

September 2011

No.53

**Reputation and Cooperation in Defence**

David Hugh Jones and Ro'i Zultan  
University of Warwick and University College London

**WORKING PAPER SERIES**

Centre for Competitive Advantage in the Global Economy

Department of Economics

# Reputation and Cooperation in Defence

David Hugh-Jones and Ro'i Zultan\*

August 30, 2011

## Abstract

Surprisingly high levels of within-group cooperation are observed in conflict situations. Experiments confirm that external threats lead to higher cooperation. The psychological literature suggests proximate explanations in the form of group processes, but does not explain how these processes can evolve and persist. We provide an ultimate explanation, in which cooperation is a rational response to an external threat. We introduce a model in which groups vary in their willingness to help each other against external attackers. Attackers infer cooperativeness of groups from members' behaviour under attack, and may be deterred by a group that bands together against an initial attack. Then, even self-interested individuals may defend each other when threatened in order to deter future attacks. We argue that a group's reputation is a public good with a natural weakest-link structure. We extend the model to cooperative and altruistic behaviour in general.

Keywords: cooperation, conflict, defence, signalling

JEL Classification: C73, C92, D74

Word count: 8,449

---

\*David Hugh-Jones: CAGE, Department of Economics, University of Warwick in Coventry CV4 7AL. D.Hugh-Jones@warwick.ac.uk. Ro'i Zultan: Cognitive, Perceptual and Brain Sciences, University College London. We thank David Myatt, Eric Maskin, Christian Ghiglino, Rene Levinsky and Morimitsu Kurino for comments.

# 1 Introduction

On August 6th, 2011, a riot started in Tottenham Hale in North London, involving arson and rampant looting. Over the next three days, riots spread to other parts of London and several other cities in the UK. Within a few days of the riots, people came together in large cooperative efforts to counter the riots and their aftermath. People who were not personally threatened by the riots voluntarily formed vigilante groups to deter further rioting in their communities, at personal cost and risk to themselves (Beaumont et al., 2011). Hundreds of volunteers arrived at riot-stricken areas to help with the clean-up efforts (BBC 2011, Davies et al. 2011). How did the riots lead to such large-scale cooperation, when people could just as easily stay home and free-ride on the effort of others?

A fundamental puzzle for rationalist explanations of group conflict is that conflicts involve individuals voluntarily cooperating, perhaps at great risk, to gain a collective benefit (Olson, 1974). Blattman and Miguel (2008), in a wide-ranging review of the civil war literature, see “the sources of armed group cohesion amid pervasive collective action problems” as a central unresolved theoretical puzzle, and designate “the complex individual motivations underlying participation in armed groups” as “an important area for future research”. Defence is a canonical example of a “public good”, whose provision benefits not only the providers, but also free-riders who contribute nothing. Accordingly, standard economic theory predicts that defence will be underprovided unless the state enforces contributions. Nonetheless, in many conflicts, people fight for their group against other groups, in the absence of state coercion. Furthermore, there is considerable laboratory and field evidence that conflict increases cooperativeness in general. Existing psychological theories, while they offer insight, can provide only proximate explanations for this effect. In this paper, we attempt to provide an ultimate explanation, in terms of the rationality and evolutionary optimality of cooperation during conflict.

We demonstrate a mechanism for the evolution of helping behaviour between individuals from the same group, when those individuals come under attack by, for example, a rival ethnic group, or a biological predator. The logic is that of reputation building (Kreps et al., 1982; Milgrom and Roberts, 1982). The argument runs as follows:

1. Attacks against one group member are less likely to be successful if the member is defended or supported by others in the group.
2. Groups vary in the willingness of their members to cooperate against attackers. This can be for many reasons. For instance, some groups may be engaged in long-term cooperative relationships (Trivers, 1971; Fudenberg and Maskin, 1986; Neyman, 1985), which will be terminated when help is not provided to attacked partners, either due to reciprocal strategies or because the partners are killed (Garay, 2008; Eshel and Shaked, 2001); other groups may be individually self-sufficient members without a direct incentive to cooperate. Or, some groups may be composed of closely related kin, with high mutual altruism, while others are made up of unrelated individuals.
3. Attackers are opportunistic: they attack so as to acquire group members' resources (or, in the case of biological predators, for food). They are therefore more willing to attack group members if they expect low levels of cooperation in defence. Conversely, if they expect a strong defence from a group, they may prefer to engage in an alternative, less risky activity, or to find a different group to attack.<sup>1</sup>
4. Because of the previous point, attackers have an interest in finding out the type of group they are facing. However, they cannot always observe a group's level of cooperativeness directly. Instead, they will find it optimal to make one or more initial attacks, in order to gauge the cooperativeness of a particular group. They can then decide whether to continue attacking or to break off.
5. As a result, even members of uncooperative groups have an interest in appearing cooperative during the initial stages of an attack. By doing so, they may deter the attacker, and prevent future attacks which would eventually fall on themselves. In the terminology of signalling games, less cooperative groups have an incentive to *pool* with more cooperative groups.

---

<sup>1</sup>We treat attackers as single self-interested agents, thus abstracting away from two-sided group conflicts. This would be an interesting extension to the theory.

6. A group's appearance of being cooperative is itself a public good, so it might seem that the collective action problem has been reintroduced at a higher level. However, this public good has a natural "weakest-link" structure. If a single group member fails to cooperate, pooling instantly fails; the attacker learns that the group is not truly cooperative, and can no longer be deterred from further attacks; other group members then have no more incentive to cooperate. This dramatic collapse of cooperation after a group member "breaks the chain" provides a strong incentive not to do so.

We model this logic in a simplified setup. Some groups (henceforth *strong* types) participate in social interaction, and will therefore help their fellows who come under attack,<sup>2</sup> whereas other groups (henceforth *normal* types) have weak intragroup connections and therefore are not motivated to help their peers. An attacker makes one or more attacks on a group; during each attack, the (randomly selected) target individual may be helped by another randomly selected individual, at a cost to the helper which the helper privately observes. After each attack, the attacker may break off and attack a new group.

When the maximum number of possible repeated attacks is large enough, this model has a unique equilibrium that survives a natural refinement. The equilibrium has the following characteristics. First, for any fixed group size, so long as individuals are patient enough, helping behaviour can be sustained, even for arbitrarily large costs of helping. These costs may even be larger than the benefit provided to the helped individual. This holds because the motivation to help is provided not by the benefit to the target, but by the deterrence effect of driving off an attacker. In fact, our results would hold even if helping purely harmed the attacker without benefiting the defender, suggesting that this model might also explain the evolution of third-party punishment. In human conflicts, seemingly trivial incidents such as insults of a group member may lead to disproportionate responses.<sup>3</sup> Second, it is irrelevant what proportion of the groups are actually strong types: this can be arbitrarily small. Third,

---

<sup>2</sup>Intuitively, helping in time of attack can be maintained in an equilibrium of a larger game that includes long-term interactions such as trading.

<sup>3</sup>Many examples can be found in Horowitz (2001). Stephan and Stephan (2000) discuss "symbolic threats" from a social psychological point of view.

cooperation among normal types becomes less likely as the number of previous attacks increases.<sup>4</sup>

Lastly, cooperation is subject to sudden collapses: if a single individual does not help, then everyone else stops helping. This is closely tied to the reputation logic of the game. An individual who doesn't help provides an unambiguous signal to the attacker that he is facing normal types, not strong types. Afterwards the attacker can no longer be deterred, and this removes the incentive for other group members to help. Thus, our theory predicts that external threats should increase not only people's cooperativeness, but also their sensitivity to each others' behaviour: they should only help if others have also helped.. This prediction goes beyond the standard social psychology claim that group identity increases in response to threat, and could be used to test our theory. It also suggests an alternative explanation for some cases of behaviour that resemble "indirect reciprocity" (Nowak and Sigmund, 1998). Individuals may condition on others' previous behaviour not so as to reward or punish them, but because others' previous play alters the *reputational* value of one's own cooperation. In addition to providing an ultimate explanation for cooperation in conflict, our paper contributes to several other streams of the literature. Signalling explanations of altruism are well-known in theoretical biology (Zahavi, 1975; Gintis, Smith and Bowles, 2001; Lotem, Fishman and Stone, 2003). In these models helping behaviour is a costly signal of individual quality, which benefits the individual helper by (e.g.) making him or her a more attractive partner for reproduction. By contrast, in our story, helping behaviour signals a fact about the group, and benefits the whole group. Theorists have also examined the effect of intergroup conflict on cooperation: Choi and Bowles (2007) show how "parochial altruism" could coevolve with intergroup conflict by providing benefits at group level. We demonstrate that, even without group-level selection, cooperation in defence may be evolutionarily stable.

The model is also of interest to economic theorists interested in reputation-building. Previous work has examined reputation-building in repeated games, either with one patient player against an infinite set of short-run players (e.g., Kreps et al., 1982; Milgrom and Roberts, 1982; Fudenberg and Levine, 1989, 1992, 1994), two players differing in patience (e.g., Schmidt, 1993; Celetani et al., 1996) or

---

<sup>4</sup>In equilibrium, the attacker moves on at once after observing a single episode of helping, so this statement holds for off-path behaviour.

with two patient players (e.g., Cripps and Thomas, 1995; Cripps, Dekel and Pesendorfer, 2005). For example, Kreps et al. (1982) show that the existence of a small proportion of cooperative types allows self-interested actors to cooperate in Prisoner's Dilemma setups. More recently, Tirole (1996) developed a theory of *collective* reputation, constructed as an aggregate of the reputation of individuals overlapping generations (see also, e.g., Bar-Isaac, 2007; Winfree and McCluskey, 2005). Healy (2007) has shown how collective reputation can build up among individuals who are only connected by their shared reputation through anonymous rematching. Here, we develop a model of short-term collective reputation in a dynamic setup that is based on an a-priori correlation of types within groups. Thus, our unique brand of collective reputation relies on group types rather than on aggregate individual types. We follow the standard modelling technique in the literature, by assuming that a (small) proportion of the reputation-building players is a "Stackelberg type" who always plays the action that gives him the long-term best response, assuming the other players best respond. Similarly, our "strong types" play so as to maximize the welfare of their group, and their proportion in the population can be arbitrarily small. A motivation for the early reputation models was to rationalize predatory pricing, in which a market incumbent might take losses so as to deter future entrants. Market entry is also a potential application here. For instance, Section 9 could be interpreted as a cartel facing an entrant and attempting to deter it by collective action.

Our paper is organized as follows. We next discuss the wide-ranging literature on cooperation in conflict. Sections 3-5 introduce our model and describe the equilibrium. The following sections develop some extensions to the basic model. In particular, Section 9 extends the basic logic to public goods games which are played among defenders before the attacker decides to attack. We can thus explain why in-group cooperativeness increases in the face of external threats. The conclusion discusses possibilities for further work.

## 2 Cooperation in conflict

Costly cooperation in an intergroup conflict has been demonstrated under laboratory conditions and in field experiments (Bornstein, 2003; Erev, Bornstein and Galili, 1993). Bornstein and Ben-Yossef (1994) showed in a laboratory experiment that group members' contributions to a public good increased when they were competing with a rival group, even though the competition did not alter the monetary incentives in any relevant way. Tan and Bolle (2007) found that competition without monetary incentive was enough to lead to increased cooperation. It appears that humans naturally respond to intergroup conflict with intragroup cooperation, somewhat mediated by the perception of in-group members as collaborators and the emotional reactions to non-cooperation once conflict is instated (Burton-Chellew, Ross-Gillespie and West, 2010; Puurtinen and Mappes, 2009).

Cooperation in conflict is particularly apparent in civil wars, where state coercion is diminished or non-existent. Admittedly, some people may be coerced into participation by other group members (Hardin, 1997; Kocher and Kalyvas, 2007). However, while we do not underestimate this aspect of the phenomenon, we do not believe that it can be a complete explanation, and in many historical episodes it seems unlikely to have played a large role. For instance, the risks from taking an active part in the French Resistance, or the Provisional IRA during the Troubles, were surely much higher than any risk one's own side might impose for not taking part.<sup>5</sup>

Cooperation in times of conflict extend beyond the conflict effort. Increased participation in pro-social behaviours was documented in Britain during World War II (Schmiedeberg, 1942; Janis, 1951, 1963). Similarly, the September 11 attacks triggered pro-social behaviour in the United States, such as volunteering and charity (Penner et al., 2005; Steinberg and Rooney, 2005) and blood donations (Glynn et al., 2003). There is a well-known "rally round the flag" effect in which expressed support for political incumbents increases after a military or terrorist attack (Baker and Oneal, 2001). Shayo and Zussman (2011) have shown that terrorist attacks in the local region lead Jewish and Arab judges in small-claims courts in Israel to rule in favour of a plaintiff of the same nationality as the judge.

---

<sup>5</sup>Weinstein (2007) provides case studies of insurgencies where material rewards and punishments played a minor role in motivating fighters.



Once more, the effect has been replicated under experimental conditions. In the classic Robbers' Cave experiments, Sherif (1958; 1961) has shown how competition between groups breeds out-group hostility and in-group solidarity. More importantly, an outside threat, common to both groups, facilitated intergroup cooperation and induced positive attitudes towards members of the out-group. Controlled experiments have similarly manipulated external threat to induce cooperation between children (Wright, 1943) and decrease prejudice towards African-American group members (Feshbach and Singer, 1957; Burnstein and McRae, 1962).<sup>6</sup> Hargreaves-Heap and Varoufakis (2002) split participants into two groups and created a situation in which one group suffered discrimination; subsequently, pairs of members of that group cooperated more often in a Prisoner's Dilemma than pairs from the other group.

Sociologists and social psychologists have long been aware of this phenomenon, and have argued that "war with outsiders... makes peace inside" (Sumner, 1906; Campbell, 1965). Social identity theorists explain that individuals' sense of group identity is increased by perceived threats to the group (Stephan and Stephan, 2000). While these theories offer insight, they give only a proximate, not an ultimate explanation. We still do not know how humans might have evolved a psychological mechanism that responds to external threats by increasing group identity (and hence encouraging altruistic behaviour, with associated costs to one's own fitness). Indeed, the same question arises in non-human biology, since some species seem to help unrelated conspecifics against predators: examples include defensive rings, mobbing of predators and alarm calls (Edmunds, 1974). Furthermore, as in humans, intergroup conflict sometimes increases within-group altruistic behaviours (Radford, 2008). Clearly, social identity theory is unlikely to explain these instances of cooperation.

In our theory, the need to deter an attacker can mitigate the within-group collective action problem and thus allow for cooperation in defence by rational, self-interested actors. We believe that this insight can extend the logic of the "security dilemma" (Posen, 1993), in which actors in a conflict are driven to fight because they fear attack from the other side, to collective settings. We also believe that attention to the within-group collective action problem will help to explain group dynamics even in

---

<sup>6</sup>For an extensive review of the classic sociological and psychological literature see Stein (1976).

the absence of overt conflict. For example, if the motivation for cooperation is given by the need to deter potential attackers, then people may be induced to cooperate by manipulating their perception of outside threats; that is, intergroup violence can be used to construct a shared social identity (cf. Fearon and Laitin 2003).

Some of the examples provided above, such as ethnocentrism in court judgments, are hard to explain as rational self-interested behaviour. However, the theory can be viewed either as a direct game-theoretic rationalization of helping behaviour within conflict, or, more indirectly, as explaining the evolution of psychological dispositions to cooperate when threatened by attack. That is, these dispositions may have evolved in strategic situations like those of the model, in which small groups faced opportunist external enemies and needed to deter them. If so, these evolved dispositions might still work the same way in larger and more specialized modern societies (Cosmides and Tooby, 1992).<sup>7</sup> Thus, our theory can be interpreted as an ultimate explanation for the proximate explanations developed by psychologists.

### 3 Model

The “defenders” are a group of size  $N$ , one of a large population of such groups. An attacker makes one or more attacks on a randomly chosen member (the “target”) of the group. Another randomly chosen member of the same group (the “supporter”) may assist the target at a cost  $c$  to its own fitness. The attack costs the defender  $A$  and gives the attacker a benefit of  $A$  if the helper does not help, and costs the defender/benefits the attacker  $a < A$  if the helper helps. We normalize defender welfare at 1 per round. Nothing in the results would change if the benefit to the defender of being helped,  $A - a$ , were decoupled from the cost to the attacker, also currently  $A - a$ ; this assumption is purely to simplify the exposition.

A proportion  $\pi$  of the groups are “strong” types, meaning that their members always help the target;

---

<sup>7</sup>Evolutionary explanations are sometimes accused of being “just-so stories”, i.e. *ex post* rationalizations of existing data. However, our model generates the novel prediction that cooperation under threat should be highly sensitive to other players’ behaviour, so it is not just just-so.

the rest are “normal”. Several different interpretations are possible. Strong types may be altruistic towards one another, perhaps because they are genetically related, while normal types are purely self-interested. Alternatively, strong types may be in long-term relationships, beyond the scope of the attack episode, and able to enforce cooperation by conditioning their future behaviour on play during the attack episode, whereas normal types do not expect to interact after the attack episode. In the animal kingdom, migratory birds may either join communities of sedentary birds who have bred together before, and may be in relationships of long-term reciprocity, or communities of other migrants who are mutually anonymous (Krams and Krama, 2002).

After every attack, the attacker may stay, or may costlessly move to a different group. (We assume that the number of groups is large enough that the chance of returning to the same group later is effectively 0, or alternatively, that the attacker can avoid groups that he has already moved away from.) However the attacker may make no more than  $T$  attacks on any one group.<sup>8</sup> Defenders and attackers share a discount rate  $\delta$ . There are  $N$  defenders in each group. The cost to the supporter of helping,  $c$ , is random and drawn independently in each round from  $\mathbf{C} \subset \mathbb{R}^+$ , with cdf  $\Phi(C) = Pr(c \leq C)$ . We assume  $\Phi$  is continuous. Only the supporter observes  $c$  in each round. For technical reasons, we assume that the cost is sometimes high, specifically:

$$\Phi(\bar{C}) < 1, \text{ where } \bar{C} = \frac{\delta}{1 - \delta} \frac{A}{N}. \quad (1)$$

The defenders and the attacker observe the history of attacks within a given group, and whether the target was helped in each case.

---

<sup>8</sup>We use finite repetitions so as to avoid folk-theorem style results where there are multiple equilibria even if the attacker does not condition on defender behaviour: we want to focus on the stark case where repeated play among defenders alone could not sustain cooperation. This also enables us to find a unique equilibrium.

## 4 Equilibrium analysis<sup>9</sup>

The set of histories of length  $t$  is  $\mathcal{H}^t = \{0, 1\}^t$ , where 1 indicates that the defender was helped, with typical element  $h_t$ . (Write  $\mathcal{H}^0 = \emptyset$ .) The set of all histories is  $\mathcal{H} = \bigcup_{t=0}^T \mathcal{H}^t$ . A strategy for the attacker is  $\zeta : \mathcal{H} \rightarrow [0, 1]$ , giving the probability of playing *stay* after each history. (We will often write  $\zeta(h) \in \{\textit{stay}, \textit{move}\}$  for clarity: i.e., define *stay* = 1 and *move* = 0.) A pure strategy for a normal type defender is  $\sigma : \mathcal{H} \times \mathbf{C} \rightarrow \{0, 1\}$ , giving the probability of helping.<sup>10</sup> (Strong types always help.) The attacker's subjective probability that he is facing a group of strong types is  $\mu : \mathcal{H} \rightarrow [0, 1]$ .

Define  $p_t$  as the  $t$ -length history of 1s, i.e. the  $t$ -length history in which supporters always helped, and let  $p_0 = \emptyset$ . Let  $\mathcal{P} = \{p_0, p_1, p_2, \dots\}$ . We call these "histories of (perfect) helping". We look for the following equilibrium strategies.

- If the defender has always been helped in the past, the attacker moves to a different group. Otherwise, the attacker attacks the same group forever. Thus  $\zeta(h) = \textit{move}$  if  $h \in \mathcal{P}$  and  $\zeta(h) = \textit{stay}$  otherwise.
- Defenders help at round  $t$  (after a history  $h_{t-1}$ ) if and only if (1) all previous defenders have helped (2)  $c$  is less than a finite cutpoint  $C_t$ . Formally,  $\sigma(h_{t-1}, c) = 1$  if  $h_t \in \mathcal{P}$  and  $c \leq C_t$ ;  $\sigma(h_{t-1}, c) = 0$  otherwise.

Notice in particular that the attacker moves after observing a single episode of helping. Because of this, histories  $p_2, p_3, \dots$  are off the equilibrium path. In order to ensure reasonable attacker beliefs at these histories, we use the sequential equilibrium concept (Kreps and Wilson, 1982).

**Proposition 1.** *For  $T$  high enough, the game has a Sequential Equilibrium of the above form (along with appropriate beliefs).*

The remainder of this section gives the proof.

---

<sup>9</sup>Since strong types always help by assumption, the following analysis deals strictly with normal types.

<sup>10</sup>The limitation to pure strategies is innocuous because defenders will only be indifferent between helping and not for a single value of  $c$ . Technically a defender could condition behaviour on his own costs of helping in previous rounds when he was a supporter. Allowing this would not affect our results.

## 4.1 Supporter behaviour

Given the attacker's strategy, and other defenders' strategies, if at round  $t$   $h_t \notin \mathcal{P}$  then a supporter's play does not affect future events in the game (future supporters will never help, and the attacker will always stay). Since  $c > 0$  it is never optimal to help.

If at round  $t$ ,  $h_t \in \mathcal{P}$ , then the supporter's behaviour determines future play. Helping will cause the attacker to move and not helping will cause the attacker to stay and all future supporters not to help.

Thus helping is optimal if

$$1 - c + \sum_{s=1}^{T-t} \delta^s \geq 1 + \sum_{s=1}^{T-t} \delta^s \left(1 - \frac{A}{N}\right)$$

equivalently

$$c \leq C_t = \frac{\delta - \delta^{T-t+1}}{1 - \delta} \frac{A}{N}. \quad (2)$$

$C_t$  is decreasing in  $t$ , and in particular,  $C_T = 0$ . Also, since  $C_t < \frac{\delta}{1-\delta} \frac{A}{N} = \bar{C}$ , there is always positive probability that the supporter does not help.

As  $T \rightarrow \infty$ ,  $C_t$  approaches  $\bar{C} = \frac{\delta}{1-\delta} \frac{A}{N}$  for any finite  $t$ . We can use the expression for  $\bar{C}$  to get a sense of the strength of the motivation to support the target. A useful benchmark is the cost a defender would be prepared to pay to prevent a single attack on him- or herself: this is exactly  $A$ . So, when  $\frac{\delta}{1-\delta} \geq N$ , supporters would bear as high a cost to protect the target as they would to avoid an attack on themselves. For example, in a group of  $N = 100$ , this will hold for  $\delta \approx 0.99$ .

## 4.2 Attacker behaviour

Given these cutpoints, we can calculate the attacker's beliefs. The initial belief  $\mu(\emptyset) = \pi$ . Since only normal types ever fail to help,  $\mu(h_t) = 0$  unless  $h_t \in \mathcal{P}$ .<sup>11</sup>

Write  $V(h_t)$  for the attacker's equilibrium value after a history  $h_t$ , and  $V = V(\emptyset)$ . Also, write

$$V_S(h_t)$$

---

<sup>11</sup>This is shown for beliefs off the path of play in Lemma 6, where the sequential equilibrium refinement is used.

for the attacker's value after  $h_t$  if he stays, and subsequently plays his equilibrium strategy.

Equilibrium strategies give

$$V(h_t) = V_S(h_t) = \sum_{s=0}^{T-t-1} \delta^s A + \delta^{T-t} V, \text{ if } h_t \notin \mathcal{P}. \quad (3)$$

In other words, after observing any non-helping, the attacker stays and receives  $A$  per round until the number of rounds is up.

Otherwise,  $V(h_t) = V$  since the attacker moves (or has just arrived). To show that these are a best response, we can apply the One-Shot Deviation Principle: to check if a strategy is a best response, we need only compare it against deviations involving a single action at one information set.<sup>12</sup> Thus, we need to show that

$$V(h_t) \geq V \text{ if } h_t \notin \mathcal{P}, \quad (4)$$

so that after observing a failure to help, it is optimal for the attacker to stay. This is true by (3) and the fact that  $V \leq \sum_{s=0}^{\infty} \delta^s A$  given that the attacker's maximum per-round payoff is  $A$ . We also need to show that

$$V \geq V_S(h_t) \text{ if } h_t \in \mathcal{P} \quad (5)$$

so that after observing helping it is optimal for the attacker to move rather than to stay. The right hand side here is the counterfactual value from staying for a further attack. This can be calculated as

$$V_S(h_t) = \mu(h_t)[a + \delta V] + (1 - \mu(h_t)) \{ \Phi(C_{t+1})[a + \delta V] + (1 - \Phi(C_{t+1})) [A + \delta V((h_t, 0))] \} \text{ if } h_t \in \mathcal{P}.$$

Here, the first term is the value if one is facing strong types: the supporter helps, so the attacker receives  $a$  and then moves at once. Similarly, if the attacker is facing normal types but the supporter's cost drawn is lower than the cutpoint, then the supporter helps, the attacker receives  $a$  and moves.

---

<sup>12</sup>Hendon, Jacobsen and Sloth (1996) prove the principle for Sequential and Perfect Bayesian Equilibrium.

Finally, if the cost is higher than the cutpoint, the attacker receives  $A$  and the game proceeds. In equilibrium, applying (3),

$$V((h_t, 0)) = V_S((h_t, 0)) = \sum_{s=0}^{T-t-2} \delta^s A + \delta^{T-t-1} V$$

and plugging this into the previous equation gives

$$\begin{aligned} V_S(h_t) = & [\mu(h_t) + (1 - \mu(h_t))\Phi(C_{t+1})][a + \delta V] \cdots \\ & + (1 - \Phi(C_{t+1})) \left[ \sum_{s=0}^{T-t-1} \delta^s A + \delta^{T-t} V \right] \text{ if } h_t \in \mathcal{P}. \end{aligned} \quad (6)$$

We now show that for  $T$  high enough, (5) holds given defender behaviour. First, we show that after enough rounds, it always holds. This is simply because the attacker's subjective probability that he is facing a strong type group becomes increasingly close to certainty after observing enough rounds of cooperation.

**Lemma 1.** *For  $M$  large enough, equation (5) holds for all  $t > M$ .*

*Proof.* First observe that  $V > a + \delta V$  since the attacker's minimum payoff in the first round is  $a$  and since the attacker receives  $A$  with strictly positive probability in equilibrium. Therefore, if  $\mu(h_t)$  is close enough to 1, (6) will be less than  $V$  and (5) will hold.

Next, write  $\mu_t \equiv \mu(p_t)$  for short (we will keep using this notation) and use Bayes' rule to write

$$\mu_t = \frac{\pi}{\pi + (1 - \pi) \prod_{s=1}^t \Phi(C_s)}. \quad (7)$$

Since  $\Phi(C_t) < \Phi(\bar{C}) < 1$ ,  $\mu_t$  is strictly increasing in  $t$  and approaches 1 for large enough  $t$ .<sup>13</sup>  $\square$

The next part of the argument demonstrates the same for early rounds. This relies on choosing  $T$  high enough that  $C_t$  is very close to  $\bar{C}$ . The logic is as follows. Observing a further round of helping has

---

<sup>13</sup>Technically a little more work is necessary to show that only the beliefs of equation (7) are possible in sequential equilibrium. See Lemma 6 in the Appendix.

three effects on the attacker. First, it increases his probability that he is facing a strong type group. This encourages him to move to a different group. Second, the end of the  $T$  rounds is now closer, and third, as a result, the defenders' cutpoint decreases somewhat (i.e.  $C_{t+1} < C_t$ ). These effects may encourage the attacker to stay. However, when  $T$  is large, they become negligible, since the end of the game is far away and (for that reason) the defenders' cutpoint changes very little. Therefore the first effect dominates.

**Lemma 2.** *For any  $M$ , for  $T$  high enough,  $V_S(\emptyset) > V_S(p_1) > \dots > V_S(p_M)$ .<sup>14</sup>*

Combining these Lemmas, along with the fact that  $V_S(\emptyset) = V$ , we can choose  $M$  and  $T$  large enough that  $V \geq V_S(h_t)$  for  $h_t \in \mathcal{P}$ , both for  $t > M$  and for  $t \leq M$  as equation (5) requires. This completes the proof of Proposition 1.

## 5 Uniqueness

Here we investigate whether there are other equilibria. We continue to write  $V$  for the value of the game to the attacker, which is also the attacker's value after choosing *move*. First, we demonstrate that behaviour for  $h_t \notin \mathcal{P}$  is always the same as in the equilibrium above. The argument is essentially by backward induction: after the attacker has become certain he is facing a normal type group, then he cannot be driven off by any further helping, and then cooperation cannot be preserved among the defenders since the game has finite periods.

**Lemma 3.** *Suppose  $\mu(h_t) = 0$ . Then in any equilibrium,  $\zeta(h_t) = \textit{stay}$  and  $\sigma(h_t, c) = 0$  for all  $c$ .*

Sequential equilibrium ensures that  $\mu(h_t) = 0$  for all  $h_t \notin \mathcal{P}$ ,<sup>15</sup> so this Lemma shows that in any equilibrium, when  $h_t \notin \mathcal{P}$ ,  $\sigma(h_t, c) = 0$  for all  $c$  and  $\zeta(h_t) = \textit{stay}$ , just as in the previous section. Therefore, the only source of variation in equilibria must be in different attacker and defender responses to a history of helping  $p_t$ .

---

<sup>14</sup>Proofs not given in the main text are in the Appendix.

<sup>15</sup>See Lemma 6 in the Appendix.



We now show that for  $T$  large enough, there is no equilibrium with  $\zeta(p_t) > 0$  for  $t \geq 1$ . Thus, the equilibrium of the previous section is the unique sequential equilibrium.<sup>16</sup>

The proof works as follows. First, we observe that for  $t$  large enough,  $\zeta(p_t) = \text{move}$  since it becomes increasingly certain that the defenders are strong types. Next, we show that when there are enough rounds, the defenders' cutpoint is higher at the end of a set of periods for which the attacker stays with positive probability even after observing helping, than at the beginning of these periods. The logic is that at the end, one's own action decides whether the attacker will leave or not. At the beginning, on the other hand, the attacker will stay until some future round and will then only leave if all other supporters have also helped. Thus, the incentive to help is greater in the later round. On the other hand, the future history of play which one can affect may be shorter in the later round; but when  $T$  is large enough, this makes little difference.

We then examine the attacker's value at round  $F$ , the last round in which  $\zeta(p_F) > 0$ , and at the last earlier period  $L - 1$  at which  $\zeta(p_{L-1}) = 0$  (or if there is none such, at the beginning of the game). At  $F$  the attacker's belief that he is facing a strong type group is strictly higher, and (as we showed) the normal types' cutpoint is also higher. Combining these facts reveals that, since the attacker is more likely to observe a further round of defense  $V_S(p_{L-1}) > V_S(p_F)$ . By our assumption that at  $L - 1$ , moving is optimal,  $V \geq V_S(p_{L-1})$ . Thus, we arrive at  $V > V(p_F)$ , which contradicts the assumption that staying is optimal at  $p_F$ .

**Proposition 2.** *For  $T$  large enough,  $\zeta(p_t) = \text{move}$  for all  $t \geq 1$ .*

## 6 Evolutionary stability

So far we have used a "rationalist" game theory approach. Given our applications to biology, and the evolutionary tone of our argument in Section 2, it is interesting to ask whether the equilibrium of Section 3 is evolutionarily stable. Technically, it is not an Evolutionarily Stable Strategy, since

---

<sup>16</sup>There may be Weak Perfect Bayesian equilibria with  $\zeta(p_1) = 0$  (i.e. *move*),  $\zeta(p_t) > 0$  for some  $t > 1$ , in which case,  $p_t$  is never reached in equilibrium. However, all Weak Perfect Bayesian equilibria have  $\zeta(p_1) = \text{move}$ .

both defenders and attackers may play differently at histories which are not on the equilibrium path (for example,  $p_t$  for  $t \geq 2$ ), without affecting their welfare. However, for  $T$  large enough, all Weak Perfect Bayesian equilibria satisfy  $\zeta((1)) = \text{move}$  (and  $C_1$  as defined in (2), and  $\zeta(h) = \text{stay}$  and  $\sigma(h, c) = 0, \forall c$ , for  $h \notin \mathcal{P}$ ). It would therefore be surprising if the equilibrium outcome given by these actions were not evolutionarily stable.

Indeed, define  $\mathcal{Q} = \{(0), (0,0), (0,0,0)\dots\}$  as the set of histories in which no defender helps, and define the following sets of strategies:

$$\begin{aligned} Z &= \{\zeta(\cdot) : \zeta((1)) = \text{move}; \zeta(h) = \text{stay for all } h \in \mathcal{Q}\} \\ S &= \{\sigma(\cdot, \cdot) : \sigma(\emptyset, c) = 1 \text{ iff } c \leq C_1; \sigma(h, c) = 0, \forall h \in \mathcal{Q}, \forall c\} \end{aligned}$$

Strategies in these sets result in the same behaviour as our equilibrium, along the path of play. Taking the game's payoff functions as a measure of fitness, we can then show the following:

**Lemma 4.** *For high enough  $T$ : if defenders are playing any  $\sigma \in S$ , then any  $\zeta \in Z$  gives the attacker strictly higher fitness than any  $\zeta' \notin Z$ ; and if the attacker is playing  $\zeta \in Z$  and other defenders are playing  $\hat{\sigma} \in S$ , then any  $\sigma \in S$  gives any defender strictly higher fitness than any  $\sigma' \notin S$ .<sup>17</sup>*

Thus, these strategy sets are evolutionarily stable in the sense that a single mutant defender or a single mutant attacker will be selected against.<sup>18</sup>

*Proof.* (1) Suppose  $\sigma \in \bar{S} = \{\sigma(\cdot, \cdot) : \sigma(\emptyset, c) = 1 \text{ iff } c \leq C_1; \sigma(h, c) = 0, \forall h \notin \mathcal{P}, \forall c\}$ . Then in equilibrium, only the histories  $\{(1)\} \cup \mathcal{Q}$  are observed by the attacker with positive probability. In each of these cases any strategy  $\zeta \in Z$  is strictly optimal. This follows simply from noticing that the argu-

<sup>17</sup>Technically, we require that, after at least one history  $h \in \mathcal{Q} \cup \emptyset$ ,  $\hat{\sigma}(h, c) \neq \sigma(h, c)$  for all  $c \in C$ , a set occurring with positive probability.

<sup>18</sup>We also expect that these sets are stable against simultaneous mutations by defenders and attackers, but showing it would be more complex. The logic is that if a small proportion of attackers becomes more aggressive in staying after a helping episode, then optimal defender cutpoints will be lower; this, however, makes helping a stronger signal that defenders are strong types, and increases the fitness of the less aggressive attackers.

ments in Lemmas 2 and 1 suffice to prove the strict versions of the inequalities in equations (4) and (5).

(2) Suppose  $\zeta \in \bar{Z} = \{\zeta(\cdot), : \zeta((1)) = \text{move}; \zeta(h) = \text{stay} \text{ for all } h \notin \mathcal{P}\}$ , and suppose that all other defenders are playing  $\hat{\sigma} \in \bar{S}$ . Then, in equilibrium, only the histories  $\emptyset \cup \mathcal{Q}$  are observed by a defender with positive probability. Defenders' payoffs from helping are strictly decreasing in cost  $c$ , so the strict optimality of  $\sigma \in S$  is trivial from the definition of  $C_1$ , and from observing that for  $h \in \mathcal{Q}$ , the attacker's and the other defenders' behaviour is unchanged by helping.

(3) The conclusion follows since  $S \subset \bar{S}$  and  $Z \subset \bar{Z}$ . □

## 7 When history is unobserved

Some readers may be concerned that our result is driven by the history-dependent behaviour of other defenders. Since future supporters will cease to help if the current supporter does not help, perhaps this is just a Folk-theorem like result albeit for finite repetitions. To show this is not so, we now assume that defenders cannot condition on others' behaviour. Instead, a normal type strategy is  $\sigma : \{1, \dots, T\} \times \mathbf{C} \rightarrow \{0, 1\}$ , where  $\sigma(t, c)$  gives the probability of helping in each round  $t$ , given a helping cost of  $c$ .

We look for an analogue of the earlier equilibrium, in which the attacker is instantly deterred by a single episode of helping on the equilibrium path.

**Proposition 3.** *If and only if  $\Phi(\frac{\delta}{1-\delta} \frac{A}{N}) < \frac{\sqrt{\pi}-\pi}{1-\pi}$ , then for large enough  $T$  there is an equilibrium of the following form:*

$\zeta(h) = \text{move}$  if and only if  $h \in \mathcal{P}$ .

*Normal defenders help during the first round if and only if  $c$  is less than  $C_1 = \sum_{t=1}^{T-1} \delta^t \frac{A}{N}$ . In subsequent rounds they never help.*

The expression  $\frac{\sqrt{\pi}-\pi}{1-\pi}$  is increasing in  $\pi$  and approaches 0 as  $\pi \rightarrow 0$ . Thus, the model's conclusions are modified somewhat when defenders cannot condition on each others' behavior. Our equilibrium

only exists when the proportion of strong types is non-negligible, compared to the probability of low costs. Also, after the first round, defenders can infer that another defender did not help and therefore cooperation collapses. Nevertheless, this result shows that cooperation does not require defenders to directly observe earlier behaviour.

## 8 Relaxing the assumptions

We now informally discuss some ways in which the model's assumptions could be relaxed. First, we have assumed that strong types always help. This gives cooperation in defense its weakest-link structure: a single episode of not helping is immediate proof that the group is normal type. Nevertheless, this structure will remain, so long as strong types help with probability close enough to 1. For, a single episode of not helping will still provide strong evidence that the group is normal type; for a fixed round  $t$ , if  $T$  is large enough, the attacker will then prefer to stay (as he preferred to stay in the previous round, and now puts a higher probability on facing a normal group). The attacker may still be deterred by observing further rounds of helping, but if this requires more than one round, then the incentive for future supporters to help will be diminished in all but the last of these rounds (since helping does not instantly deter the attacker). Thus, not helping will continue both to alter the attacker's and future supporters' behaviour.

Second, suppose that the attacker faces some cost in moving to a new group (e.g. search costs). The main difference this makes is that  $\pi$  now becomes relevant. In the model, the probability of the existing group being strong type is exactly balanced by the probability that any other group is strong type. Introducing fixed costs of moving would drive a wedge between the values of moving and staying. However, if moving costs are low, a single episode of helping will remain sufficient to deter the attacker, and defender behaviour will be unchanged.

Lastly, we have assumed that defenders are *harmed* but not *killed* by the attack. Killing is more than an extreme loss of fitness; it also alters the strategic structure of future rounds, by removing some actors. In particular allowing defenders to be killed would bring the partner effect into play

(Eshel and Shaked, 2001): each death shrinks the group, and therefore increases the probability that an individual survivor will be targeted in a given round. At large group sizes this effect is negligible (i.e.  $\frac{1}{N} \approx \frac{1}{N-1}$ ), but at smaller group sizes it would strengthen the incentive to help.

## 9 Cooperation before conflict

In the introduction we mentioned the evidence that cooperative and helping behaviour seems to increase when there is an attack, or the threat of an attack, from the outside. We can extend the model to give a natural explanation for this. The setup is kept as simple as possible to focus on the intuition.

Suppose now that the attacker must commit before the game to attacking for all  $T$  periods, or moving. This resembles an irrevocable decision to launch a war. In the period before making his choice, the attacker observes  $K$  randomly selected group members playing a one-shot Prisoner's Dilemma. Each player may cooperate or defect; a player's cooperation gives  $R \in (1/K, 1)$  to each of these  $K$  players, at a cost of  $q$  to the player. The value of  $q$  is common knowledge among defenders, but is not known by the attacker; it is drawn from a distribution with pdf  $\Psi(\cdot)$ , supported on  $(R, 1)$ . After observing play in the Prisoner's Dilemma, the attacker either attacks, or does not, earning a payoff of  $P$ . This could be the expected payoff from attacking a different group, or the payoff from some other activity.

We assume that strong types always cooperate, and, as before, always support each other against attacks.<sup>19</sup> Normal types never help during the attack itself, since the attacker cannot be deterred. We assume

$$\sum_{t=1}^T \delta^t \frac{a}{N} < P < \sum_{t=1}^T \delta^t \frac{A}{N}.$$

---

<sup>19</sup>The Prisoner's Dilemma itself may be the basis for the differentiation between group types. For example, strong types can be engaging in the game repeatedly with the same partners, and condition their cooperation on helping during the attacks as well as on cooperation in previous rounds of the Prisoner's Dilemma. Conversely, normal types often reconstruct new groups with strangers, and therefore have no incentives to cooperate in the absence of an imminent attack. The attacker observes only one period of the repeated game, and therefore cannot distinguish between partner and stranger groups.

The expected loss to each defender from facing an attack is:

$$\sum_{t=1}^T \delta^t \frac{A}{N}.$$

There is always an equilibrium in which normal types do not cooperate. However, there may also be cooperation in equilibrium, for the same signalling reason as before. We seek an equilibrium in which all normal types cooperate if  $q$  is below some level  $\bar{q}$ .

It must be the case that such cooperation (and only such cooperation) deters the attacker. The attacker's belief after observing full cooperation is

$$\mu = \frac{\pi}{\pi + (1 - \pi)\Psi(\bar{q})} \quad (8)$$

and he is deterred if

$$\mu \sum_{t=1}^T \delta^t \frac{a}{N} + (1 - \mu) \sum_{t=1}^T \delta^t \frac{A}{N} \leq P. \quad (9)$$

If he observes any non-cooperation he learns for sure that the defenders are normal types, and attacks (since  $\sum_{t=1}^T \delta^t \frac{A}{N} > P$ ).

Since  $\mu$  in (8) is decreasing in  $\bar{q}$ , (9) provides an upper limit for  $\bar{q}$ . Above this upper limit, cooperation is not convincing enough since too many normal types are doing it. Call this the “attacker deterrence constraint”.

If the attacker is deterred by full cooperation, and  $q \leq \bar{q}$  so that other defenders will cooperate, then it is optimal for each defender to join in cooperating if

$$R - q \geq - \sum_{t=1}^T \delta^t \frac{A}{N},$$

equivalently if

$$q \leq R + \sum_{t=1}^T \delta^t \frac{A}{N}.$$

This provides another upper limit on  $\bar{q}$ . Call it the “reward constraint”, since it requires that the reward

from cooperation be large enough to justify the cost. Of course,  $\bar{q}$  may be lower than these, since no defender will cooperate if, for a given value of  $q$ , he or she expects the others not to cooperate. To sum up, there is a set of equilibria in which normal type defenders cooperate for  $q \leq \bar{q}$  where

$$0 \leq \bar{q} \leq \min\left\{R + \sum_{t=1}^T \delta^t \frac{A}{N}, \hat{q}\right\},$$

where

$$\hat{q} \equiv \Psi^{-1}\left(\frac{\pi}{1-\pi} \left(\frac{\sum_{t=1}^T \delta^t \frac{A-a}{N}}{\sum_{t=1}^T \delta^t \frac{A}{N} - P}\right)\right)$$

is the solution to (8) and (9).

Examining the upper bound for  $\bar{q}$  reveals the following. (1) If only the attacker's deterrence constraint is binding, so that the upper bound is given by  $\hat{q}$ , then it is weakly increasing in  $P$  and  $\pi$ . An increase the value of the outside option, or in the probability the attacker puts on the defenders being strong types, will make him easier to deter. Also, in this case the upper bound is decreasing in  $A^{20}$  and  $a$ : a greater benefit for the attacker from finding either kind of group makes him harder to deter. Finally, the upper bound increases if  $\Psi$  increases (in the sense of first order stochastic dominance): when average costs get higher, then cooperation up to a higher cost level will still persuade the attacker that he is facing a strong group. (2) If only the reward constraint is binding then the upper bound is increasing in  $R$  and  $A$ : cooperation is sustainable at higher levels when it is more efficient in itself, and when the cost of an attack is high.

It is clear that this logic could be extended to many different game forms, including episodes of pairwise cooperation or altruism – any behaviour that correlates with the desire to cooperate in an actual attack.

---

<sup>20</sup>To show this, differentiate  $\hat{q}$ , recalling that  $P > \sum_{t=1}^T \delta^t \frac{a}{N}$ .

## 10 Conclusion

Economists, political scientists and biologists have puzzled over the problem of cooperation in group conflict. This paper demonstrates one possibility: if there is even some small uncertainty regarding the cohesiveness of the group, then a group consisting of selfish unrelated individuals may cooperate against outside attackers so as to deter them by appearing cohesive. The resulting cooperation levels decrease in group size, but can be arbitrarily high if the time horizon of the attack is long enough and defenders are patient enough.

The collaborative efforts that followed the 2011 riots in England can be thus explained by an effort to signal to rioters that they stand to face cooperative resistance from communities. Activists made statements to convey that efforts were collaborative as part of a cohesive community rather than individual charitable helping. A dedicated website set up to coordinate efforts was reported to state that “This is not about the riots. This is about the clean up – Londoners who care, coming together to engender a sense of community” (BBC England, 2011, August 9). In our theoretical framework, both vigilante actions and clean-up efforts can be seen as a way to signal that people in the community are willing to sacrifice in order to help their neighbours, by that reducing the incentives to riot and loot. In line with this reasoning, empathy and helping effort declined once the deterrence effect was achieved.<sup>21</sup>

Our analysis provides an ultimate explanation for the proximate mechanisms identified in the psychological literature for cooperation in conflict. Those proximate mechanisms should generally lead to the rational behavior identified in our model. Therefore, our analysis can be instrumental in directing future empirical research, as it points at the necessary conditions for cooperation to be selfishly beneficial in the long run. Our results heavily rely on a strategic attacker, who can condition his decisions on observed cooperation within the group. Cooperation in conflict is thus predicted to be reduced when these conditions are not satisfied. For example, cooperation should diminish if it is not visible to enemies; cooperation should be higher in the face of threats from other groups than of a natural

---

<sup>21</sup>The website [www.riotcleanup.com](http://www.riotcleanup.com), for example, stopped publishing calls for donations immediately after the riots ended.



threat. Although proximate mechanisms sometimes generalize beyond the context for which they are adapted, our analysis raises interesting new questions that merit empirical investigation, and may lead to new insights regarding human behavior in conflict.

We see scope for further theoretical work in the following areas. First, can the uniqueness result be generalized to a wider class of games with group reputation? Second, extending the model to multiple groups, and/or differentiating between leaders and followers within groups, would help us to understand how leaders can manipulate followers' willingness to take part in group conflict. Lastly, in our theory, defensive cooperation is due to group members' expectations of further attacks. In the model, groups are exogenously given. However, a group might also be defined by the attacker's (perhaps arbitrary) choice of targets. This could provide a model of "violence and the social construction of identity" (Fearon and Laitin, 2003).

## References

- Baker, W. D. and J. R. Oneal. 2001. "Patriotism or Opinion Leadership?: The Nature and Origins of the "Rally Round the Flag" Effect." *Journal of Conflict Resolution* 45(5):661–687.
- Bar-Isaac, H. 2007. "Something to prove: reputation in teams." *The RAND Journal of Economics* 38(2):495–511.
- BBC England. 2011, August 9. England riots: Twitter and Facebook users plan clean-up. url: <http://www.bbc.co.uk/news/uk-england-london-14456857>, retrieved on August 27, 2011.
- Beaumont, P., J. Coleman and S. Laville. 2011, August 10. London riots: 'People are fighting back. It's their neighbourhoods at stake'. url: <http://www.guardian.co.uk/uk/2011/aug/09/london-riots-fighting-neighbourhoods>, retrieved on August 27, 2011.
- Blattman, C. and E. Miguel. 2008. "Civil war." *forthcoming in Journal of Economic Literature* .
- Bornstein, G. 2003. "Intergroup conflict: Individual, group, and collective interests." *Personality and Social Psychology Review* 7(2):129–145.
- Bornstein, G. and M. Ben-Yossef. 1994. "Cooperation in inter-group and single-group social dilemmas." *Journal of Experimental Social Psychology* 30:52–52.
- Burnstein, E. and A.V. McRae. 1962. "Some effects of shared threat and prejudice in racially mixed groups." *Journal of Abnormal and Social Psychology* 64(4):257–263.
- Burton-Chellew, M.N., A. Ross-Gillespie and S.A. West. 2010. "Cooperation in humans: competition between groups and proximate emotions." *Evolution and Human behavior* 31(2):104–108.
- Campbell, D. T. 1965. Ethnocentric and other altruistic motives. In *Nebraska symposium on motivation*. Vol. 13 pp. 283–311.
- Celetani, M., D. Fudenberg, D.K. Levine and W. Pesendorfer. 1996. "Maintaining a reputation against a long-lived opponent." *Econometrica* 64(3):691–704.

- Choi, J. K and S. Bowles. 2007. "The coevolution of parochial altruism and war." *Science* 318(5850):636.
- Cosmides, L. and J. Tooby. 1992. Cognitive adaptations for social exchange. In *The adapted mind: Evolutionary psychology and the generation of culture*, ed. L. Cosmides, J. Tooby and J. H Barkow. Vol. 163 Oxford: Oxford University Press p. 228.
- Cripps, M.W., E. Dekel and W. Pesendorfer. 2005. "Reputation with equal discounting in repeated games with strictly conflicting interests." *Journal of Economic Theory* 121(2):259–272.
- Cripps, M.W. and J.P. Thomas. 1995. "Reputation and commitment in two-person repeated games without discounting." *Econometrica* 63(6):1401–1419.
- Davies, L., A. Topping, J. Ball and I. Sample. 2011, August 9. London riots: hundreds answer appeal to clean up streets. url: <http://www.guardian.co.uk/uk/2011/aug/09/london-riots-cleanup-appeal>, retrieved on August 27, 2011.
- Edmunds, M. 1974. *Defence in animals: a survey of anti-predator defences*. Longman Harlow.
- Erev, I., G. Bornstein and R. Galili. 1993. "Constructive intergroup competition as a solution to the free rider problem: A field experiment." *Journal of Experimental Social Psychology* 29(6):463 – 478.
- Eshel, I. and A. Shaked. 2001. "Partnership." *Journal of Theoretical Biology* 208(4):457–474.
- Fearon, J. D. and D. D. Laitin. 2003. "Violence and the social construction of ethnic identity." *International Organization* 54(04):845–877.
- Feshbach, S. and R. Singer. 1957. "The effects of personal and shared threats upon social prejudice." *Journal of Abnormal and Social Psychology* 54(3):411–416.
- Fudenberg, D. and D. Levine. 1994. "Efficiency and observability with long-run and short-run players." *Journal of Economic Theory* 62(1):103–135.

- Fudenberg, D. and D.K. Levine. 1989. "Reputation and equilibrium selection in games with a patient player." *Econometrica: Journal of the Econometric Society* pp. 759–778.
- Fudenberg, D. and D.K. Levine. 1992. "Maintaining a reputation when strategies are imperfectly observed." *The Review of Economic Studies* 59(3):561.
- Fudenberg, D. and E. Maskin. 1986. "The folk theorem in repeated games with discounting or with incomplete information." *Econometrica: Journal of the Econometric Society* pp. 533–554.
- Garay, J. 2008. "Cooperation in defence against a predator." *Journal of Theoretical Biology* .
- Gintis, H., E. A. Smith and S. Bowles. 2001. "Costly signaling and cooperation." *Journal of Theoretical Biology* 213(1):103–119.
- Glynn, Simone A., Michael P. Busch, George B. Schreiber, Edward L. Murphy, David J. Wright, Yongling Tu and Steven H. Kleinman. 2003. "Effect of a National Disaster on Blood Supply and Safety: The September 11 Experience." *JAMA* 289(17):2246–2253.  
**URL:** <http://jama.ama-assn.org/cgi/content/abstract/289/17/2246>
- Hardin, R. 1997. *One for all: The logic of group conflict*. Princeton Univ Pr.
- Hargreaves-Heap, S. and Y. Varoufakis. 2002. "Some experimental evidence on the evolution of discrimination, co-operation and perceptions of fairness." *Economic Journal* pp. 679–703.
- Healy, P. J. 2007. "Group reputations, stereotypes, and cooperation in a repeated labor market." *The American Economic Review* pp. 1751–1773.
- Hendon, Ebbe, Hans J?rgen Jacobsen and Birgitte Sloth. 1996. "The One-Shot-Deviation Principle for Sequential Rationality." *Games and Economic Behavior* 12(2):274–282.  
**URL:** <http://www.sciencedirect.com/science/article/B6WFW-45V7FNN-7/2/6f9d051e8e6e8968539ce284d3c7ad5d>
- Horowitz, D. L. 2001. *The deadly ethnic riot*. University of California Press.

- Janis, I.L. 1963. "Group identification under conditions of external danger." *The British journal of medical psychology* 36:227–238.
- Janis, M. 1951. *Air war and emotional stress: Psychological studies of bombing and civilian defense*. New York: McGraw-Hill.
- Kocher, M. A and S. N Kalyvas. 2007. "How "Free" Is Free Riding in Civil Wars?: Violence, Insurgency, and the Collective Action Problem." *World Politics* 59(2):177–216.
- Krams, I. and T. Krama. 2002. "Interspecific reciprocity explains mobbing behaviour of the breeding chaffinches, *Fringilla coelebs*." *Proceedings of the Royal Society of London. Series B: Biological Sciences* 269(1507):2345.
- Kreps, D. M., P. Milgrom, J. Roberts and R. Wilson. 1982. "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma." *Journal of Economic Theory* .
- Kreps, D.M. and R. Wilson. 1982. "Sequential equilibria." *Econometrica: Journal of the Econometric Society* pp. 863–894.
- Lotem, A., M.A. Fishman and L. Stone. 2003. "From Reciprocity to Unconditional Altruism through Signalling Benefits." *Proceedings: Biological Sciences* 270(1511):199–205.
- Milgrom, P. and J. Roberts. 1982. "Predation, reputation, and entry deterrence\* 1." *Journal of economic theory* 27(2):280–312.
- Neyman, A. 1985. "Bounded complexity justifies cooperation in the finitely repeated prisoners' dilemma." *Economics Letters* 19(3):227–229.
- Nowak, M. A and K. Sigmund. 1998. "Evolution of indirect reciprocity by image scoring." *Nature* 393(6685):573–577.
- Olson, M. 1974. *The logic of collective action: Public goods and the theory of groups*. Harvard University Press.

- Penner, L., M.T. Brannick, S. Webb and P. Connell. 2005. "Effects on Volunteering of the September 11, 2001, Attacks: An Archival Analysis." *Journal of Applied Social Psychology* 35(7):1333–1360.
- Posen, B. R. 1993. "The security dilemma and ethnic conflict." *Survival* 35(1):27–47.
- Puurtinen, M. and T. Mappes. 2009. "Between-group competition and human cooperation." *Proceedings of the Royal Society B: Biological Sciences* 276(1655):355–360.
- Radford, Andrew N. 2008. "Type of threat influences postconflict allopreening in a social bird." *Current Biology* 18(3):R114–R115.  
**URL:** <http://www.sciencedirect.com/science/article/pii/S0960982207024220>
- Schmidt, K.M. 1993. "Reputation and equilibrium characterization in repeated games with conflicting interests." *Econometrica* 61(2):325–351.
- Schmiedeberg, M. 1942. "Some observations on individual reactions to air raids." *International Journal of Psycho-analysis* 23:146–175.
- Shayo, M. and A. Zussman. 2011. "Judicial Ingroup Bias in the Shadow of Terrorism." *Quarterly Journal of Economics* .
- Sherif, M. 1958. "Superordinate goals in the reduction of intergroup conflict." *American Journal of Sociology* 63(4):349–356.
- Sherif, M. 1961. *Intergroup conflict and cooperation: The Robbers Cave experiment*. University Book Exchange Norman, OK.
- Stein, A.A. 1976. "Conflict and cohesion." *Journal of Conflict Resolution* 20(1):143–172.
- Steinberg, K.S. and P.M. Rooney. 2005. "America gives: A survey of Americans' generosity after September 11." *Nonprofit and voluntary sector quarterly* 34(1):110–135.
- Stephan, W. G. and C. W. Stephan. 2000. An Integrated Threat Theory of Prejudice. In *Reducing prejudice and discrimination: The Claremont symposium on applied social psychology*. Lawrence Erlbaum p. 23.

- Sumner, W. G. 1906. *Folkways: A study of the sociological importance of usages, manners, customs, mores, and morals*. Ginn.
- Tan, J.H.W. and F. Bolle. 2007. "Team competition and the public goods game." *Economics Letters* 96(1):133–139.
- Tirole, J. 1996. "A theory of collective reputations (with applications to the persistence of corruption and to firm quality)." *The Review of Economic Studies* 63(1):1.
- Trivers, R. L. 1971. "The Evolution of Reciprocal Altruism." *The Quarterly Review of Biology* 46(1):35–57.
- Weinstein, J. M. 2007. *Inside rebellion: The politics of insurgent violence*. Cambridge Univ Pr.
- Winfree, J.A. and J.J. McCluskey. 2005. "Collective reputation and quality." *American Journal of Agricultural Economics* 87(1):206–213.
- Wright, M.E. 1943. "The influence of frustration upon the social relations of young children." *Journal of Personality* 12(2):111–122.
- Zahavi, A. 1975. "Mate selection—a selection for a handicap." *Journal of theoretical Biology* 53(1):205–214.

# Appendix

## Proof of Lemma 2

*Proof.* Rewrite (6) as

$$V_S(p_t) = [\mu_t + (1 - \mu_t)\Phi(C_{t+1})][a + \delta V] + [(1 - \mu_t)(1 - \Phi(C_{t+1}))][\sum_{s=0}^{T-t-1} \delta^s A + \delta^{T-t} V].$$

Now,  $\sum_{s=0}^{T-t-1} \delta^s A + \delta^{T-t} V$  is strictly decreasing in  $t$  and is greater than  $a + \delta V$ . Therefore, to show the above is strictly decreasing in  $t$ , it will suffice if

$$(1 - \mu_t)(1 - \Phi(C_{t+1})) \tag{10}$$

is decreasing in  $t$ . Rewrite this expression, using the definition of  $\mu(h_t)$  in (7), as

$$\left(1 - \frac{\pi}{\pi + (1 - \pi) \prod_{s=1}^t \Phi(C_s)}\right) (1 - \Phi(C_{t+1})).$$

Observe from the definition of  $C_t$  in (2) that, for any  $t$ ,  $C_t \rightarrow \bar{C}$  as  $T \rightarrow \infty$ . Since  $\Phi$  is continuous, the above expression approaches

$$(1 - \bar{\mu}_t)(1 - \Phi(\bar{C})) \text{ where } \bar{\mu}_t \equiv \frac{\pi}{\pi + (1 - \pi)\Phi(\bar{C})^t} \tag{11}$$

as  $T \rightarrow \infty$ . This expression is strictly decreasing in  $t$ , since  $\bar{\mu}_t$  is strictly increasing in  $t$ . Define  $\varepsilon = \min_{t \in \{0, \dots, M-1\}} (1 - \bar{\mu}_{t+1})(1 - \Phi(\bar{C})) - (1 - \bar{\mu}_t)(1 - \Phi(\bar{C}))$  and note that  $\varepsilon > 0$ . Now, by selecting  $T$  large enough, we can ensure that

$$\left| \frac{(1 - \pi) \prod_{s=1}^t \Phi(C_s)}{\pi + (1 - \pi) \prod_{s=1}^t \Phi(C_s)} (1 - \Phi(C_{t+1})) - (1 - \bar{\mu}_t)(1 - \Phi(\bar{C})) \right| < \frac{\varepsilon}{2} \text{ for all } t,$$

and this, combined with our definition of  $\varepsilon$ , ensures that (10) is decreasing.  $\square$

**Lemma 5.** *In any equilibrium, after any history  $h_t$ , normal types do not help with probability of at*



least  $1 - \Phi(\bar{C}) > 0$ .

*Proof.* Normal types help if

$$1 - c + \delta W \geq 1 + \delta W'$$

where  $W$  and  $W'$  are continuation values from helping and not helping respectively. These are bounded below by  $\sum_{s=0}^{T-t} \delta^s (1 - \frac{A}{N})$  and above by  $\sum_{s=0}^{T-t} \delta^s$ . The above bound is reached if the attacker leaves; the lower bound holds because the defender can achieve at least this payoff by never helping. The maximum difference between  $\delta W$  and  $\delta W'$  is thus  $\delta \sum_{s=0}^{T-t-1} \delta^s \frac{A}{N} = \frac{\delta - \delta^{T-t+1}}{1-\delta} \frac{A}{N} < \bar{C}$ ; so for  $c \geq \bar{C}$  the inequality above will not be satisfied.  $\square$

**Lemma 6.** *In any sequential equilibrium, beliefs  $\mu(p_t)$  must be as given in equation (7), while  $\mu(h_t) = 0$  for  $h_t \notin \mathcal{P}$ .*

*Proof.* First, observe that in any equilibrium, defender play  $\sigma(p_t, c)$  can be characterized by a (perhaps infinite) cutpoint  $C_t$ , because if  $\sigma(p_t, c) = 1$  is optimal, then  $\sigma(p_t, c')$  must be strictly optimal for  $c' < c$ . Since  $p_t$  may be off the equilibrium path of play, permissible beliefs must be derived by constructing a sequence of equilibria of perturbed games in which (1) defenders' probability of helping at  $h_t$ ,  $\sigma_n(h_t, c)$  is bounded within a subinterval of  $(0, 1)$ , with the interval approaching  $[0, 1]$  as  $n \rightarrow \infty$ , for all  $h_t$  and  $c$ ; (2)  $\sigma_n(h_t, c) \rightarrow \sigma(h_t, c)$  as  $n \rightarrow \infty$  (to avoid complications we assume that this convergence is uniform across all  $c$ ) and (3) attacker's probability of leaving or staying is similarly bounded between 0 and 1 and converges to 0 or 1 according to  $\zeta(h_t) \in \{stay, move\}$ . We also assume that normal defenders help with probability  $1 - \eta_n(h_t, c) \rightarrow 1$  as  $n \rightarrow \infty$ . We then apply Bayes' rule to give the attacker's beliefs. For  $p_t$ , this results in

$$\mu_n(p_t) = \frac{\pi \prod_{s=1}^t \{\int (1 - \eta_n(p_s, c)) d\Phi(c)\}}{\pi \prod_{s=1}^t \{\int (1 - \eta_n(p_s, c)) d\Phi(c)\} + (1 - \pi) \prod_{s=1}^t \{\int \sigma_n(p_s, c) d\Phi(c)\}}.$$

As  $n \rightarrow \infty$  we arrive at the limit

$$\mu(p_t) = \frac{\pi}{\pi + (1 - \pi) \prod_{s=1}^t \{\int \sigma_n(p_s, c) d\Phi(c)\}}$$

and in the equilibrium of Section 3, since  $\sigma_n(p_s, c) \rightarrow 1$  for  $c \leq C_s$ ,  $\sigma_n(p_s, c) \rightarrow 0$  otherwise, this must reduce to

$$\mu(p_t) = \frac{\pi}{\pi + (1 - \pi) \prod_{s=1}^t \Phi(C_s)}$$

as in (7).

For  $h_t \notin \mathcal{P}$ , in any equilibrium, write  $h_t = (r_1, r_2, \dots, r_t)$ , with  $r_s \in \{0, 1\}$  for  $s \in \{1, \dots, t\}$ . Bayes' rule gives

$$\mu_n(h_t) = \frac{\pi \prod_{s=1}^t \{r_s \int (1 - \eta_n(h_s, c)) d\Phi(c) + (1 - r_s) \int \eta_n(h_s, c) d\Phi(c)\}}{D}$$

with

$$D = \pi \prod_{s=1}^t \left\{ r_s \int (1 - \eta_n(h_s, c)) d\Phi(c) + (1 - r_s) \int \eta_n(h_s, c) d\Phi(c) \right\} \\ + (1 - \pi) \prod_{s=1}^t \left\{ r_s \int \sigma_n(p_s, c) d\Phi(c) + (1 - r_s) \int (1 - \sigma_n(p_s, c)) d\Phi(c) \right\}.$$

Since  $r_s = 0$  for at least one  $s$ , the numerator of the above expression goes to 0 as  $n \rightarrow \infty$ , and the denominator  $D$  remains bounded above 0 since, after any history, normal types sometimes fail to help (Lemma 5). Thus  $\mu(h_t) = 0$ .  $\square$

**Lemma 7.** *Suppose that  $\zeta((h_t, 0, h_+)) = \zeta((h_t, 1, h_+))$  for all continuation histories  $h_+$  of length 0 or more. Then in any equilibrium,  $\sigma(h_t, c) = 0$  for all  $c$ .*

*Proof.* We prove by backwards induction over the  $T$  periods. First, in a final period history  $h_{T-1}$ ,  $\sigma(h_{T-1}, c) = 0$  for all  $c$ , since supporter behaviour cannot affect future play. Next, at  $T - 2$ ,  $\sigma(h_{T-2}, c) = 0$  for all  $c$ , since the supporter cannot affect either future supporter play (as we have just shown) or the attacker's future play (by assumption). Then at  $T - 3$ ,  $\sigma(h_{T-3}, c) = 0$  for all  $c$  for the same reason, and so on.  $\square$

### Proof of Lemma 3

*Proof.* Again, start at the end. Since  $\mu(h_{T-1}) = 0$ , the attacker is certain that the defenders are normal types, and since  $\sigma(h_{T-1}, c) = 0$  for all  $c$ , the attacker will gain his maximum per-round payoff of  $A$  next round by staying, giving a continuation value of  $A + \delta V > V$  (since there is positive probability of receiving  $a$  in the first round,  $V < A/(1 - \delta)$ ). Thus  $\zeta(h_{T-1}) = \text{stay}$  is strictly optimal.

Now consider  $\zeta(h_{T-2})$ . Since  $\mu(h_{T-2}) = 0$ , the attacker's belief will stay at 0 for any continuation history. Thus,  $\zeta((h_{T-2}, 0)) = \zeta((h_{T-2}, 1)) = \text{stay}$  as we have just shown. Therefore, the assumption of Lemma 7 holds for histories of length  $T - 2$ . Applying Lemma 7, we conclude that  $\sigma(h_{T-2}, c) = 0$  for all  $c$ . Therefore  $\zeta(h_{T-2}) = \text{stay}$ . For, given that  $\sigma(h_{T-2}, c) = \sigma((h_{T-2}, 0), c) = \sigma((h_{T-2}, 1), c) = 0$  for all  $c$ , and that  $\mu(h_{T-2}) = 0$ , the continuation value for staying is  $A + \delta A + \delta^2 V > V$ . We have now proved the conclusion of the Lemma for histories of length  $T - 2$ .

At  $h_{T-3}$ , if  $\zeta(h_{T-3}) = \text{stay}$  then the previous paragraph shows that  $\zeta((h_{T-3}, h_+)) = \text{stay}$  for any positive-length continuation history  $h_+$ . Again this allows us to apply Lemma 7 and shows that  $\sigma(h_{T-3}, c) = 0$  for any  $c$ , and again this shows that  $\zeta(h_{T-3}) = \text{stay}$ . This plus the previous paragraph proves the conclusion of the Lemma for histories of length  $T - 3$ . Continuing thus, we prove it for histories of any length.  $\square$

**Lemma 8.** *There is some  $\bar{t}$  such that in any equilibrium for a game of any length  $T$ ,  $\zeta(p_t) = \text{move}$  for all  $t \geq \bar{t}$ .*

*Proof.* Applying (7), Lemma 5 shows that in any equilibrium  $\mu(p_t)$  is strictly increasing in  $t$ , and so approaches 1. Furthermore, in any equilibrium, since the probability of helping is no more than  $\Phi(\bar{C})$ ,  $\mu(p_t) \geq \bar{\mu}_t$  as defined in (11). Therefore, the set of beliefs  $\mu(p_t)$ , defined over all equilibria, approaches 1 *uniformly* as  $t \rightarrow \infty$ : for any  $\varepsilon > 0$ , there is some  $\bar{t}_\varepsilon$  such that  $\mu(p_{\bar{t}_\varepsilon}) \geq \bar{\mu}_{\bar{t}_\varepsilon} > 1 - \varepsilon$  in any equilibrium.

Now, the value to the attacker of staying in equilibrium can be written

$$V_S(p_t) = \mu(p_t)[a + \delta V'] + (1 - \mu(p_t))V'' \quad (12)$$

where  $V'$  is the continuation value conditional on the defenders being strong types, and  $V''$  is the value if the defenders are normal types. Since strong types always help, the best response when faced with them is to leave; therefore  $a + \delta V' \leq a + \delta V$ . Furthermore,

$$V \geq (\pi + (1 - \pi)\Phi(\bar{C}))a + (1 - \pi)(1 - \Phi(\bar{C}))A + \delta V = a + \delta V + (1 - \pi)(1 - \Phi(\bar{C}))(A - a),$$

since (1) the probability of normal types helping is no more than  $\Phi(\bar{C})$ , and (2) the attacker can achieve at least the payoff on the RHS, by leaving after the first round. Therefore, in any equilibrium,  $a + \delta V' \leq V - \varepsilon_2$  where  $\varepsilon_2 = (1 - \pi)(1 - \Phi(\bar{C}))(A - a)$ . Plugging this into (12), and using the fact that  $V''$  is bounded above by  $\sum_{s=0}^{\infty} \delta^s A$ , gives for any  $\varepsilon$  some  $\bar{t}_\varepsilon$  such that

$$\begin{aligned} V_S(p_{\bar{t}_\varepsilon}) &\leq (1 - \varepsilon)(V - \varepsilon_2) + \varepsilon \sum_{s=0}^{\infty} \delta^s A \\ &\leq V - (1 - \varepsilon)\varepsilon_2 + \varepsilon \sum_{s=0}^{\infty} \delta^s A \end{aligned}$$

Choosing  $\varepsilon$  so that the right hand side is strictly less than  $V$  for any equilibrium value of  $V$ , we can set  $\bar{t} = \bar{t}_\varepsilon$ . Then, it is sequentially rational to leave after  $p_{\bar{t}}$ , so  $\zeta(p_{\bar{t}}) = \textit{leave}$ .  $\square$

## Proof of Proposition 2

*Proof.* Suppose false, so that  $\zeta(p_t) > 0$  for some  $t > 0$ . If  $T \geq \bar{t}$ ,  $\zeta(p_t) = 0$  (i.e. *leave*) for  $t$  high enough, as Lemma 8 shows. So, for  $T$  large enough we may take  $F$  such that  $\zeta(p_F) > 0$ , but  $\zeta(p_{F+1}) = 0$ . Now, define  $L = \min\{t \geq 1 : \zeta(p_{t'}) > 0 \text{ for all } t \leq t' \leq F\}$ . Observe that if  $\zeta(p_t) = 0$  for all  $t < F$ , then  $L = F$ ; if  $\zeta(p_t) > 0$  for all  $t < F$ , then  $L = 1$ .

First we show that  $C_L < C_{F+1}$ . After  $p_F$ , the attacker will condition on the next round, staying until  $T$  if he observes no helping and leaving otherwise. Thus,

$$C_{F+1} = \frac{\delta - \delta^{T-F} A}{1 - \delta} \frac{A}{N},$$

just as in (2). Observe that for any  $T, F < \bar{t}$ , by Lemma 8. Therefore as  $T$  becomes large,

$$C_{F+1} \rightarrow \bar{C} = \sum_{t=1}^{\infty} \delta^t \frac{A}{N}. \quad (13)$$

Now examine the supporter's problem in round  $L$ . The benefit of not helping is

$$1 + \sum_{t=L+1}^T \delta^{t-L} \left[ 1 - \frac{A}{N} \right]. \quad (14)$$

The benefit of helping is

$$1 - c + \sum_{t=L+1}^F \delta^{t-L} \left[ 1 - \text{Nohelp}_t \frac{A}{N} - \text{Attack}_t \left\{ \frac{1}{N} \int_0^{C_t} \hat{c} d\Phi(\hat{c}) + \frac{1}{N} [\Phi(C_t)a + (1 - \Phi(C_t))A] \right\} \right] + \sum_{t=F+1}^T \delta^{t-L} 1 - \left[ N \right] \quad (15)$$

where  $\text{Nohelp}_t$  gives the probability that at least one defender failed to help between rounds  $L+1$  and  $t-1$ , and  $\text{Attack}_t$  gives the probability that the attacker is still present at time  $t$  even though all defenders helped. That is, until round  $F$ , the attacker may still be present even after observing helping. If so, the defender bears the expected cost in curly brackets, which includes the expected cost of being a supporter and helping if  $c \leq C_t$ , and the expected cost of being attacked and perhaps helped. From round  $F+1$  onwards, either the attacker has observed perfect helping and left, or  $h \notin \mathcal{P}$ , the attacker is staying forever and no defenders help.

We can calculate  $\text{Attack}_t$  as

$$\prod_{s=L+1}^{t-1} \Phi(C_s) \zeta(p_s)$$

which is positive by definition of  $L$ , and  $\text{Nohelp}_t$ , recursively, as

$$\text{Nohelp}_{p_{t-1}} + (1 - \text{Nohelp}_{p_{t-1}}) \zeta(p_{t-2}) (1 - \Phi(C_{t-1}))$$

with  $\text{Nohelp}_{p_{L+1}} = 0$  since by assumption the current supporter helped. I.e. even if every supporter helped up till  $t-2$ , if the attacker continued to stay then at  $t-1$  the supporter may have failed to help.

All that matters is that both  $Attack_t$  and  $Nohelp_t$  are positive, since  $\zeta(p_t)$  is positive for  $L \leq t \leq F$ .

Rearranging (15) and (14), and taking  $T \rightarrow \infty$ , gives

$$C_L \xrightarrow{T \rightarrow \infty} \sum_{t=L+1}^F \delta^{t-L} \left[ (1 - Nohelp_t) \frac{A}{N} - Attack_t \left\{ \frac{1}{N} \int_0^{C_t} \hat{c} d\Phi(\hat{c}) + \frac{1}{N} [\Phi(C_t)a + (1 - \Phi(C_t))A] \right\} \right] + \sum_{t=F+1}^{\infty} \delta^{t-L} (1 - Nohelp_{p_{F+1}})$$

Comparing this with 13 shows  $C_L < C_{F+1}$ , since each term of the above sum is less than  $\frac{A}{N}$ .

Now,

$$V_S(p_{L-1}) = [\mu_{L-1} + (1 - \mu_{L-1})\Phi(C_L)](a + \delta V(p_L)) + (1 - \mu_{L-1})(1 - \Phi(C_L))(A + \delta A + \dots + \delta^{T-L}A + \delta^{T-L+1}V)$$

where the first term in brackets gives the probability of the supporter helping, and  $V(p_L)$  is the value after  $p_L$ . Observe that

$$a + \delta V(p_L) < A + \delta A + \dots + \delta^{T-L}A + \delta^{T-L+1}V$$

since  $V(p_L)$  involves a sequence of no more than  $T - L$  attacks which can give no more than  $A$ , followed by  $V$ , and since  $V < A + \delta V$  implies  $V < A + \delta A + \dots + \delta^{t-1}A + \delta^t V$  for any  $t \geq 1$ . Therefore we can write

$$\begin{aligned} V_S(p_{L-1}) &> [\mu_F + (1 - \mu_F)\Phi(C_{F+1})](a + \delta V(p_L)) + (1 - \mu_F)(1 - \Phi(C_{F+1}))(A + \delta A + \dots + \delta^{T-L}A + \delta^{T-L+1}V) \\ &\quad (\text{by } \mu_F > \mu_{L-1} \text{ and } C_L < C_{F+1}, \text{ and } a + \delta V(p_L) < A + \delta A + \dots + \delta^{T-L}A + \delta^{T-L+1}V) \\ &> [\mu_F + (1 - \mu_F)\Phi(C_{F+1})](a + \delta V) + (1 - \mu_F)(1 - \Phi(C_{F+1}))(A + \delta A + \dots + \delta^{T-F-1}A + \delta^{T-F}V) \\ &\quad (\text{since } V(p_L) \geq V, \text{ as must always hold given that leaving is an option,} \\ &\quad \text{and } V < A + \delta V \Rightarrow \delta^{T-F}V < \delta^{T-F}A + \delta^{T-F+1}A + \dots + \delta^{T-L}A + \delta^{T-L+1}V) \\ &= V(p_F). \end{aligned}$$

But since, by definition of  $L$ , either  $\zeta(p_{L-1}) = 0$ , or  $V_S(p_{L-1}) = V$  if  $L = 1$ , it must be that  $V \geq V_S(p_{L-1})$ . We therefore arrive at  $V > V(p_F)$  which contradicts  $\zeta(p_F) > 0$ .  $\square$

### Proof of Proposition 3

*Proof.* First consider defender behaviour. Since  $\zeta((1)) = \text{move}$ , if  $t \geq 2$  then the attacker must have observed not helping and will stay forever. Therefore it is not optimal to bear any cost to help. Now suppose that  $t = 1$ . Helping gives expected welfare of

$$1 - c + \sum_{t=1}^{T-1} \delta^t$$

and not helping gives

$$1 + \sum_{t=1}^{T-1} \delta^t (1 - A/N)$$

giving a cutpoint

$$C_1 = \sum_{t=1}^{T-1} \delta^t \frac{A}{N}.$$

Next consider attacker behavior. Write  $p_t = (1, 1, \dots, 1)$  for a  $t$ -length history of helping, so that  $p_t \in \mathcal{P}$ . Clearly since only related helpers help in the second and subsequent periods,  $v(p_t) = \text{move}$  is optimal for  $t \geq 2$ . The interesting question is  $\zeta(p_1)$ , the optimal strategy after a single episode of helping. The benefit of attacking is

$$\mu_1(a + \delta V) + (1 - \mu_1) \left( \sum_{t=0}^{T-2} \delta^t A + \delta^{T-1} V \right)$$

with

$$\mu_1 = \frac{\pi}{\pi + (1 - \pi)\Phi(C_1)}$$

while the benefit of moving is

$$V = [\pi + (1 - \pi)\Phi(C_1)](a + \delta V) + (1 - \pi)(1 - \Phi(C_1)) \left( \sum_{t=0}^{T-1} \delta^t A + \delta^T V \right)$$

We wish to show conditions when the benefit of moving is greater than that of attacking. Taking  $T$  to infinity, the relevant inequality becomes

$$\mu_1(a + \delta V) + (1 - \mu_1) \sum_{t=0}^{\infty} \delta^t A \leq [\pi + (1 - \pi)\Phi(C_1)](a + \delta V) + (1 - \pi)(1 - \Phi(C_1)) \sum_{t=0}^{\infty} \delta^t A.$$

Since  $a + \delta V < \sum_{t=0}^{\infty} \delta^t A$ , this will hold in the limit whenever  $\mu_1 > \pi + (1 - \pi)\Phi(C_1)$ , equivalently when

$$\frac{\pi}{\pi + (1 - \pi)\Phi(C_1)} > \pi + (1 - \pi)\Phi(C_1).$$

This results in a quadratic, but we can observe at once that it holds for  $\Phi(C_1) \rightarrow 0$ , does not hold for  $\Phi(C_1) \rightarrow 1$ , and has a single crossover point in terms of  $\Phi(C_1)$ . Intuitively, when  $\Phi(C_1)$  is small enough, the fact that the supporter helped is strong evidence that the defenders are indeed strong types. Taking  $T \rightarrow \infty$  gives  $C_1 \rightarrow \sum_{t=1}^{\infty} \delta^t \frac{A}{N} = \frac{\delta}{1 - \delta} \frac{A}{N}$ . Solving the quadratic for  $\Phi(C_1)$  gives

$$\Phi(C_1) = \frac{\sqrt{\pi} - \pi}{1 - \pi}$$

as the upper bound for  $\Phi(C_1)$  for the equilibrium to exist. □