moreover, if the first best equilibrium for C has the maximum possible cutoff, then the equilibrium with $b^* = B^*_{pure}$ is a social optimum. The first best equilibrium for C is the focus of the next two sections.

6.3. The Truster's Expected Payoff

Consider a pure equilibrium $(b^*, \{x^*(b), b = 0, ..., b^*\})$. Recall that $w(b) \equiv P[X > x^*(b)]$ is the probability that player D betrays when she has b black marks. When C interacts with a D who has b black marks, his expected payoff is $1 - 2 \cdot w(b)$ if $b < b^*$, and 0 if $b = b^*$. In light of Proposition 1, this payoff increases in b when $b \in \{0, ..., b^* - 1\}$. C's payoff thus depends heavily on the number of black marks of the D that he interacts with.

As far as C is concerned, the number of black marks of any D evolves according to a Markov chain. The state is the number of black marks, which either increases by 1 (if C trusts D and a betrayal is recorded), or remains the same (if either C trusts D and a reward is recorded or C does not trust D), or becomes 0 (if D dies and thus is replaced with a new player with a blank reputation).[8] Let π be the steady state distribution of this Markov chain. The following lemma gives C's expected payoff and provides a way to compute π.

[8] If D dies, she is replaced with a different D, but that does not affect C's payoff.

LEMMA 3. *Let* $(b^*, \{x^*(b), b = 1, 2, ..., b^*\})$ *be an equilibrium. The expected payoff of C is equal to*

$$\sum_{b=0}^{b^*-1} \pi(b)(1 - 2w(b)), \tag{5}$$

where $\pi(b)$ *satisfies*

$$\pi(0) = \frac{1-s}{1 - s((1-q)(1-w(0)) + rw(0))}; \tag{6}$$

and

$$\pi(b) = s\frac{q(1-w(b-1)) + (1-r)w(b-1)}{1 - s((1-q)(1-w(b)) + rw(b))}\pi(b-1), b \in \{1, ..., b^* - 1\}. \tag{7}$$

Note that we make the Cs do a long-term maximization, which removes their lifespan from the optimization. An alternative formulation leading to the same result would have the Cs start against a long-term equilibrium distribution of Ds. If we do not do this, the first generation of Cs might put forth an equilibrium that would be superior for them, but inferior in the long run for Cs in general, since they start in an anomalous situation where all Ds have $b = 0$.

In the next section, we demonstrate that $b^*(C)$ can involve the minimum (non-trivial) equilibrium cutoff, the maximum equilibrium cutoff, or an interior solution. C does not care how long D lives, since he is guaranteed to meet a new D each period. However, he is interested in influencing the behavior of D, and the subtleties of that influence can yield these distinctive outcomes.

6.4. Numerical Examples

Proposition 2 shows that $b^*(C) \leq b^*(D)$. In this section we give some numerical examples to demonstrate that both $b^*(C) < b^*(D)$ and $b^*(C) = b^*(D)$ are possible. For simplicity, we consider perfect monitoring (i.e., $r = q = 0$) and assume that the random variable X has a uniform distribution.

Suppose X is uniform on $[\gamma, \delta]$ for some $\gamma < \delta$. Then $f(x) = 1/(\delta - \gamma)$ for $x \in [\gamma, \delta]$, $m = (\gamma + \delta)/2$, and

$$\mathbb{E}[(X - x^*)^+] = \int_{x^*}^{\infty} (x - x^*)f(x)dx = \frac{1}{\delta - \gamma}\int_{x^*}^{\delta} (x - x^*)dx = \frac{(\delta - x^*)^2}{2(\delta - \gamma)}.$$

Under perfect monitoring, (3) simplifies to

$$v(b) = v(b+1) + x^*(b)/s,$$

and (3) with (4) give

$$\frac{1-s}{s}x^*(b) + (1-s)v(b+1) = 1 + \frac{(\delta - x^*(b))^2}{2(\delta - \gamma)}.$$

The probability that player D betrays when she has b black marks is

$$w(b) = \frac{\delta - x^*(b)}{\delta - \gamma},$$

and the steady-state reputation distribution of the D population can be computed from the following equations (which follow from (6) and (7)).

$$\pi(0) = \frac{1-s}{1-s+sw(0)};$$

$$\pi(b) = s\frac{w(b-1)}{1-s-w(b)}\pi(b-1), b \in \{1, ..., b^* - 1\}.$$

We use these equations to compute the expected payoffs of C and D at the equilibria in the following examples.

EXAMPLE 1. Assume that X is uniform on $[0, 20]$ and $s = 0.95$. Then $B^*_{pure} = 4$. C's payoff is maximized at $b^*(C) = 1$. That is, the one-betrayal-and-you-are-out strategy is best for Cs. This is the strategy that many societies have employed to deal with marital infidelities, particularly those of women. D's payoff on the other hand is maximized at $b^*(D) = 4$. Thus, in this case, $1 = b^*(C) < b^*(D) = 4$.

EXAMPLE 2. Assume that X is uniform on $[0, 30]$ and $s = 0.95$. Then, $b^*(C) = 3$ and $b^*(D) = B^*_{pure} = 4$.

EXAMPLE 3. Assume that X is uniform on $[0, 1000]$ and $s = 0.95$. Then, for any $\alpha \geq 0$, $b^*(C) = b^*_\alpha(S) = b^*(D) = B^*_{pure} = 3$.

7. Mixed-Strategy Equilibria

Our analysis thus far has only considered pure-strategy equilibria. However, it is also possible to study mixed-strategy equilibria in the same framework. In this section, we consider mixed strategies for player C, where C trusts D with some probability that depends on her reputation. We consider the same model as in section 3 and allow for imperfect monitoring, assuming that *D's reputation does not change in periods that she did not interact with a C because she was not trusted by the C she was matched to.*

We consider settings where C may use a mixed strategy: C's mixed strategy consists of the *probability* he trusts D as a function of her reputation. This is a generalization of the pure strategy where C trusts D with either probability 1 or probability 0. We do not consider mixed strategies for player D here.[9] Thus, D's strategy shows (as in the pure equilibrium case) whether she rewards as a function of her reputation.

[9] The results of the paper can be easily extended to consider mixed strategies for player D; such strategies would only be relevant if X follows a discrete distribution. In particular, D would only randomize if $X = x^*(b)$. If X is drawn from a continuous distribution (as we are assuming in this paper), X will be equal to $x^*(b)$ with zero probability, and thus what D does at $x^*(b)$ does not affect the players' payoffs. If X is drawn from a discrete distribution, she could mix optimally at most at one point (i.e., $x^*(b)$).

In Section 7.1, we characterize mixed-strategy equilibria. In particular, we give generalizations of Lemmas 2 and 3 that allow for the truster to play a mixed strategy. In Section 7.2, we then consider a special class of mixed equilibria, at which C trusts D strictly more than at any pure equilibrium. We show that the temptee strictly prefers such an equilibrium to a pure equilibrium. On the other hand, we conjecture that the truster is better off at a pure equilibrium.

7.1. Characterization of Mixed-Strategy Equilibria

C's mixed strategy represents the *probability* he trusts D as a function of her reputation. Thus, C's strategy is summarized by $\{p^*(b), b = 0, 1, ...\}$, where $p^*(b)$ is the probability that C trusts D when her reputation consists of b black marks. On the other hand, we do not consider mixed strategies for player D; thus the set $\{x^*(b), b = 0, 1, ...\}$ represents D's strategy (as in the pure equilibrium case)

Similarly to Lemma 2 we can characterize mixed equilibria.

LEMMA 4. *The sets* $\{x^*(b), b = 0, 1, ...\}$, $\{p^*(b), b = 0, 1, ...\}$ *constitute an equilibrium if the following conditions are satisfied:*

1. $\{x^*(b), b = 0, 1, ...\}$ *is a best response of D to C's strategy, i.e., there exists a function* $v(b)$ *such that*

$$x^*(b) = s \cdot (1 - r - q) \cdot (v(b) - v(b+1)); \tag{8}$$

$$(1 - s(1 - q \cdot p^*(b)))v(b) = p^*(b) \left(1 + s \cdot q \cdot v(b+1) + \mathbb{E}[(X - x^*(b))^+]\right). \tag{9}$$

2. $p^*(b)$ *is a best response of C to D's strategy, i.e.,*

- *if* $x^*(b) > m$ *then* $p^*(b) = 1$
- *if* $x^*(b) = m$ *then* $p^*(b) \in [0, 1]$
- *if* $x^*(b) < m$ *then* $p^*(b) = 0$.

Pure equilibria are a special case of the mixed equilibria discussed here. In particular, if C either trusts D with probability 1 or does not trust D (i.e., $p^*(b) = 1$ for $b < b^*$), then Lemma 4 gives the pure equilibrium conditions of Lemma 2. In particular, if $p^*(b) = 1$, then Equation (9) yields (4). We note that Equation (8) is the same as (3), rewritten here for convenience.

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 C does not trust D. If D is never trusted when her reputation consists of b^* black marks, then she never has the chance to reach more than b^* black marks.

We already know that a finite cutoff (after which C stops trusting D) is associated with every pure equilibrium. The following lemma shows that this is also the case for mixed equilibria.

LEMMA 5. *For every equilibrium* $\{(x^*(b), p^*(b)), b = 0, 1, ...\}$ *there exists a finite cutoff* b^* *such that* $p^*(b) > 0$ *for* $b < b^*$ *and* $p^*(b^*) = 0$.

Proposition 1 shows that for pure equilibria, $x^*(b)$ is strictly increasing and convex 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 $x^*(b) = m$; this is discussed in Example 4 in Section 7.2.

We next show how C's expected payoff is computed in a mixed equilibrium, by providing a generalization of Lemma 3.

LEMMA 6. *Let* $(b^*, \{p^*(b), b = 1, 2, ..., b^*\}, \{x^*(b), b = 1, 2, ..., b^*\})$ *be an equilibrium. The expected payoff of C is equal to*

$$\sum_{b:p^*(b)=1} \pi(b)(1 - 2w(b)),$$

where $\pi(b)$ *satisfies*

$$\pi(0) = \frac{1 - s}{1 - s((1 - q)(1 - w(0)) + rw(0))p^*(0) - (1 - p^*(0))};$$

and

$$\pi(b) = s\frac{q \cdot (1 - w(b - 1)) + (1 - r)w(b - 1)}{1 - s((1 - q)(1 - w(b)) + rw(b))p^*(b) - (1 - p^*(b))} p^*(b - 1)\pi(b - 1), b \in \{1, ..., b^* - 1\}.$$

7.2. Dominant Extend Equilibria

In this section we study a particular category of mixed equilibria that (when they exist) prolong trust compared to pure equilibria. Recall that B^*_{pure} is the maximum cutoff that arises in a *pure* equilibrium. Then, at the best possible pure equilibrium for D, C trusts her until she has B^*_{pure} black marks (by Proposition 2). Often, there exist mixed equilibria at which C always trusts D until she has B^*_{pure} black marks and then at B^*_{pure} black marks he trusts her with some probability that is strictly between 0 and 1. We call such equilibria *dominant extend equilibria*. A precise definition is given below.

DEFINITION 1. A dominant extend equilibrium satisfies

- $p^*(b) = 1$ for $b = 0, 1, ..., B^*_{pure} - 1$
- $p^*(B^*_{pure}) \in (0, 1)$

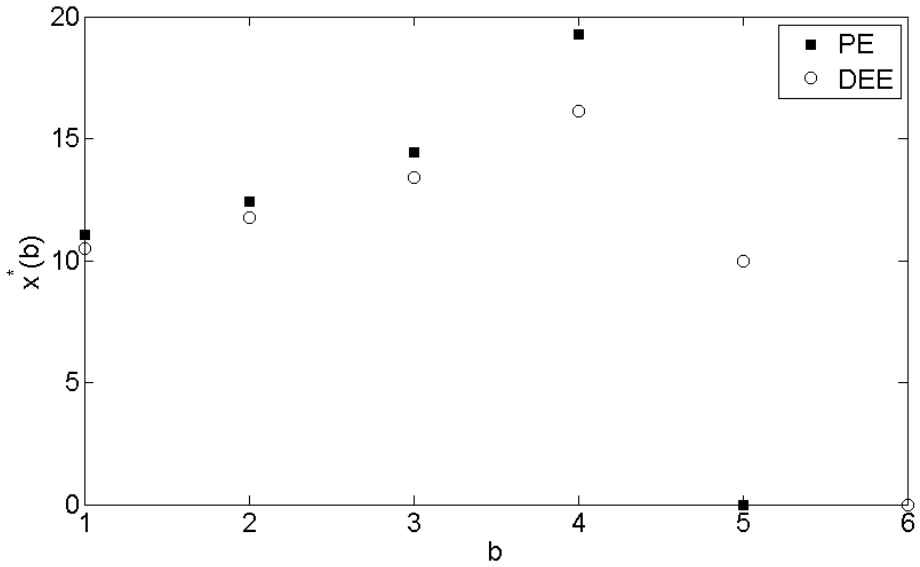

Figure 3 **D's strategy at her preferred pure equilibrium (PE, shown with solid squares) and the dominant extend equilibrium (DEE, shown with open circles) for perfect monitoring with $s = 0.95$ and X following the uniform distribution on $[0, 20]$. Details are given in Example 4.**

- $p^*(B^*_{pure} + 1) = 0$

Definition 1 and Lemma 4 imply that at a dominant extend equilibrium, $x^*(b) \geq m$ for $b < B^*_{pure}$ and $x(B^*_{pure}) = m$. Moreover, for $b < B^*_{pure}$, condition (9) of Lemma 4 simplifies to (4) as in the pure equilibrium case.

A dominant extend equilibrium gives D a longer expected lifetime, and, as the following proposition shows, D always prefers a dominant extend equilibrium to any pure equilibrium.

PROPOSITION 3. *D strictly prefers a dominant extend equilibrium to a pure equilibrium.*

Dominant extend equilibria are always attractive to D, because D's expected payoff is maximized the longer she can expect to live. Thus, rather than being expelled for sure at $b = B^*_{pure}$, she would prefer to have a probabilistic chance there, with expulsion at $b = B^*_{pure} + 1$.

On the other hand, we conjecture that player C always prefers a pure equilibrium to a dominant extend equilibrium. Our intuition for this is as follows. Consider the pure equilibrium with the maximum cutoff and the dominant extend equilibrium. We use the subscripts PE and DEE to denote the two equilibria. By Lemma 6 it suffices to show that

$$\sum_{b=0}^{B^*_{pure}-1} \pi_{PE}(b)(1 - w_{PE}(b)) \geq \sum_{b=0}^{B^*_{pure}-1} \pi_{DEE}(b)(1 - w_{DEE}(b)). \tag{10}$$

Lemma 8 in the Appendix shows that $w_{PE}(b) \leq w_{DEE}(b)$. On the other hand, the number of states is larger at the dominant extend equilibrium, suggesting that (10) is plausible. We conducted

extensive simulations that support this claim.

If player C always prefers a pure equilibrium, and C is the one that specifies which equilibrium will be played, then we will have a pure equilibrium. On the other hand, pure equilibria often tend to arise as focal points because of their simplicity (Myerson 1997).

The following example illustrates the discussion on dominant extend equilibria.

EXAMPLE 4. Assume that $s = 0.95$ and X follows the uniform distribution on $[0, 20]$. We have seen in Example 1 that the maximum possible cutoff at a pure equilibrium is $B^*_{pure} = 4$. At the best pure equilibrium for D, C trusts her when she has fewer than 4 black marks, and D's expected payoff is 60.18. However, there also exists a dominant extend equilibrium where C trusts D with probability 0.15 when she has 4 black marks, yielding an expected payoff of 65.07 for D.

On the other hand, C's expected payoff at the dominant extend equilibrium is 0.15, which is significantly smaller than his expected payoff at any (non-trivial — with a non-zero cutoff) pure equilibrium. In particular, C's payoff ranges between 0.45 and 0.54 at the pure equilibria. Thus, if the truster is the one that chooses which equilibrium will be played, then a pure equilibrium will be chosen.

The values of $x^*(b)$ are shown in Figure 3 for both the pure and the dominant extend equilibria. We observe that $x^*(b)$ is increasing and convex for $b = 0, 1, ..., B^*_{pure} - 1$. We already know that this is true for pure equilibria by Proposition 1; a similar proof shows that this is always the case for dominant extend equilibria.

8. Mechanisms that also Track the Number of Transactions

This paper has focused thus far on black mark reputation mechanisms, that is, mechanisms that only keep track of the number of recorded betrayals. In this section, we extend the reputation mechanism to also include information on the number of transactions. We show that if the number of transactions is recorded, the maximum number of black marks that C allows D at a pure equilibrium does not increase compared to the case where the number of transactions is not recorded. In particular, at a pure equilibrium, C will never trust D once she has B^*_{pure} black marks — regardless of whether C knows the number of transactions of D.

We assume that D's reputation consists of the number of black marks b and the number of transactions (which we denote by n) that she has completed. We thus denote D's reputation by (b, n). A strategy of player C in this more general model consists of a cutoff for each number of transactions. With a slight abuse of notation, let $b^*(n)$ be the cutoff for n transactions, that is, C does not trust D at (b, n) if $b \geq b^*(n)$. On the other hand, D's strategy will consist of a threshold

$x^*(b,n)$ for every possible reputation (b,n). That is, when D has b black marks in n transactions, then she betrays if the strength of her temptation to betray (i.e., the realization of X) exceeds the threshold $x^*(b,n)$.

We next give equilibrium conditions[10] for this two-dimensional reputation model.

LEMMA 7. *The sets* $\{b^*(n), n=0,1,...\}$ *and* $\{x^*(b,n), b=0,1,...,b^*(n), n=0,1,...\}$ *constitute a pure equilibrium of the temptation game if the following conditions are satisfied:*

1. $x^*(b,n)$ *is a best response of D to C's strategy, i.e., there exists a function* $v(b,n)$ *such that* $x^*(b,n)$ *and* $v(b,n)$ *satisfy:*

$$v(b,n) = 1 + s\cdot(1-q)\cdot v(b,n+1) + s\cdot q\cdot v(b+1,n+1) + \mathbb{E}[(X - x^*(b,n))^+] \text{ for } b < b^*(n); \quad (11)$$

$$x^*(b,n) = s\cdot(1-r-q)\cdot(v(b,n+1) - v(b+1,n+1)) \text{ for } b < b^*(n); \quad (12)$$

and

$$v(b^*(n),n) = 0 \text{ for all } n. \quad (13)$$

2. $b^*(n)$ *is a best response of C to D's strategy, i.e.,*

- $x^*(b,n) \geq m$ *for* $b < b^*(n)$
- $x^*(b^*,n) \leq m$

We note that Lemma 7 is a generalization of Lemma 2. In particular, if we set $b^*(n) = b^*$ where $b^* \leq B^*_{pure}$, then we get an equilibrium of the two-dimensional reputation mechanism, which corresponds to an equilibrium of the one-dimensional mechanism; in this case, $b^*(n)$ is a constant that does not vary with the number of transactions n. However, in the two-dimensional case, there also exist equilibria where $b^*(n)$ varies with n.

We next show two fundamental properties of $b^*(n)$. First, $b^*(n)$ is non-decreasing in n. This is an intuitive property: the larger the number of transactions of D, the larger the number of black marks that C will tolerate. Second, $b^*(n)$ is upper-bounded by B^*_{pure}, i.e., the maximum cutoff for which there exists a pure equilibrium when reputation only consists of the number of black marks. Even though the number of transactions is recorded in the two-dimensional reputation case, C does not allow D to commit more betrayals than in the one-dimensional case (at a pure equilibrium).

PROPOSITION 4. *If reputation consists of both the number of black marks* b *and the number of transactions* n*, then at any pure equilibrium of the temptation game*

[10] The derivation of the equilibrium conditions of Lemma 7 is similar to the derivation of the conditions of Lemma 2 in section 4.

(i) $b^*(n)$ *is non-decreasing*

(ii) $b^*(n) \leq B^*_{pure}$ *for all* n

The fact that $b^*(n)$ is upper-bounded by B^*_{pure} may at first seem counterintuitive. One could expect that C would allow D more black marks when he knows that she has completed a very large number of transactions than when he has no information on the number of transactions. However, if C tolerated a large number of transactions, then D would betray with high probability, since one more betrayal would make a small difference in her future expected payments. This would be similar to a reputation mechanism that aggregates ratings over the lifetime of the temptee; an approach that has been shown to be ineffective in a number of settings (e.g., Fan, Tan, and Whinston 2005, Aperjis and Johari 2010a, Cripps, Mailath, and Samuelson 2004).

We get some additional intuition for why $b^*(n)$ is upper-bounded by B^*_{pure} by considering the steps of the proof. We first observe that $b^*(n)$ cannot be very large; if it were, D would not be properly incentivized at $(0, n)$. In particular, even though betraying at $(0, n)$ would bring her closer to expulsion, expulsion would be so far in the future that D would not reward with a sufficiently large probability (i.e., she would use a small $x^*(b, n)$), and C on his side would not trust her. Now, given that $b^*(n)$ is upper bounded and increasing, we conclude that $b^*(n)$ is constant for all large n, which means that the number of transactions does not play any role after some point. Then, after a sufficiently large number of transactions, this two-dimensional problem is equivalent to the setting where only the number of black marks is included in D's reputation. As a result, the number of black marks that C tolerates does not exceed B^*_{pure}.

As mentioned above, after a certain number of transactions $b^*(n)$ becomes constant. Then, the thresholds $x^*(b, n)$ correspond to thresholds of a one-dimensional equilibrium. Thus, after that point D is less likely to betray when she has more betrayals ($x^*(b, n)$ is increasing in b). 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.[11]

9. Reputations in Long-Term Relationships

Our analysis above is cast in terms of situations where one contracts with another party for just one period, and then moves on. Reputations, possibly imperfect, carry all available information about the temptee. The analysis can be extended to a situation where one contracts with the same party, period after period. Here too reputations may be imperfect. Thus, in a romantic relationship, one

[11] 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.

may not know whether one's partner has cheated,[12] and any instance of betrayal may only come to light probabilistically. Even if infidelity is revealed, the relationship may not terminate, as anyone who has read advice columns — which are peppered with questions on whether one should leave a cheating partner — knows well. And the rules for terminating a relationship may depend solely on the number of black marks, with players receiving a warning after one or more betrayals has come to light. Business relationships may have much the same character. A may continue with B as a supplier if B fails to deliver on a contract once, or even twice, but he surely has a breaking point where he goes out to seek a new source.

To analyze situations with long-term relationships, we need to specify what happens after a relationship is terminated. In particular, what happens to C after he stops trusting D? Depending on the setting, C may not be able to interact with another D in the future, or may be able to find a new partner, possibly by incurring a search cost. On the other hand, D may also be able to find a new partner once C terminates their relationship.

Here, we briefly illustrate how the results of this paper can be extended to long-term relationships. Suppose that neither C nor D is able to find a new partner once their relationship is terminated. Then, D's expected payoff is given by the dynamic program considered in this paper (Equation (4) for pure equilibria and Equation (9) for mixed equilibria). On the other hand, C trusts D as long as $x^*(b) \geq m$. Thus, the pure equilibria of this game with long-term relationships are again given by Lemma 2, and the mixed equilibria are given by Lemma 4. The only thing that changes for long-term relationships is C's expected payoff, which is no longer given by Lemma 3, since C interacts with the same partner in every period.

10. Conclusion

This paper studies how reputations work when the truster's decision to trust is based only on the number of recorded betrayals of the temptee (black mark reputation mechanism). We find that at a pure equilibrium, the greater the number of black marks, the less likely the temptee is to betray. This insight applies to a broad range of situations including agencies for consumer protection and online review sites. Our analysis assumes that all players on one side of the market are identical in terms of morals, self-control, and payoff structure; thus setting aside any information about types. However, we expect our insights to carry on to situations with players of different types as long as the differences across types are relatively small.

[12] We recognize that, contrary to our model, people do differ on propensities to cheat, and there therefore would also be learning about types.

Black mark reputations approximate settings where buyers focus on the number of negatives — even if more reputation information is provided. The Availability Heuristic (Tversky and Kahnenman 1973) leads individuals to judge the frequency of an event by how easily they can bring an instance to mind. This heuristic leads individuals to give significant weight to extreme bad outcomes. Recognizing this, we have been told that in the venture capital industry, executives work extremely hard to prevent portfolio companies from going bankrupt, even though that same time devoted to profitable ventures would yield greater benefits to both the executives and the investors. The VC executives recognize that their world is an approximation of the betrayals world, where embarrassing strikeouts are remembered, and batting averages are sometimes forgotten.

We recognize that to accurately describe some interactions between trusters and temptees, the model of this paper would have to be elaborated. We show in Section 8 that the same qualitative results apply in the long run when the number of transactions is also recorded as part of a temptee's reputation. Two further extensions suggest themselves immediately. First, some relationships have a natural termination or sunset date quite apart from black marks. Thus, for a college and a student, rules infractions — plagiarism or disorderly behavior — would be the equivalent of betrayals. But once graduation occurs, the relationship is ended no matter what. Second, many long-term relationships — and some one-time-only relationships — have both parties trusting and both parties tempted. Thus, the business and its supplier or the husband and the wife may both rely on each other; each has a reputation 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, referring to his own shortcomings. And George W. Bush, not known for being soft on crime, observed: "America is the land of the second chance — and when the gates of the prison open, the path ahead should lead to a better life." 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.

Acknowledgements

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

References

Aperjis, C., and R. Johari (2010a): "Designing Reputation Mechanisms for Efficient Trade," Discussion paper, Stanford University.

——— (2010b): "Optimal windows for aggregating ratings in electronic marketplaces," *Management Sci.*, 56(5), 864–880.

Bar-Isaac, H. (2003): "Reputation and Survival: Learning in a Dynamic Signalling Model," *The Review of Economic Studies*, 70(2), 231–251.

Bhattacharjee, R., and A. Goel (2005): "Avoiding ballot stuffing in eBay-like reputation systems," in *P2PECON*, pp. 133–137.

Bolton, G., B. Greiner, and A. Ockenfels (2009): "Engineering Trust - Reciprocity in the Production of Reputation Information," Working Paper Series in Economics 42, University of Cologne, Department of Economics.

Cabral, L., and A. Hortacsu (2010): "Dynamics of Seller Reputation: Theory and Evidence from eBay," *J. of Industr. Econom.*, 58(1), 54–78.

Chwelos, P., and T. Dhar (2008): "Differences in "Truthiness" across Online Reputation Mechanisms. Working Paper, Sauder School of Business," .

Crawford, V. P., and J. Sobel (1982): "Strategic Information Transmission," *Econometrica*, 50(6), 1431–1451.

Cripps, M., G. Mailath, and L. Samuelson (2004): "Imperfect Monitoring and Impermanent Reputations," *Econometrica*, 72, 407–432.

Dellarocas, C. (2005): "Reputation Mechanism Design in Online Trading Environments with Pure Moral Hazard," *Inform. Systems Res.*, 16(2), 209–230.

Dellarocas, C., and C. Wood (2008): "The Sound of Silence in Online Feedback: Estimating Trading Risks in the Presence of Reporting Bias," *Management Sci.*, 54(3), 460–476.

Ekmekci, M. (2010): "Sustainable Reputations with Rating Systems," *Journal of Economic Theory (to appear)*.

Fan, M., Y. Tan, and A. Whinston (2005): "Evaluation and Design of Online Cooperative Feedback Mechanisms for Reputation Management," *IEEE Trans. on Knowl. and Data Eng.*, 17(2), 244–254.

Farrell, J. (1987): "Cheap Talk, Coordination, and Entry," *The RAND Journal of Economics*, 18(1), 34–39.

Farrell, J., and M. Rabin (1996): "Cheap Talk," *The Journal of Economic Perspectives*, 10(3), 103–118.

Fudenberg, D., and D. K. Levine (1992): "Maintaining a reputation when strategies are imperfectly observed," *The Review of Economic Studies*, 59(3), 561–579.

Ghose, A., P. Ipeirotis, and A. Sundararajan (2005): "Reputation premiums in electronic peer-to-peer markets: Analyzing textual feedback and network structure," in *P2PECON*, pp. 150–154.

Houser, D., and J. Wooders (2006): "Reputation in Auctions: Theory, and Evidence from eBay," *J. Econom. & Management Str.*, 15(2), 353–369.

Kalyanam, K., and S. McIntyre (2001): "Return on reputation in online auction markets. Working Paper, Santa Clara University," .

Kandori, M. (1992): "Social Norms and Community Enforcement," *The Review of Economic Studies*, 59(1), 63–80.

Lucking-Reiley, D., D. Bryan, N. Prasad, and D. Reeves (2007): "Pennies from eBay: The Determinants of Price in Online Auctions," *J. Industrial Econom.*, 55(2), 223–233.

Mailath, G. J., and L. Samuelson (2001): "Who Wants a Good Reputation?," *The Review of Economic Studies*, 68(2), 415–441.

——— (2006): *Repeated Games and Reputations*. Oxford University Press.

Miller, N., P. Resnick, and R. Zeckhauser (2005): "Eliciting Informative Feedback: The Peer-Prediction Method," *Management Sci.*, 51(9), 1359–1373.

Myerson, R. B. (1997): *Game Theory: Analysis of Conflict*. Harvard University Press.

——— (2009): "Learning from Schelling's *Strategy of Conflict*," *Journal of Economic Literature*, 47(4), 1109–1125.

Pavlou, P., and A. Dimoka (2006): "The Nature and Role of Feedback Text Comments in Online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation," *Inform. Systems Res.*, 17(4), 392–414.

Resnick, P., R. Zeckhauser, J. Swanson, and K. Lockwood (2006): "The value of reputation on eBay: A controlled experiment," *Experimental Econom.*, 9(2), 79–101.

Rosenthal, R. W., and H. J. Landau (1979): "A Game-Theoretic Analysis of Bargaining with Reputations," *Journal of Mathematical Psychology*, 20, 233–255.

Schelling, T. C. (1980): *The Strategy of Conflict*. Harvard University Press, Cambridge, MA.

Shapiro, C. (1983): "Premium for High Quality Products as Returns to Reputations," *Quart. J. Econom.*, 98, 659 – 679.

Sobel, J. (1985): "A Theory of Credibility," *The Review of Economic Studies*, 52(4), 557–573.

Tversky, A., and D. Kahnenman (1973): "Availability: A heuristic for judging frequency and probability," *Cognitive Psychology*, (5), 207–232.

Appendix

Proof of Lemma 1: Consider some $\beta \in \{1, ..., b^* - 1\}$. The optimal strategy of D when starting at β is $\{x^*(b), b = \beta, ..., b^* - 1\}$. We now consider $\beta - 1$, and assume that D uses the following strategy: at $\beta - 1$ she rewards if $X < x^*(\beta)$, at β she rewards if $X < x^*(\beta + 1)$,..., at $b^* - 1$ she rewards if $X < x^*(b^* - 2)$, and at $b^* - 1$ she rewards for any X. That is, she uses the optimal strategy for β starting at $\beta - 1$ and then always rewards if her reputation consists of $b^* - 1$ betrayals. Up to (and including) state $b^* - 2$, this yields payoff $v(\beta)$ since everything is the same as when starting at β. Then, D reaches state $b^* - 1$ with positive probability and gets a strictly positive payment once she is there. Therefore, when starting at $\beta - 1$ the described strategy yields a strictly higher payoff than $v(\beta)$; thus $v(\beta - 1) > v(\beta)$.

Proof of Proposition 1: We first show that $x^*(b)$ is strictly increasing in b for $b \in \{0, ..., b^* - 1\}$. (3) can be rewritten as

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

Substituting in (4) we have

$$\frac{1 - s(1-q)}{s(1-r-q)} x^*(b) + (1-s)v(b+1) = 1 + \mathbb{E}[(X - x^*(b))^+]. \tag{14}$$

Let $b_1 < b_2$, and let $x_1 = x^*(b_1)$ and $x_2 = x^*(b_2)$ be the corresponding solutions of (14). Then,

$$\frac{1 - s(1-q)}{s(1-r-q)} x_1 + (1-s)v(b_1+1) = 1 + \mathbb{E}[(X - x_1)^+]. \tag{15}$$

$$\frac{1 - s(1-q)}{s(1-r-q)} x_2 + (1-s)v(b_2+1) = 1 + \mathbb{E}[(X - x_2)^+]. \tag{16}$$

Suppose that $x_1 \geq x_2$. Then,

$$\begin{aligned} &\frac{1 - s(1-q)}{s(1-r-q)} x_2 + (1-s)v(b_2+1) < \\ &\frac{1 - s(1-q)}{s(1-r-q)} x_1 + (1-s)v(b_1+1) = \\ &1 + \mathbb{E}[(X - x_1)^+] \leq \\ &1 + \mathbb{E}[(X - x_2)^+], \end{aligned}$$

which contradicts (16). We note that the first inequality follows because v is decreasing in b (by Lemma 1) and $s < 1$; the equality follows from (15), 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 $b \in \{0, ..., b^* - 1\}$. Let

$$g(y) \equiv 1 + \mathbb{E}[(X - y)^+] = 1 + \int_y^\infty (x - y) f(x) dx.$$

We observe that $g'(y) = -(1 - F(y))$, where $F(y) \equiv \int_{-\infty}^{y} f(x)dx$ is the cumulative distribution function of the random variable X. This implies that $g'(y)$ is negative and increasing in y, and thus g is decreasing and convex. From (14) 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 (3) 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

$$\begin{aligned}&\frac{1-s(1-q)}{s(1-r-q)}(x^*(b_1) - x^*(b_1-1)) + (g(x^*(b_1-1)) - g(x^*(b_1))) < \\ &\frac{1-s(1-q)}{s(1-r-q)}(x^*(b_2) - x^*(b_2-1)) + (g(x^*(b_2-1)) - g(x^*(b_2)))\end{aligned} \tag{17}$$

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 (17). 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: We first show that $b^*(D) = B^*_{pure}$. 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 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 $b^*_\alpha(S)$ which optimizes the weighted return of C and D? Because the return for D decreases when b^* decreases, it is not possible that $b^*_\alpha(S) < b^*(C)$, because both players would be better off with $b^*_\alpha(S) = b^*(C)$. It is also not possible that $b^*_\alpha(S) > b^*(D)$, because $b^*(D)$ is the highest b^* possible in the equilibrium set. Then we have $b^*(C) \leq b^*_\alpha(S) \leq b^*(D)$ for all $\alpha \geq 0$.

Proof of Lemma 3: Let $\pi(i)$ be the probability that a randomly chosen D player has i betrayals. For $i = 0$ the balance equation is:

$$\pi(0) = s((1-q)(1-w(0)) + rw(0))\pi(0) + (1-s).$$

In particular, when D has no betrayals, she rewards with probability $1-w(0)$ and betrays with probability $w(0)$. If she rewards, her reputation remains at $b=0$ with probability $1-q$; if she betrays, her reputation remains at $b=0$ with probability r. Finally, from any state, there is a transition to 0 with probability $1-s$.

Similarly, for $b\in\{1,...,b^*-1\}$ we have

$$\pi(b)=s(q(1-w(b-1))+(1-r)w(b-1))\pi(b-1)+s((1-q)(1-w(b))+r\cdot w(b))\pi(b)$$

Solving for $\pi(0)$ and $\pi(b)$ we get (6) and (7) respectively.

Proof of Lemma 4: Let $p^*(b)$ be the probability that C trusts D when D has b betrayals. As before, let $v(b)$ be the maximum expected infinite horizon payoff to D when she has b betrayals. Then,

$$\begin{aligned}v(b)=&p^*(b)\left(1+\mathbb{E}[\max\{X+s((1-r)v(b+1)+r\cdot v(b)),s((1-q)v(b)+q\cdot v(b+1))\}]\right)\\&+(1-p^*(b))sv(b)\end{aligned}$$

In particular, with probability $p^*(b)$, C trusts D and then D chooses whether to reward or betray. On the other hand, with probability $1-p^*(b)$, C does not trust D. In that case, D receives zero payment in this period and her reputation remains the same.

Straightforward calculations show that

$$v(b)=p^*(b)\left(1+s(1-q)v(b)+s\cdot q\cdot v(b+1)+\mathbb{E}[(X-x^*(b))^+]\right)+(1-p^*(b))s\cdot v(b),$$

where

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

These equations yield (8) and (9).

We next consider player C. As before, C trusts D if $\mathbb{P}[X>x^*(b)]>1/2$; C does not trust D if $\mathbb{P}[X>x^*(b)]<1/2$; and C is indifferent between trusting and not trusting if $\mathbb{P}[X>x^*(b)]=1/2$. Thus, if $x^*(b)>m$, then C plays $p^*(b)=1$; if $x^*(b)=m$, any $p^*(b)\in[0,1]$ is a best response for C; and if $x^*(b)<m$, then C plays $p^*(b)=0$.

Proof of Lemma 5: Suppose that $p^*(b)>0$ for all b. Then, (8) implies that

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

because $m>0$.

On the other hand, by substituting (8) in (9), we have that

$$(1-s)v(0) = p^*(0)\left(1+\mathbb{E}[(X-x^*(b))^+] - \frac{q}{1-r-q}x^*(b)\right),$$

which cannot hold if $v(0)=\infty$, since $\mathbb{E}X<\infty$ and $p^*(0)\leq 1$. We conclude that there always exists a finite cutoff.

Proof of Lemma 6: Let $\pi(b)$ be the probability that a randomly chosen D player has b betrayals. For $b=0$ the balance equation is:

$$\pi(0) = s(((1-q)(1-w(0))+rw(0))p^*(0)+(1-p^*(0)))\pi(0)+(1-s).$$

In particular, when D has no betrayals, C trusts her with probability $p^*(0)$. If D is trusted, then she rewards with probability $1-w(0)$ and betrays with probability $w(0)$. If she rewards, her reputation remains at $b=0$ with probability $1-q$; if she betrays, her reputation remains at $b=0$ with probability r. Finally, from any state, there is a transition to 0 with probability $1-s$.

Similarly, for $b\in\{1,...,b^*-1\}$ we have

$$\begin{aligned}\pi(b) =& s(q(1-w(b-1))+(1-r)w(b-1))p^*(b-1)\pi(b-1)\\ &+s(((1-q)(1-w(b))+rw(b))p^*(b)+(1-p^*(b)))\pi(b)\end{aligned}$$

Solving for $\pi(0)$ and $\pi(b)$, we get the equations of the lemma.

C's expected payoff is then equal to

$$\sum_{b=0}^{b^*-1} p^*(b)\pi(b)(1-2w(b))$$

If $p^*(b)$ is strictly between 0 and 1, then $x^*(b)=m$ and $w(b)=1/2$, and thus C's expected payoff in that period is equal to 0. This observation implies that C's expected payoff can be equivalently written as

$$\sum_{b:p^*(b)=1} \pi(b)(1-2w(b)).$$

Proof of Proposition 3: We show that D's payoff is strictly greater at a dominant extend equilibrium. Consider the pure equilibrium with $b^*=B^*_{pure}$, and suppose that D's strategy is $\{x^*(b), b=0,1,...,b^*-1\}$. Let $v_{PE}(0)$ be D's expected payoff at this pure equilibrium. Now suppose there exists a dominant extend equilibrium with cutoff $b^*=B^*_{pure}+1$. D's expected payoff in the dominant extend equilibrium will be at least as much as the payoff she would get by using $x^*(b)$ for $b=0,1,...,B^*_{pure}-1$ and always rewarding at B^*_{pure}. This strategy yields payoff $v_{PE}(0)$ up to state $B^*_{pure}-1$. Then, D reaches a reputation of B^*_{pure} betrayals with positive probability; and once

she has B^*_{pure} betrayals, C trusts her with positive probability. Thus, in expectation she gets a strictly positive payoff once she has B^*_{pure} betrayals. Therefore, D's payoff at the dominant extend equilibrium is strictly greater than her payoff at a pure equilibrium.

LEMMA 8. *For* $b \in \{0, 1, ..., B^*_{pure} - 1\}$, $w_{PE}(b) \leq w_{DEE}(b)$.

Proof: It suffices to show that $x^*_{PE}(i) \geq x^*_{DEE}(i)$, since w is decreasing in x^*.

We rewrite (14) from the proof of Proposition 1.

$$\frac{1-s(1-q)}{s(1-r-q)}x^*(b) + (1-s)v(b+1) = 1 + \mathbb{E}[(X - x^*(b))^+].$$

Since the left-hand side is increasing in $x^*(b)$ and the right-hand side is decreasing in $x^*(b)$, the solution $x^*(b)$ decreases as $v(b+1)$ increases. (The formal proof of this is similar to the first part of the proof of Proposition 1.) We use this fact to show that $x^*_{PE}(i) > x^*_{DEE}(i)$ for $b \in \{0, 1, ..., B^*_{pure} - 1\}$ by induction. The induction hypothesis is that $x^*_{PE}(i) \geq x^*_{DEE}(i)$ and $v_{PE}(i+1) < v_{DEE}(i+1)$.

- Basis: $i = B^*_{pure} - 1$. We have that $v_{PE}(B^*_{pure}) = 0$, while $v_{DEE}(B^*_{pure}) > 0$. Thus $v_{PE}(B^*_{pure}) < v_{DEE}(B^*_{pure})$ and $x^*_{PE}(B^*_{pure} - 1) > x^*_{DEE}(B^*_{pure} - 1)$.
- Induction Step: Suppose that $x^*_{PE}(i) > x^*_{SE}(i)$ and $v_{PE}(i+1) < v_{DEE}(i+1)$. Then, by (4), $v_{PE}(i) < v_{DEE}(i)$, which in turn implies that $x^*_{PE}(i-1) > x^*_{DEE}(i-1)$ and completes the induction step.

Proof of Proposition 4: Suppose that $b^*(n) > b^*(n+1)$. Suppose that D has $b^*(n+1)$ betrayals and n transactions. C will trust D in this state, because $b^*(n) > b^*(n+1)$. However, D knows that whatever she does in this period, C will not trust her in the next period. So then, it is optimal for her to betray. But if D betrays, C has 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,

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

Consider an equilibrium where C's strategy is given by $\{b^*(n)\}$ and D's strategy is $\{x^*(b, n)\}$ 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 D 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 D's current reputation is $(0, n)$ and let x be the current realization of X. If D betrays now, her current payment will increase by x. We next upper bound the amount that D will lose by betraying now. First observe that the earlier time that D may be expelled is K periods later. This is a very

conservative estimate, because D will probably not always betray and $b^*(n)$ may be increasing. We next observe that D misses at most $1+\mathbb{E}X$ in expectation for each period after she is expelled. Considering that D 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 D betrays now, her future payment will decrease by at most

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

We conclude that D 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

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

So D is better off using $x(0,n) < m$.

Thus, for any given problem there exists some constant 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 $\bar{n}$ the exact number of transactions does not affect the cutoff (which is constant). Without loss of generality, we can restrict the state-space to

$$\{(b,n) : b \leq b^*(n), n < \bar{n}\} \cup \{(b,\bar{n}) : b \leq \bar{b}\},$$

which is finite.

For $n < \bar{n}$, equations (11), (12), and (13) need to be satisfied. For $n = \bar{n}$, we have

$$v(b,\bar{n}) = 1 + s\cdot(1-q)\cdot v(b,\bar{n}) + s\cdot q\cdot v(b+1,\bar{n}) + \mathbb{E}[(X - x^*(b,\bar{n}))^+] \text{ for } b < \bar{b};$$

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

$$v(\bar{b},\bar{n}) = 0.$$

Observe that $\bar{n}$ is just a dummy variable in the previous equations. More importantly, if we ignore $\bar{n}$ these are exactly equations (1), (3), and (4), that is, the equations we had in the one-dimensional case, where D's reputation consisted only of the number of betrayals. This observation implies that $\bar{b} \leq B^*_{pure}$, and also $b^*(n) \leq B^*_{pure}$, which concludes the proof for (ii).