On the Viability of Vengeance

by Daniel Friedman

and

Nirvikar Singh

Economics Department

UC Santa Cruz

May 21, 1999

Abstract

Using a simple symmetric game to illustrate, we point out shortcomings in previous theoretical accounts of how the human vengeance motive survives despite a free rider problem. We offer a new theoretical explanation involving the coevolution of genes that determine the capacity for vengeance and memes that regulate its expression. The main result, illustrated in a simple parametric example, is that coevolution in a fixed environment circumvents the free rider problem and that the prevailing vengeance level will efficiently serve groups of individuals.

Acknowledgements

The first author is grateful to CES Munich University for hospitality while writing the first fragments in May 1997. We have benefited from the comments of audiences at Indiana, Purdue, UCLA, and UCSC.

Comments of all sorts are very welcome; please email dan@cats.ucsc.edu.

Introduction

Vengeance is a powerful human motive. We become angry when someone wrongs us or our compatriots, and we often try to harm the culprit in return even at some personal cost. Seeking vengeance, Ahab pursued the great white whale halfway around the world and lost his fictional life. Many non-fictional people in the mid-East, the Balkans and elsewhere ruin their lives and their countries for the sake of vengeance. We are all personally familiar vengeance on a smaller scale, as we (or at least our colleagues!) indulge in office intrigue and pay at least a moderate cost to settle scores.

Vengeance can bring social benefits. As negative reciprocity, it complements and extends positive reciprocity, the desire to help others who have helped us. The folk theorem of game theory explains positive reciprocity as an individually rational (indeed, a subgame perfect) way to support efficient exchange of favors, as long as the discount factor exceeds the ratio of personal cost to social benefit. Vengeance can increase social value in two ways. It can deter opportunistic behavior that would otherwise undermine positive reciprocity. And it can support efficient exchange even when the discount factor is low, as for example when repeat interaction is sporadic.

Of course, vengeance is sometimes extremely dysfunctional, as in the Balkans. For this reason it is discouraged by moral and religious teachings. "An eye for an eye, a tooth for a tooth" is an Old Testament plea for moderation of vengeance, and "turn the other cheek" is a New Testament plea to abandon vengeance altogether.

The existence of vengeance is empirically obvious, but theoretically mysterious. As far as existence is concerned, it is beside the point whether vengeance is helpful or harmful to society. The crucial theoretical issue is whether vengeful traits convey a selective advantage. Apparently the answer is no, because vengeance is not individually rational. We will show that vengeance is weakly dominated by otherwise similar behavior that shirks on the personal cost. Therefore it is a theoretical puzzle how vengeance ever established itself in the repertoire of human motives, and how it sustains itself. Until the puzzle is solved, theory will offer no guidance on how to regulate vengeance to maximize its social value and to minimize its devastation.

In this paper we offer an evolutionary account of vengeance. The account draws on the perspectives of both selfish genes and cultural memes in the context of a standard normal form game that captures, in the simplest and most direct manner, a personal cost incurred to reap social gains. The game illustrates how a taste for vengeance realigns incentives and supports a socially efficient equilibrium, but demonstrates that vengeance is not evolutionarily viable. After reviewing several unconvincing earlier treatments of the viability problem, we propose an evolutionary model with individual learning and evolution as well as meme selection for groups of individuals. We argue that groups of individuals can use sanctions (or simply status changes) to enforce a particular norm on the proper degree of vengeance. Actual behavior typically will fall short of the norm, but selection across groups will adjust the norm so that actual behavior maximizes the fitness of group members, and the free rider problem is overcome.

Introductory Remarks for Economists. In most sections we include at the end some more technical remarks for interested readers. This material can be skipped without loss of continuity. Here we simply position ourselves in recent strands of economics literature.

Economists traditionally assume that people are self-interested. This assumption is safe in perfect competition, since perfect competition offers no opportunity to help or harm others. Recently it has become clear that the assumption is not safe, and indeed is systematically wrong, in some imperfectly competitive environments. Important applications, such as pricing dynamics in oligopolistic industries or outcomes of labor or trade bargaining, are not well explained by folk-theorem style arguments. Cooperation doesn't necessarily break down when tit-for-tat (or grim trigger or intermediate) punishments become ineffective, and Pareto optimal agreements are often missed when theoretically attainable (e.g., Blinder and Choi, 1990; Steers and Porter, 1991). Laboratory studies have all but eliminated alternative explanations to the common-sense observation that people like to help people who have helped them and like to punish people who have harmed them or their friends, and are willing to incur moderate personal costs to do so (e.g., Fehr and Gaechter, 1998; Camerer and Thaler, 1995).

Several authors recently have constructed models to deal with the issue. Rabin (1993) adapts the formalism of psychological games (defined on beliefs as well as actions) to analyze theoretically the impact of positive and negative reciprocation motives in two-person normal form games of complete information. Andreoni and Miller (1996), Fehr and Schmidt (1998) and Bolton and Ockenfels (1998) propose simpler models with preferences defined on others' material payoffs as well as one's own. Levine (1998) considers an incomplete information model where preferences over outcomes depend on one's perception of others' degree of altruism or spite. These models are constructed to fit laboratory data while not departing too far from traditional self-interest models, and they are reasonably successful in doing so.

Our purposes are a bit different. Positive or negative regard for others, we assume, is not innate spite but rather is contingent on others' behavior. In our model, the desire to harm someone comes entirely from negative experience with that person. Also, we are not yet concerned with fitting laboratory or field data. Our focus is entirely theoretical – how might preferences evolve that are not entirely self-interested? Using basic techniques of game theory and evolutionary theory, we show how such preferences can be shaped by genetic and cultural selection processes.

1. The Underlying Game

We begin by demonstrating how a taste for vengeance can convert a standard prisoner's dilemma problem to a simple coordination problem with a Pareto efficient equilibrium. The idea is simply that given a vengeance motive, cooperative behavior is no longer is dominated and can become part of a Nash equilibrium even when there is no repeat interaction. Subsequent analysis will build on this game, the simplest game we know in which individual self interest diverges from the social interest.

The basic underlying game is a simple symmetric 2-player prisoner's dilemma with a cooperator payoff of 1, a temptation payoff of 2, a sucker payoff of -1, and an all-defect payoff of 0. In other words, the benefits of full cooperation of 2 are evenly split and the benefit of one-sided cooperation of 1 is very unevenly split at (2,-1), relative to the no cooperation payoff which is normalized to (0,0). The specific numbers are for the sake of illustration; the same points could be made with any other parameterization of the prisoner's dilemma.

Table 1: Fitness with No Vengeance

(*v* =0)	C	D
C	1 , 1	-1, 2
D	2 , -1	0, 0

Payoffs so far are material and can be thought of as describing both fitness and utility. The personal cost (or personal fitness reduction) to cooperating is 1 and the social gain (or increase in the fitness sum) is also 1. The game has a unique Nash equilibrium in which each player chooses the dominant strategy D and achieves fitness 0.

To this underlying game we add a punishment technology and a punishment motive parameterized by its incurred cost *v*. We hypothesize that a player can inflict harm (fitness loss) *h* on the other player at personal fitness cost *ch*. The marginal cost *c* is a constant parameter between 0 and 1 that captures the technological opportunities for punishing others. We further hypothesize that inflicting harm *h* yields the player a utility bonus of *v* ln *h* (but no fitness bonus) when he is the victim of the sucker payoff and no bonus in other circumstances. Thus the motive isn't spite but rather is vengeance for damage personally experienced. The motivational parameter *v* is subject to evolutionary forces and is intended to capture an individual's temperament, e.g., his susceptibility to anger. See R. Frank (1988) for an extended discussion of such traits.

The objective function when victim of a sucker payoff now is $v \ln h - ch - 2$. The utility-maximizing degree of vengeance h^* to inflict on a culprit (the beneficiary of the temptation payoff) is the unique solution of the first order condition, so $h^*=v/c$ is the inflicted damage and $ch^*=v$ indeed is the incurred cost. The game now has the same fitness payoffs as before on the main diagonal, but the sucker payoff is reduced by the cost of vengeance and the temptation payoff is reduced by the amount of harm inflicted, as in Table 2.

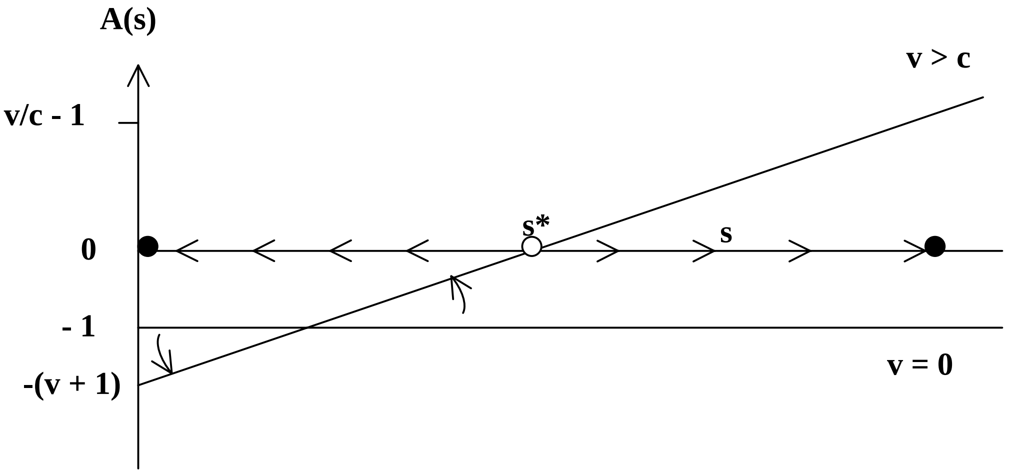
Table 2: Fitness with Vengeance

(*v* >0)	C	D
C	1 , 1	-1-*v*, 2 - *v/c*
D	2 - *v/c* , -1-*v*	0, 0

For *v* >*c*, the transformed game no longer has D as a dominant strategy. When population fraction *s* plays C, the expected fitness of C is W(C)=1*s* - (1+*v*)(1-*s)* and the expected fitness of D is W(D)=(2 -*v/c*)*s*. The two expressions are equal at *s** = (1+1/*v*)/(1+1/*c*). For *s*<*s** the expected utility is higher for D and we can expect cooperation to disappear as play converges to the inefficient (fitness 0) all-D equilibrium, as in the basic game. But for *s*>*s** the expected utility is higher for C and we can expect vengeance to drive out defection, resulting in the Pareto efficient all-C equilibrium. Thus for *v* >*c* we have a coordination game with two locally stable pure Nash equilibria and an unstable mixed Nash equilibrium at *s**<1, as illustrated in Figure 1.

Figure 1: The Advantage of Cooperating.

The fitness advantage A(s)=W(C)-W(D) is graphed as a function of the population fraction s playing C for two values of the vengeance parameter v. The graph of A rotates counterclockwise as v increases.



Technical Notes. Of course, efficient all-C behavior can also be sustained as a repeated game Nash equilibrium even in the original (*v*=0) version if culprits can be detected and identified, and if all players have a discount factors that exceed 0.5. One uses standard Tit-for-Tat or grim trigger strategies. But it may well be the case that repeat meetings are infrequent or culprits are hard to track, so the discount factor will be too small to sustain the efficient outcome. Thus the fear of vengeance can support efficient social outcomes that cannot be sustained by standard repeated game strategies.

The analysis above assumed linearity of payoff in the population fraction s that chooses C. A justification is that every individual interacts pairwise with each other individual equally often, or, alternatively, that the joint interaction in the entire group happens to be linear. Other assumptions regarding encounters (within a given group) lead to more complicated nonlinear expressions in s.

To expand on this point slightly, note that perturbations of the basic (v =0) game solely within the class of symmetric bimatrix games yield (as the only generic bifurcation) the coordination game above. But nonlinear perturbations of the basic game's payoff advantage function A(s) = W(C| s) - W(D| s) yield a generic bifurcation to a game with 3 NE--an all D and an unstable interior NE s^* as before, but with an asymptotically stable third NE at a point s^{**} between s^* and 1. The stable interior NE of course must award equal fitness to C and D. It is efficient (awarding higher fitness than the other equilibria) if v is not too large relative to c. (Similarly, s^* in the linear game above is more efficient than the all-D equilibrium if $2c > v > c$.) The intuition is that at s^{**} culprits are rare and usually escape harm, and the monitoring/vengeance behavior indexed by the bifurcation parameter v becomes too expensive at low incidence to pursue the last few culprits. There appears to be a somewhat more complicated analysis of this case that parallels the analysis to follow.

2. The Viability Problem

There is a flaw in the argument so far. The vengeance motive v itself is subject to evolutionary forces, perhaps slower forces than those determining the prevalence s of cooperation, but real forces nonetheless. Recall that the expected fitness of a cooperator is $W(\mathrm{C}|\ s,v) = 2s\text{-}1\text{-}v(1\text{-}s)$, which is a strictly decreasing function of v for any fixed s<1. Only when there are no culprits left to punish at s=1 is the expected fitness independent of v. Thus the fitness of player v is weakly dominated by that of player v' whenever $0<v'<v$. Assuming that players occasionally encounter culprits (an assumption we shall develop later), the vengeance preference parameter v will be driven towards 0 under any plausible evolutionary dynamics. We have a variant of the classic free rider or chiseling problem, and it seems that vengeance is not viable.

Several authors have encountered the viability problem in one form or another, and have found ways to finesse it. Rosenthal (1996) considers a limited form of vengeance in which a player can detect culprits and shun them after the first encounter. The payoffs of such players (called TBV for "trust but verify") are all reduced by verification costs. Rosenthal begins with a basic stage game like ours and then modifies it by expressing payoffs as present values of the continuing relationship. The harm a TBV player inflicts on a culprit is the present value of payoffs the culprit foregoes after the initial temptation payoff. The punishment cost is the present value of verification less the present value of the avoided (sucker payoff) loss, which for relevant parameter values is negative. Thus punishment brings a net personal *benefit* and the all-C strategy (corresponding to our v=0 player) does not weakly dominate the TBV strategy. Rosenthal finds several NE for his 3x3 symmetric game, and all-D need not be the only stable equilibrium. For certain parameter configurations, there is an interior NE that is stable under some (but not all) monotone dynamics. Unfortunately, no such stable equilibrium would exist under our maintained assumption that vengeance is costly and can not reduce the sting of the sucker payoff.

Sethi and Somanathan (1996) offer two attempts to get around the viability problem. First, they define stability to include neutral stability, not requiring convergence back to an equilibrium point following a small perturbation (ie, they do not require local asymptotic stability). In their model there is a continuum of neutrally stable equilibria with no culprits. Following a perturbation (a small invasion of culprits) the state moves along the continuum away from a vertex. Eventually, following sufficiently many such perturbations, the state leaves the equilibrium set and ultimately converges back to the all-D equilibrium. Thus, from a long-run evolutionary viewpoint, their other equilibria really are not stable, and their vengeful strategies are not viable. Implicitly recognizing the problem, Sethi and Somanathan refer in an appendix to a second approach due to Binmore and Samuelson (1996), in which the evolutionary dynamic is perturbed by a continuing stream of mutants in fixed positive proportions. The perturbed dynamic has a single asymptotically stable equilibrium point instead of the continuum of neutrally stable equilibria, but it has a very shallow basin of attraction and is supported by an arbitrary convention on the composition of mutants.

Huck and Oechssler (1996) deal with the problem in a richer context than ours. In the "ultimatum game" they study, players interact in small groups and have two roles, each played half the time. In one role ("responder") they can pursue a costly vengeance strategy that involves a fixed, common threshold for vengeance. However, they are not allowed in the other role (“proposer”) to exploit the common threshold by making just acceptable offers. If this kind of “shading” were allowed, it seems that cooperation would unravel.

The solution we shall propose to the viability problem is related to the two-level model for the evolution of cooperation as exposited in Sober and Wilson (1998) and S. Frank (1998). These authors note that, using a tautology known as the Price equation (G. R. Price, 1970)[1], one can demonstrate the possibility that a socially beneficial but dominated strategy (call it C) might survive in evolutionary equilibrium when group interactions are important. The idea in their analysis is that groups with a high proportion of C players have higher average fitness and thus grow faster than groups with a smaller proportion, and this effect may more than offset C’s decline in relative prevalence within each particular group. The necessary conditions for C to survive (it can never eliminate D but may be able to coexist in equilibrium) are rather stringent. Besides the obvious condition that the group effect favoring C must be stronger than the individual effect favoring the dominant strategy D, it must also be the case that the groups dissolve and remix sufficiently often, and that the new groups have sufficiently variable proportions of C and D players. These special conditions may be met for some parasites, but seem quite implausible as a genetic explanation of human cooperation. Indeed, Sober and Wilson invoke memetic evolution and discuss the importance of cultural norms for rewarding cooperative behavior and punishing uncooperative behavior. They avoid the viability problem by assuming in essence that c is 0; see p151 for the most explicit discussion of this point.

Bowles and Gintis (1998) consider the genetic evolution of vengeance in the context of a voluntary contribution game. Like Huck and Oechssler, they assume a direct tie between two discrete traits, a preference for punishing shirkers (analogous to our v) and a preference for helping a team of cooperators. Their argument is a version of two-level selection as in Sober and Wilson and again is rather delicate. In an essay on the rise of the nation state in the last

[1] The Price equation uses the definition of covariance to decompose the change in prevalence of a trait into two components, e.g., the direct effect from individual fitness and an indirect effect incorporating the spillovers within the group.

millennium, Bowles (1998) uses a version of the same model that allows for cultural and genetic coevolution.

Technical Notes. Another possible way to avoid the viability problem is to assume that individuals with higher values of *v* encounter D play less frequently. R. Frank (1987) discusses this possibility informally and formally models the evolution of a visible altruistic (rather than vengeful) trait. It is not hard to show under some specifications of how *s* depends on *v* that there is a positive level of *v* that maximizes fitness; the argument is precisely the same as used in section 4 below to maximize group fitness $W^g(\bar{v})$ as a function of group average vengefulness $\bar{v}$.

We resort to the more complex resolution of the viability problem because we believe that the relation between *s* and *v* arises mainly at the group level rather than at the individual level. Within well-functioning groups, D behavior is rare and dealing with it is not an important source of fitness differences. Presumably D behavior is more frequently encountered with partners outside one's own group, and we believe that here group reputations are the key, not individual signals or individual reputations. The distinction drawn in the next section is that within the group encounters are frequent, everyone knows everyone, and the all-C equilibrium is a good approximation of behavior. Across-group encounters are also frequent, but a given individual will encounter a specific non-group member only very sporadically. An individual in such encounters can't reliably signal her true *v* because outward signs can be mimicked at low cost, but neither (due to the large numbers of sporadic personal encounters) can she easily establish a reputation for her true *v*.

A specific assumption that captures these considerations is that the perceived vengeance parameter of one's opponent v^e is equal to the true value v in encounters within the group, but in encounters outside the group $v^e = \lambda\bar{v} + (1\text{-}\lambda)\mathrm{E}\bar{v} + \varepsilon$, an idiosyncratic error plus the weighted average of the partner's group average $\bar{v}$ and overall population average $\mathrm{E}\bar{v}$, with the weight λ on the group average an increasing function of group size. The idea is that v^e is a Bayesian posterior, with sample information on any individual overwhelming priors for internal matches and sample information on the relevant group being important for external matches. Implicit in this formulation is a theory of group size. Very large groups would violate the assumptions that everyone knows everyone well and monitors the all-C equilibrium, so there are diseconomies of scale. At the margin, these diseconomies should balance the economies arising from the dependence of λ on group size. We shall not attempt to develop such a theory[2] here, but simply will assume the existence of moderate size groups.

The viability problem is attenuated for social creatures that form groups of closely related individuals, such as slime molds (r=1- ε) or ants (r=2/3). But we are interested in humans, whose groups traditionally consist of individuals who are not necessarily closely related (say r= 0 to .25). As Sober and Wilson emphasize, there is a close abstract connection between multi-level selection and inclusive fitness arguments for related individuals, but the multi-level selection process we present below is so simple that such considerations seem irrelevant.

3. Coevolution

We propose a simple coevolution model to resolve the viability problem. Coevolution refers to the interaction of individual level ("gene") selection and group level ("meme") selection; it can be regarded as a variant of two-level selection. Primary references include Boyd and Richerson (1985), Dawkins (1976), and Durham (1991). In this section we motivate key elements of our model, and put them together in the next section.

During the vast majority of its evolutionary history, *Homo sapiens*, like other social primates, presumably lived in small groups of individuals who interacted with other individuals within the group on a daily basis and who interacted much less frequently with individuals outside the group. We do not model group membership, or changes in the size of groups, but take these as given. As we shall see below, groups are important in our analysis as carriers of reputation. This is empirically plausible for many kinds of encounters. For example, if one of the authors met a stranger on a train in India, the stranger might try to ascertain the author's family village and his last name, as ways of assigning him to a group with a particular reputation. The questioner is likely to find such information more useful than observable personal details, which are easier to disguise.

All known groups of humans maintain social norms, or memes, that prescribe appropriate behavior towards fellow group members and typically prescribe different appropriate behavior

[2] We are grateful to Bryan Ellickson for suggesting this approach to endogenous group size.

towards individuals outside the group. Sober and Wilson (1998) offer a full chapter on this point and an overview of the anthropological literature, but for us a representative quote on the Sikh Jats from Pettigrew (1975) will suffice.

> Relationships of extreme friendship and hostility between families were actively involved with the philosophy of life embodied in the concept of *izzat* -- the complex of values regarding what was honourable. …
> That aspect of *izzat* according to which the relationships between families were supposed to be ordered emphasized the principle of equivalence in all things, i.e., not only equality in giving but also equality in vengeance. *Izzat* was in fact the principle of reciprocity of gifts, plus the rule of an eye for an eye and a tooth for a tooth….*Izzat* enjoined aid to those who had helped one. It also enjoined that revenge be exacted for personal insults and damage to person or property. (p58)

The success of the meme, as with any other adaptive unit, is measured by its ability to displace alternatives, i.e., its fitness. There are several distinct reasons why one meme can displace others, as noted in Durham and in Boyd and Richerson, but we will focus on the most fundamental reason, enhanced individual fitness. Thus in the analysis to follow, a meme prescribing a particular pattern of vengeful behavior is fitter than existing alternatives when it brings higher average fitness to group members.

We summarize the relevant memes using two parameters: v^n for the group's normative vengeance level and a for the rigor with which the group enforces that norm. For example, Izzat applied to the basic game calls for rather strict enforcement of $v^n = 2/c$, since the culprit causes a loss of 2 (relative to the cooperative outcome of 1) and therefore h=2 is enjoined.

Enforcement is modeled by a loss function $\rho(x)$, where $x = v^n - v$ is the deviation of an individual's vengeful behavior from the group norm. The group imposes the fitness loss ρ on a deviator by lowering that individual's status or reputation within the group. (Note that this is not reputation in the sense it is used in modeling repeated games of incomplete information. Our modeling is more in the spirit of Akerlof (1983), though our focus and model details are quite distinct. See also Sober and Wilson for a discussion of low cost enforcement within groups.)

We have explored various specifications for ρ, including asymmetries for positive and negative deviations and with first order losses for first order small deviations; see the technical note in section 4. Fortunately, the main results hold for any convex loss function with a minimum at 0. The basic analysis uses the simplest possible quadratic specification, $\rho(x; a) = x^2/(2a)$, where enforcement is more rigorous the smaller is the parameter a>0.

Enforcement could affect the fitness of nondeviators as well as deviators. Indeed, since status is relative, a decrease in one individual's status will increase the status of some other group members and hence increase their fitness. To illustrate this possibility, consider Catanzaro's (1992) relativistic view of status in writing about Sicily and the Mafia: "... the men who usurped honor did so at the expense of others who stood to lose it to the same degree... Ultimately, honor has been described as a system of stratification [by Davis, 1980] ..." (pp. 46-47). In a technical comment in the next section we will discuss other possible effects of enforcement external to the individual but internal to the group, including the possibility that there are no spillovers from the loss of status. In the meantime we shall let R denote the fitness increment (possibly zero or even negative) an individual receives due to the deviations of other group members from v^n.

Each individual is characterized by two parameters: his actual vengeance level v, and the maximum possible value $v^{\max}$ that any meme could induce. The *capacity* for feeling anger and expressing it by damaging others as summarized in $v^{\max}$ may well be genetically transmitted, but the actual v of an individual probably is best regarded as learned from personal experience.

The last element of our model describes the frequency f with which an individual encounters culprits, and incorporates the idea that external reputation is carried by the group as a whole. Consider a group of individuals with average vengeance level $\bar{v} > c$. In external interactions these individuals may encounter D-players. Such external D players may play C within their own group but be hostile to outsiders, or may belong to groups whose internal interactions are described by the inefficient all-D equilibrium. For sufficiently high values of $\bar{v}$ any such outsider will be deterred from playing D against a member of the given group, if only because the outsider has learned to avoid them. For $\bar{v}$ sufficiently close to 0 outsiders can be expected to play D against the group members. These considerations can be summarized in a smooth decreasing function $f(\bar{v})$ describing the overall frequency of encounters with culprits, so that $f(\bar{v}) = 1$ at $\bar{v} = 0$ and approaches 0 as $\bar{v}$ becomes very large. For present purposes it will suffice to take the function f as exogenous and note that it will be shifted by changes in the group's environment, including the composition of neighboring groups. A convenient parameterization that we adopt is $f(v) = \exp(-v/b)$, where b is a positive parameter representing the hostility of the environment.

The next section derives the uniform level v^o that is optimal for the group given the encounter function $f(\bar{v})$. Derivation of v^o is conceptually and technically straightforward, but its relevance is not immediately obvious, due to the basic viability problem. In the next section we will show that v^n mediates a close connection of v^o to the individual optimum and hence to the group average $\bar{v}$.

Technical Notes. A complete specification of a group's meme should not only prescribe the norm v^n for outgroup culprits and the enforcement ρ imposed on deviators, but also prescribe behavior towards fellow group members and towards outgroup non-culprits. For example, the group may be able to enforce internal cooperation more cheaply by using status enforcement than by using the punishment technology. It turns out that those additional prescriptions don't affect our conclusions, with the following exception. In conflicts between a *v* group and a $v' < v$ group, suppose that members of both groups have an equal probability of meeting D behavior from the other group. Then the *v* group has higher fitness because (since $c < 1$) their extra cost is more than compensated by the extra damage they inflict. The argument is straightforward.

The function *f* may be viewed as a simplification of a more general model of interactions between group members and outsiders in which *f* is a function of other variables besides the average vengeance parameter of the group. In the Appendix, however, we show that these complications are all finessed for quite a general specification of how individuals behave in encounters outside the group, and that the simple analysis in the next section carries over neatly to the more general specification.

A more complete model would derive proportion $f(\bar{v})$ of culprits from the state variable in the population game that describes the distribution of memes across groups. Friedman and Yellin (1997) shows that it is possible to do so using a system of coupled non-linear partial differential equations. The simpler static analysis of Continuously Stable Strategies as in Eshel (1983) might also suffice.

A few remarks may be in order about fitness, monotone dynamics and time scales. We shall assume that individual levels of *v* adjust rapidly within $[0, v^{max}]$; the idea is that people learn and accommodate themselves to the group's meme within a short period, say weeks or months. (According to stories in the media, kids raised in Belfast and Lebanon brought to the US have no problem adapting with a few months to the US norm and then adapting back when they return.) Memes also adjust, but in the medium run of years to decades. By definition, v^{max} is innate, but it too can adjust in the long run, over several generations. Thus for simplicity we assume that, at any given time scale, only a single (scalar) variable is adapting. In this case, the direction of change is immediate from the definition of fitness: values of v that bring higher fitness become more prevalent in the population at the expense of values that bring lower fitness.

4. Results

Recall that $v \in [0, v^{max}]$ denotes the level of vengeance currently used by a given individual, where v^{max} is the maximum capacity for vengeance under the most extreme provocation. The current average level in a given group is $\bar{v}$, and the normative level is v^n. The objective social optimum in the current environment is v^o. The first result shows that short-run learning dynamics will drive v and $\bar{v}$ toward some individually optimal level v^*, and the second result shows how medium run meme selection aligns v^n with the social optimum v^o.

Recall that a v–cooperator encountering a defector will receive fitness loss $(1+v+\rho(v^n -v))$, the sucker payoff plus the cost of wreaking vengeance plus the social loss from departing from the norm. The same individual will receive a fitness gain of 1 in encounters with cooperators. Recall that the proportion of encounters with defectors is $f(\bar{v})$. Thus the individual's expected fitness is

$$W(v \mid \bar{v}, v^n) \ = \ 1(1- f(\bar{v})) – (1+v+\rho(v^n -v)) f(\bar{v})+ \mathrm{R} \ = \ 1 - f(\bar{v})(2+v+ \rho(v^n -v)) + \mathrm{R},$$

where R is the base-level fitness including the (positive) effect on one's status from other group members' deviations from the norm v^n. The fitness function does not account for the possibility that the individual will ever play D, but this omission is harmless as pointed out in the technical notes.

For given $\bar{v}$ and v^n, short run selection (or learning dynamics) will drive v towards values that increase fitness W, or equivalently, that decrease the simpler expression $v + \rho(v^n – v)$. The first-order condition is $1 = \rho'(v^n – v) = (v^n – v)/a$, with solution $v^* = v^n – a$. It is easy to see that W is single peaked at v^*, so short run dynamics push the individual's parameter towards this optimum. The optimum will be attained as long as the value is within the allowable range;

otherwise v^* is truncated below[3] at 0 and above at v^{max}. Since learning dynamics are assumed rapid, we conclude that v^* is a good approximation of all individuals and an even better approximation of their average. Thus we have our first result: in short run equilibrium, $\bar{v} = v^* = [v^n - a]_{[0,\ vmax]}$, maximizing individual fitness for given a meme v^n and a.

Of course, this individual optimum does not necessarily maximize the group's fitness. Recall from the previous section that R cancels the mean contribution of ρ, so the group average fitness is simply

$$W^g(\bar{v}) \ = \ 1(1- f(\bar{v})) - (1+ \bar{v})\, f(\bar{v}) = 1 - f(\bar{v})(2+ \bar{v}).$$

The group optimum v^o is the value that maximizes this expression on (c , v^{max}]. Inserting the parameterization $f(v) = \exp(-v/b)$, the first order condition reduces to $2+v = -f/f' = b$, and v^o then is b-2, truncated to (c , v^{max}]. While the solution here is particularly simple, the technical notes show that similar conclusions hold quite generally.

What then is the relation between the group optimum v^o and the individual optimum v^* ? Assume for the moment that the both are interior, so $v^* = v^n - a$ and $v^o = b$-2. The group meme, embodied in the parameters a and v^n, is subject to selective pressures in the medium run, and W^g is again a single-peaked function. Any group whose memes bring $v^* = v^n - a$ closer to $v^o = b - 2$ has a selective advantage. Hence we have our second result: in the medium run equilibrium, b-2 $= v^n - a$, so $v^n = a + b - 2$. (No truncation is necessary for memes.)

The main conclusion, a simple corollary of result 2, is that $v^o = v^*$, i.e., the vengeance level is socially optimal in medium run evolutionary equilibrium. The memes that support this efficient equilibrium, however, do not typically include the actual optimum value v^o but rather an exaggerated version $v^n = v^o + a$. We show in the technical notes that this main conclusion holds under conditions far more general than the simple parametric model used here.

A few words on corner solutions may be in order. If v^* or v^o is c, then D is a (weakly) dominant strategy and the viability problem becomes moot. When the environment is so hostile that either v^* or v^o is v^{max} then there is a fitness gain from increasing v^{max}, and presumably biological evolution will do so in the long run. The point is that it is the coevolution of the meme

[3] In view of the fact established in the basic model that efficient group equilibrium requires $\bar{v} > c$, we can truncate below at c instead of 0. On the other hand, truncation at 0 would be appropriate in an extention of the basic model along the lines discussed at the beginning of the previous technical notes, with cheap ingroup enforcement of cooperation. This point doesn't affect our results.

(v^n and perhaps a) with the gene (v^{max}) that enables actual behavior to track optimal behavior as the environment changes. See Durham (1991) for several examples of such coevolution at work, including sickle-cell anemia in yam-growing areas, or lactose tolerance in herding communities.

On the longer time scale, there can be shifts in the environment (as captured in the parameter b) and in the punishment technology (as captured in c). These shifts will affect the encounter function f and hence the group optimum v^o. Our main conclusion implies that memes will adjust under selective pressure in the medium run so that individual behavior $v*$ will track the new group optimum.

Technical Notes.

*Alternative Loss Functions***.** Consider the example $\rho = \exp(k\,|v^n - v|\,) - 1$, where k is a positive parameter that measures the severity of the enforcement of the norm. The kink in ρ at 0 implies a first order loss for first order small deviations.

The first order condition $\rho'(v^n - v) = 1$ is now $k\exp[k\,(v^n - v)] = 1$, with solution $v* = v^n + \ln k/k$. If $k \leq 1$ then $v* \leq v^n$ and the solution is still of the form $v = v^n - a$, and the previous analysis therefore carries over to this case. If $k > 1$, we have a corner solution, given by $v* = v^n$, which is a limiting case of $v^n - a$ as a approaches 0. In the medium run equilibrium in this case, $v^n = v^o$, that is, the memes that support this group-optimal equilibrium include the actual optimum value v^o. Thus the analysis proceeds as in the main text, with a treated as 0.

Asymmetry can be introduced by setting $\rho = 0$ for $v > v^n$, or by using different values of k for positive and negative deviations. Since $v* \leq v^n$ is the relevant range for solutions, such asymmetries will have no effect on the subsequent analysis.

Alternative Assumptions about Status. Recall the expression for individual fitness $W(v \mid \bar{v}, v^n) = 1 - f(\bar{v})(2+v+\rho(v^n - v)) + \mathrm{R}$. Suppose now that status is not completely relative, so that R only partially cancels out $\rho(v^n - v)$. We can model this by introducing a parameter $t\ \varepsilon\ [0, 1]$ that measures the net loss of average fitness due to deviations from the norm. Group average fitness becomes $W^g(\bar{v}) = 1 - f(\bar{v})(2+\bar{v} + t\rho(v^n - \bar{v}))$. With f and ρ as specified in the main text, the first order condition for the medium run equilibrium is now

$[1 - t(v^n - v)/a]\exp(-v/b) = -[2 + v + t(v^n - v)^2/2a]\,(-1/b)\exp(-v/b)$.

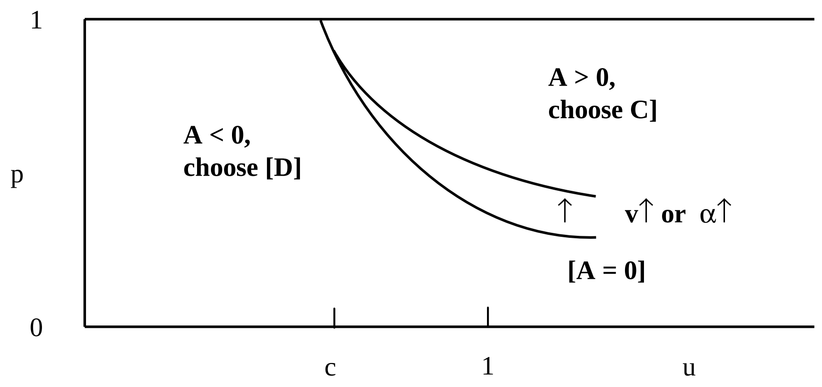
Canceling the exponential terms, multiplying through by b, and substituting v^n -v with a, yields $b(1 - t) = (2 + v^n - a + at/2)$, or $v^n = a(1 - t/2) + b(1 - t) - 2$.

If $t = 0$, we have the case analyzed in the text. At the other extreme, $t = 1$, only absolute status matters. In that case, $v^n = a/2 – 2$, independent of the parameter b. In general, greater weight on absolute rather than relative status (i.e., a higher t) decreases the equilibrium norm v^n, since the derivative $dv^n /dt = -a/2 - b$ is negative. The comparative statics for v^n with respect to a and b are qualitatively the same for all values of t in the unit interval, i.e., v^n increases as either a or b increases. In words, if enforcement is less stringent (higher a) or the environment is more hostile (higher b), then the norm of vengeance in the medium run equilibrium will be higher.

Endogenous encounter function. A mathematical appendix sketches the derivation of the encounter function $f(\bar{v})$ and, under much more general conditions than used above, obtains the first-order condition characterizing v^*. It first derives consistent estimates of the probability that two strangers will choose C and D given imperfect observation of each others' v parameters. It then identifies regions in the perceived characteristic space where the individual will choose C or D as in Figure 2 below. Now individual fitness is given by a sum of integrals over the choice regions. The encounter function f and the first order condition $1=\rho'(v^n -v)$ turn out to arise naturally in this general setting.

Finally, one can consider a whole population of groups, each with possibly different v's, rather than a given group adapting to a fixed environment. The asymptotic behavior in such an adaptive landscape perhaps can analyzed using a proposition of Friedman and Yellin (1997) on monotonous landscapes.

Figure 2: The Decision Rule.



The appropriate choice of C or D is given by the sign of the advantage function A(p,u), where p is the probability that the partner will choose C and u is an unbiased estimate of her vengeance parameter. The A=0 locus shifts up with increases in the the decision maker's direct (v) or full (α) vengeance cost.

6. Discussion

Our argument can be summarized briefly. Vengeance, or a taste for negative reciprocity, is an important part of the human emotional repertoire. We model its important role in sustaining cooperative behavior but highlight an intrinsic free-rider problem: the fitness benefits of vengeance are dispersed throughout the entire group but the fitness costs are borne personally. Evolutionary forces tend to unravel people's willingness to bear the personal cost of punishing culprits. In our model, the countervailing force that sustains vengeance is a group norm together with low-powered (and low-cost) group enforcement of the norm. Such memes coevolve with personal tastes and capacities so as to produce the optimal level of vengeance.

One could object to our account on several grounds. First, it is too simple. The underlying social dilemma was modeled as a specific and very simple prisoner's dilemma game. In reality, the stakes and complexity of social interactions vary considerably, and actual memes are more complex and variable than in our model. Ours is the usual response: insight is clearest with an appropriate simple model, and the model can be extended as necessary to deal with specific complexities that are essential in specific applications.

One could also object that the model is too complicated, especially if the main goal is to explain cooperation. Norms of cooperative behavior and their enforcement could be modeled directly. The same apparatus should suffice: preferences that offer a utility gain (but not a fitness gain) for positive reciprocity together with a social norm from which deviations lead to fitness loss. Vengeance thus seems redundant.[4] Our response is twofold. First, our primary goal is to explain vengeance, not cooperation per se. Second, since culprits are rare and cooperators are ubiquitous in successful society, the fitness cost of a meme that relies entirely on positive reciprocation might be excessive. Although their model is somewhat different, ultimately this is the problem faced by Sethi and Somanathan's analysis, in that they are not quite able to get a long-run stable cooperative equilibrium based solely on a norm of cooperation. Our suggestion, therefore, is that social norms of vengeance, in taking advantage of biological capacities in that direction, are able to relieve direct social norms of cooperation of at least some of the burden of sustaining cooperative behavior. Thus, the existence of direct social norms of cooperation does not, in our view, make vengeance redundant.

A third objection to our account is that it is too powerful: all sorts of behavior, including behavior that has never been seen and never will, could be described as coevolutionary equilibria. We concede this point, but have been unable to find a simpler account that explains the viability of vengeful preferences and their role in supporting cooperation. Of course one needs additional principles to get a reasonably sharp theory, and here we have relied on basic anthropological facts. There are indeed many ways to capture the potential gains to cooperation. Social insects, for example, rely on close genetic kinship. Likewise, bipedalism isn't the only (or even necessarily the best) form of locomotion: it is worth studying because it is the one humans use. We claim nothing more (or less) than this for our focus on vengeance as a means of reaping the gains of cooperation.

It is reasonable to speculate how our model applies in different societies. The application to hunter-gatherer bands or villagers is perhaps clearest; here the parameter *b* reflects directly the uncooperative tendencies of people from neighboring bands or villages, and *c* reflects the opportunities to identify, track down and inflict harm on them. In more highly structured societies, vengeance may be exacted by specialists delegated to carry out such roles, rather than

[4] We are grateful to our colleague, Donald Wittman, for raising this important point.

the solely by the vengeful individual. The marginal cost *c* of vengeance then is lower to the extent that some of the cost is transferred from the individual to others in society, but the relevant *c* presumably is still positive.[5] Our model therefore still applies to more complex societies, but it is incomplete in that it takes as exogenous the institutional mechanisms that alter the vengeance technology.

What are the empirical implications and applications of our model? One can easily imagine laboratory experiments that would distinguish a taste for negative reciprocity from the egalitarian preferences hypothesized by recent writers. The comparative statics of the model are also clear in principle and testable with anthropological data: norms of vengeance and actual vengeful behavior should vary systematically with the hostility of the environment, the technology for harming culprits, and the technology for enforcing group norms. If our approach fares well in such tests, one could look for economic applications. The connections to industrial relations (wage bargaining and strikes, etc.) seem clear. International trade negotiations might also be a fruitful area of application.[6] If the model is on the right track, there is reason to hope that extremely disfunctional vengeful behavior (as in the Balkans) might improve in coming decades as the relevant memes evolve.

[5] The delegation of vengeance to specialists presumably reduces total costs because of economies of scale, comparative advantage, learning by doing, and so on, and a large portion of the total cost is borne by society at large. Even so, the relevant individual costs of venegeance remain positive: reporting harm and pursuing vengeance through the legal system can still be quite costly. In fact, legal systems that are inefficient and costly tend to encourage the pursuit of vengeance by extra-legal means.

[6] International trade negotiations are a bit difficult to understand on a purely rational basis: a country benefits unilaterally from liberalizing trade, but negotiations proceed as if some sort of exchange of concessions is required for mutual gain. Trade sanctions can be regarded as vengeance. Another possible international application is labor or environmental standards. One can pose a race-to-the-bottom prisoner’s dilemma model for such standards set nationally, and here vengeance (in the form of trade sanctions) might create a more efficient "cooperative" regime.

Bibliography

Akerlof, George (1983), *An Economist's Book of Tales,* NY: Cambridge University Press.

Andreoni, James and John H. Miller (1996), "Giving According to GARP: An Experimental Study of Rationality and Altruism," University of Wisconsin working paper.

Binmore, Kenneth and Larry Samuelson (1996), "Evolutionary Drift and Equilibrium Selection," University of Wisconsin working paper; forthcoming, *Review of Economic Studies*.

Blinder, Alan and Don Choi (1990), "A Shred of Evidence on Theories of Wage Stickiness," *Quarterly Journal of Economics* 105, 1003-1016.

Bolton and Ockenfels (1998). "ERC: A Theory of Equity, Reciprocity and Fairness," Penn State University manuscript.

Bowles, Samuel (1998), "Cultural Group Selection and Human Social Structure: The effects of segmentation, egalitarianism and conformism," University of Massachusetts Amherst working paper.

Samuel Bowles and Herbert Gintis (1998), "The Evolution of Strong Reciprocity," University of Massachusetts Amherst working paper.

Boyd, Robert and Peter J. Richerson (1985), *Culture and the Evolutionary Process,* University of Chicago Press.

Catanzaro, Raimondo (1992), *Men of Respect: A Social History of the Sicilian Mafia*, New York: The Free Press.

Davis, J. (1980), *Antropologia della societa mediterranee: un'analisi comparata,* Turin: Rosenberg & Sellier.

Dawkins, Richard (1976), *The Selfish Gene*, NY: Oxford University Press.

Durham, William H. (1991), *Coevolution : genes, culture, and human diversity,* Stanford, Calif.: Stanford University Press.

Eshel, Ilan (1983). "Evolutionary and Continuous Stability," *Journal of Theoretical Biology* 103, 911-111.

Fehr, Ernst and Simon Gaechter(1998), "Cooperation and Punishment," University of Zurich manuscript, September.

Fehr, Ernst and Klaus Schmidt (1998) "A Theory of Fairness, Competition and Cooperation," University of Zurich manuscript, December; forthcoming QJE.

Frank, Robert (1987), “If *Homo Economicus* Could Choose His Own Utility Function, Would He Want One with a Conscience?” *American Economic Review* 77:4, 593-604.

Frank, Robert (1988), *Passions within Reason*: *The Strategic Role of the Emotions,* NY: WW Norton.

Frank, Steven (1998), *Foundations of Social Evolution*, Princeton NJ: Princeton University Press.

Friedman, Daniel and Joel Yellin (1997), “Evolving Landscapes for Population Games,” UC Santa Cruz manuscript.

Huck, Steffen and Jorg Oechssler (1996), “The Indirect Evolutionary Approach to Explaining Fair Allocations", Humboldt University Berlin manuscript, forthcoming in *Games and Economic Behavior*.

David K. Levine (1998) “Modeling Altruism and Spitefulness in Experiments,” *Review of Economic Dynamics* 1, 593-622.

Pettigrew, Joyce (1975), *Robber Noblemen: A Study of the Political System of the Sikh Jats*, London: Routledge & Kegan Paul

Price, George R. (1970), “Selection and Covariance,” *Nature* 227(5257, August 1), 520-521.

Rabin, Mathew (1993), “Incorporating Fairness into Game Theory and Economics,” *American Economic Review* 88:5, 1281-1302.

Rosenthal, Robert W. (1996), “Trust and Social Efficiencies,” Boston University manuscript.

Sethi, Rajiv and Eswaran Somanathan (1996), The Evolution of Social Norms in Common Property Resource Use,” *American Economic Review*, 86:4, 766-788.

Elliott Sober and David Sloan Wilson (1998). *Onto others: The evolution and psychology of unselfish behavior*, Harvard University Press.

Steers, Richard and Lyman Porter (1991), *Motivation and Work Behavior, Fifth Edition*, NY: McGraw-Hill.

Appendix A: Derivations

These notes derive several key constructs from more general assumptions than those used in the body of the paper. We solve a decision problem faced by an individual encountering a new partner, or "stranger". The encounter function f and the characterization of the individual optimum emerge endogenously.

Probabilities of Cooperation. Let i=1 index the given individual and i=2 index the stranger. Their true degrees of vengefulness (v^1, v^2) are imperfectly perceived by the other person; 1's perception of 2's v is $\hat{v}^2 = v^2 + e^2$, and similarly (replacing 2 by 1) for 2's perception of 1. It is common knowledge that the perception errors (e^1, e^2) have mean zero and joint cumulative distribution function $G(e^1, e^2)$.

The expected payoffs to cooperation $W^i(C|...)$ and to defection $W^i(D|...)$ can be expressed in terms of i's perceptions of $j = 3 - i$ and i's own characteristics as follows. Let $p^i \in I = [0, 1]$ be j's estimate of the probability that i will play C; for the moment it is arbitrary, but we shall derive it shortly. Let $\alpha^i = v^i + \rho(v^{n(i)} - v^i)$ be the full cost of vengeance to i, taking into account (if applicable) the loss ρ that his group imposes when he deviates from their norm $v^{n(i)}$. (If applicable, denote the induced estimation error of α^i by $\tilde{e}^i$.) Then $W^i(C) = (1)p^j + (-1 - \alpha^i)(1 - p^j) = -(1 + \alpha^i) + p^j(2 + \alpha^i)$, and $W^i(D) = (2 - v^j/c)p^j + (0)(1 - p^j) = p^j(2 - v^j/c)$. Either person i will choose C when the perceived advantange $A^i(p^j, v^j, \alpha^i) = W^i(C) - W^i(D)$ is positive and choose D when A^i is negative.

Now we need some second-order reasoning. Write j's perception of i's perceived advantage function as $A^i(p^j, v^j + e^j, \alpha^i + \tilde{e}^i)$. The error $\tilde{e}^i$ reflects the fact that j knows i's vengeance cost α^i imperfectly, and the error e^j is included because j realizes that i knows j's own v imperfectly. (The error e^j was omitted from the $W^i(D)$ expression above because it has mean zero, covariances can be relevant.) The probability p^j is still arbitrary, but now we have the machinery in place to enforce consistency.

The construction of consistent (i.e., Bayesian Nash equilibrium) probability estimates uses a best response map (p^1, p^2) into an updated choice (q^1, q^2), and looks for a fixed point. The idea is that the tentative choice probabilities plugged into the decision function A imply new choice probabilities, and the probabilities are internally consistent at a fixed point. Formally, the first component of $B(p^1, p^2)$ is $q^1 = m[A^1(p^2, v^2 + e^2, \alpha^1 - \tilde{e}^1)|G(e^1, e^2)]$, where the expression $m[a(x)|F(x)]$ denotes the measure (i.e., the probability mass) of the set of x's such that $a(x) \geq 0$, given that x has distribution function F. The second component of B is $q^2 = m[A^2(p^1, v^1 + e^1, \alpha^2 + \tilde{e}^2)|G(e^1, e^2)]$.

One can show that the mapping $B : (p^1, p^2) \mapsto (q^1, q^2)$ of the positive unit square I^2 into itself, satisfies the assumptions of the Brouwer theorem and therefore has a fixed point. This conclusion holds for particular choice of (v^1, v^2); indeed, the mapping B depends smoothly on (v^1, v^2) if G has a density function. Therefore one can assign (not necessarily uniquely) fixed point (p^1, p^2) probability estimates as a function of (v^1, v^2). Thus we have the mapping we sought, call it $P : [0, v^{max}]^2 \to I^2,\ (v^1, v^2) \mapsto (p^1, p^2)$. One can verify (although it is not necessary for our purposes) that P is the assessment component of a Bayesian Nash equilibrium.

In practice, a nice way to implement P is to begin with initial estimates $p^1 = p^2 = 0.5$ and to iterate using the B map (for the actual values of the v's) until convergence.[1] The intuition is not that people actually do the iteration or the calculation, but rather that a stable convention emerges on how likely you (as member of a group with a particular value of v) are to encounter C play from a stranger with given apparent v.

The Individual Optimum and the Encounter Function. The next exercise is to derive general expressions for fitness functions and to characterize the individual optimum. We focus on a particular individual (i=1 in the last subsection) whose vengeance parameter v is to be shaped by the learning process. Others' perceptions of him have mean $\bar{v}$ and remain constant during this process; the interpretation in the text was that the others perceive his group affiliation but have no other credible information about him.

The individual faces an environment defined by a distribution function $F(u)$ for strangers' vengeance parameters $v^2 = u$. The distribution $F(u)$, together with the mapping P derived above, induces a distribution function $H(p, u|\bar{v})$, where p denotes the first component p^1 of $P(\bar{v}, u)$. The distribution H summarizes the fitness-relevant data for the individual: the probability p that the stranger will play C and her (correlated) vengeance parameter u. Monotonicity properties of the mapping P imply an ordering by $\bar{v}$ of the distributions H via first-order stochastic dominance. (Verifying this ordering is an exercise not yet completed.)

Consider the possible values of (p, u) in the rectangle $I \times [0, v^{max}]$, as in

[1] We leave as an exercise for the interested reader with access to Matlab or similar software to graph the countours of final (p^1, p^2) in (v^1, v^2)-space.

Figure 2 of the text. Simplifying the notation of the previous subsection, the individual's decision function $A^1(p, u, \alpha^1(v)) = A(p, u, \alpha) = -(1+\alpha) + p(u/c+\alpha)$. The locus $A(p, u, \alpha) = 0$, which is the graph of the relation $p = \frac{1+\alpha}{u/c+\alpha}$, separates the rectangle into two regions, denoted [C] and [D] to indicate the individual's choice. The measure (or probability mass, using the distribution H) of these regions gives the overall probabilities of C and D play by an individual whose imperfectly perceived vengeance parameter is $\bar{v}$.

The individual's fitness is the expectation (with respect to the distribution H) of the fitness payoff to C or D over the possible new partners. It is given by the Stieltjes integral

$$\begin{aligned} w(v|\bar{v}, H, \rho) = \int_{p=0}^{1} \int_{u=0}^{v^{max}} \max\{W(C), W(D)\} H(dp, du|\bar{v}) = \quad (1) \\ = \iint_{[C]} W(C) H(dp, du|\bar{v}) + \iint_{[D]} W(D) H(dp, du|\bar{v}). \end{aligned}$$

The key calculation is the fitness gradient. Taking the derivative in (1) with respect to v we obtain

$$\begin{aligned} \frac{dw}{dv} = \iint_{[C]} \frac{dW(C)}{dv} H(dp, du|\bar{v}) + \iint_{[D]} \frac{dW(D)}{dv} H(dp, du|\bar{v}) \quad (2) \\ + \oint_{[A=0]} (W(C) - W(D)) \cdot (dA/dv) H(dp, du|\bar{v}). \end{aligned}$$

The last term in (2) is a line integral over the locus A=0. It comes from the relevant generalization of the fundamental theorem of calculus (or a special case of Stokes' Theorem) because the locus moves when v changes. Conveniently, zero because $W(C) = W(D)$ precisely on the locus A=0 where C and D are equally fit.

Recall that $W(D) = p(2-u)$ depends on the stranger's vengeance parameter u but is independent of the individual's own value of v, so the middle term in (2) also vanishes. That leaves only the first term, whose integrand is the derivative of $W(C) = -(1+\alpha(v)) + p(2+\alpha(v))$ with respect to v. Hence

$$\frac{dw}{dv} = -(d\alpha/dv) \iint_{[C]} (1-p) H(dp, du|\bar{v}) = [\rho'(v^n - v) - 1] f(\bar{v}), \quad (3)$$

where the encounter function used in the text is now seen to be precisely the probability $f(\bar{v}) = \iint_{[C]}(1-p)H(dp, du|\bar{v})$ that the individual is the victim of the sucker payoff. This probability is independent of v, so the shape of the payoff function w depends only on the group's enforcement function ρ.

It now is clear that the simple argument in the text applies directly since it was based on the same first order condition $\rho'(v^n - v) = 1$ that emerges here. We conclude as before that individuals will adapt monotonically towards a point v^* somewhat below the group norm v^n, with the size of the gap depending on the rigor with which the norm is enforced.