# The Evolution of Cooperation under Cheap Pseudonyms

Michal Feldman John Chuang

School of Information Management and Systems
University of California, Berkeley

## Abstract

*A wide variety of interactions on the Internet are characterized by the availability of* cheap pseudonyms*, where users can obtain new identities freely or at a low cost. Due to the availability of cheap pseudonyms, incentive schemes that are based on reward and punishment are vulnerable to the* whitewashing *attack, where users continuously discard their old identity and acquire a new one to escape the consequences of their bad behavior. In this paper, we study the implications of the whitewashing attack from an* evolutionary *perspective. Not surprisingly, the whitewashing attack degrades the evolutionary stability of strategies that are otherwise stable. In particular, the Tit-for-Tat strategy and its variant, probabilistic TFT, are not stable against whitewashers, unless identity costs are sufficiently large. In addition, we extend the* indirect reciprocity *model and find that discriminators can defeat whitewashers only if the probability to cooperate with strangers is small enough, which in turn degrades social welfare.*

## 1. Introduction

The performance of many distributed systems rely on voluntary resource sharing between individual peers. Some examples are contribution of files in file-sharing systems, packet forwarding in wireless ad hoc networks or Internet routing, and more. Alas, in many cases, the contributors may incur significant communication and computation costs without deriving any direct utility from contributing. Rational users, who attempt to maximize their own welfare, may thus attempt to "free-ride" on the other users – benefiting from the resources of others without offering their own resources in exchange. The inherent tension between individual rationality and collective welfare threatens to degrade the system's performance. Hardin has coined the phrase "the tragedy of the commons" [5] to refer to this phenomenon.

The problem of free-riding has been extensively studied via a game theoretic approach [13]. Perhaps one of the most celebrated demonstrations of the social dilemma is the *Prisoner's Dilemma* (PD) [7]. PD is a two-player game (See Figure 1) in which each player chooses whether to cooperate with, or defect on, the other. The dominant strategy of each player is to defect, resulting in an outcome of mutual defection, where both users yield lower payoffs than under mutual cooperation. Since defection is the unique Nash equilibrium (NE) in the one-shot PD game, defecting in all periods is also the unique subgame perfect NE in the finitely repeated game. It is only in the infinitely repeated game that cooperation can be sustained in equilibrium.

*Evolutionary game theory* studies equilibria of games played by a populations of players, where the "fitness" of the players is derived from their success in playing the game. It provides a tool for describing and analyzing situations where a number of agents interact and change their strategies at the end of any particular interaction. The Tit-for-tat ($tft$) strategy has been proven to be an *evolutionary stable strategy* (formal definition follows in Section 2) in the PD game both analytically and through simulations [1]. $tft$ bases its decision on the notion of *direct reciprocity*; it always cooperates on the first move, and reciprocates what the other player did on the previous move thereafter.

While $tft$ performs well in environments with many repeated transactions, it does not perform as well if there is only a small probability to interact repeatedly with the same opponent. In these cases, defectors can exploit the generosity of $tft$ towards strangers. Nowak and Sigmund [12] have introduced the *Image* strategy that is based on the notion of *indirect reciprocity*. The Image strategy uses the experience of other players to discriminate between cooperators and defectors, thereby can defeat defectors even in games with large populations and few repeated transactions.

However, one of the main challenges in attempting to transform strategies that are based on reciprocity (either direct or indirect) into protocols in online computational environments is the problem of cheap pseudonyms [4]. The availability of low-cost identities enables the *whitewashing* attack [2], in which defecting nodes continuously change

Proceedings of the Seventh IEEE International Conference on E-Commerce Technology (CEC'05)
1530-1354/05 $20.00 © 2005 **IEEE**

identities to escape the consequences of their behavior.

In this paper, we examine the effect of the whitewashing attack on the evolutionary stability of strategies that are based on direct and indirect reciprocity. Not surprisingly, we find that the stability of strategies degrades as a result of the whitewashing attack. In particular, $tft$ and probabilistic $tft$ can be defeated by whitewashers. Only by imposing a sufficiently large identity cost can $tft$ remain stable in the presence of whitewashers. Furthermore, the *image* strategy that is based on indirect reciprocity can defeat whitewashers only if the probability to cooperate with a stranger is smaller than $0.5$, but setting a low probability necessarily degrades social welfare in the residual whitewasher-free population.

The rest of this paper is organized as follows. Section 2 reviews the concepts of evolutionary stability and replication dynamics. In Section 3, we examine the evolutionary stability of $tft$ and probabilistic $tft$ against whitewashers in the prisoner's dilemma, and Section 4 studies the evolutionary stability of the image strategy against whitewashers. We present related work in Section 5, and Section 6 concludes the paper.

## 2. Evolutionary Game Theory

In this section, we provide a brief background to evolutionary game theory and review the notions of *evolutionary stable strategy* (ESS) and *replication dynamics*. For more details on evolutionary game theory, see [1, 6, 14]. In later sections, we use the tools introduced in this section to study the effect of cheap pseudonyms on the stability of cooperative strategies.

### 2.1. Evolutionary Stable Strategies (ESS)

A *strategy* is a mapping from the game's history into an action in the current move. A strategy can be either *pure* or *mixed*, where a mixed strategy consists of possible actions and a probability distribution that corresponds to the weight of each action.

In evolutionary games, the population consists of players playing various strategies. The score of a strategy in any round determines the relative number of "offsprings" in the next round. Thus, over time, the lower scored strategies decrese in number, and the higher scoring strategies increase. The *fitness* of a strategy is the strategy's expected score. Let $V(A|B)$ denote the fitness of strategy $A$ when interacting with strategy $B$.

**Invasion** Strategy $A$ is said to *invade* a population of strategy $B$ players if $V(A|B) > V(B|B)$. If no strategy can invade a population of strategy $B$ players, $B$ is said to be *collectively stable*, or an *evolutionary stable strategy* (ESS). Formally, in order for strategy $B$ to be stable, it must hold that for all $A$,

$$V(B|B) > V(A|B)$$

or

$$V(B|B) = V(A|B) \text{ and } V(B|A) > V(A|A)$$

**Invasion in clusters** The above notion of invasion refers to invasion by individuals. A different kind of invasion is invasion by *clusters*. Invasion by clusters refers to scenarios in which the invaders can control to some extent whom they interact with. Strategy $A$ is said to invade strategy $B$ in clusters if the $A$'s (invaders) can provide a significant part of each other's environment, but a negligible part of the $B$'s (natives) environment. An $x$-cluster of $A$ is said to invade $B$ if

$$xV(A|A) + (1-x)V(A|B) > V(B|B)$$

This definition assumes that pairing in the interactions is not random. $x$ is the proportion of $A$'s interactions with another $A$, whereas interactions of $B$'s with $A$'s are negligible. Some of the collectively stable strategies, while cannot be invaded by individuals, are invadeable by clusters.

**Random mixing** With *random mixing*, the proportion needed for newcomers to invade natives is $q$, such that:

$$qV(A|A)+(1-q)V(A|B) > qV(B|A)+(1-q)V(B|B)$$

We also denote the *discount factor* by $w$. Scores are discounted by $w$ as time passes. An equivalent interpretation of $w$ is the probability to have an additional round in the repeated game with uncertain number of rounds.

### 2.2. Replication dynamics

According to *replication dynamics*, the initial population is represented by a set of pairs $(s_1, p_1), \ldots, (s_n, p_n)$, where $s_i$ and $p_i$ denote the strategies and their respective proportion in the population. The score of a strategy in each round determines the relative number of offsprings in the next round. Under the assumption that the number of entities in the population is fixed over time, the proportion $p_i$ of each strategy $s_i$ in the successor round is given by

$$p_i^{t+1} = p_i^t \frac{V_i^t}{V^t}$$

where $V_i^t$ is the score of strategy $s_i$ in round $t$ and $V^t$ is the average score in the population. Thus, strategies that score above the average increase over time, and those that score below the average decrease over time.

## 3. Direct Reciprocity

Direct reciprocity suggests that agent $i$ should base his choice of action toward agent $j$ on $j$'s previous behavior to $i$. Tit-for-tat ($tft$) is based on the idea of direct reciprocity, where agent $i$ reciprocates to $j$ what $j$ has done to $i$ in the last round. In the remainder of this section, we focus on the PD game, whose payoff matrix is presented in Figure 1. As illustrated in the matrix, mutual cooperation results in payoff $R$ (Reward) to both players, mutual defection results in $P$ (Punishment) to both players, and if only one of the agents cooperates, he gets $S$ (Sucker) while his opponent gets $T$ (Temptation). The payoffs satisfy the relation

$$S < P < R < T \tag{1}$$

A strategy in the one-shot PD game is a decision whether to cooperate or defect (and can be also probabilistic), and a strategy in the repeated PD game is a mapping from every possible history of the game into a probability of cooperation. Defection is the unique dominant strategy in the one-shot PD. That is, no matter what the other player does, defection always yields a higher payoff than cooperation. As such, cooperation cannot be sustained in the finitely repeated PD, no matter how many rounds it is repeated and how patient agents are. However, based on the Folk theorem [9], cooperation can be sustained in equilibrium in the infinitely repeated PD (or equivalently, if there is uncertainty about the number of rounds).

It has been demonstrated by Axelrod in [1] that $tft$ is evolutionary stable in the PD game if identities are permanent. In Section 3.1 we review Axelrod's results about the stability of $tft$ in PD, and in Section 3.2 we show the effect of the whitewashing attack on the stability of $tft$ and its variant, probabilistic $tft$.

| | | Player 2 | |
|---|---|---|---|
| | | Cooperate | Defect |
| Player 1 | Cooperate | $R = 3$ / $R = 3$ | $S = 0$ / $T = 5$ |
| | Defect | $T = 5$ / $S = 0$ | $P = 1$ / $P = 1$ |

**Figure 1. Payoff matrix for the Prisoner's Dilemma. The condition $S < P < R < T$ must be met to satisfy the social dilemma.**

### 3.1. Tit-for-Tat in Environments with Permanent Identities

Axelrod [1] has taken an evolutionary approach to the study of the PD game. He has found that if the discount factor, $w$, is sufficiently large, there is no strategy that is best independent of the strategy used by the other players. However, if everybody in the population plays $tft$, then no one can do better by switching to any other strategy. The *fitness* (score) of strategy $A$ when playing against strategy $B$ in the infinitely repeated game is

$$V(A|B) = \sum_{t=0}^{t=\infty} V(A|B)_t w^t$$

where $V(A|B)_t$ is the payoff of $A$ when playing against $B$ at time $t$ , and $w$ is the discount factor.

**Proposition 3.1** *(Axelrod [1]) $tft$ is collectively stable in the PD game if and only if*

$$w \geq max(\frac{T-R}{T-P}, \frac{T-R}{R-S})$$

The proof of this proposition is presented in [1].

As mentioned in section 2.1, some of the collectively stable strategies, while cannot be invaded by individuals, are invadeable by clusters. A *nice* strategy is a strategy that is never the first to defect. It is shown in [1] that if a *nice* strategy cannot be invaded by a single individual, it cannot be invaded by any cluster of individuals either. $tft$ is obviously a nice strategy. Therefore, unlike $allD$, $tft$ is uninvadeable even by clusters.

### 3.2. Tit-for-Tat in the Presence of Whitewashers

The above findings demonstrate the stability of $tft$ against invasion of both individuals and clusters. However, its stability relies on two conditions:

- permanent identities.
- traceable actions (both cooperation and defection).

If both conditions are satisfied, a player who defects in one round can be punished in later rounds. However, both conditions are challenged in online environments. Identities may not be permanent due to cheap pseudonyms [4, 3], and in many applications defection cannot be traced. In this paper we focus on the first problem (lack of permanent identities) and leave the analysis of untraceable defections to future work. We first define the *whitewashing* strategy and then study its effect in situations where identities are completely free (Section 3.2.1) and in situations where identities can be replaced, but at some positive cost (Section 3.2.2).

**Whitewashing** The whitewashing strategy is denoted by $ww$. A $ww$ player always defects, but changes identity after each interaction.

### 3.2.1. Free Identities

It is easy to see that under free identities $tft$ is not stable against $ww$, since $ww$ can always exploit the generosity of $tft$ to strangers. The fitness of $ww$ against $tft$ is:

$$V(ww|tft) = \frac{T}{1-w}$$

which is the maximal possible score in the game, and is clearly greater than $V(tft|tft) = R/(1-w)$. Therefore, $tft$ is invadeable by $ww$, even individually, and is not an ESS in the presence of whitewashing.

**Probabilistic** $tft$ In an attempt to deal with whitewashers, we consider a variant of $tft$, namely *probabilistic tft*, or $tft_p$. Like $tft$, $tft_p$ reciprocates to a player what he did on the previous move. However, unlike $tft$ that always cooperates with a stranger, $tft_p$ cooperates with a stranger randomly with probability $p$.

**Proposition 3.2** *$ww$ invades $tft_p$ for all $w$ and all $p \neq 0$*[1].

**Proof** In order to prove this claim, we have to show that

$$V(ww|tft_p) > V(tft_p|tft_p) \quad \forall p, w$$

The fitness of $ww$ and $tft_p$ against $tft_p$ is expressed as follows:

$$\begin{aligned} V(ww|tft_p) &= \frac{pT + (1-p)P}{1-w} \\ V(tft_p|tft_p) &= p^2\frac{R}{1-w} + p(1-p)\frac{S+T}{1-w} + (1-p)^2\frac{P}{1-w} \end{aligned}$$

Therefore:

$$\begin{aligned} & V(ww|tft_p) > V(tft_p|tft_p) \\ \Leftrightarrow\ & p(p(R-T) + (1-p)(S-P)) < 0 \end{aligned}$$

but $R - T < 0$ and $S - P < 0$ for all $S, P, T, R$ (by inequation 1. Therefore, $ww$ invades $tft_p$ for all $w$ and $p \neq 0$.

**Proposition 3.3** *$tft_p$ cannot invade $ww$ with random mixing.*

**Proof** In order for $tft_p$ to invade $ww$ with random mixing, there need to exist $p$ and $q$ satisfying:

$$\begin{aligned} & qV(tft_p|tft_p) + (1-q)V(tft_p|ww) > \\ & \quad qV(ww|tft_p) + (1-q)V(ww|ww) \end{aligned}$$

but, as shown above:

$$V(tft_p|tft_p) < V(ww|tft_p)$$

and

$$V(tft_p|ww) < V(ww|ww)$$

Therefore, $tft_p$ cannot invade $ww$ with random mixing.

**Proposition 3.4** *An $x$-cluster of $tft_p$ invades $ww$ if*

$$x > \frac{P-S}{T-P+p(R-S-T+P)}$$

**Proof** An $x$-cluster of $tft_p$ invades $ww$ if

$$\begin{aligned} & xV(tft_p|tft_p) + (1-x)V(tft_p|ww) > V(ww|ww) \\ \Leftrightarrow\ & x(p^2\frac{R}{1-w} + p(1-p)\frac{S+T}{1-w} + (1-p)^2\frac{P}{1-w}) \\ & + (1-x)(\frac{pS+(1-p)P}{1-w}) > \frac{P}{1-w} \\ \Leftrightarrow\ & x > \frac{P-S}{T-P+p(R-S-T+P)} \end{aligned}$$

Therefore, in order for an $x$-cluster of $tft_p$ to invade $ww$, the following condition must hold:

$$P - S < T - P + p(R - S - T + P)$$

When the above condition holds, the effect of $p$ on the necessary cluster size depends on the proportions between the intermediate ($R$, $P$) and extreme ($T$, $S$) payoffs.

- If $R+P > S+T$, then as $p$ increases, the denominator increases, and the necessary cluster size decreases.
- If $R+P < S+T$, then as $p$ increases, the denominator decreases, and the necessary cluster size increases.

In summary, we find that $ww$ invades $tft_p$ for all $w$ and $p$, and $tft_p$ cannot invade $ww$ with random mixing, but can invade $ww$ with a large enough cluster.

### 3.2.2. Positive Identity Costs

So far, we have assumed that whitewashers can *freely* acquire a new identity. However, in most cases, identities are cheap, but not completely free. In addition, it would be interesting to study the effect of a positive identity cost since one may consider applying an artificial identity cost into the system in order to discourage whitewashing behavior. In what follows, we analyze the effect of positive identity costs on the dynamics of the game. Let $C$ denote the identity cost. Then, the payoffs of the different interactions become:

- $V(ww|tft) = \frac{T-C}{1-w}$
- $V(tft|tft) = \frac{R}{1-w} - C$
- $V(tft|ww) = \frac{S}{1-w} - C$
- $V(ww|ww) = \frac{P-C}{1-w}$

[1] if $p = 0$, $V(i|i) = P/(1-w)$ for $i \in tft, ww$.

If $C$ is large enough, $tft$ is stable against invasion of $ww$. In particular, if $C > \frac{T-R}{w}$, $ww$ cannot invade $tft$[2]. Thus, imposing a large identity cost helps in preventing invasion by whitewashers.

**Sophisticated whitewasher** Once identities are costly, a more sophisticated $ww$ may change identity only every $I$ iterations instead of every single iteration. A $ww_I$ player always defects and whitewashes every $I$ iterations. Interactions between $ww_I$ and $tft$ players yield the following scores:

- $V(ww_I|tft) = T + \frac{Pw}{1-w} - \frac{Pw^{I-1}}{1-w^I} + \frac{(T-C)w^{I-1}}{1-w^I}$
- $V(tft|tft) = \frac{R}{1-w}$

Therefore, $tft$ is stable against invasion by $ww$ if

$$C > T - P + \frac{1-w^I}{w^{I-1}}(\frac{Pw}{w^{I-1}} - \frac{R}{1-w} + T)$$

A $ww_I$ player who plays against a $tft$ player attempts to maximize

$$V(ww_I|tft) = T + \frac{Pw}{1-w} + \frac{w^{I-1}}{1-w^I}(T - P - C)$$

What is the optimal number of iterations after which to change identity? It is only the last element in the expression that depends on $I$. Since $w < 1$, the fraction $\frac{w^{I-1}}{1-w^I}$ decreases in $I$. Therefore:

- if $C < T - P$, $V(ww_I|tft)$ decreases in $I$ and it is optimal for $ww_I$ to whitewash every iteration. In this case $ww_I$ is equivalent to $ww$ and we have shown that $tft$ is stable against invasion by $ww$ if $w > \frac{T-R}{C}$. Therefore, $tft$ is stable against $ww_I$ if

$$w > max\left(\frac{T-R}{C}, \frac{T-R}{T-P}\right)$$

- if $C > T - P$, $V(ww_I|tft)$ increases in $I$ and it is optimal for $ww_I$ to never whitewash. In this case, $ww_I$ is equivalent to $allD$, and we have shown that $tft$ is stable against invasion of $allD$ if

$$w > \frac{T-R}{T-P}$$

- if $C = T - P$, $V(ww_I|tft)$ is independent of $I$, thus the choice of $I$ has no effect on its payoff. In this case, $ww_I$ invades $tft$ if

$$w > \frac{T-R}{T-P}$$

In summary, a sufficiently large identity cost may discourage whitewashing behavior. If the identity cost is too small, whitewashers defeat $tft$, and for intermediate costs, whitewashing may still be worthwhile, but $tft$ will defeat the whitewashers.

## 4. Indirect Reciprocity

After illustrating that the stability of $tft$ is limited in the presence of whitewashers, we present the effect of whitewashers on strategies that are based on *indirect reciprocity*. Indirect reciprocity means that agent $x$ cares not only about the last action of $y$ toward $x$, but also the last action of $y$ toward a third agent, $z$.

Nowak and Sigmund [12] have proposed to alleviate the problem of cooperation in environments with large populations by maintaining shared history that aggregates information from all players. They present a model that uses the *replication dynamics* evolutionary rule (see section 2) to study the dynamics of systems with discriminators and defectors. We extend their model to study the effect of whitewashers on the system.

### 4.1. Model

In the game, users are paired up at random in every round such that one is the *donor* and the other is the *recipient*. The donor has to decide whether to cooperate or to defect. If she cooperates, the donor gets payoff of $-c$, the recipient gets payoff of $b$ (such that $b > c$; otherwise, the socially desired outcome is one in which all players defect), and the image score of the donor is 1. If she defects, they both get payoff of 0, and the image score of the donor is 0.

The population consists of *discriminators* and *whitewashers*.

- Discriminators always cooperate with players with image score 1, and with strangers with probability $p$. The image score of discriminators is known with probability $q$.
- Whitewashers never cooperate and continuously change identity, such that the probability to know their image score is zero.

We use the following notation:

- $x$: the fraction of discriminators in the population.

[2] In the above analysis, we assumed that the initial identity is also costly. However, the same condition holds if the initial identity is considered sunk cost. In this case,

- $V(ww|tft) = T + \frac{(T-C)w}{1-w}$
- $V(tft|tft) = \frac{R}{1-w}$
- $V(tft|ww) = \frac{S}{1-w}$
- $V(ww|ww) = P + \frac{(P-C)w}{1-w}$

and the condition for the stability of $tft$ against $ww$ remains $C > \frac{T-R}{w}$.

- $y$: the fraction of whitewashers in the population.
- $x_0, y_0$: the fraction of discriminators and whitewashers with image 0, respectively.
- $x_1, y_1$: the fraction of discriminators and whitewashers with image 1, respectively.

We assume that $x$ is initially equally divided between $x_0$ and $x_1$, and the same for $y$.

### 4.2. Evolutionary Stability of Discriminators

In each round, the payoff to the individual types is:

$$P(x_0) = \frac{1}{2}(-c)(qx_1 + p((1-q)x + y)) + \frac{1}{2}bx(1-q)p$$

$$P(x_1) = \frac{1}{2}(-c)(qx_1 + p((1-q)x + y)) + \frac{1}{2}bx(q + (1-q)p)$$

$$P(y_0) = \frac{1}{2}bxp$$

$$P(y_1) = \frac{1}{2}bxp$$

The frequencies of players of image 0 and 1 change from round to round according to the following difference equations:

$$\begin{aligned} x_0(k+1) &= \frac{x_0(k)}{2} + \frac{x_0(k)^2 q}{2} \\ &+ \frac{x(k)}{2}((1-q)x(k) + y(k))(1-p) \\ &+ \frac{x_1(k)x_0(k)q}{2} \\ x_1(k+1) &= x(k) - x_0(k+1) \\ y_1(k+1) &= \frac{y_1(k)}{2} \\ y_0(k+1) &= y(k) - y_1(k+1) \end{aligned}$$

Solving the difference equations yields the fraction of the different types of users as a function of the round number, $k$, in closed forms:

$$\begin{aligned} x_0(k) &= x\left[(\frac{xq+1}{2})^k(p - \frac{1}{2}) + 1 - p\right] \\ x_1(k) &= x\left[p - (\frac{xq+1}{2})^k(p - \frac{1}{2})\right] \\ y_1(k) &= y(\frac{1}{2})^{k+1} \\ y_0(k) &= y(1 - (\frac{1}{2})^{k+1}) \end{aligned}$$

Using these expressions, we can express the expected payoffs of discriminators and whitewashers in the $k^{th}$ round:

$$\begin{aligned} P_{disc}(k) &= \frac{x_0(k)}{x}p(x_0) + \frac{x_1(k)}{x}p(x_1) \\ &= (\frac{xq+1}{2})^k \frac{1}{2}xq(\frac{1}{2} - p)(b - c) - \frac{1}{2}p(c - bx) \\ P_{ww}(k) &= \frac{y_0(k)}{y}p(y_0) + \frac{y_1(k)}{y}p(y_1) \\ &= \frac{1}{2}bxp \end{aligned}$$

Assuming that there exists a fixed probability $w$ for a further round[3], the total payoffs to whitewashers and discriminators are:

$$\begin{aligned} P(disc) &= \sum_{k=1}^{\infty} w^{k-1} P_{disc}(k) \\ &= \frac{qx(c-b)(p-\frac{1}{2})(xq+1)}{2(2 - w(xq+1))} + \frac{bxp}{2(1-w)} - \frac{cp}{2(1-w)} \\ P(ww) &= \sum_{k=1}^{\infty} w^{k-1} P_{ww}(k) \\ &= \frac{pbx}{2(1-w)} \end{aligned}$$

Modeling the change in frequency of discriminators and whitewashers from one generation to the next by replication dynamics [6], discriminators win if and only if $P(disc) > P(ww)$. That is, discriminators are evolutionary stable against whitewashers if and only if

$$P(disc) - P(ww) = \frac{qx(c-b)(p-\frac{1}{2})(xq+1)}{2(2 - w(xq+1))} - \frac{cp}{2(1-w)} > 0$$

We make the following observations[4]:

- There exist parameter values for which discriminators win. This is an important observation since one could have thought that discriminators are hopeless when confronting whitewashers.
- If $p \geq 1/2$, then $P(disc) - P(ww) < 0$ and whitewashers win. That is, a necessary (but not sufficient) condition for discriminators to win is that $p < 1/2$. Note that in the absence of whitewashers (as in [12]), if indirect reciprocity works at all, then discriminators with large $p$ outcompete the others.

[3] If $w$ is the probability for another round, there are on average $1/(1-w)$ rounds per generation. An equivalent interpertation of $w$ is as a discount factor.

[4] While these results hold for an equal initial distribution of players with image scores 0 and 1, the quantitative results are sensitive to the initial distriubtion. For example, if all players are initially with image score 0, whitewashers always win, and if all players are initially with image score 1, discriminators can win for a higher range of values. For example, discriminators can win even if $p > \frac{1}{2}$.

- Moreover, if $p = 0$, $P(disc) - P(ww) > 0$ always. The interpretation of setting $p$ to 0 is always defecting on strangers. Since whitewashers are always strangers, they do not gain any benefit and necessarily lose. Indeed, defecting on strangers is an effective method to handle whitewashers in evolutionary terms. However, as demonstrated below, it incurs some social loss (see [2, 3] for additional analysis of behavior to strangers in the context of whitewashers).
- Smaller values of $w$ increase the chances of discriminators to win. Thus, as the average number of rounds increase, it is more likely that whitewashers will win. This is, again, in contrast to the results in [12]. The reason for the difference is that the number of defectors with image score 1 decreases over time while that with image score 0 increases over time, so defectors get less and less cooperation over time, which increases the chances of discriminators to win. In contrast, the proportion of whitewashers with image scores 0 and 1 remains static over time since they are indistinguishable from one another.

### 4.3. The social cost of defecting on strangers

While it is important to keep the value of $p$ low in order for discriminators to be evolutionary stable against whitewashers, it has a social cost. To see this, consider the case in which discriminators win and the resulting population is "whitewashers-free". The dynamics of discriminators with image scores 0 and 1 over time is given by the difference equation:

$$\begin{aligned} x_0(k+1) &= \frac{x_0(k)}{2} + \frac{x_0(k)^2 q}{2} \\ &+ \frac{x(k)}{2}((1-q)x(k))(1-p) + \frac{x_1(k)x_0(k)q}{2} \end{aligned}$$

Solving the difference equation, we get that the system reaches a stationary state when:

$$x_0 = x(1-p) = 1-p$$

$$x_1 = xp = p$$

Substituting these values into the payoffs, $p(x_0)$ and $p(x_1)$, we get:

$$p(x_0) = -\frac{1}{2}cp + \frac{1}{2}b(p - pq)$$

$$p(x_1) = -\frac{1}{2}cp + \frac{1}{2}b(q + p - pq)$$

and the average payoff in the population is:

$$\begin{aligned} E[P] &= x_0 P(x_0) + x_1 P(x_1) \\ &= (1-p)(-\frac{1}{2}cp + \frac{1}{2}b(p - pq)) \\ &+ p(-\frac{1}{2}cp + \frac{1}{2}b(q + p - pq)) \\ &= \frac{1}{2}p(b-c) \end{aligned}$$

We find that the average payoff in the population, which is actually the social welfare, increases linearly in $p$. Thus, the tradeoff is apparent. On the one hand, if $p$ is set too high, whitewashers win and discriminators vanish. On the other hand, setting $p$ too low results in a low social welfare once whitewashers are gone.

## 5. Related Work

Axelrod [1] has presented the *iterative prisoner's dilemma* model, and concluded that $tft$ was the most successful strategy, growing at a faster rate than other strategies. A later simulation study, done by Nowak and Sigmund [10], has used a different model that allowed for new "mutant" to enter the game at any stage. Their arrived at different results, where $tft$-like strategies were necessary in order to eliminate the exploiters, but the strategy generous $tft$ ($Gtft$) gained the highest profit. $Gtft$ is more forgiving of defections that the original $tft$ and cooperates with defectors with probability $min\,(1-(T-R)/(R-S), (R-P)/(T-P))$. A later series of simulations run by Nowak and Sigmund [11] has led them to conclude that the best strategy in evolutionary terms is a strategy that conditions the action on the previous realized payoff; it cooperates after receiving $R$ or $T$ and defects after receiving $P$ or $S$.

To study the stability of strategies in large populations that exhibit few repeated transaction, Nowak and Sigmund [12] have introduced the notion of *indirect reciprocity*, and proposed the *image* strategy that bases its decision on the experiences of others. They have presented the interaction between agents as a game, in which users cooperate with or defect on each other based on a globally observed *image* score, characterizing their past contribution. They conclude that indirect reciprocity helps in sustaining cooperation in environments with large populations.

The problem of cheap pseudonyms was first studied by Friedman and Resnick [4], where they study the effect of cheap pseudonyms on the emerging cooperation and social welfare. They have found that a large degree of cooperation can be achieved by a convention in which newcomers accept poor treatment from high-reputable players. This identity model, where users can easily discard their identity and acquire a new one is studied in other works in the context of attacks on reputation systems [2, 8].

## 6. Conclusions

In this paper, we have analyzed the evolutionary stability of strategies in the presence of whitewashers, and presented results for strategies that are based on direct and indirect reciprocity. We find that while $tft$ is evolutionary stable in

the prisoner's dilemma under permanent identities, it is no longer stable in the presence of whitewashers. We further find that whitewashers always invade probabilistic $tft$, but probabilistic $tft$ can only invade whitewashers in a sufficiently large cluster. Yet, imposing a sufficiently large identity cost can turn $tft$ into an evolutionary stable strategy and may even discourage whitewashing behavior altogether. In a game of whitewashers and discriminators, who base their decisions on indirect reciprocity, a necessary (but not sufficient) condition for discriminators to win is to cooperate with strangers with a probability of less than $0.5$. However, setting this probability too low results in a social loss.

In future work, we intend to extend the models presented in this paper to broaden our understanding of the effect of cheap pseudonyms on the interactions between various strategies in a population. In particular, we are interested in learning the evolutionary stability of other strategies that have been proposed in the literature and gaining better understanding of the dynamics in populations that consist of more than two strategies. In addition, we wish to study the effect of untraceable defections on the stability of strategies that are based on direct and indirect reciprocity.

## 7. Acknowledgments

This work is supported in part by the National Science Foundation under ITR awards ANI-0085879 and ANI-0331659.

## References

[1] R. Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.

[2] M. Feldman, K. Lai, I. Stoica, and J. Chuang. Robust Incentive Techniques for Peer-to-Peer Networks. In *ACM Conference on Electronic Commerce (EC'04)*, May 2004.

[3] M. Feldman, C. Papadimitriou, I. Stoica, and J. Chuang. Free-Riding and Whitewashing in Peer-to-Peer Systems. In *Proc. SIGCOMM workshop on Practice and Theory of Incentives and Game Theory in Networked Systems*, 2004.

[4] E. Friedman and P. Resnick. The Social Cost of Cheap Pseudonyms. *Journal of Economics and Management Strategy*, 10(2):173–199, 1998.

[5] G. Hardin. The Tragedy of the Commons. *Science*, 162:1243–1248, 1968.

[6] Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.

[7] S. Kuhn. Prisoner's Dilemma. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer edition, 2003.

[8] S. Marti and H. Garcia-Molina. Limited Reputation Sharing in P2P Systems. In *ACM Conference on Electronic Commerce (EC'04)*, May 2004.

[9] A. Mass-Colell, M. Whinston, and J. Green. *Microeconomic Theory*. Oxford University Press, 1995.

[10] M. A. Nowak and K. Sigmund. Tit For Tat in Heterogeneous Populations. *Nature*, 355:250–253, 1992.

[11] M. A. Nowak and K. Sigmund. A Strategy of Win-stay, Lose-shift that Outperforms Tit-for-tat in the Prisoner's Dilemma Game. *Nature*, 364:56–58, 1993.

[12] M. A. Nowak and K. Sigmund. Evolution of Indirect Reciprocity by Image Scoring. *Nature*, 393:573–577, 1998.

[13] T. Palfrey and Rosentral. Private Incentives in Social Dilemmas: The Effects of Incomplete Information and Altruism. In *Journal of Public Economics*, volume 35, 309-332, 1988.

[14] J. M. Smith. *Evolution and the Theory of Games*. Cambridge University Press, 1982.