The Social Cost of Cheap Pseudonyms

Eric J. Friedman*
Department of Economics, Rutgers University
New Brunswick, NJ 08903.

Paul Resnick
University of Michigan School of Information
550 East University Avenue
Ann Arbor, MI 48109-1092

August 11, 1999

Abstract

We consider the problems of societal norms for cooperation and reputation when it is possible to obtain "cheap pseudonyms", something which is becoming quite common in a wide variety of interactions on the Internet. This introduces opportunities to misbehave without paying reputational consequences. A large degree of cooperation can still emerge, through a convention in which newcomers "pay their dues" by accepting poor treatment from players who have established positive reputations. One might hope for an open society where newcomers are treated well, but there is an inherent social cost in making the spread of reputations optional. We prove that no equilibrium can sustain significantly more cooperation than the dues-paying equilibrium in a repeated random matching game with a large number of players in which players have finite lives and the ability to change their identities, and there is a small but nonvanishing probability of mistakes.

Although one could remove the inefficiency of mistreating newcomers by disallowing anonymity, this is not practical or desirable in a wide variety of transactions. We discuss the use of entry fees, which permits newcomers to be trusted but excludes some players with low payoffs, thus introducing a different inefficiency. We also discuss the use of free but unreplaceable pseudonyms, and describe a mechanism which implements them using standard encryption techniques, which could be practically implemented in electronic transactions.

*The authors would like to thank Roger Klein, Jeff MacKie-Mason, Rich Mclean, Hiroki Tsurumi, and workshop participants at Michigan, Rutgers and Stonybrook for helpful conversations and comments. Email: friedman@econ.rutgers.edu and presnick@umich.edu. The latest version is available at Δhttp://econ.rutgers.edu/home/friedman/research.htm#wpapers

1 Introduction

One of the fundamental questions of social theory is the conditions which facilitate cooperation. Repetition and reputation are two of the most useful features. Repetition causes people to cooperate in the present in order to avoid negative consequences in future interactions with the same people. Reputations spread information about people's behavior, so that expectations of future interactions can influence behavior even if the future interactions may be with different people than those in the present. The ways in which reputations spread can affect their ability to influence behavior, and it is especially interesting to consider situations where people exercise some control over the spread of their own reputations, a situation that is common on the Internet.

The Internet has spawned numerous social and business environments that allow frequent and meaningful interactions among people who are relative strangers. This leads to many problems and properties that do not usually arise in other social settings. However, the pliability of the Internet as a social structure also allows for a large degree of "engineering" which is also more difficult in standard social settings, allowing for the solution of many of these problems, and providing a fertile ground for the application of many tools from economics and game theory.[1]

The key aspect of reputation on the Internet that does not typically arise in non-electronic settings is the ability to easily change one's identity; whereas in real life this is a complex process (often involving national governments and cosmetic surgery) on the Internet an identity change may require just a few keystrokes.[2] Thus, a person has a choice of interacting

[1]For example, recent applications include the economics of information (Varian, 1997), economic aspects of evaluations (Avery, Resnick and Zeckhauser, 1999), aspects of competition (Bakos and Brynjolfsson, 1998), cost sharing (Moulin and Shenker (1992), Herzog, Shenker and Estrin, (1997)) and various properties of learning (Friedman and Shenker, 1998).

[2]Many games, auction sites, and interactive forums allow users to choose a pseudonym when they register. Even services that identify users based on their email addresses do not prevent identity changes, since users can easily acquire new email addresses through free services like Hotmail. Beyond name changes, the Internet enables completely anonymous interactions. For example, anonymizing intermediaries such as remailers and proxy servers can exchange messages between parties without revealing either one's identity to the other (Goldschlag, Reed, and Syverson 1999, Rubin and Reiter 1999). There are even techniques using cryptography that allow for electronic payments where the buyer's identity is untraceable (Schneier, 1996, p. 139-147).

anonymously (by changing identifiers constantly) or maintaining a persistent identity. This case is intermediate between persistent identities and totally anonymous interactions. In effect, the option of anonymity turns the transfer of reputation information into a strategic variable, controlled by each player, in contrast to previous work (Kandori (1992), and Milgrom, North and Weingast (1990)) where reputation transfer is limited but not under players' control.[3]

With name changes, people can easily shed negative reputations.[4] This makes it natural to distrust newcomers, since they may really be people who have just changed identifiers. There can still be a fair amount of cooperation, however, as people will want to develop positive reputations. For example, the on-line auction service eBay (www.ebay.com) maintains a "Feedback Forum" for buyers and sellers to make comments about each other, after a trade is completed.[5] As analyzed by Peter Kollock (1998), people go out of their way to accumulate positive comments and, once they have accumulated them, to avoid negative comments.[6]

Newcomers can overcome initial distrust by accepting bad treatment for a while, a form of dues paying that is sufficient to discourage participants from changing identifiers. But suspicion of newcomers is socially inefficient, especially in free-flowing environments with lots of newcomers: it would be more efficient to trust newcomers until they proved untrustworthy, if that did not provide incentives for participants to misbehave and then change identifiers. The distrust can be eliminated entirely through a subtle punishment strategy, where newcomers are distrusted only if a veteran player in the previous period did some-

[3]Tadelis (1999) considers an interesting model where reputation transfer is a strategic variable but where performance is not. In that model, names may be traded from higher skill to lower skill players, decreasing but not eliminating the signaling value of reputations. By contrast, we are interested in situations where reputations serve not as signals of underlying skill but as motivators for good performance.

[4]On the Internet, nobody knows that yesterday you were a dog, and therefore should be in the doghouse today.

[5]Recognizing the importance of persistent reputations, eBay offers an easy name-changing facility, but a person's feedback comments follow such name changes. This attempt to limit reputation shedding may be futile, however, since a person can easily acquire a new email address and then re-register with no trace of the earlier comments.

[6]One participant reported that after an accidental snafu, he received a check from the seller for more than the purchase price of the item he had bought, along with a request not to enter a negative comment. [David Richardson, personal communication, January 1998.]

thing wrong. That strategy is quite brittle, however, in the face of either a few malicious participants[7] or occasional mistakes (trembles), such as typing the wrong key by accident. In fact, with either malicious players or occasional trembles, we prove that there is no way to achieve substantially more cooperation in equilibrium than that achieved by distrusting all newcomers. Thus, the distrust of newcomers is an inherent social cost of easy identity changes.[8] Recently, many health support forums have been shaken by people who pretend to have severe illnesses and other problems. Once found out such people often reappear on the same or a different forum with a new identity and repeat the process. Some groups have developed defenses involving distrust of newcomers, but these interventions impose other costs. Alice Grady (1998) reports:

> In another scam that dragged on for months last year, a girl [Kim] who said she was 15 communicated on line with parents of premature infants. The 400 or so members in the virtual support group had babies who were or had been critically ill and had spent months in the hospital. Some of the infants died, and some who survived were expected to suffer lifelong disabilities...
>
> Regardless of what drove Kim, her behavior had a chilling effect on a group that had been trusting and closely knit. Some parents expressed feelings of betrayal, and many stopped posting messages. People in the group agreed to provide information so a coordinator could verify that they really were parents of preemies. Some newcomers were put off by the atmosphere of suspicion.

Unfortunately, the obvious solution of disallowing anonymity is problematic for a variety of reasons, from questions of civil liberties to the practical effects on information exchange. On-line support forums for diseases such as AIDS could not function without some guarantees of anonymity, in order to avoid negative consequences from people they know in real life finding out about their condition. On a lighter side, many users of gaming network sites such as chess (chess.onenet.net), backgammon (www.netgammon.com), bridge (www.okbridge.com),

[7] If it were possible to collapse an entire social order with a single malicious act, then it is hard to imagine that some player would not topple the system for 'fun.' Consider such common entities on the Internet as viruses and worms.

[8] This is analogous to a model of long-term relationships where individuals have a choice at each period whether to continue the current relationship or start a new one (the analog of adopting a new identifier in our analysis). Watson (1999) describes a slow-start equilibrium where partners play for low stakes initially and gradually trust more over time. The disparity between the expected future trade value with an existing partner and a new partner encourages good behavior with existing partners.

go (igs.nuri.net) or quake (www.idsoftware.com) prefer anonymity, while role playing games depend fundamentally on the disassociation between real identities and roles played. Despite the disassociation with real identities, information about the past behavior of players is important in choosing partners who play at a similar skill level, have a compatible sense of sportsmanship, and have fast network connections. Thus, we must encourage players to maintain a persistent identifier within each social arena without relying on the verification and revelation of true identities.

An obvious candidate is the use of entry fees (associated with each personal identifier). One commonly used procedure is a time consuming registration process; while such a procedure may encourage cooperation, it is clearly wasteful. Using monetary registration fees entails no such loss of efficiency, as they are pure transfers, but in a heterogeneous environment may prevent players from using the system, which is clearly inefficient as these systems are essentially public goods with zero or effectively negative marginal cost (due to network externalities).

The conventional wisdom is that there is an inherent tradeoff between anonymity and accountability. For example, several articles in a special issue of *The Information Society* emphasize that there are real benefits to anonymity, despite the costs that come from reduced accountability (Kling et al 1999, Marx 1999, Nissenbaum 1999). The consensus of a AAAS sponsored conference, as reported by Teich et al (1999), was that the tradeoffs should sometimes be resolved in favor of anonymity and thus that regulatory regimes should strive to preserve the possibility of anonymity. We show, however, that it is not necessary to choose: there is an intermediate form of anonymity that minimizes the social costs from loss of accountability.

We propose a system of anonymous certificates in which, for each different social arena, a person is given a single identifier that is unrelated to the person's true identity; however, the certificate provider guarantees that each person will only be granted a single certificate (in each arena). We call these once-in-a-lifteime identifiers. A player using a once-in-a-lifetime identifier effectively commits to having his reputation spread through the arena. Given the

option, players would choose to make such commitments and thus achieve the same level of cooperation that would be achieved playing under their true identifiers. If, for example, a collection of support groups defined itself as a single arena, then a malicious intruder could disrupt only one group; exclusion from that group could lead to exclusion from the others. We show that such certificates can be constructed with a large degree of security using standard encryption techniques.

The paper is organized as follows. In the next section we present the basic model. Section 3 considers the effects of a fraction ϵ of malicious players who thrive on sowing discord among the other players and the related scenario where each player trembles with probability ϵ. Section 4 discusses monetary entry fees. Finally, section 5 presents the certificate mechanism for once-in-a-lifetime identifiers.

2 The Basic Model

We consider an infinite, synchronously repeated game with periods $t \in T = \{0, 1, \ldots\}$. In each period, there are M active players. At the end of each period αM exit (e.g., their interests change and they no longer visit the web site or participate in that newsgroup) and the same number of new players enter. (Assume that $\alpha \in (0, 1)$ and αM is an integer.) Players are labeled by $i \in Z_+$, where players $1..M$ enter at $t = 0$. In each period, current players are matched at random (uniformly) to play a prisoner's dilemma[9] with payoffs:

	C	D
C	$1, 1$	$-1, 2$
D	$2, -1$	$0, 0$

At the beginning of each period, active players may have the choice of continuing to play under their current identifiers or getting new ones. (When obtaining a new identifier is possible, it is costless; we discuss entry fees in Section 4.) When players are paired, each is

[9]While some of the real-world interactions we discuss, such as health support forums, are not pure prisoner's dilemmas, they do contain opportunities for either selfish behavior that hurts others or for more cooperative, mutually benefical behavior. The prisoner's dilemma is a useful model because it places the incentives for choosing between these behaviors in stark relief.

told only the identifier currently being used by the other. Thus, a player who acquires a new identifier is indistinguishable (to all the other players) from a new entrant to the system.

We assume that the system keeps a public history of which identifiers were paired in previous periods and the actions taken by the players controlling those identifiers.[10] Thus, when two players meet, each can see the opponent's complete history, which includes not only the actions played by the opponent, but also those by the opponent's opponents, ad infinitum. To model this simply, we assume that in period t the entire history of play, $h_s^t \in H_s^t$, is common knowledge, where h_s^t is the pairing of identifiers and the actions taken in time periods prior to t.[11] Each player also knows her own personal history of name changes, $h_i^t \in H_i^t$, where h_i^t is the history of identifiers used by player i in periods prior to t.

We will also assume that there is an exogenous signal q, which is uniformly distributed on $[0, 1]$. This signal is revealed at the beginning of period t before players choose their actions.[12] Player i's strategy in period t is a mapping $s_i^t : \ H_s^t \times H_i^t \times H_E^t \to \Delta(\{C, D\})$, where H_E^t is the history of exogenous signals up to and including q^t. Let S be the set of all such (mixed behavioral) strategies.

A player's payoff for a strategy s_i is given by the total (undiscounted) expected payoff. For example, if player i enters in period $b(i)$ then her payoff from strategy vector s, which includes her own strategy and strategies for each of the other players, is given by $u_i(s) = \sum_{t=b(i)}^{b(i)+l(i)} u_i^t$ where u_i^t is her payoff in period t and $l(i)$ is the "age" of player i when she exits the system.

[10] In our scenario, it is not possible to have explicit norms of behavior that are centrally enforced. It may be fairly easy, however, to publish the history, and leave the enforcement up to the actual players. In practice, this history is captured either by monitoring play, such as on the Internet Go Server (igs.nuri.net), or by gathering explicit feedback from participants about each other, as on the Internet Auction site eBay (www.ebay.com). Wherever possible, explicit and centrally enforceable rules of behavior should be applied, and then one can interpret our analysis as modeling the part of the interaction for which such rules are not centrally enforceable.

[11] This assumption is made to simplify notation; the equilibrium strategies that we are interested in will use far less information. Also, this assumption allows us to disentangle the effects caused by name changes from those generated by imperfect transfer of information about the history of play. The question of the reliability of player reported information is quite complex and beyond the scope of this paper. For example, eBay founder Pierre Omidyar exhorted users to give negative feedback when it was warranted (Omidyar, 1998), apparently because users hesitated to give negative feedback in fear that it would be reciprocated.

[12] Once we introduce trembles (Section 3) there will be no need for exogenous signals, as players would be able to correlate on the history of trembles. Nonetheless, we will maintain the exogenous signals to simplify the presentation.

Note that the expected lifetime of a player is $1/\alpha$ so we define the normalized (per-period) payoff to be $\alpha E[u_i(s)]$.[13] We will consider only sequential equilibria, for which we use the standard definition (e.g., Fudenberg and Tirole (1991)).

Our benchmark for the amount of cooperation will be the average among all the players of the expected per-period payoff,

$$V(s) = \liminf_{N\to\infty} \frac{\sum_{i=0}^{N} \alpha E[u_i(s)]}{N}$$

Thus, if every player cooperates in every period then $V = 1$ while if every player defects in every period then $V = 0$.

2.1 Fixed identifiers

If players are unable to change their identifiers (eg., if they must reveal their real-world names), the public history makes total cooperation a sustainable equilibrium. For example, suppose every player adopts the following localized punishment strategy (LPS). LPS calls for a player to play C against a newcomer or against a veteran who complied with LPS in the previous period, and D against a veteran who deviated in the previous period. For $\alpha \leq 1/2$, $V(LPS) = 1$ when identifiers are fixed.

2.2 Free identifier changes

If players can change their identifiers freely, LPS is no longer an equilibrium, because a player can defect, then acquire a new identifier and be treated as a newcomer, against whom other players cooperate. Another strategy, however, does lead to total cooperation in equilibrium. That strategy, the "public grim trigger strategy" (PGTS – a generalized punishment strategy), has every player defect if there has ever been a defection in an earlier period, and otherwise cooperate. As long as $\alpha \leq 1/2$, cooperation is the best strategy while everyone else is cooperating and defection is the best strategy if a defection triggers everyone else to start defecting. Thus, PGTS is a sequential equilibrium. For $\alpha \leq 1/2$, $V(PGTS) = 1$.

[13] Note also that $1 - \alpha$ plays a role analogous to a discount factor, although this model does not include an explicit discount factor.

Intuitively, however PGTS seems fragile and unrealistic. We now introduce a new element into the model that highlights the fragility of PGTS.

3 Malicious Players and Trembles

Suppose there are a few malicious players who like to see others suffer and thus will choose actions that cause a general increase in the level of defection. Malicious players make up a small but non-zero fraction of the population, ϵ. Alternatively, consider the problem of occasional mistakes by well-meaning players. These may occur from errors of judgment, unstable network connections (a player who is faring badly in a backgammon game may quit the game because of a lost network connection, which could appear to the opponent as poor sportsmanship) or simply because a person mistakenly hits the wrong key on a keyboard. Let ϵ be the probability that when a player attempts to choose an action, that she trembles and plays the other action. In the presence of trembles, a strategy defines the deliberate choices that players make, conditioned on the observed actions of others. Trembles are randomly determined after their deliberate choices, so that a player who deliberately chooses D will actually play C with probability ϵ, and vice-versa.[14]

In our model, we need some exogenous variability in the number of new identifiers each period, to eliminate unrealistic strategies that trigger based on the number of new identifiers. The variability could come from variation in the number of players leaving the game (so long as the number of arrivals matches the number of departures). Instead, we assume that at the end of each period each player 'loses' his identifier by accident with probability ϵ and must start again as an entrant with a new identifier. There is no reason why the probability of losing one's identifier should be equal to that of trembling; we simply assume this to reduce notation.

The effects of malicious players are similar to those for trembles, since each introduces a few defections that are not chosen by normal players. In the remainder of this paper we focus

[14]Note that when there are finite trembles, this game is essentially a repeated game with imperfect public information (Green and Porter 1984, Abreu, Pearce, and Stachetti, 1990), i.e., players cannot always tell whether defections were deliberate or caused by trembles.

on the model with trembles. The analysis for the game with malicious players is analogous.[15]

We will be interested in the social welfare for fixed ϵ and large populations. Let $V^*(\epsilon, M)$ be the supremum of $V(s)$ over all sequential equilibria, with population M in each period and the probability of a tremble ϵ and probability of a lost identifier ϵ. Define the "stable value" of the game to be

$$SV = \lim_{\epsilon \to 0} \lim_{M \to \infty} V^*(\epsilon, M).$$

Thus, the stable value is the maximal expected per-period social welfare when the population is large in relation to the error rate.[16]

To simplify presentation and analysis, we will rely on order notation. Thus, the statement $g(\epsilon) = O(f(\epsilon))$ implies that there exists some $c > 0$ such that for ϵ sufficiently small, $|g(\epsilon)| \leq cf(\epsilon)$. Similarly, for M, where we are interested in large values, $g(M) = O(f(M))$ implies that there exists some $c > 0$ such that for M sufficiently large, $|g(M)| \leq cf(M)$.

3.1 Fixed Identifiers

First, consider the case in which players can not change their identifiers. The LPS strategy, where players defect against players who deviated in the previous period, is an equilibrium, with no deliberate defection.

Proposition 1 *For all $\alpha < .3$, $M > 1$ and $\epsilon < .1$, LPS is an equilibrium with $V(s) = 1 - O(\epsilon)$. More precisely, $V(s) \geq 1 - 2\epsilon$.*

[15] That a small probability of trembles or malicious players can have important effects is well known. For example, the evolutionary behavior of the prisoner's dilemma is very different with trembles than without (Nowak and Sigmund, 1993) while the effect of a few "atypical players" can have dramatic effects on the set of equilibria (Kreps et. al., 1982).

[16] Note that the stable value differs from some analyses of games with trembles in which the order of the limits is reversed. For example, with the order reversed, Ellison's (1994) analysis shows that the prisoner's dilemma with anonymous random matching attains

$$\lim_{M \to \infty} \lim_{\epsilon \to 0} V(\epsilon, M) = 1$$

using randomized versions of contagion strategies, while Friedman (1997) has shown that the stable value for that game is 0. The SV order of limits is more appropriate for our analysis, where we expect a large enough population so that at least one error per period is expected. The reversed order would effectively take a limit where the number of errors, $M\epsilon$, approaches 0.

Thus, we get the standard result of full cooperation (except for trembles) analogous to that for a prisoner's dilemma of iterated play with the same partner.

Corollary 1 *For the game with persistent identities* $SV = 1$.

3.2 Free Identifier Changes

When players can freely change identifiers, malicious players or trembles ruin the PGTS equilibrium. A single tremble or malicious player causes mass defection in future periods. For any $\epsilon > 0$, PGTS has an expected average per-period payoff of 0; $V(PGTS) = 0$.

As discussed by Ellison (1994), PGTS can be replaced by a "public forgiving trigger strategy" (PFTS) that works for (very) small $\epsilon > 0$. In PFTS a player cooperates until the first time someone defects. Then she chooses D for a finite number of periods (the punishment phase) after which she goes back to cooperating. For fixed ϵ, however, as M gets large, there will be a tremble in almost every period and this will not be an equilibrium.[17]

The point of the punishment phase is to deter non-malicious players from defecting. An alternative way to do that is through a "paying your dues" strategy (PYD), which makes it much less attractive to have a new identifier than one that has a history of cooperative action. Essentially, it rewards positive reputations rather than punishing negative reputations. Under PYD, when an entrant meets a compliant veteran (non-entrant) the entrant chooses C and the veteran chooses D. Thus, 'dues' are transferred from the entrant to the veteran, although at a cost to overall efficiency. The dues act as an endogenous entry fee, discouraging a veteran from deviating since he must then change his name, behave as an entrant and pay more dues.

While our prisoner's dilemma model suggests that dues be paid in the form of "defection" against cooperative newcomers in a simultaneous game, in practice newcomers' dues may take several forms. In fraternity initiation, newcomers perform work or accept humiliation. At eBay, newcomers may have to accept shipping delays (for example, a seller may wait for

[17]Even if trembles are very rare, an environment where PFTS operates may attract a malicious player, since such a player can create a large disruption. Thus in (the perhaps more realistic) case when the number of malicious players is endogenous our discussion is also valid.

a newcomer's check to clear before sending goods, but send goods immediately to veteran buyers). Kollock (1998) reports that newcomers accept more transaction risks in the on-line environment for trading playing cards for the game *Magic*. After people who have never met each other agree to exchange cards (or sell cards for money), the person without an established reputation has to send his card first; the veteran should reciprocate after receiving the newcomer's card, but the newcomer accepts the risk that this might not happen.

Formally, the PYD strategies are as follows. Identifiers are divided into two types, entrants (those that have no history of previous actions), and veterans. Identifiers are said to be "in compliance" if all their past actions conform to the PYD strategy (entrants are trivially in compliance). Note that an identifier can remain in compliance even after defecting, so long as the defection was called for in the PYD strategy. A player always cooperates if both she and her opponent are the same type and "in compliance". If a compliant veteran meets an entrant then the entrant chooses C and, if $q < \hat{q}(\alpha, \epsilon, M)$, the veteran chooses D (otherwise C) where

$$\hat{q}(\alpha, \epsilon, M) = \frac{1 - 1/M}{(1-\alpha)\,(2 - \alpha - 2/M - \epsilon + \epsilon/M + \epsilon\,\alpha)\,(1 - 2\,\epsilon)}$$

(Note that to improve efficiency we only require dues to be paid part of the time. This is the reason for introducing the exogenous signal q.) If either player is not in compliance, then both players choose D. Finally, the strategy calls for a player whose identifier is not in compliance to take on a new identifier and begin again as an entrant.

The function $\hat{q}(\alpha, \epsilon, M)$ is precisely the minimal punishment probability to prevent a player from deliberately deviating and then returning as an entrant in the following period. In the absence of trembles, for M going to infinity, this equilibrium has expected payoff per player of $\frac{1}{\alpha} - \frac{1}{2-\alpha}$. Some of a player's first-period dues may be recovered in later periods if the player is allowed to defect against a newcomer, but there is a net loss of between .5 and 1 utils per player. As the following proposition shows, for small ϵ, PYD is still an equilibrium with approximately the same payoffs.

Proposition 2 *For $\alpha < .3$, $\epsilon < .1$ and $M > 11$, and $\hat{q}(\alpha, \epsilon, M) \leq 1$, PYD is an equilibrium*

of the game with impersistent identities, where $V(s) = 1 - \frac{\alpha}{2-\alpha} - O(\epsilon) - O(1/M)$.

From this we get a lower bound for the stable value, resulting from the dues paid by newcomers (essentially, the $\frac{\alpha}{2-\alpha}$ term). As epsilon go to zero in computing the stable value limit, the losses by veterans who tremble goes to 0. (Note that the condition $\hat{q}(\alpha, \epsilon, M) \leq 1$ is automatically satisfied when $\alpha < .24$ or $\epsilon < .05$.)

Corollary 2 *For the game with impersistent identities* $SV \geq 1 - \frac{\alpha}{2-\alpha}$.

Note that for small ϵ, the PYD equilibrium implies an average loss in (unnormalized) payoffs to each player of $(2-\alpha)^{-1}$, which is approximately $1/2$ for α close to zero.

Although compliant veterans never deliberately deviate from the PYD equilibrium, the equilibrium includes defections. There is dues paying by newcomers and by veterans who trembled in the previous period, leading to some inefficiency. It is logical that trembling players be punished, else other players will misbehave and claim to have trembled. It seems wasteful, however, to punish the true newcomers, who have done nothing wrong. If, somehow, the trembling players were usually punished but the true newcomers usually were not, such an outcome would have value $V(s) = 1 - O(\epsilon)$ and a stable value of 1 even though previous period deviants cannot be distinguished from this period's true newcomers.

It is, in fact, possible, to do somewhat better than the PYD equilibrium. For example, consider a variant that omits the dues for newcomers in any period following one where there are no deviations. This strategy yields an equilibrium, but for fixed ϵ, as M gets large there will almost always be at least one deviation, so the improvement over PYD is only $O(e^{-\epsilon M})$.

More generally, there may be ways to condition the payoffs for newcomers in the next period on the collective behavior of veterans in the previous period (more dues next period if more deviation this period). This increases the fraction of the total dues paid by tremblers (as opposed to true newcomers). Our next proposition, however, shows that no equilibrium yields significantly higher payoffs than PYD. Thus, while there can be improvements over the PYD equilibrium, the improvements are slight and the bound for the stable value given in Corollary 2 is tight.

The key ideas in the proof are:

1) Although an equilibrium can have unusual behavior for special periods or special players, on average, veterans must receive expected payoffs that are sufficiently larger than the entrants' payoffs to prevent someone from defecting and then returning in the following period as a new entrant. (See lemma 2 in the proof).

2) The "most efficient" (i.e., with the fewest defections) way to create a differential between the value of being a veteran rather than an entrant is by having a veteran defect against an entrant, since this "transfers" utility from the entrants to the veterans. (See lemma 5 in the proof).

Proposition 3 *Fix $\alpha < .3$. There exists some $\beta > 0$ such that for any $\overline{v} > 1 - \frac{\alpha}{2-\alpha}$ there exists an $\overline{\epsilon}$ such that for all $\epsilon < \overline{\epsilon}$ and $M > \beta/\epsilon$ there is no equilibrium, s^*, of the game with impersistent identities with $V(s^*) \geq \overline{v}$.*

Thus, the stable value is precisely what was obtained from the PYD equilibrium.

Corollary 3 *For the game with impersistent identities $SV = 1 - \frac{\alpha}{2-\alpha}$.*

This shows that there is no fully efficient "stable" equilibrium when identities are not persistent and PYD has the highest payoffs (to within $O(\epsilon) + O(1/M)$) of any equilibrium strategy. One further implication of the two ideas behind the proof is that any equilibrium with approximately as much cooperation as PYD must have almost all its defections be by veterans against entrants. Thus, although the PYD equilbrium is not unique, all other equilibria that achieve near maximum efficiency must operate in the same spirit that PYD does, with veterans defecting against entrants. In particular, slow-start schemes where newcomers initially play low stakes games until they build reputations would be less efficient than schemes that transfer utility from newcomers to veterans.

4 Payments for Identifiers

The simplest method to attain full efficiency in the game with impersistent identities and either malicious players or trembles is to make dues paying explicit, such as with the im-

position of an entry fee.[18] It is easy to see that if such a fee is chosen appropriately then players will have a sufficient incentive not to defect from the equilibrium and begin again with a new identifier, as they would then incur a new entry fee.

Suppose that the entry fees collected in period t+1 are distributed evenly among all the players who participated in period t. Since such a fee is purely a transfer it does not impact efficiency. If each player uses, in expectation, the same number of identifiers, then each player will, in expectation, collect back exactly the amount of her entry fee. Players who change identifiers deliberately would increase the amount that other players collect in distributions. Thus, if an equilibrium strategy calls for a player not to change identifiers deliberately, the entry fee would not impact that player's willingness to participate.

While attractive, this scheme suffers from two problems. First, the redistribution payments may introduce incentives for players to stay in the game beyond the time when their natural interest or life circumstances change. Thus, redistribution of entry fees would invalidate our modeling of the exit process as exogenous. Even without this problem, this solution does not work if players' expected lifetimes are heterogeneous. For example, some players may know that they have a short attention span and thus don't expect to be in the system long enough to recoup their entry fee.

These problems can be eliminated if entry fees are not redistributed to the players (perhaps they are given to charity, or kept as profit by an entity running the environment). If, however, player payoffs are heterogeneous, such fees will introduce inefficiency: some players will choose not to participate.

To make this argument explicitly, consider a variation of the game with impersistent identities in which players' varying wealth causes them to value money differently, as modeled by a parameter $\lambda \in (0, 1]$. The expected payoff for a player with intensity λ is $\frac{V(s)}{\alpha} - \lambda F$, where F is the entry fee. Our point is easily made when $\alpha = .1$, and players' intensities are i.i.d. with $\lambda = 1$ (the poor players) with probability p and $\lambda = 0.01$ (the wealthy

[18] An alternative is to require posting of a bond for each new identifier, to be forfeited if some central authority determines that the player has deviated from acceptable behavior. The advantage of straight entry fees over bonds is that no central authority is needed, which is important on the internet where there is often strong distrust of such authorities.

players) with probability $1-p$. It is clear that in this case, the entry fee must be sufficiently large to prevent the wealthy players from deviating, but this will deter the other players from entering, thereby leading to efficiency losses. More generally, the optimal entry fee will often exclude some players yet still be insufficient to deter the wealthiest players from defecting. A similar problem occurs if players have heterogeneous payoffs in the game rather than heterogeneous value for money; in that case, the optimal dues for a PYD equilibrium would also exclude some players from the game yet be insufficient to deter some others from defecting regularly. The problem of large fixed costs deterring some entrants is well known in the economics literature, but standard solutions such as price discrimination or two-part tariffs are not applicable here.

5 Identifier Commitments

We now describe an implementable system which achieves full efficiency even in the presence of heterogeneous payoffs, by allowing players to credibly commit not to change identifiers, still without revealing their true identities. As a starting point, suppose that there were an intermediary, trusted by all players. The intermediary assigns identifiers to players when they request them, but promises never to reveal which players received which identifiers. Suppose that the intermediary also offers a special class of identifiers, which we call *once-in-a-lifetime identifiers* but for each social arena will issue at most one such identifier to each player.[19] A player with a once-in-a-lifetime identifier is not prevented from returning with a regular identifier, although regular identifiers may be viewed with suspicion by other players.

Any equilibrium strategy vector for the game where identifiers are fixed can be extended to a strategy vector for the game where players have the option of using once-in-a-lifetime identifiers. Players choose D against regular identifiers and follow the original strategy

[19] eBay recently introduced Equifax as an optional service provider to authenticate the registration information (name, address, etc.) of users, suggesting that a third-party registration service may be viable. The Equifax service, however, does not quite match our proposal for once-in-a-lifetime pseudonyms. First, it is not clear under what conditions Equifax would reveal the mapping between real identities and pseudonyms. Second, Equifax does not currently advertise a policy of one eBay pseudonym per person, so that it would still be possible to escape a negative reputation by re-registering.

against once-in-a-lifetime identifiers. Since regular identifiers are treated so poorly, use of a once-in-a-lifetime identifier effectively signals a commitment to keep using that identifier rather than returning anonymously. Conversely, a player who does not use a once-in-a-lifetime identifier (i.e., does not make an identifier commitment) signals that she is not trustworthy. In equilibrium, no one uses regular identifiers.

In particular, LPS (defect against anyone who deviated in the previous period) extends to an equilibrium with nearly complete cooperation (only trembles are punished). Note also that even if players differ in the intensity of their payoffs, this remains an equilibrium with full participation and full cooperation, unlike entry fees, which might exclude some players from participating. Thus we note that in this game the stable value is 1.

In this scenario, the players have to trust the intermediary not to reveal their true identities, even though the intermediary knows the mapping between players and identifiers. We can reduce the trust requirement somewhat through a cryptographic technique known as blind signatures (Schneier, p. 112-114).[20] The protocol, though it would actually be implemented using encrypted electronic communications, is easiest to describe with an analogy to carbon paper and envelopes. Player A signs the outside of an envelope with her true signature. A then types up a letter specifying a new once-in-a-lifetime identifier for herself and puts it in the envelope together with a piece of carbon paper. She sends the envelope to the intermediary, who checks A's signature on the envelope without opening it. After checking that A has not previously requested a once-in-a-lifetime identifier, the intermediary signs the outside of the envelope; because of the carbon paper, the signature bleeds through onto the letter. The intermediary sends the unopened envelope back to A, who removes the letter, now signed by the intermediary, and presents it to other players in the game as proof of her once-in-a-lifetime identifier.

[20]One of the strengths of the Internet is the ease with which complicated encryption and verification mechanisms can be implemented. For example, Eudora Lite, a standard email program, is distributed free with Pretty Good Privacy, an encryption program which provides a large degree of security against eavesdroppers. It is easy to use even for the novice as the program does most of the work. Thus, it is possible for ordinary people to use sophisticated encryption programs, something that is quite difficult for non-electronic transactions.

The intermediary never learns what identifier A is using, since it was sealed in the envelope, although the intermediary knows that A acquired some once-in-a-lifetime identifier. This protocol is still subject to a timing attack, however: the intermediary can watch to see what new once-in-a-lifetime identifier is used in the game, and associate it with the last player who requested one. If players wish to avoid this, they need to acquire their identifiers and hold onto them for a random length of time before they use them.

The envelope and carbon paper protocol described above can be implemented quite practically if identifiers correspond to private-public encryption key pairs. Encryption keys are just long strings of bits; the private portion of the key pair is known only to the key's owner, while the public key is available to everyone. A private key is used to "sign" a string of bits by computing a function of the bits and the private key. The function works such that anyone with the corresponding public key can verify that the private key was used to make the signature, but no one can forge a signature by computing the function's output without knowing the private key.[21]

Each player is assumed to start with a private key associated with her true identity.[22] To establish a once-in-a-lifetime identifier for some arena, player A first constructs a brand new key pair (a new pseudonym). A sends the public half of the new pair to the intermediary, but blinds it by multiplying by a randomly chosen number (the equivalent of sealing it in an envelope with carbon paper). The player uses the private key for her true identity to sign the request, so that the intermediary can verify that it came from A (only someone knowing A's private key could have generated the signature). If the intermediary has never previously certified a pseudonym for A, the intermediary uses its own private key to sign the new blinded public key that A provided. A receives the blinded signed key and is able to remove the blinding factor (the equivalent of opening the envelope), leaving a certificate, signed by the

[21]Private-public encryption and signature handling software is already built into the major Web browsers and is routinely used for establishing private communication (URLs that begin https:// usually cause this feature to be invoked) and for assessing the safety of downloaded code.

[22]There are some practical difficulties to be surmounted in setting up an infrastructure for establishing key pairs for individuals and publicizing the public portion. A few companies, most notably Verisign, have established a foothold in this business, and there is also speculation that governments may provide such services.

intermediary, that attests that the new public key is valid as a once-in-a-lifetime identifier. The intermediary knows that A has acquired a once-in-a-lifetime identifier, but does not know which one.

Subsequently, player A can participate in the game without revealing her true identity. She presents the certificate and signs communications with the pseudonym's private key (not the private key associated with her true identity). Other players can verify that the certificate is authentic, using the intermediary's public key to verify the intermediary's signature. They can verify that the communications are signed by whoever owns the once-in-a-lifetime identifier, using the identifier's public key. But no one, not even the intermediary, can tell that the identifier belongs to A.

There can be different intermediaries for different social arenas, or a single intermediary can handle several arenas simultaneously, enforcing a restriction of one once-in-a-lifetime identifier per arena. For game servers or support groups, this process will prevent players returning over and over again with new pseudonyms, while protecting their true identities. There is still a danger that a person can acquire several once-in-a-lifetime identifiers for a single arena, if she uses several people's true identifiers to acquire the certificates. If a robust cryptography infrastructure develops, however, most people will be very reluctant to allow another to use their true identifiers. In any case, the need to use a true identifier to acquire a once-in-a-lifetime identifier will impose almost no cost on individuals who wish to acquire just one, but will impose a significant cost on those who try to acquire several.

How should the intermediary for an arena be selected? One possibility would be for the official intermediary to be allocated according to some public auction. Once chosen, the intermediary will be a monopolist (we cannot have competition unless the competitors share information about which players have already been issued committed identifiers). The initial auction, however, can compete away the monopolists' rents, at least for those services that can be specified by contract. Thus, for example, the winner of the auction may have to agree to provide identifiers for a fixed fee, and within a specified turnaround time, or else lose its franchise. The intermediary may also be required to submit to regular audits, to make sure

that it issues only one once-in-a-lifetime identifier per player.

Note that there is an inherent tradeoff between anonymity and accountability in the choice of how broad a set of activities to define as a single arena. Should the arena in which a person commits to a single identifier consist of eBay's Beanie Baby auctions, all eBay auctions, or all auctions at any on-line service? A broader arena increases accountability, both because there will be more historical data available to assess any individual's reputation, and because a bad reputation follows an individual to more places.

However, the broader the arena, the more opportunities there are for correlating behavior between activities that an individual would like to keep separate. For example, a participant in an education discussion group may not want other participants to know what she has said in a discussion group on politics. In the most extreme case, there would be just one arena for all of the the Internet and hence just one Internet identifier per person. We would expect more narrowly defined arenas, however, in those sensitive areas where people care more about anonymity.

6 Concluding Remarks

Even in the physical world, name changes have always been possible as a way to erase one's reputation. The Internet highlights the issue, by making name changes almost cost-free. This creates a situation where positive reputations are valuable, but negative reputations do not stick. It is natural to ask how much cooperation can be sustained relying only on positive reputations. The answer is, "quite a lot", but not complete cooperation. A natural convention is to distrust or even mistreat strangers until they establish positive reputations.

Suspicion of strangers is costly to society. It is especially costly on the Internet, since the great potential of the medium is to allow people to expand their horizons, to sample a variety of interest groups and to trade with people they have never met. It would be nice to create environments where strangers were trusted until proven otherwise. Unfortunately, obvious strategy vectors involving cooperation with strangers are not stable, and we proved that no strategy vector can do substantially better than punishing all newcomers.

Thus, there is an inherent social cost to free name changes. We can mitigate this cost by charging for name changes, but this also requires charging for names in the first place. That may exclude poor people or those who are just exploring and not yet sure whether the payoffs from participation would justify the entry fee. A better solution is to give people the option of committing not to change identifiers. We described cryptographic mechanisms that enable credible commitment to a single pseudonym within some arena, without revealing one's true identity. We expect both techniques for limiting name changes, entry fees and pseudonym commitments, to blossom in Internet arenas.

A Proofs of Propositions

A.1 Proposition 1

For all $\alpha < .3$, $M > 1$ and $\epsilon < .1$, LPS is an equilibrium with $V(s) = 1 - O(\epsilon)$. More precisely, $V(s) \geq 1 - 2\epsilon$.

The proof of Proposition 1 is similar to standard equilibrium proofs with some complications due to the existence of finite tremble probabilities.

First, consider a single deviation from the asserted equilibrium. The only possibly profitable deviation is to try to defect when LPS calls for cooperation. When the defection is carried out, the gain is 1 as compared to cooperation (for either action by the opponent) but the player's opponent will try to defect in the next period. This increases by $1 - 2\epsilon$ (due to trembles) the probability of the opponent actually defecting in the next period, which would impose a penalty of 2 (for either action by the player). Thus, a decision to deviate will be profitable only if $1 > 2(1 - \alpha)(1 - 2\epsilon)$. But for the given parameters, this is never true: $2(1 - \alpha)(1 - 2\epsilon) > 2(1 - .3)(1 - .2) = 1.12 > 1$. Thus, LPS is an equilibrium.

Next, we compute the per-period average payoff for each player. In any period, some players may have deviated (unintentionally) from LPS. When 2 non-deviators meet, they (attempt to) cooperate and each has an expected payoff of $(1 - \epsilon)^2(1) + \epsilon(1 - \epsilon)(2) + \epsilon(1 - \epsilon)(-1) + \epsilon^2(0) = 1 - \epsilon$. When 2 deviators meet they (attempt to) defect the expected payoff is

$(1-\epsilon)^2(0)+\epsilon(1-\epsilon)(-1)+\epsilon(1-\epsilon)(2)+\epsilon^2(1) = \epsilon$. Similarly, a deviator meeting a non-deviator gets $-1+3\epsilon$ and the opposite yields $2-3\epsilon$. If there were k deviations in the previous period, the average payoff per player in the current period will be:

$$(k/M)^2\epsilon + (1-k/M)(k/M)(-1+3\epsilon) + (1-k/M)(k/M)(2-3\epsilon) + (1-k/M)^2(1-\epsilon).$$

Using the fact that $E[k/M] = \epsilon$ in equilibrium, this is $1 - 2\epsilon + 2\epsilon^2$, which is larger than $1-2\epsilon$.

A.2 Proposition 2

For $\alpha < .3$, $\epsilon < .1$ and $M > 11$, and $\hat{q}(\alpha, \epsilon, M) \leq 1$, PYD is an equilibrium of the game with impersistent identities, where $V(s) = 1 - \frac{\alpha}{2-\alpha} - O(\epsilon) - O(1/M)$.

As in the previous proof, the gain for defecting when the equilibrium strategy calls for cooperating is 1, while the loss arises because the expected payoff in the next period is reduced, since the player must return as an entrant. This loss only occurs when $q_{t+1} \leq \hat{q}$. If the player is matched with a veteran, then the loss is due to the veteran choosing defect (which leads to a loss of 2 utils when it happens) with probability $1-\epsilon$ instead of probability ϵ. If the player is matched with an entrant then the loss is due to the player not defecting (which loses 1 util when it happens) with probability $1-\epsilon$ instead of ϵ. The probability of being matched with an entrant in the next period is the same whether the player deviates or not, and can be calculated from the expected number of trembles this period and the expected number of true newcomers in the next period, $p_e = (M\alpha + (M(1-\alpha)-1)\epsilon)/(M-1)$.

The expected loss, then, from a defection when the strategy calls for cooperation, is $(1-\alpha)(1-2\epsilon)\hat{q}(p_e + 2(1-p_e))$. Thus, players will not try to deviate, which increases the probability of actually deviating, if $1 \leq (1-\alpha)(1-2\epsilon)\hat{q}(p_e + 2(1-p_e))$, which is satisfied with equality for the value of $\hat{q}$ in the proposition.

To compute the expected payoff for this equilibrium, we note that by stationarity and anonymity of PYD we need only compute the average payoff for a period. To do this we note that the total payoff in a period is $M - \hat{M}$ where $\hat{M}$ is the number of defections in

that period, since every defection costs 1 util in total payoffs. Thus $V = 1 - p_D$ where p_D is the probability that a randomly chosen player will defect. The only type of player who attempts to defect in equilibrium is a veteran who is matched with an entrant in a period in which $q_t \leq \hat{q}$. Let p be the probability that a veteran is matched with an entrant. Then $p_D = (1-\epsilon)\hat{q}p + \epsilon(1-p))$. Since in any period there are (on average) $M\alpha + M(1-\alpha)\epsilon$ entrants, $p = \alpha(1-\alpha) + O(\epsilon + 1/M)$. Since $\hat{q} = ((1-\alpha)(2-\alpha))^{-1} + O(\epsilon + 1/M)$ this implies that

$$p_D = \frac{\alpha}{2-\alpha} + O(\epsilon + 1/M)$$

and thus

$$V = 1 - p_D = 1 - \frac{\alpha}{2-\alpha} - O(\epsilon + 1/M)$$

proving the proposition. □

A.3 Proposition 3

Fix $\alpha < .3$. There exists some $\beta > 0$ such that for any $\overline{v} > 1 - \frac{\alpha}{2-\alpha}$ there exists an $\overline{\epsilon}$ such that for all $\epsilon < \overline{\epsilon}$ and $M > \beta/\epsilon$ there is no equilibrium, s^*, of the game with impersistent identities with $V(s^*) \geq \overline{v}$.

We will show that no strategy vector with average expected payoffs greater than $1/\alpha - 1/(2-\alpha)$ can provide sufficient incentives to prevent entrants from defecting. An equilibrium can include unusual behavior in selected periods or by selected players. We establish, however, a minimal difference, on average, between the payoffs of newcomers and veterans. If this minimum is not met, then there will be at least one player (at least one newcomer, in fact) who in some period would deviate from the strategy. This is sufficient to infer a minimal number of defections in any equilibrium strategy.

First note that for any equilibrium, there is a payoff equivalent equilibrium in which no player ever intentionally gets a new identity. Let s be a set of strategies. Define s' to be the set of strategies which are identical with s except for the following. 1) If s tells a player to intentionally get a new identity, then s' has the player maintain her current identity. 2) In

s' when playing against a player who would have gotten a new identity in s treat them as if they were an entrant in the most recent period when they should have gotten a new identity. Clearly, such a change will not affect any player's payoffs or incentives and thus s' is still an equilibrium and is payoff equivalent (along every sample path) to s. Thus, if there is an equilibrium strategy with payoffs greater than our bound, there is also one that involves no deliberate name changes (except after name trembles). Without loss of generality, we assume for the remainder of the proof that strategies involve no deliberate name changes.

Define V_i to be the expected per-period payoff to player i conditional on the history of play before she enters, (note that we are suppressing the explicit notation for histories, for ease of presentation). Note that V_i will be the same for all new identifiers in the period that i begins. Thus we will abuse the notation slightly by writing V_t for the expected per-period payoff to any newcomer in period t, or $V_{b(i)}$ for the expected per-period payoff to any newcomer in i's first period. Define W_i as the expected per-period payoff for player i starting in the second period of participation $(b(i)+1)$, conditional on the fact that the player actually conformed to the strategy in the previous period, conditional on not exiting after the first period, and conditional on all information available at the time of their action choice in the period in which they enter. Define V_i' as the expected per-period payoff to player i starting in the second period of participation, conditional on player i actually deviating and not exiting.

First we note that $V_{b(i)+1}$ is a good approximation for V_i' for 'most' players.

Lemma 1 *For all $\psi, \phi > 0$ there exists some $\beta > 0$ such that for all $M > \beta/\epsilon$ the following holds: Given any $t > 0$ let Z be the set of entrants in period $t-1$; then the set $Z' = \{i \in Z \mid V_i' - V_t > \psi\}$ satisfies $|Z'| < \phi|Z|$.*[23]

We will refer to Z' as the "trigger" players, the ones whose deviations trigger big changes in the payoffs in the next period. The proof is by contradiction. Suppose there exists $\psi, \phi > 0$ such that for any $\beta > 0$ there is a strategy vector s, with $M > \beta/\epsilon$, such that $|Z'| \geq \phi|Z|$.

[23]This result closely parallels the main lemma in Fudenberg, Levine and Pesendorfer (1998), in a different setting.

Let $x \in \{0,1\}^Z$ where $x_i = 0$ if player $i \in Z$ deviates in period $t-1$ and 0 otherwise. Let $V(x)$ be the expected value of V_t under s if x is the actual pattern of deviations by entrants in period $t-1$. Let Σ be the set of all permutations of Z which respect Z', i.e., $\sigma \in \Sigma$ is a mapping $Z \to Z$ such that $\sigma(Z') = Z'$. With a slight abuse of notation let $\sigma(x)$ be the permutation of the vector x by σ, e.g., $\sigma(x)_{\sigma(i)} = x_i$.

Now consider a new function $\hat{V}(\cdot)$ which is defined as follows, $\hat{V}(x) = \sum_{\sigma\in\Sigma} V(\sigma(x))/|\Sigma|$. Define $\hat{V}_i' = E[\hat{V} \text{ in period } t \mid i \text{ deviates in period t-1}]$ and $\hat{V}_t = E[\hat{V} \text{ in period } t]$. Note that since deviations by all trigger players are equally likely that $\hat{V}_t = V_t$. Moreover, if $i \in Z \setminus Z'$, $\hat{V}_i'$ is the average among non-trigger players of V_i' (again, because in equilibrium deviations by all trigger players are equally likely), but $V_i' - V_t \leq \psi$ for all such players, so that $\hat{V}_i' - V_t \leq \psi$ for such players. Similarly, if $i \in Z'$, $\hat{V}_i' - V_t > \psi$. Thus, $\hat{V}$ has the same set of trigger players as V but $\hat{V}(x)$ depends only on the number of deviations by each type of entrant (trigger and non-trigger) and not on which particular players deviate.

We will now show that when there are enough trigger players, each can have only a limited impact on the distribution of the number of deviations, and hence on $\hat{V}$, which contradicts the definition of being a trigger player. Define $\hat{V}^k = E[\hat{V}|k \text{ deviations by trigger players}]$. Thus, $V_t = E[\hat{V}^k]$ while for a trigger player $i \in Z'$, $V_i' = E[\hat{V}^k|$ i actually deviates]. Let $m = |Z'|$, the number of trigger players. Then the probability of k deviations by trigger players is given by the formula for a binomial distribution, $P_k^m = \frac{m!}{k!(m-k)!}\epsilon^k(1-\epsilon)^{m-k}$, while for $k \geq 1$ the probability, contingent on i deviating, is P_{k-1}^{m-1}. This implies that

$$\hat{V}_i' - \hat{V}_t = -\hat{V}^0 P_0^m + \sum_{k=1}^{m} \hat{V}^k [P_{k-1}^{m-1} - P_k^m].$$

Since $\hat{V}^k \in [-1, 2]$ we see that

$$\frac{1}{2}|\hat{V}_i' - \hat{V}_t| \leq P_0^m + \sum_{k=1}^{m} |P_{k-1}^{m-1} - P_k^m|.$$

The sum on the r.h.s. of this equation is equal to

$$\sum_{k=1}^{\lfloor m\epsilon \rfloor} P_k^m - P_{k-1}^{m-1} + \sum_{k=\lfloor m\epsilon \rfloor+1}^{m} P_{k-1}^{m-1} - P_k^m$$

and thus the r.h.s. of that equation is equal to the sum of $|Pr[k \leq \lfloor m\epsilon \rfloor] - |Pr[k \leq \lfloor m\epsilon \rfloor|$i deviates] and $Pr[k > \lfloor m\epsilon \rfloor] - Pr[k > \lfloor m\epsilon \rfloor|$i deviates]$|$ assuming that $m\epsilon$ is not an integer; but both of these terms are small since $m\epsilon$ is the mode of both distributions, and all the probabilities converge to $1/2 + O((m\epsilon)^{-1/2})$ by the central limit theorem (Hoeffding, 1994). Thus, it is easy to show that the r.h.s. of that equation is $O((m\epsilon)^{-1/2})$ for fixed α and is $O((M\epsilon)^{-1/2})$ for fixed ϕ, since, by assumption, $m > \phi M$. Thus for $M\epsilon$ sufficienty large this implies that $\hat{V}_i' - \hat{V}_t \leq \psi$ providing the required contradiction.

Intuitively, the assumption of a constant fraction of trigger players means that, as M gets large, there are a large number of trigger players. But when there are large number of them, and the payoffs are controlled only by the quantity who deviate (and not which ones), one player's deliberate decision to deviate can have only a minor impact on the total payoffs. But that contradicts what it means to be trigger player. $\diamond$

Now, consider an entrant i who in equilibrium chooses C in her first period of play. The immediate benefit from a deviation is 1 while the future cost of returning the next period as an entrant is $\frac{W_i - V_i{}'}{\alpha}$ with probability $(1-\alpha)$. Thus, to maintain equilibrium, we must have $\frac{(1-\alpha)}{\alpha}(W_i - V_i') > 1$.

For an entrant i who in equilibrium chooses D in her first period of play, there is no immediate benefit from a deviation. However, if $W_i - V_i' < 0$ she will choose to get a new name in the following period. Thus, to maintain an equilibrium with no deliberate name changes, we must have $W_i - V_i' > 0$.

Fix $\alpha, T > 0$ and define $V(T)$ to be the expected value of V_i averaged over all players entering before period T, i.e., all i such that $b(i) < T$. Note that since the same number of entrants are expected in each period, $V(T)$ is also the average of V_t over all $t < T$. Similarly let $W(T)$ be the expected average over W_i for all i such that $b(i) < T$.

For $r, s \in \{e, v\}$ let p^{rs} be the empirical probability in periods 1 through T that player of type r defects against a player of type s, while p^r is the empirical probability that type r defects and p is the empirical probability of any player defect. Note that since players never deliberately change names, to $O(\epsilon + 1/M)$, $p^r = \alpha p^{re} + (1-\alpha)p^{rv}$ and $p = \alpha p^e + (1-\alpha)p^v$.

Lemma 2 *If s is an equilibrium then* $(1-\alpha)(W-V)/\alpha \geq (1-p^e) - O(\psi+\phi+1/T+\epsilon+1/M)$.

Proof: On all sample paths $\frac{(1-\alpha)}{\alpha}(W_i - V_i') \geq 1$ for players who choose C in their first period in the system and $\frac{(1-\alpha)}{\alpha}(W_i - V_i') \geq 0$ for those who choose D. Note that $(1-p^e)$ of the entrants are of the first type and p^e are of the second type. By lemma 1, only a fraction ϕ are trigger players, and their payoffs are bounded, and of the remaining players, their V_i' values are within ψ of $V_{b(i)+1}$. Finally, note that V is within a constant of the average of the $V_{b(i)+1}$ (a few V_0 values are replaced by V_T values). Taking the expectation and combining these proves the result. $\diamond$

Now we will show that this can not occur for any equilibrium with payoffs larger than PYD. First we compute V and W.

Lemma 3 $V = 1 - p + O(1/T + 1/M + \epsilon)$.

Proof: This can be computed directly, but it is most easily seen by noting that every defection removes one util from the total payoff to the players. $\diamond$

Lemma 4 $W = 1 + \alpha(p^{ve} - 2p^{ev}) - (1-\alpha)p^{vv} + O(1/T + 1/M + \epsilon)$.

Proof: This follows since a defection by a veteran against another veteran costs the set of veterans 1 util, a defection of a veteran against an entrant gains 1 util, and a defection of an entrant against a veteran loses 2 utils. $\diamond$

We now show that if V is large, there are not enough defections overall to keep $W - V$ sufficiently large.

Lemma 5 *If* $V \geq 1-\alpha/(2-\alpha)+\delta$ *then* $\frac{(1-\alpha)}{\alpha}(W-V) \leq (1-p^e-\frac{(1-\alpha)}{\alpha}\delta)+O(1/M+1/T+\epsilon)$, *for any* $\delta > 0$.

Proof: Let $Y = \frac{(1-\alpha)}{\alpha}(W-V) - (1-p^e-\frac{(1-\alpha)}{\alpha}\delta)$. Applying the formulas for W and V yield $Y = \frac{(1-\alpha)}{\alpha}[\alpha(p^{ve} - 2p^{ev}) - (1-\alpha)p^{vv} + \delta + p] - (1-p^e) + O(1/M + 1/T + \epsilon)$. Thus, we need to show that $Y_{max} = \{\max\ Y \mid V \geq 1 - \alpha/(2-\alpha) + \delta\} \leq 0$.

Since $V \geq 1-p$, $Y_{max} \leq \{\max\ Y \mid p \leq \alpha/(2-\alpha) - \delta\} \leq \{\max\ Y + [\alpha/(2-\alpha) - \delta - p]\frac{(1-\alpha)}{\alpha} \mid p \leq \alpha/(2-\alpha) - \delta\}$. Ignoring the term $O(1/M + 1/T + \epsilon)$, we get $Y_{max} \leq \{\max\ \frac{(1-\alpha)}{\alpha}[\alpha(p^{ve} - 2p^{ev}) - (1-\alpha)p^{vv} + \delta + \alpha/(2-\alpha) - \delta] - (1-p^e) \mid p \leq \alpha/(2-\alpha) - \delta\}$. It is easy to see that both p^{ev} and p^{vv} will be 0 at the maximum, so $Y_{max} \leq \{\max\ (1-\alpha)p^{ve} - 1/(2-\alpha) + \alpha p^{ee} \mid \alpha^2 p^{ee} + \alpha(1-\alpha)p^{ve} \leq \alpha/(2-\alpha) - \delta\}$. The constraint implies that $Y_{max} \leq -1/(2-\alpha) + 1/(2-\alpha) - \delta/\alpha = -\delta/\alpha$. This is strictly negative and thus remains negative when the order terms, $O(1/M + 1/T + \epsilon)$, are included. $\diamond$

Proof of Proposition: By the assumption on V and Lemma 5 we know that $\frac{(1-\alpha)}{\alpha}(W - V) \leq (1 - p^e - \frac{(1-\alpha)}{\alpha}\delta) + O(1/M + 1/T + \epsilon)$, but for any $\delta > 0$, this contradicts Lemma 2 when ϕ, ψ, and $1/T$ are sufficiently small. By Lemma 1, choosing β sufficiently large and letting T go to infinity makes those values arbitrarily small and thus yields a contradiction. $\square$

References

[1] D. Abreu, D. Pearce, and E. Stachetti. Toward a theory of discounted repeated games with imperfect monitoring. *Journal of Economic Theory*, 58:1041–1064, 1990.

[2] C. Avery, P. Resnick, and R. Zeckhauser. The market for evaluations. *American Economic Review*, 89(3), 1999.

[3] Y. Bakos and E. Brynjolfsson. Bundling information goods: Pricing, profits and efficiency. MIT Sloan School working paper, 1998.

[4] G. Ellison. Cooperation in the prisoner's dilemma with anonymous random matching. *Review of Economic Studies*, 61:567–588, 1994.

[5] E. J. Friedman. On social norms in noisy environments. mimeo, 1999.

[6] E. J. Friedman and S. Shenker. Learning and implementation in the Internet. mimeo, 1998.

[7] D. Fudenberg, D. Levine, and W. Pesendorfer. When are nonanonymous players negligible? *Journal of Economic Theory*, 79(1):46–71, 1998.

[8] D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, Cambridge, Massachusetts, 1991.

[9] David Goldschlag, Michael Reed, and Paul Syverson. Onion routing for anonymous and private internet connections. *Communications of the ACM*, 42(2):39–41, 1999.

[10] D. Grady. Faking pain and suffering on the Internet. *The New York Times*, 1998. April 23.

[11] E. Green and R. Porter. Noncooperative collusion under imperfect price information. *Econometrica*, 52:87–100, 1984.

[12] S. Herzog, S. Shenker, and D. Estrin. Sharing the cost of multicast trees: an axiomatic analysis. *Transactions on Networks*, 5(6):847–860, 1997.

[13] W. Hoeffding. Probability inequalities for sums of bounded random variables. In N. Fisher and P. Sen, editors, *The Collected Works of Wassily Hoeffding*. Springer-Verlag, 1994.

[14] M. Kandori. Social norms and community enforcement. *Review of Economic Studies*, 59:63–80, 1992.

[15] R. Kling, Y. Lee, A. Teich, and M. Frankel. Assessing anonymous communication on the internet: Policy deliberations. *The Information Society*, 15(2), 1999.

[16] P. Kollock. The production of trust in online markets. To appear in: Advances in Group Processes (Vol. 16), eds E.J. Lawler, M. Macy, S. Thyne, and H. A. Walker. Greenwich, CT: JAI Press. 1999, 1998.

[17] D. Kreps, P. Milgrom, J. Roberts, and R. Wilson. Rational cooperation in the finitely repeated prisoner's dilemma. *Econometrica*, 27:245–252, 1982.

[18] G. Marx. What's in a name? some reflections on the sociology of anonymity. *The Information Society*, 15(2), 1999.

[19] P. Milgrom, D. North, and B. Weingast. The role of institutions in the revival of trade: the law merchant, private judges, and the champaign fairs. *Economics and Politics*, 2:1–23, 1990.

[20] H. Moulin and S. Shenker. Serial cost sharing. *Econometrica*, 60:1009–1037, 1992.

[21] H. Nissenbaum. The meaning of anonymity in an information age. *The Information Society*, 15(2), 1999.

[22] M. Nowak and K. Sigmund. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, 364(64):56, 1993.

[23] P. Omidyar. Regarding the feedback forum, letter from founder pierre omidyar to the ebay community. *http://pages.ebay.com/aw/letter-060998-feedback.html*, 1998. June 9.

[24] Michael Reiter and Aviel Rubin. Anonymous web transactions with crowds. *Communications of the ACM*, 42(2):32–38, 1999.

[25] B. Schneier. *Applied Cryptography*. John Wiley & Sons, Inc., New York, 2nd edition, 1996.

[26] S. Tadelis. What's in a name? Reputation as a tradeable asset. *American Economic Review*, 89(3), 1999.

[27] A. Teich, M. Frankel, R. Kling, and Y. Lee. Anonymous communication policies for the internet: Results and recommendations of the AAAS. *The Information Society*, 15(2), 1999.

[28] H. Varian. Buying, sharing and renting information goods. mimeo, 1994.

[29] J. Watson. Starting small and renegotiation. *Journal of Economic Theory*, 85(1), 1999.