It's on Me! The Benefit of Altruism in BAR Environments

UTCS Technical Report TR-10-08

Edmund L. Wong, Lorenzo Alvisi
The University of Texas at Austin
{elwong,lorenzo}@cs.utexas.edu

Abstract

Cooperation, a necessity for any peer-to-peer (P2P) cooperative service, is often achieved by rewarding good behavior now with the promise of future benefits. However, in most cases, interactions with a particular peer or the service itself eventually end, resulting in some last exchange in which departing participants have no incentive to contribute. Without cooperation in the last round, cooperation in any prior round may be unachievable.

In this paper, we propose leveraging altruistic participants that simply follow the protocol as given. We show that altruism is a simple, necessary, and sufficient way to incentivize cooperation in a realistic model of a cooperative service's last exchange, in which participants may be Byzantine, altruistic, or rational and network loss is explicitly considered. By focusing on network-level incentives in the last exchange, we believe our approach can be used as the cornerstone for incentivizing cooperation in any cooperative service.

1 Introduction

Decentralized services in which peers belonging to multiple administrative domains (MAD) provide content for one another instead of relying on designated servers are, in principle, more scalable, robust, and flexible than traditional client-server approaches. Experience with deployed MAD services, however, shows that establishing and maintaining cooperation between peers is hard [15, 18]: because participants may be selfish and withhold resources unless contributing is in their best interest, cooperative services must provide sufficient incentives for participants to contribute. These incentive structures must of course be resilient against buggy or malicious peers; however, they must also be robust against a more subtle threat: an overabundance of goodwill from the correct and unselfish peers who simply follow protocol run by the service. It is, after all, the unselfishness of correct peers—as codified in the protocol they obediently follow—that allows selfish peers to continue receiving service without contributing their fair share. Yet, the efforts of well-meaning peers alone may be insufficient to sustain the service. Further, asking these peers to increase their contribution to make up for free-riders may backfire: even well-meaning peers, if blatantly taken advantage of, may give in to the temptation of joining the ranks of the selfish, leading in turn to more defections and to the service's collapse.

The impact of correct and unselfish peers on the incentive structure of MAD services is not well understood. The BAR model [3] does explicitly account for these peers—they are the *altruistic* peers, who, together with the selfish *rational* peers and the potentially disruptive *Byzantine* peers, give the model its acronym—but existing BAR-tolerant systems have essentially sidestepped the challenge of altruism by designing protocols that neither depend on nor leverage the presence of altruistic peers.[1] While this design decision ensures the cooperation of rational peers, it does so at the cost of making altruistic peers simply act like their selfish counterparts. It is hard not to feel the sense of a lost opportunity: real MAD systems do include a sizable fraction of altruistic peers [2] who are willing to continue to follow their assigned protocol despite the presence of selfish peers. Can we leverage their good will and still give rational participants the incentive to cooperate?

In this paper we show that not only is altruism not antithetical to rational cooperation, but that, in a fundamental way, rational cooperation can only be achieved in the presence of altruism. To do so, we distill the issue to a rational peer's *last* opportunity to cooperate with a service.

The last exchange. Rational peers are induced to cooperate with another peer (or, more generally, with a service) by the expectation that, if they cooperate, they will receive future benefit. However, in most cases, interaction with a particular peer or with the service itself eventually comes to an end. In this last exchange, rational peers do not have incentive to contribute, as doing so incurs cost without any future benefit. Unfortunately, rational cooperation throughout the protocol often hinges on this critical last exchange, and the lack of incentive to cooperate at the end may, in a sort of reverse domino effect, demotivate rational peers from cooperating in *any* prior exchange.

Most current systems address this problem in one of three ways (or some combination of them). Some systems [3, 14] assume that rational peers interact with the service forever, and thus future incentives always exist; others [11, 12, 13] strengthen the condition under which rational peers depart from their assigned protocol, so that, instead of deviating as soon as it is in their interest to do so, they deviate only if their increase in utility is above a certain threshold; others [14], finally, try to force rational peers into cooperation by threatening them with the possibility of losing utility

[1] Gossip-based BAR-tolerant streaming protocols [13, 14] do rely on an altruistic source for seeding the stream but otherwise model the gossiping peers as either rational or Byzantine.

if they deviate. For instance, in BAR Gossip [14], peers that do not receive the data they expect proceed to *pester* the guilty peer by repeatedly requesting the missing contribution.

Unfortunately, each of these approaches relies on somewhat unrealistic assumptions. Few relationships in life are infinite in length; worse, as we will show later in this paper, in many real-world environments, in which the network is lossy and peers may be arbitrarily faulty (*Byzantine*), incentivizing cooperation may be impossible even in an infinite-length protocol. Assuming, on the other hand, that rational players only deviate if their increase in utility is significant leaves open the real possibility of "penny pinching" rational players shirking their responsibilities, thus causing both cooperation and the service that relies on it to collapse. Finally, threats such as pestering are effective only when they are credible, *i.e.*, to feel threatened, a peer must believe that it will be rational for the other party to pester. Unfortunately, since pestering incurs cost for both receiver and initiator, it is hard to motivate rational peers to pester in the first place. In BAR Gossip, pestering is a credible threat only under the somewhat implausible assumption that rational peers always blame another peer's lack of contribution on the unreliability of the underlying network and, lacking a response, are willing to pester forever.

Our contributions. In this paper, we show that leveraging the presence of altruistic peers in MAD systems offers a simple and realistic way to elicit cooperation from rational peers in the last exchange without relying on any of the aforementioned unrealistic assumptions.

The approach that we propose models the last exchange as a finite-round game between two players that communicate through an unreliable channel: the first player chooses to contribute or do nothing; the second replies by either pestering or doing nothing. Because the channel is lossy, players do not necessarily share the same view of the ongoing game; for instance, the first player may have contributed, but the second player may not have received it. Although our approach, like BAR Gossip, is based on pestering, we do not require implausible network assumptions or the specter of never-ending pestering to motivate rational peers to contribute in the last exchange. Instead, we show that the presence of altruistic nodes is both necessary and sufficient to cause cooperation. In particular, we prove that there exists no equilibrium strategy where rational peers contribute if all peers are either rational or Byzantine—even if we allow for an infinite number of pestering rounds. Further, we show that the presence of altruistic participants is sufficient to transform pestering in a credible threat. Intuitively, if rational participants believe with sufficiently high beliefs that they may be interacting with an altruistic peer, they are motivated to pester, making it in turn preferable for rational peers to contribute.

The fraction of altruistic participants sufficient to sustain rational contribution depends on several system parameters, including the probability of network loss, the fraction of Byzantine peers in the system, and the behavior that rational peers expect from altruistic and Byzantine peers. Exploring this space through a simulator we find that:

- Altruistic peers make rational cooperation easy to achieve under realistic conditions. In particular, we find that even if less than 10% of the population is altruistic, rational participants are incentivized to cooperate in a system where the network drops 5% of all packets and Byzantine participants make up over 50% of the remainder of the population.
- Overly profligate altruistic peers, on the other hand, are indeed harmful to rational cooperation: if altruistic peers contribute every time they are pestered, then we cannot always achieve rational cooperation; when we do, it requires an implausibly high fraction of altruistic peers. This is good news: the less foolishly generous the behavior of the altruistic participants

sufficient to incentivize rational contribution, the more feasible it becomes to design systems with a sustainable population of altruistic peers.

- The uncertainty introduced by network loss is both a bane and a boon. On one hand, it prevents each player involved in the last exchange from acquiring common knowledge about what the other has observed and therefore significantly complicates the analysis of a player's optimal strategy. On the other, uncertainty lowers the bar for rational cooperation by leaving open some possibility that the other player may be altruistic even when the observed behavior suggests otherwise.

2 Formalizing the last exchange problem

We consider two peers in a cooperative service, $\mathcal{P}_1$ and $\mathcal{P}_2$, communicating through an unreliable channel. We assume both peers believe to be engaged in their last exchange—neither peer expects it will interact with the other beyond this exchange. We assume that $\mathcal{P}_1$ holds a contribution (*e.g.*, some information) that is of value to $\mathcal{P}_2$. We are interested in studying under what conditions it is possible to induce a selfish $\mathcal{P}_1$ to contribute. We call a non-Byzantine $\mathcal{P}_2$ *destitute* if $\mathcal{P}_2$ has not observed $\mathcal{P}_1$'s contribution.

We model the last exchange as a game, in which $\mathcal{P}_1$ and $\mathcal{P}_2$ are the players. The game lasts $T+1$ rounds starting with round T and ending with round 0.[2] In each round, $\mathcal{P}_1$ moves first by either contributing (denoted by c) to $\mathcal{P}_2$ or doing nothing (denoted by n); $\mathcal{P}_2$ then responds by either pestering (denoted by p) $\mathcal{P}_1$ or doing nothing (n). Each player initially is assigned a strategy which, as we will see, the player may or may not end up following. We discuss a specific assigned strategy in detail in Section 4.

Network loss and signals. To model the unreliable channel through which $\mathcal{P}_1$ and $\mathcal{P}_2$ communicate we adopt from game theory the concept of *private signals*: for every action a played by some player, both players privately observe some (possibly different) resulting *signal*.

Specifically, let ρ, $0 < \rho < 1$, be the rate of network loss, and assume that ρ is common knowledge. When a player plays a, the other observes a with probability $1-\rho$ and n with probability ρ. Thus, although players observe perfectly their own actions, they do not observe perfectly their peer's action, and a player may not know what the other player has observed.[3]

The sequence of signals observed by a player defines that player's *history*. A round-t history h^t consists of, for every completed round i, $t < i \leq T$, a pair of signals a_1^i and a_2^i (corresponding to $\mathcal{P}_1$ and $\mathcal{P}_2$'s actions) and, for round t, the proper prefix of the pair, which may be empty (indicating $\mathcal{P}_1$'s move) or a_1^t (indicating $\mathcal{P}_2$'s move).[4] If it is $\mathcal{P}_1$'s move, given a round-t history h^t and a sequence of signals $a_1^t a_2^t, \ldots$, $(h^t, a_1^t a_2^t, \ldots)$ is the history that results from observing h^t followed by the specified sequence of signals. Similarly, if it is $\mathcal{P}_2$'s move, then given a round-t history h^t and a sequence of signals $a_2^t, a_1^{t-1}a_2^{t-1}, \ldots$, $(h^t, a_2^t, a_1^{t-1}a_2^{t-1}, \ldots)$ is the resulting history.

Player types and beliefs. We consider three different *types* of players:

- Byzantine (**B**): These players play an arbitrary strategy, independent of the assigned strategy.

[2] It is important not to confuse an exchange and a round. An exchange is an application-level transaction between two peers, and the protocol required to achieve it may be comprised of multiple rounds of communication.

[3] A player who plays n knows what the other will observe; a player who observes c or p knows what the other played.

[4] Throughout the paper, we use a_i^j and b_i^j to denote two signals from $\mathcal{P}_i$ in round j.

- Altruistic (**A**): These players follow the assigned strategy.
- Rational (**R**): These players follow the assigned strategy only if deviating unilaterally does not increase their utility.

$\mathcal{P}_1$ and $\mathcal{P}_2$ know their own type but can only make probabilistic guesses as to the type of their peer. $\mathcal{P}_i$'s *beliefs* are the probabilities $\mu_i(\theta)$ that $\mathcal{P}_i$ assigns to the statement that the other player is of type θ. For simplicity, we assume that, if $\mathcal{P}_1$ and $\mathcal{P}_2$ are non-Byzantine, then initially $\mu_1(\theta) = \mu_2(\theta)$ for all θ.

Utilities, strategies, and equilibrium. Sending and receiving (whether contributions or pesters) incur positive costs: respectively, s_c and r_c for contributions and s_p and r_p for pesters. Receiving a contribution, though, also yields a one-time benefit of b_c. Doing nothing has neither a cost nor a benefit. We assume $\mathcal{P}_2$ strongly prefers to receive a contribution and will pester to get it ($b_c - r_c - s_p \gg 0$).

We define $\mathcal{P}_1$ and $\mathcal{P}_2$'s utilities as follows:

$$u_1(C, \hat{P}) = -\Big(|\hat{P}|r_p + |C|s_c\Big) \qquad\qquad u_2(P, \hat{C}) = H[|\hat{C}| - 1]b_c - \Big(|P|s_p + |\hat{C}|r_c\Big)$$

where C and $\hat{P}$ are, respectively, the sets of rounds in which $\mathcal{P}_1$ contributed and observed pestering from $\mathcal{P}_2$; P and $\hat{C}$ are the sets of rounds in which $\mathcal{P}_2$ respectively pestered and observed $\mathcal{P}_1$ contribute; and $H[n]$ is the unit step function.[5] Intuitively, rational players aim to minimize contributing and pestering; additionally, $\mathcal{P}_2$ prefers to receive the contribution.

Let $\sigma = (\sigma_1, \sigma_2)$ denote the strategy profile that describes the strategies σ_1 and σ_2 of $\mathcal{P}_1$ and $\mathcal{P}_2$. Similarly, let $\mu = (\mu_1, \mu_2)$ denote the belief profile that describes, for each player type θ, the beliefs μ_1 and μ_2 held by the two players. In general, a player's strategy and beliefs depend on the signals the player observed; for a given history h^t, let $\mu_i(\theta|h^t)$ denote $\mathcal{P}_i$'s conditional beliefs computed using Bayes rule, and $\sigma_i(a|h^t)$ denote the conditional probability that a is played by $\mathcal{P}_i$.

We are interested in perfect Bayes equilibrium, *i.e.*, in a strategy profile σ^* and a set of beliefs μ^* such that for all $i \in \{1, 2\}$, for all h^t, and for all σ' and corresponding beliefs μ':

$$E^{\sigma^*,\mu^*}[u_i|h^t] \geq E^{\sigma',\mu'}[u_i|h^t]$$

where $E^{\sigma,\mu}[u_i|h^t]$ is, for each $\mathcal{P}_i$, the expected utility of playing a strategy σ_i and holding beliefs μ_i, both conditional on history h^t.

Note that σ^* and μ^* must incorporate the expected Byzantine and altruistic strategies. We assume that rational players expect that the strategy chosen by Byzantine players is independent of the observed history, *i.e.*, rational players do not expect that they can influence the strategy of Byzantine players, and Byzantine players play independently of how they played in the past. We denote the probability that a player expects to observe a Byzantine peer doing nothing in round t as $\beta_t \geq \rho$ (to account for network loss). We define the altruistic strategy later in Section 4.

We assume that all players are limited to actions in the strategy space. This can be accomplished in practice if actions outside of the strategy space generate a proof of misbehavior [3, 9] and if the associated punishments (*e.g.*, exclusion from the service or financial penalties) are sufficient to deter rational players and limit the damage that can be caused by any one Byzantine player.

[5] $H[n] = 0$ if $n < 0$; else $H[n] = 1$.

3 The need for altruism

Our first result is to show that, in a bounded game, incentivizing contribution during the last exchange is impossible without altruism, *i.e.*, with only rational and Byzantine players.

In our proofs, we only consider a player's expected utility when the other player is rational: since a Byzantine player's strategy is independent of the actions of the other player, one is always better off doing nothing when facing a Byzantine player.

LEMMA 3.1. *There exists no equilibrium where a rational $\mathcal{P}_2$ pesters with any positive probability if a rational $\mathcal{P}_1$ will not subsequently contribute.*

Intuition. $\mathcal{P}_2$ incurs cost by pestering—with no chance of future contribution from $\mathcal{P}_1$.

Proof. Suppose such an equilibrium (σ^*, μ^*) exists where after some history h^t, $\mathcal{P}_2$ pesters with probability $\gamma > 0$ during some round t. Consider an alternate strategy/belief pair $((\sigma_1^*, \sigma_2'), \mu')$ in which $\mathcal{P}_2$ plays exactly as in σ_2^* until round t, after which $\mathcal{P}_2$ never pesters again. Since $\mathcal{P}_1$ will never contribute again, then

$$\begin{aligned} E^{\sigma^*,\mu^*}[u_2|h^t, \mathbf{R}] &\leq \gamma\left(-s_p + E^{\sigma^*,\mu^*}[u_2|(h^t,p),\mathbf{R}]\right) + (1-\gamma)\left(E^{\sigma^*,\mu^*}[u_2|(h^t,n),\mathbf{R}]\right) \\ &< 0 = E^{(\sigma_1^*,\sigma_2'),\mu'}[u_2|(h^t,n),\mathbf{R}] \end{aligned}$$

Following σ_2' instead of σ_2^* improves $\mathcal{P}_2$'s utility. Contradiction. □

LEMMA 3.2. *There exists no equilibrium where a rational $\mathcal{P}_1$ contributes with any positive probability if a rational $\mathcal{P}_2$ will not subsequently pester.*

Intuition. $\mathcal{P}_1$ incurs cost by contributing, yet there is no threat of pestering from $\mathcal{P}_2$.

Proof. Suppose such an equilibrium (σ^*, μ^*) exists where, after some history h^t, $\mathcal{P}_1$ expects to contribute with probability $\gamma > 0$ during some round t. Consider an alternate strategy/belief pair $((\sigma_1', \sigma_2^*), \mu')$ in which $\mathcal{P}_1$ plays exactly as in σ_1^* until round t, after which $\mathcal{P}_1$ never contributes again. Since $\mathcal{P}_1$ is not pestered again from round t on,

$$E^{\sigma^*,\mu^*}[u_1|h^t, \mathbf{R}] \leq \gamma\left(-s_c + E^{\sigma^*,\mu^*}[u_1|(h^t,c),\mathbf{R}]\right) + (1-\gamma)E^{\sigma^*,\mu^*}[u_1|(h^t,n),\mathbf{R}] < E^{(\sigma_1',\sigma_2^*),\mu'}[u_1|(h^t,n)]$$

Following σ_1' instead of σ_1^* improves $\mathcal{P}_1$'s utility. Contradiction. □

THEOREM 3.3. *There exists no equilibrium in which a rational $\mathcal{P}_1$ contributes or a rational $\mathcal{P}_2$ pesters.*

Proof. Suppose such an equilibrium (σ^*, μ^*) exists. Then there exists some last round $t_p \leq T$ during which $\mathcal{P}_2$ pesters with some positive probability and some round $t_c \leq T$ after which $\mathcal{P}_1$ never contributes again. By Lemma 3.2, a rational $\mathcal{P}_1$ never contributes after round t_p and so $t_c \geq t_p$.[6] However, by Lemma 3.1, a rational $\mathcal{P}_2$ only pesters until round $t_c + 1$; thus, $t_p > t_c$. Contradiction. □

A weaker, but in practice still crippling, result holds also for the unbounded version of the game.

THEOREM 3.4. *If there exist* any *positive proportion of Byzantine peers who never contribute or always pester, then there exists no equilibrium in which a rational $\mathcal{P}_1$ contributes or a rational $\mathcal{P}_2$ pesters.*

[6] Recall that we count rounds in reverse.

Proof. (Sketch) Suppose there exists some Byzantine $\mathcal{P}_2$ who always pesters, despite $\mathcal{P}_1$'s contributions, then $\mathcal{P}_1$'s belief that $\mathcal{P}_2$ is Byzantine eventually grows arbitrarily close to 1. Similarly, if a Byzantine $\mathcal{P}_1$ never contributes despite $\mathcal{P}_2$'s incessant pestering, then $\mathcal{P}_2$ becomes increasingly certain $\mathcal{P}_1$ is Byzantine. It can be shown that, once that belief grows to the point that the expected future benefit is lower than the cost of, respectively, contributing and pestering, $\mathcal{P}_1$ and $\mathcal{P}_2$ stop cooperating, reducing the problem to the bounded case covered by Theorem 3.3. □

4 Altruism to the rescue

Although altruism is necessary to motivate rational cooperation, computing a tight bound for the fraction of altruistic peers needed to sustain cooperation is hard because of the uncertainty introduced by the unreliable channel through which players communicate. For instance, when $\mathcal{P}_2$ tries to determine whether to pester $\mathcal{P}_1$ in the current round, $\mathcal{P}_2$ generally does not have full knowledge of the actions that $\mathcal{P}_1$ has actually played or the signals that $\mathcal{P}_1$ has observed. Analyzing the game quickly becomes intractable.

We narrow the design space by considering strategies that result in a rational $\mathcal{P}_2$ pestering in every round of the game (except for the last one). We feel that this is a reasonable design point, since we expect that typically the benefit that $\mathcal{P}_2$ expects from $\mathcal{P}_1$'s contribution will significantly outweigh the cost of a few rounds of pestering. We then prove that the threat of both altruistic and rational pestering for every round but the last is sufficient to motivate $\mathcal{P}_1$ to contribute.

We assume that the cooperative service specifies the following strategy:

- $\mathcal{P}_1$ contributes during round T with probability 1. During round $t < T$, $\mathcal{P}_1$ contributes, if pestered, with probability $(1-\alpha)/(1-\rho)^2$, where α is some known parameter such that $0 \leq \alpha < 1$. Intuitively, if during round t $\mathcal{P}_2$ pesters a $\mathcal{P}_1$ following this strategy, $\mathcal{P}_2$ expects to observe a contribution in round $t-1$ with probability $1-\alpha$.

- $\mathcal{P}_2$ pesters until a contribution is received or round 0.

As stated in Section 2, altruistic players follow this strategy throughout the entire game.

4.1 Conditions for pestering

We derive the condition under which a rational $\mathcal{P}_2$ is motivated to pester in every round (except the last) under three assumptions.

1. For any round $t < T$, a non-Byzantine $\mathcal{P}_1$ contributes only if pestered. We prove in Section 4.2 that rational $\mathcal{P}_1$ never contributes otherwise.
2. A rational $\mathcal{P}_1$ contributes in round t iff $\mathcal{P}_1$'s belief that $\mathcal{P}_2$ is destitute exceeds some threshold $\bar{\mu}_1^t$. We prove in Section 4.2 that this is indeed the case.
3. We assume that a rational $\mathcal{P}_1$ expects a non-Byzantine $\mathcal{P}_2$ to always pester (except in the last round). The main result of this section is to prove that a rational $\mathcal{P}_2$ has no incentive to unilaterally deviate from this expectation.

We start by making two simple observations that are easy to prove.

LEMMA 4.1. *If a rational $\mathcal{P}_2$ has received a contribution, $\mathcal{P}_2$ prefers to do nothing.*

Intuition. If $\mathcal{P}_2$ already has the contribution, $\mathcal{P}_2$ receives no further benefit from receiving another contribution. In fact, pestering and receiving another contribution only incurs cost.

Proof. By contradiction. Let (σ^*, μ^*) be some equilibrium in which in round t, $\mathcal{P}_2$ pesters with some probability $\gamma > 0$ after observing some history h^t. Construct a strategy/belief pair $((\sigma_1^*, \sigma_2'), \mu')$ in which $\mathcal{P}_2$ plays the same as in σ_2^* but does nothing starting from h. The expected difference in utility between σ_2' versus σ_2^* is at least

$$E^{(\sigma_1^*,\sigma_2'),\mu'}[u_2|h^t] - E^{\sigma^*,\mu^*}[u_2|h^t] \geq \gamma s_p > 0$$

$\mathcal{P}_2$ is better off following σ_2', a contradiction. □

Lemma 4.2. *A destitute rational $\mathcal{P}_2$ prefers to pester an altruistic $\mathcal{P}_1$ if in expectation the benefit exceeds the cost,* i.e. $s_p \leq (1-\alpha)(b_c - r_c)$.

Proof. If $\mathcal{P}_1$ is altruistic and is pestered, then $\mathcal{P}_1$ contributes with probability $(1-\alpha)/(1-\rho)^2$. The condition above guarantees that the cost of pestering is at most that of the expected benefit. □

We now show that a rational $\mathcal{P}_2$ is no less likely to get a contribution from a rational $\mathcal{P}_1$ if $\mathcal{P}_2$ pesters more frequently. To prove this, we need to show hat $\mathcal{P}_1$ is as likely to contribute after being pestered than not. As this involves $\mathcal{P}_1$'s beliefs, we use a lemma regarding $\mathcal{P}_1$'s beliefs and behavior (Lemma 4.10) that we will later prove. Intuitively, because $\mathcal{P}_1$ expects a destitute $\mathcal{P}_2$ to always pester whereas Byzantine $\mathcal{P}_2$ may not, $\mathcal{P}_1$ is more convinced that $\mathcal{P}_2$ is destitute if pestering is observed.

Lemma 4.3. *If a destitute $\mathcal{P}_2$ pesters following some history h^t, $\mathcal{P}_2$ is no less likely to receive a contribution from a rational P_1 than if $\mathcal{P}_2$ instead did nothing.*

Proof. Let h_1^t be the history that $\mathcal{P}_1$ has observed when $\mathcal{P}_2$ observed h^t. If $\mathcal{P}_2$ pesters and $\mathcal{P}_1$ does not observe it, then in either case, $\mathcal{P}_1$ starts from history (h_1^t, n) and plays exactly the same. As a result, the likelihood of $\mathcal{P}_1$ contributing and $\mathcal{P}_2$ receiving said contribution is exactly the same.

Thus, suppose that $\mathcal{P}_1$ observes the history (h_1^t, p) if $\mathcal{P}_2$ pesters and (h_1^t, n) otherwise. Consider any two complete histories $h_1 = (h_1^t, p, a_1^{t-1}a_2^{t-1}, \ldots, a_1^0 a_2^0)$ and $h_1' = (h_1^t, n, b_1^{t-1}a_2^{t-1}, \ldots, b_1^0 a_2^0)$; let m_p and m_n be the number of rounds following round t in which $\mathcal{P}_1$ observes $\mathcal{P}_2$ pestering and doing nothing. By Lemma 4.10, we know that h_1 must contain at least as many c as h_1'; let m_c and m_c' be the number of rounds in which $\mathcal{P}_1$ contributes following h_1^t.

Recall that $\mathcal{P}_1$ expects that a destitute $\mathcal{P}_2$ will pester every round except the last. The probability that, starting from (h_1^t, p), play will evolve as specified in h_1 yet $\mathcal{P}_2$ will not end up with the contribution is simply $\rho^{m_c+m_n-1}(1-\rho)^{m_p}$ (since a destitute $\mathcal{P}_2$ never pesters in the last round), whereas the probability that, starting from (h_1^t, n), play will evolve as specified in h_1' yet $\mathcal{P}_2$ does not have the contribution is

$$\rho^{m_c'+m_n-1}(1-\rho)^{m_p} \geq \rho^{m_c+m_n-1}(1-\rho)^{m_p}$$

since $m_c \geq m_c'$. As the probability of $\mathcal{P}_2$ not getting the contribution starting from (h_1^t, p) and (h_1^t, n) is simply the sum of the probabilities of all possible complete histories h_1 and h_1', it is obvious that $\mathcal{P}_2$ is no less likely to get the contribution starting from (h_1^t, p). □

We are now ready to derive sufficient conditions under which a rational $\mathcal{P}_2$ prefers to pester.

Theorem 4.4. *Let $t > 0$ be the current round and h^t be the current history. A rational and destitute $\mathcal{P}_2$ prefers to pester*[7] *if*

$$\mu_2(\boldsymbol{A}|h^t) \geq \frac{s_p}{\alpha^{t-1}(1-\alpha)(b_c - r_c) + (1-\alpha^{t-1})s_p} \tag{1}$$

Proof. By contradiction. Consider some equilibrium (σ^*, μ^*) in which $\mathcal{P}_2$ prefers not to pester in round t after observing h^t. Construct another strategy/belief pair $((\sigma_1^*, \sigma_2'), \mu')$ in which $\mathcal{P}_2$ does instead pester in t after h^t. If $\mathcal{P}_2$ receives a contribution in round $t-1$, $\mathcal{P}_2$ does nothing for the remainder of the game; otherwise, for the remaining rounds σ_2' and σ_2^* are identical.

Consider $\mathcal{P}_2$'s difference in expected utility between playing σ_2' and σ_2^*. There are three cases. If $\mathcal{P}_1$ is Byzantine, the expected difference in utility between σ_2^* and σ_2' is s_p. If $\mathcal{P}_1$ is rational, by Lemma 4.3, $\mathcal{P}_2$ has a better chance of receiving the contribution in future rounds if $\mathcal{P}_2$ pesters in round t. Hence, if at the end of round t $\mathcal{P}_2$ is still destitute, then the expected difference in utility between σ_2' and σ_2^* is at most s_p. Finally, if $\mathcal{P}_1$ is altruistic, then the expected utility from playing an altruistic $\mathcal{P}_1$ starting from round $t-1$ from σ^* and σ' are the same; let $V(\mathbf{A}, t-1)$ represent this utility. Thus, the expected difference in utility between σ_2^* and σ_2' is

$$E^{\sigma^*,\mu^*}[u_2|h^t, \mathbf{A}] - E^{(\sigma_1^*,\sigma_2'),\mu'}[u_2|h^t, \mathbf{A}] = s_p - (1-\alpha)(b_c - r_c - V(\mathbf{A}, t-1))$$

By Lemma 4.2, pestering an altruistic $\mathcal{P}_1$ until $\mathcal{P}_2$ gets the contribution or $t = 0$ is in $\mathcal{P}_2$'s best interest, and thus $V(\mathbf{A}, t-1) \leq -s_p + (1-\alpha)(b_c - r_c) + \alpha V(\mathbf{A}, t-2)$, where $V(\mathbf{A}, 0) = 0$. Solving the recursion, we have $V(\mathbf{A}, t-1) \leq \frac{1-\alpha^{t-1}}{1-\alpha}(-s_p + (1-\alpha)(b_c - r_c))$. Using condition (1) we get:

$$E^{\sigma^*,\mu^*}[u_2|h^t] - E^{(\sigma_1^*,\sigma_2'),\mu'}[u_2|h^t] \leq s_p - \mu_2(\mathbf{A}|h^t)((1-\alpha^{t-1})s_p + \alpha^{t-1}(1-\alpha)(b_c - r_c)) \leq 0$$

This implies that $\mathcal{P}_2$ prefers to play σ_2' over σ^*. Contradiction. □

4.2 Conditions for contributing

In every round, $\mathcal{P}_1$ must make a choice:

- Pay the cost of contributing now (s_c), hoping to stop a non-Byzantine $\mathcal{P}_2$ from pestering in the future. The savings are a function of the remaining rounds and the beliefs about $\mathcal{P}_2$.
- Delay contributing, at the risk of being pestered (with cost at most $(1-\rho)r_p$), hoping to glean more about $\mathcal{P}_2$'s type.

Procrastination has its lure. Since we are considering strategies where a non-Byzantine $\mathcal{P}_2$ always pesters (minus the last round) whereas a Byzantine $\mathcal{P}_2$ may not, every action and signal can drastically affect future play and expected utilities, and possibly save $\mathcal{P}_1$ the cost of contributing. Moreover, doing nothing now does not preclude $\mathcal{P}_2$ from contributing in the future.

In this section we show that if $\mathcal{P}_1$ can muster a sufficiently strong belief that $\mathcal{P}_2$ is destitute, procrastination is something best put off until tomorrow: we prove that for every round t sufficiently removed from the end of the game, there exists a belief threshold $\bar{\mu}_1^t$ beyond which contributing yields a higher expected utility for $\mathcal{P}_1$. Furthermore, we prove Lemma 4.10, which we previously

[7] Technically, $\mathcal{P}_2$ *weakly* prefers to pester (*i.e.*, non-strict inequality), as the expected utility of pestering and doing nothing may be the same. For simplicity, we assume for the remainder of the paper that weak preference for pestering/contributing is sufficient and that players do nothing only if the preference is strong (*i.e.*, strict inequality).

used to prove Lemma 4.3. We show these results under the assumption that a destitute $\mathcal{P}_2$ pesters in every round (minus the last) and non-destitute $\mathcal{P}_2$ never pester.

"Sub-typing" non-Byzantine $\mathcal{P}_2$: D and $\neg$D. Recall that Lemma 4.1 states that a non-Byzantine $\mathcal{P}_2$, upon receiving a contribution, stops pestering. Thus, a rational $\mathcal{P}_1$ is only interested in contributing if there is a sufficient pestering threat from a destitute $\mathcal{P}_2$. Because destitute players are effectively responsible for motivating $\mathcal{P}_1$ to contribute, we divide the group of non-Byzantine players into two "sub-types": destitute ($\mathbf{D}$) and non-destitute ($\neg\mathbf{D}$).

These sub-types simplify the expected strategy of $\mathcal{P}_2$: if $\mathcal{P}_2$ is of type $\neg\mathbf{D}$, $\mathcal{P}_2$ always does nothing (by Lemma 4.1), whereas a $\mathcal{P}_2$ of type $\mathbf{D}$ always pesters except in round 0. In addition, if $\mathcal{P}_1$ contributes, a $\mathcal{P}_2$ of type $\mathbf{D}$ has a $1-\rho$ probability of observing the contribution and becoming a type $\neg\mathbf{D}$:

$$\mu_1(\mathbf{D}|(h^t,c)) = \rho\mu_1(\mathbf{D}|h^t) \qquad\qquad \mu_1(\neg\mathbf{D}|(h^t,c)) = \mu_1(\neg\mathbf{D}|h^t) + (1-\rho)\mu_1(\mathbf{D}|h^t)$$

Since $\mathcal{P}_1$ may observe a non-Byzantine $\mathcal{P}_2$ play either n or p, $\mathcal{P}_1$ can never choose an action knowing for certain that $\mathcal{P}_2$ is Byzantine;[8] thus, for any history h^t, $\mu_1(\mathbf{B}|h^t) < 1$ and $\mu_1(\mathbf{D}|h^t) > 0$.

Result: $\mathcal{P}_1$ contributes only if pestered. We now prove that $\mathcal{P}_1$ contributes only if pestered. This prevents a rational $\mathcal{P}_2$ from waiting for free contributions that may come without pestering.

We begin by stating some basic results that are easily proven.

Lemma 4.5. *$\mathcal{P}_1$ prefers to do nothing for rounds $t \le \tau$, where*

$$\tau = \left\lceil \frac{1}{(1-\rho)^2}\frac{s_c}{r_p} \right\rceil \tag{2}$$

Lemma 4.6. *Let h^t and $(h')^t$ be two histories observed by $\mathcal{P}_1$. Then $\mu_1(\boldsymbol{D}|h^t) = \mu_1(\boldsymbol{D}|(h')^t)$ if*

1. *For every round, $\mathcal{P}_2$'s signals in h^t and $(h')^t$ are the same;*
2. *The number of contributions played are equal in h^t and $(h')^t$; and*
3. *The last contribution in both h^t and $(h')^t$ are followed by a pester at some future round.*

Proof. Since pestering has been observed after the last contribution and a non-destitute $\mathcal{P}_2$ would never pester (Lemma 4.1), $\mu_1(\neg\mathbf{D}|h^t) = \mu_1(\neg\mathbf{D}|(h')^t) = 0$. Letting C represent the rounds in which $\mathcal{P}_1$ contributed and $\hat{P}$ and $\hat{N}$ be the rounds in which $\mathcal{P}_2$ is observed to pester and do nothing,

$$\mu_1(\mathbf{D}|h^t) = \frac{\mu_1(\mathbf{D})(1-\rho)^{|\hat{P}|}\rho^{|\hat{N}|+|C|}}{\mu_1(\mathbf{B})\prod_{i\in\hat{P}}(1-\beta_i)\prod_{i\in\hat{N}}\beta_i + \mu_1(\mathbf{D})(1-\rho)^{|\hat{P}|}\rho^{|\hat{N}|+|C|}} = \mu_1(N|(h')^t) \qquad \square$$

We now show that $\mathcal{P}_1$ prefers to contribute only if pestering has been observed since the last contribution. In other words, given the bound in condition (3), $\mathcal{P}_1$ is never better off trying to contribute unless $\mathcal{P}_1$ knows for certain that $\mathcal{P}_2$ is destitute.

Lemma 4.7. *Let $t < T$ be the current round, where*

$$T < \frac{1-\rho+\rho^2}{\rho^2(1-\rho)^2}\frac{s_c}{r_p} \tag{3}$$

If $\mathcal{P}_1$ has contributed and has not been pestered since, then $\mathcal{P}_1$ prefers not to contribute in round t.

[8] $\mathcal{P}_1$ can become certain in the last round if a Byzantine $\mathcal{P}_2$ decides to pester, but at that point the game is over.

Intuition. After contributing, $\mathcal{P}_1$'s belief that $\mathcal{P}_2$ is destitute is reduced by a factor of $1/\rho$. A sufficiently lossy network and high pestering cost could potentially motivate $\mathcal{P}_1$ to send an unsolicited contribution to avoid further pestering. Condition (3) ensures that $\mathcal{P}_1$ is better off waiting for $\mathcal{P}_2$ to pester .

Proof. By Lemma 4.5, $\mathcal{P}_1$ never contributes starting from round τ (as defined in (2)); thus, assume $t > \tau$. We first prove, by contradiction, the lemma if after round t, $\mathcal{P}_1$ contributes only after being pestered again. Let m_p and m_c be the number of rounds since $\mathcal{P}_1$ has observed pestering and contributing, *i.e.*, $\mathcal{P}_1$ has done nothing and observed $\mathcal{P}_2$ doing nothing in the past m_c and m_p rounds. By assumption, $m_p > m_c \geq 0$. Let the current history be $h^t = (h^{t+m_c+1}, cn, a_1^{t+m_c}a_2^{t+m_c}, \ldots, a_1^{t+1}a_2^{t+1})$ where $a_1^i = a_2^i = n$ for all $t < i \leq t + m_c$.

Let (σ^*, μ^*) be some equilibrium such that $\mathcal{P}_1$ prefers to contribute in some round t. Let $V(\theta)$ be the continuation payoff after $\mathcal{P}_1$ contributes in round t given that $\mathcal{P}_2$ is of type θ. Construct an alternate strategy/belief pair (σ', μ') such that $\mathcal{P}_1$ does not contribute in round t but for the remaining rounds is identical to σ^* (as if $\mathcal{P}_1$ had contributed in round t). Thus, the continuation payoff from playing σ' after doing nothing in round t is also $V(\theta)$. Note that $V(\neg\mathbf{D}) = 0$ since a non-Byzantine $\mathcal{P}_2$ who has the contribution never pesters by Lemma 4.1.

Following σ^* and contributing during round t results in an expected payoff of

$$-s_c + \mu_1^*(\mathbf{D}|(h^t, c))V(\mathbf{D}) + \mu_1^*(\mathbf{B}|(h^t, c))V(\mathbf{B})$$

whereas following σ' and doing nothing earns

$$\mu_1'(\mathbf{D}|(h^t, n))V(\mathbf{D}) + \mu_1'(\mathbf{B}|(h^t, n))V(\mathbf{B})$$

Observe that $\mu_1(\mathbf{B}|(h^t, n)) = \mu_1(\mathbf{B}|(h^t, c))$. Since σ^* is an equilibrium strategy and $\mu_1^*(\theta|h^t) = \mu_1'(\theta|h^t)$, contributing results in a higher utility only if

$$-s_c \geq (1-\rho)\mu_1^*(\mathbf{D}|h^t)V(\mathbf{D})$$

Observe that by Bayes rule,

$$\begin{aligned}\mu_1^*(\mathbf{D}|h^t) &= \frac{\mu_1^*(\mathbf{D}|h^{t+m_c+1})\rho^{m_c+1}}{\mu_1^*(\mathbf{D}|h^{t+m_c+1})(\rho^{m_c+1} + (1-\rho)) + \mu_1^*(\neg\mathbf{D}|h^{t+m_c+1}) + \prod_{i=0}^{m_c}\beta_i\mu_1^*(\mathbf{B}|h^{t+m_c+1})} \\ &\leq \frac{\rho^2}{1-\rho+\rho^2}\end{aligned}$$

Note that $\mathcal{P}_1$, in any optimal strategy, can do no worse than simply being pestered for the remainder of the game. Thus, $V(\mathbf{D}) \geq -T(1-\rho)r_p$, giving us

$$-s_c \geq -\frac{\rho^2(1-\rho)^2}{1-\rho+\rho^2}Tr_p$$

which contradicts (3).

We have proven the lemma if we assume that $\mathcal{P}_1$ will not contribute in the future unless pestered again. We now finish the proof of the original lemma using induction; all we need is to show that $\mathcal{P}_1$ does not contribute in the future unless pestering has been observed since $\mathcal{P}_1$'s last contribution.

Base case: $t = \tau + 1$. By Lemma 4.5, $\mathcal{P}_1$ never contributes in the future.

Inductive step. Assume true for round t, $\tau + 1 \leq t \leq t_0$; we prove our lemma for round $t = t_0 + 1$. By the inductive hypothesis, $\mathcal{P}_1$ will not contribute in future rounds unless pestered. □

We now prove that a rational $\mathcal{P}_1$ prefers to contribute only if pestered in the prior round.

Lemma 4.8. *Let $t < T$ be the current round and h^t be the current history. Furthermore, suppose that from round t on, $\mathcal{P}_1$ contributes iff $\mathcal{P}_1$'s belief that $\mathcal{P}_2$ is destitute is at least some threshold $\bar{\mu}_1^t$. If $\mathcal{P}_1$ was not pestered in round $t+1$ and condition (3) holds, then $\mathcal{P}_1$ prefers to do nothing.*

Intuition. The belief that $\mathcal{P}_2$ is destitute is strictly non-decreasing when $\mathcal{P}_1$ observes $\mathcal{P}_2$ do nothing, and the number of expected pesters also drops as the number of remaining rounds decreases. Thus, if $\mathcal{P}_1$ intends to contribute in the next round despite observing n, then $\mathcal{P}_1$ is better off contributing in the current round.

Proof. By Lemma 4.5, $\mathcal{P}_1$ never contributes starting from round τ (as defined in (2)). Also, if $\mathcal{P}_1$ has contributed and not been pestered since $\mathcal{P}_1$'s last contribution, then this follows from Lemma 4.7. Thus, assume that $t > \tau$ and that either $\mathcal{P}_1$ has never contributed or has been pestered since the last contribution. Let t_p be the last round in which $\mathcal{P}_1$ observed pestering.

We prove Lemma 4.8 by first assuming for rounds $t' < t$, $\mathcal{P}_1$ contributes only if pestered in round $t'+1$. For the sake of contradiction, assume that there exists some equilibrium (σ^*, μ^*) such that for $t_p > t+1$ (*i.e.*, $\mathcal{P}_1$ last observed pestering more than one round ago), $\mathcal{P}_1$ does prefer to contribute following some history (h^t, nn). $\mathcal{P}_1$'s beliefs in $\mathcal{P}_2$ being non-Byzantine must be at least as high following some history (h^t, np); thus, $\mathcal{P}_1$ must also prefer to contribute following some history (h^t, np). Consider an alternate strategy/belief pair (σ', μ') in which $\mathcal{P}_1$ instead contributes in round $t+1$ after h^t but always does nothing in round t; the strategy played in subsequent rounds is the optimal one.

If $\mathcal{P}_2$ is destitute, then starting from some history h^t, $\mathcal{P}_1$ expects to earn

$$E^{\sigma^*,\mu^*}[u_1|h^t, \mathbf{D}] = -s_c + (1-\rho)(-r_p + E^{\sigma^*,\mu^*}[u_1|(h^t, np, c), \mathbf{D}]) + \rho E^{\sigma^*,\mu^*}[u_1|(h^t, nn, c), \mathbf{D}]$$

following σ^* versus

$$E^{\sigma',\mu'}[u_1|h^t, \mathbf{D}] = -s_c + \rho((1-\rho)(-r_p + E^{\sigma',\mu'}[u_1|(h^t, cp, n), \mathbf{D}]) + \rho E^{\sigma',\mu'}[u_1|(h^t, cn, n), \mathbf{D}])$$

following σ'. By assumption, after round t, $\mathcal{P}_1$ contributes only if pestered in the prior round. If $\mathcal{P}_1$ is never pestered again, then $\mathcal{P}_1$'s continuation utility from round t on is 0. Suppose then that $\mathcal{P}_1$ is pestered again in some round $k < t$; let $h^k = (h^t, np, b_1^{t-1}a_2^{t-1}, \ldots, a_1^k b_2^k)$ and $(h')^k = (h^t, cp, a_1^{t-1}a_2^{t-1}, \ldots, a_1^k b_2^k)$, where $b_1^{t-1} = c$, $b_2^k = p$, and $a_i^j = n$ for all $i \in \{1, 2\}$ and $k \leq j < t$.

By Lemma 4.6, $\mu_1^*(\mathbf{D}|h^k) = \mu_1^*(\mathbf{D}|(h')^k)$. It follows that $\mathcal{P}_1$, in maximizing utility, plays the same actions and expects the same responses following σ^* and σ'. Thus, the continuation utility from either strategy is the same, and thus $E^{\sigma^*,\mu^*}[u_1|h^k, \mathbf{D}] = E^{\sigma',\mu'}[u_1|(h')^k, \mathbf{D}]$.

A similar argument gives us

$$E^{\sigma^*,\mu^*}[u_1|(h^t, nn, b_1^{t-1}a_2^{t-1}, \ldots, a_1^k b_2^k), \mathbf{D}] = E^{\sigma',\mu'}[u_1|(h^t, cn, a_1^{t-1}a_2^{t-1}, \ldots, a_1^k, b_2^k), \mathbf{D}]$$

Thus, comparing $E^{\sigma^*,\mu^*}[u_1|h^t, \mathbf{D}]$ and $E^{\sigma',\mu'}[u_1|h^t, \mathbf{D}]$, we have

$$E^{\sigma^*,\mu^*}[u_1|h^t, \mathbf{D}] - E^{\sigma',\mu'}[u_1|h^t, \mathbf{D}] \leq -(1-\rho)^2 r_p < 0$$

Finally, it can be easily verified through similar arguments that against a Byzantine player, the expected utility of playing either σ^* or σ' is exactly the same. Thus, we have $E^{\sigma^*,\mu^*}[u_1|h^t] < E^{\sigma',\mu'}[u_1|h^t]$, contradicting the assumption that (σ^*, μ^*) is an equilibrium.

We have proven the lemma if we assume that $\mathcal{P}_1$ does not contribute in some future round $t' > t$ unless pestered in round $t' + 1$. We now finish the proof of the original lemma using induction; all we need is to show is that this assumption holds.

Base case: $t = \tau + 1$. By Lemma 4.5, $\mathcal{P}_1$ never contributes in the future.

Inductive step. Assume true for round t, $\tau + 1 \leq t \leq t_0$; we prove our lemma for round $t = t_0 + 1$. By the inductive hypothesis, $\mathcal{P}_1$ will not contribute in future rounds unless pestered in the prior round. □

We can now prove that $\mathcal{P}_1$ contributes only when pestered.

THEOREM 4.9. *Let t be the current round, where $t < T$. Suppose that $\mathcal{P}_1$ observed nothing from $\mathcal{P}_2$ in round $t + 1$. Then $\mathcal{P}_1$ prefers to do nothing in round t if* (3) *holds.*

Proof. By Theorem 4.14 (proved later) and Lemma 4.8. □

Result: $\mathcal{P}_1$ is as likely to contribute when pestered. We prove that a rational $\mathcal{P}_1$ is more likely to contribute when $\mathcal{P}_2$ pesters. This result is previously used in proving Lemma 4.3.

LEMMA 4.10. *Suppose that $\mathcal{P}_1$ contributes iff $\mathcal{P}_1$'s belief that $\mathcal{P}_2$ is destitute is at least some threshold $\bar{\mu}_1^t$. Let $h^k = (h^t, a_1^t p, \ldots, a_1^{k+1} a_2^{k+1})$ and $(h')^k = (h^t, a_1^t n, b_1^{t-1} a_2^{t-1}, \ldots, b_1^{k+1} a_2^{k+1})$ be two possible round-k histories from $\mathcal{P}_1$'s perspective, where $k \leq t$. Then either:*

1. *$\mu_1(\boldsymbol{D}|h^k) \geq \mu_1(\boldsymbol{D}|(h')^k)$ and h^k contains at least as many c's as $(h')^k$,* i.e., *for $k < i \leq t$, $|\{a_1^i | a_1^i = c\}| = |\{b_1^i | b_1^i = c\}|$; or*
2. *h^k contains more c's than $(h')^k$,* i.e., *for $k < i \leq t$, $|\{a_1^i | a_1^i = c\}| > |\{b_1^i | b_1^i = c\}|$.*

Proof. By induction on k.

Base case: $k = t$. Then since $\mathcal{P}_1$ expects that destitute $\mathcal{P}_2$ always pester, whereas Byzantine $\mathcal{P}_2$ may not, then

$$\begin{aligned}
\mu_1(\mathbf{D}|(h^t, a_1^t p)) &= \frac{(1-\rho)\mu_1(\mathbf{D}|(h^t, a_1^t))}{(1-\rho)\mu_1(\mathbf{D}|(h^t, a_1^t)) + (1-\beta_t)\mu_1(\mathbf{B}|(h^t, a_1^t))} \\
&\geq \frac{\rho\mu_1(\mathbf{D}|(h^t, a_1^t))}{\rho\mu_1(\mathbf{D}|(h^t, a_1^t)) + \mu_1(\neg\mathbf{D}|(h^t, a_1^t)) + \beta_t\mu_1(\mathbf{B}|(h^t, a_1^t))} = \mu_1(\mathbf{D}|(h^t, a_1^t n))
\end{aligned}$$

Inductive step. Assume true for all $k = t_0 \leq t$; we prove the lemma for $k = t_0 - 1$. By the inductive hypothesis, we know that either:

$\mu_1(\mathbf{D}|h^{t_0}) \geq \mu_1(\mathbf{D}|(h')^{t_0})$ and h^{t_0} contains at least as many c's as $(h')^{t_0}$. If $\mathcal{P}_1$ prefers to contribute following $(h')^{t_0}$, then $\mathcal{P}_1$ must also prefer to contribute following h^{t_0}. Similarly, if $\mathcal{P}_1$ prefers to do nothing following h^{t_0}, $\mathcal{P}_1$ must also prefer to do nothing following $(h')^{t_0}$. In either case, $\mu_1(\mathbf{D}|(h^{t_0}, a_1^{t_0} a_2^{t_0})) \geq \mu_1(\mathbf{D}|((h')^{t_0}, a_1^{t_0} a_2^{t_0}))$.

If $\mathcal{P}_1$ only prefers to contribute following h^{t_0}, then $(h^{t_0}, c a_2^{t_0})$ has more contributions than $((h')^{t_0}, n a_2^{t_0})$.

h^{t_0} contains more c's than $(h')^{t_0}$. If the number of c's in h^{t_0} exceeds the number in $(h')^{t_0}$ by more than one, then even if $\mathcal{P}_1$ contributes following $(h')^{t_0}$ and not h^{t_0}, then $((h')^{t_0}, c a_2^{t_0})$ still has fewer contributions than $(h^{t_0}, n a_2^{t_0})$. Also, if $\mathcal{P}_1$ prefers to contribute (or do nothing) following both

h^{t_0} and $(h')^{t_0}$, then $(h^{t_0}, a_1^{t_0} a_2^{t_0})$ still contains more contributions than $((h')^{t_0}, a_1^{t_0} a_2^{t_0})$. Either way, the inductive step is trivially proven.

Thus, suppose that h^{t_0} has only one more c than $(h')^{t_0}$ and $\mathcal{P}_1$ prefers to contribute after $(h')^{t_0}$ but not h^{t_0}. Since $\mathcal{P}_1$ prefers to contribute in $(h')^{t_0}$, by Lemma 4.7, this implies that $\mu_1(\neg\mathbf{D}|h^{t_0}) = \mu_1(\neg\mathbf{D}|(h')^{t_0}) = 0$. Let $\bar{t}$ be the minimum (latest) round such that $|\{a_1^i | a_1^i = c\}| = |\{b_1^i | b_1^i = c\}|$ for $\bar{t} < i \leq t$, and let $h^{\bar{t}} = (h^t, a_1^t p, a_1^{t-1} a_2^{t-1}, \ldots, a_1^{\bar{t}+1} a_2^{\bar{t}+1})$ and $(h')^{\bar{t}} = (h^t, a_1^t n, b_1^{t-1} a_2^{t-1}, \ldots, b_1^{\bar{t}+1} a_2^{\bar{t}+1})$.

Since $\bar{t} > t_0$, by the inductive hypothesis, $\mu_1(\mathbf{D}|h^{\bar{t}}) \geq \mu_1(\mathbf{D}|(h')^{\bar{t}})$ since $h^{\bar{t}}$ and $(h')^{\bar{t}}$ have the same number of contributions. Let $\hat{P}$ be the rounds in which pestering is observed in h^{t_0} following $h^{\bar{t}}$, $\hat{N}$ be the rounds in which doing nothing is observed in h^{t_0} following $h^{\bar{t}}$, and m_c be the number of rounds in which $\mathcal{P}_1$ contributes in h^{t_0} following $h^{\bar{t}}$. Then we know that

$$\mu_1(\mathbf{D}|(h^{t_0}, n)) = \rho^{m_c} \frac{\mu_1(\mathbf{D}|h^{\bar{t}})(1-\rho)^{|\hat{P}|}\rho^{|\hat{N}|}}{\mu_1(\mathbf{D}|h^{\bar{t}})(1-\rho)^{|\hat{P}|}\rho^{|\hat{N}|} + \mu_1(\mathbf{B}|h^{\bar{t}}) \prod_{i \in \hat{P}}(1-\beta_i) \prod_{i \in \hat{N}} \beta_i}$$

whereas

$$\mu_1(\mathbf{D}|((h')^{t_0}, c)) = \rho^{m_c} \frac{\mu_1(\mathbf{D}|(h')^{\bar{t}})(1-\rho)^{|\hat{P}|}\rho^{|\hat{N}|}}{\mu_1(\mathbf{D}|(h')^{\bar{t}})(1-\rho)^{|\hat{P}|}\rho^{|\hat{N}|} + \mu_1(\mathbf{B}|(h')^{\bar{t}}) \prod_{i \in \hat{P}}(1-\beta_i) \prod_{i \in \hat{N}} \beta_i}$$

It can be shown that $\mu_1(\mathbf{D}, (h^{t_0}, n a_2^{t_0})) \geq \mu_1(\mathbf{D}, ((h')^{t_0}, c a_2^{t_0}))$ follows. □

Result: $\mathcal{P}_1$ (sometimes) contributes if pestered. Up to now, we have proven that if a belief threshold exists, then $\mathcal{P}_1$ contributes only if pestered in the previous round. We now prove that such a threshold exists in every round except near the end of the game (Lemma 4.5 and Theorem 4.14); one simply needs to check whether beliefs fall above or below the threshold to determine whether it is in $\mathcal{P}_1$'s best interest to contribute.

We first observe that Bayes rule, which maps prior beliefs to posterior beliefs, is continuous: given some prior belief and its posterior beliefs, after any series of signals, we can find (infinitely many) other prior beliefs which, after the same sequence of signals, map to posterior beliefs that are arbitrarily close to the original posterior belief.

LEMMA 4.11. *Let h^t be the current history and $\mu_1(\theta|h^t)$ be $\mathcal{P}_1$'s belief. Then for all $\epsilon > 0$, there exists some $\delta > 0$ such that for all beliefs $\mu_1'(\theta|(h')^t)$ where $0 \leq \mu_1'(\boldsymbol{D}|(h')^t) - \mu_1(\boldsymbol{D}|h^t) < \delta$ and $0 < \mu_1'(\neg\boldsymbol{D}|(h')^t) - \mu_1(\neg\boldsymbol{D}|h^t) < \delta$,*

$$0 \leq \mu_1'(\boldsymbol{D}|((h')^t, a_1^t a_2^t)) - \mu_1(\boldsymbol{D}|(h^t, a_1^t a_2^t)) < \epsilon$$

and

$$0 \leq \mu_1'(\neg\boldsymbol{D}|((h')^t, a_1^t a_2^t)) - \mu_1(\neg\boldsymbol{D}|(h^t, a_1^t a_2^t)) < \epsilon$$

Proof. Recall that a destitute $\mathcal{P}_2$ always pesters, a non-destitute $\mathcal{P}_2$never pesters, and a Byzantine $\mathcal{P}_2$ pesters with some probability independent of the history. Letting $\hat{\sigma}_\theta(a_2^t|(h^t, a_1^t))$ be the probability of observing a type-θ $\mathcal{P}_2$ play a_2^t after history (h^t, a_1^t), we have

$$\begin{aligned} \epsilon &> \mu_1'(\mathbf{D}|((h')^t, a_1^t a_2^t)) - \mu_1(\mathbf{D}|(h^t, a_1^t a_2^t)) \\ &= \frac{\mu_1'(\mathbf{D}|((h')^t, a_1^t))\hat{\sigma}_{\mathbf{D}}(a_2^t|((h')^t, a_1^t))}{\sum_{\theta \in \{\mathbf{D}, \neg\mathbf{D}, \mathbf{B}\}} \mu_1'(\theta|((h')^t, a_1^t))\hat{\sigma}_\theta(a_2^t|((h')^t, a_1^t))} - \frac{\mu_1(\mathbf{D}|(h^t, a_1^t))\hat{\sigma}_{\mathbf{D}}(a_2^t|(h^t, a_1^t))}{\sum_{\theta \in \{\mathbf{D}, \neg\mathbf{D}, \mathbf{B}\}} \mu_1(\theta|(h^t, a_1^t))\hat{\sigma}_\theta(a_2^t|(h^t, a_1^t))} > 0 \end{aligned}$$

Consider the two signals that $\mathcal{P}_1$ may observe from $\mathcal{P}_2$.

Case 1. $a_2 = n$. Then we need

$$\epsilon > \frac{\mu_1'(\mathbf{D}|((h')^t, a_1^t))\rho}{\mu_1'(\mathbf{D}|((h')^t, a_1^t))\rho + \mu_1'(\neg\mathbf{D}|((h')^t, a_1^t)) + \mu_1'(\mathbf{B}|((h')^t, a_1^t)))\beta_t} - \frac{\mu_1(\mathbf{D}|(h^t, a_1^t))\rho}{\mu_1(\mathbf{D}|(h^t, a_1^t))\rho + \mu_1(\neg\mathbf{D}|(h^t, a_1^t)) + \mu_1(\mathbf{B}|(h^t, a_1^t)))\beta_t}$$

and

$$\epsilon > \frac{\mu_1'(\neg\mathbf{D}|((h')^t, a_1^t))}{\mu_1'(\mathbf{D}|((h')^t, a_1^t))\rho + \mu_1'(\neg\mathbf{D}|((h')^t, a_1^t)) + \mu_1'(\mathbf{B}|((h')^t, a_1^t)))\beta_t} - \frac{\mu_1(\neg\mathbf{D}|(h^t, a_1^t))}{\mu_1(\mathbf{D}|(h^t, a_1^t))\rho + \mu_1(\neg\mathbf{D}|(h^t, a_1^t)) + \mu_1(\mathbf{B}|(h^t, a_1^t)))\beta_t}$$

Since $\mu_1'(\mathbf{D}|(h')^t) \geq \mu_1(\mathbf{D}|h^t)$ and $\mu_1'(\neg\mathbf{D}|(h')^t) \geq \mu_1(\neg\mathbf{D}|h^t)$, it suffices to show that

$$\epsilon > \frac{\beta_t}{\rho}(\mu_1'(\mathbf{D}|(h')^t) - \mu_1(\mathbf{D}|h^t))$$

and

$$\epsilon > \frac{\beta_t}{\rho^2}(\mu_1'(\neg\mathbf{D}|(h')^t) - \mu_1(\neg\mathbf{D}|h^t))$$

Let $\delta_n = \epsilon\rho^2/\beta_t$.

Case 2. $a_2^t = p$. Since a non-destitute player never pesters, $\epsilon > \mu_1'(\neg\mathbf{D}|((h')^t, a_1^t p)) - \mu_1(\neg\mathbf{D}|(h^t, a_1^t p)) = 0$ which is trivially satisfied for any δ. For the destitute type, we need

$$\epsilon > \frac{\mu_1'(\mathbf{D}|((h')^t, a_1^t))(1-\rho)}{\mu_1'(\mathbf{D}|((h')^t, a_1^t))(1-\rho) + (1 - \mu_1'(\mathbf{D}|((h')^t, a_1^t)))(1-\beta_t)} - \frac{\mu_1(\mathbf{D}|(h^t, a_1^t))(1-\rho)}{\mu_1(\mathbf{D}|(h^t, a_1^t))(1-\rho) + (1 - \mu_1(\mathbf{D}|(h^t, a_1^t)))(1-\beta_t)}$$

If $\beta_t = 1$, we end up with $\epsilon > 0$, which is trivially satisfied. Thus, assume that $\beta_t < 1$. It is sufficient to show that

$$\epsilon > \frac{1-\rho}{1-\beta_t}(\mu_1'(\mathbf{D}|(h')^t) - \mu_1(\mathbf{D}|h^t))$$

Let $\delta_p = \epsilon(1-\beta_t)/(1-\rho) > \mu_1'(\mathbf{D}|(h')^t) - \mu_1(\mathbf{D}|h^t)$.

It is easy to show that $\delta = \min(\delta_n, \delta_p)$ satisfies the necessary conditions. □

Lemma 4.12. *Let t be the current round, h^t be the current history, and $\mu_1(\theta|h^t)$ be $\mathcal{P}_1$'s belief. Then for all $\epsilon > 0$ and k, $0 < k \leq t$, there exists some $\delta > 0$ such that for all beliefs $\mu_1'(\theta|(h')^t)$, where $0 \leq \mu_1'(\boldsymbol{D}|(h')^t) - \mu_1(\boldsymbol{D}|h^t) < \delta$ and $0 \leq \mu_1'(\neg\boldsymbol{D}|(h')^t) - \mu_1(\neg\boldsymbol{D}|h^t) < \delta$,*

$$0 \leq \mu_1'(\boldsymbol{D}|((h')^t, a_1^t a_2^t, \ldots, a_1^k a_2^k)) - \mu_1(\boldsymbol{D}|(h^t, a_1^t a_2^t, \ldots, a_1^k a_2^k)) < \epsilon$$

and

$$0 \leq \mu_1'(\neg\boldsymbol{D}|((h')^t, a_1^t a_2^t, \ldots, a_1^k a_2^k)) - \mu_1(\neg\boldsymbol{D}|(h^t, a_1^t a_2^t, \ldots, a_1^k a_2^k)) < \epsilon$$

Proof. By induction on k.

Base case: $k = t$. By Lemma 4.11.

Inductive step. Assume true for all $k = t_0 \leq t$; we now prove it true for $k = t_0 - 1$. By Lemma 4.11, we know that in round t_0, for any history h^{t_0}, associated belief $\mu_1(\theta|h^{t_0})$, and $\epsilon > 0$, we can find a ϵ' such that for all beliefs $\mu_1'(\theta|(h')^{t_0})$ where $0 \leq \mu_1'(\mathbf{D}|(h')^{t_0}) - \mu_1(\mathbf{D}|h^{t_0}) < \epsilon'$ and $0 \leq \mu_1'(\neg\mathbf{D}|(h')^{t_0}) - \mu_1(\neg\mathbf{D}|h^{t_0}) < \epsilon'$,

$$0 \leq \mu_1'(\mathbf{D}|((h')^{t_0-1}, a_1^{t_0-1}a_2^{t_0-1})) - \mu_1(\mathbf{D}|(h^{t_0-1}, a_1^{t_0-1}a_2^{t_0-1})) < \epsilon$$

and

$$0 \leq \mu_1'(\neg\mathbf{D}|((h')^{t_0-1}, a_1^{t_0-1}a_2^{t_0-1})) - \mu_1(\neg\mathbf{D}|(h^{t_0-1}, a_1^{t_0-1}a_2^{t_0-1})) < \epsilon$$

By the inductive hypothesis, there exists some $\delta > 0$ such that for all beliefs $\mu_1'(\theta|h^t)$ where $0 \leq \mu_1'(\mathbf{D}|h^t) - \mu_1(\mathbf{D}|h^t) < \delta$,

$$0 \leq \mu_1'(\mathbf{D}|((h')^t, a_1^t a_2^t, \ldots, a_1^{t_0}a_2^{t_0})) - \mu_1(\mathbf{D}|(h^t, a_1^t a_2^t, \ldots, a_1^{t_0}a_2^{t_0})) < \epsilon'$$

and

$$0 \leq \mu_1'(\neg\mathbf{D}|((h')^t, a_1^t a_2^t, \ldots, a_1^{t_0}a_2^{t_0})) - \mu_1(\neg\mathbf{D}|(h^t, a_1^t a_2^t, \ldots, a_1^{t_0}a_2^{t_0})) < \epsilon'$$

Chaining these two together gives us a δ and ϵ that fulfill the needed conditions. □

We need one more important result: if $\mathcal{P}_1$ plays an equilibrium strategy and $\mathcal{P}_2$ is Byzantine, $\mathcal{P}_1$'s expected utility from contributing is no more than that from doing nothing.

Lemma 4.13. *Let t be the current round and h^t be $\mathcal{P}_1$'s current history, such that $t > \tau$. Assume that there exists a belief threshold for rounds $t-1$ through 0. Then for any equilibrium (σ^*, μ^*),*

$$-s_c + E^{\sigma^*,\mu^*}[u_1|(h^t, c), \boldsymbol{B}] \leq E^{\sigma^*,\mu^*}[u_1|(h^t, n), \boldsymbol{B}]$$

Intuition. Since a Byzantine $\mathcal{P}_2$'s likelihood of pestering is independent of contribution, $\mathcal{P}_1$ never does better contributing if $\mathcal{P}_1$'s peer is Byzantine.

Proof. Suppose that $-s_c + E^{\sigma^*,\mu^*}[u_1|(h^t, c), \mathbf{B}] > E^{\sigma^*,\mu^*}[u_1|(h^t, n), \mathbf{B}]$. Since a Byzantine $\mathcal{P}_2$'s actions are independent of $\mathcal{P}_1$'s actions, this implies that if $\mathcal{P}_1$ expects to contribute n times in continuation from contributing in round t, $\mathcal{P}_1$ expects to contribute more than $n+1$ times in continuation from doing nothing in round t. It follows that there must exist two histories $h^k = (h^t, b_1^t a_2^t, a_1^{t-1}a_2^{t-1}, \ldots a_1^{k+1}, a_2^{k+1})$ and $(h')^k = (h^t, a_1^t a_2^t, \ldots, a_1^{k+1}a_2^{k+1})$ such that $k < t$, $a_1^t = n$, $b_1^t = c$, and $\mathcal{P}_1$ prefers to contribute following $(h')^k$ but not h^k.

Following this contribution, by Lemma 4.8, $\mathcal{P}_1$ does not prefer contributing again unless pestered. If $\mathcal{P}_1$ is never pestered again, then $\mathcal{P}_1$ never contributes again in either history. Otherwise, if $\mathcal{P}_1$ is next pestered in some round $\ell \leq k$, then we know by Lemma 4.6 that

$$\mu_1(\mathbf{D}|((h')^k, b_1^k a_2^k, a_1^{k-1}a_2^{k-1}, \ldots, a_1^\ell p) = \mu_1(\mathbf{D}|(h^k, a_1^k a_2^k, \ldots, a_1^\ell p))$$

where $b_1^k = c$ and $a_i^j = n$ for $i \in \{1, 2\}$ and $\ell \leq j \leq k$. It follows that $\mathcal{P}_1$ plays the same exact strategy from this point on in either history. Consequently, $\mathcal{P}_1$, starting from any history (h^t, n), contributes at most once before returning to the same strategy that would have been played if

$\mathcal{P}_1$ had started from (h^t, c). It follows that the expected number of additional contributions after doing nothing cannot exceed 1, thus

$$-s_c + E^{\sigma^*,\mu^*}[u_1|(h^t, c), \mathbf{B}] \leq E^{\sigma^*,\mu^*}[u_1|(h^t, n), \mathbf{B}]$$

contradicting the original assumption. □

We now show the existence of a belief threshold.

Theorem 4.14. *Let τ be as defined in (2); t be the current round, where $\tau < t \leq T$; and h^t be the current history. Then there exists some threshold $\bar{\mu}_1^t \leq 1$ such that:*

1. *If $\mu_1(\boldsymbol{D}|h^t) \geq \bar{\mu}_1^t$, $\mathcal{P}_1$ prefers to contribute; and*
2. *If $\mu_1(\boldsymbol{D}|h^t) < \bar{\mu}_1^t$, $\mathcal{P}_1$ prefers to do nothing.*

Proof. By induction.

Base case: $t = \tau + 1$. Let (σ^*, μ^*) be some equilibrium. $\mathcal{P}_1$ prefers to contribute iff

$$-s_c + E^{\sigma^*,\mu^*}[u_1|(h^t, c)] \geq E^{\sigma^*,\mu^*}[u_1|(h^t, n)]$$

Since $\mathcal{P}_1$ never prefers to contribute afterwards, we have

$$-s_c \geq E^{\sigma^*,\mu^*}[u_1|(h^t, n)] - E^{\sigma^*,\mu^*}[u_1|(h^t, c)] = -(1-\rho)^2 \mu_1^*(\mathbf{D}|h^t) \tau r_p$$

Solving for $\mu_1^*(\mathbf{D}|h^t)$, we have

$$\mu_1^*(\mathbf{D}|h^t) \geq \frac{1}{(1-\rho)^2 \tau} \frac{s_c}{r_p} = \bar{\mu}_1^{\tau+1}$$

Inductive step. Assume true for all t, $\tau < t \leq t_0$; we prove $t = t_0 + 1$ by contradiction. If a threshold does not exist, there must be some beliefs (in $\mathcal{P}_2$ being destitute) in which $\mathcal{P}_1$ prefers to contribute and higher beliefs in which $\mathcal{P}_1$ prefers to do nothing, *i.e.*, there must exist some η_1^t and $\delta > 0$ such that either:

1. For $\mu_1(\mathbf{D}|h^t) = \eta_1^t$, $\mathcal{P}_1$ prefers to do contribute, and for $\mu_1(\mathbf{D}|h^t)$ such that $\eta_1^t < \mu_1(\mathbf{D}|h^t) < \eta_1^t + \delta$, $\mathcal{P}_1$ prefers to do nothing; or

2. For $\mu_1(\mathbf{D}|h^t)$ such that $\eta_1^t - \delta < \mu_1(\mathbf{D}|h^t) < \eta_1^t$, $\mathcal{P}_1$ prefers to contribute, and for $\mu_1(\mathbf{D}|h^t) = \eta_1^t$, $\mathcal{P}_1$ prefers to do nothing.

We only consider the first case, as the proof of the other case is very similar. Let (σ^*, μ^*) be an equilibrium such that $\mu_1^*(\mathbf{D}|h^t) = \eta_1^t$. By the inductive hypothesis, there exists a threshold $\bar{\mu}_1^i$ for all rounds $i < t$. Let

$$\epsilon = \min_{0 \leq i < t} \left\{ \bar{\mu}_1^i - \mu_1^*(\mathbf{D}|(h^t, a_1^t a_2^t, \ldots, a_1^{i+1} a_2^{i+1})) \mid \bar{\mu}_1^i > \mu_1^*(\mathbf{D}|(h^t, a_1^t a_2^t, \ldots, a_1^{i+1} a_2^{i+1})) \right\}$$

ϵ represents the minimum difference between $\mathcal{P}_1$'s belief and the threshold at any future round when $\mathcal{P}_1$'s belief in $\mathcal{P}_2$ being destitute is less than that round's threshold.

By Lemma 4.12, we know that we can find some belief $\mu_1'(\theta|(h')^t)$ such that for $0 \leq \mu_1'(\mathbf{D}|(h')^t) - \mu_1^*(\mathbf{D}|h^t) < \delta$, $0 \leq \mu_1'(\neg\mathbf{D}|(h')^t) - \mu_1^*(\neg\mathbf{D}|h^t) < \delta$, and $0 \leq i < t$,

$$0 \leq \mu_1'(\mathbf{D}|((h')^t, a_1^t a_2^t, \ldots, a_1^{i+1} a_2^{i+1})) - \mu_1^*(\mathbf{D}|(h^t, a_1^t, a_2^t, \ldots, a_1^{i+1} a_2^{i+1})) < \epsilon$$

Thus, for any round $i < t$, $\mu_1^*(\mathbf{D}|h^t, a_1^t a_2^t, \ldots, a_1^{i+1} a_2^{i+1})) < \bar{\mu}_1^i$ iff $\mu_1'(\mathbf{D}|((h')^t, a_1^t a_2^t, \ldots, a_1^{i+1} a_2^{i+1})) < \bar{\mu}_1^i$. It follows that a rational $\mathcal{P}_1$ with belief $\mu_1^*(\theta|h^t)$, upon observing some non-empty sequence of signals after h^t, prefers to play the same action as if $\mathcal{P}_1$ held the belief $\mu_1'(\theta|(h')^t)$ and observed the same non-empty sequence of signals after $(h')^t$. Given that $\mathcal{P}_2$ is of type θ, $\mathcal{P}_1$'s expected utility of playing action a_1^t followed by the optimal strategy σ^* with either belief $\mu_1^*(\theta|h^t)$ or $\mu_1'(\theta|(h')^t)$ must be equal; let $V(a_1^t, \theta)$ be this expected continuation utility:

$$V(a_1^t, \theta) = E^{\sigma^*, \mu^*}[u_1|(h^t, a_1^t), \theta] = E^{\sigma^*, \mu'}[u_1|((h')^t, a_1^t), \theta]$$

Given belief μ_1^*, $\mathcal{P}_1$ prefers to contribute during round t; by Lemma 4.7, this implies that $\mu_1^*(\neg\mathbf{D}|h^t) = 0$ and thus $\mu_1^*(\mathbf{B}|h^t) = 1 - \mu_1^*(\mathbf{D}|h^t)$. However, since $\mu_1^*(\mathbf{D}|h^t) = \eta_1^t < \mu_1'(\mathbf{D}|(h')^t) < \eta_1^t + \delta$, then given belief μ_1', $\mathcal{P}_1$ prefers to do nothing during round $t_0 + 1$:

$$-s_c + E^{\sigma^*, \mu^*}[u_1|(h^t, c)] \geq E^{\sigma^*, \mu^*}[u_1|(h^t, n)]$$
$$-s_c + E^{\sigma^*, \mu'}[u_1|((h')^t, c)] < E^{\sigma^*, \mu'}[u_1|((h')^t, n)]$$

We know that

$$E^{\sigma^*, \mu}[u_1|(h, a_1^t)] = \mu_1(\mathbf{B}|h)V(a_1^t, \mathbf{B}) + \mu_1(\mathbf{D}|h)V(a_1^t, \mathbf{D}) + \mu_1(\neg\mathbf{D}|h)V(a_1^t, \neg\mathbf{D})$$

for $\mu \in \{\mu^*, \mu'\}$, $h \in \{h^t, (h')^t\}$, and $a_1^t \in \{c, n\}$. By Lemma 4.1, a non-destitute $\mathcal{P}_2$ stops pestering if $\mathcal{P}_2$ has the contribution. By Lemma 4.8 and the inductive hypothesis (there exists a threshold starting from round t_0), $\mathcal{P}_1$ never contributes unless pestered; this gives us $V(a_1^t, \neg\mathbf{D}) = 0$. Combining the last two groups of expressions and moving terms around, we have

$$\mu_1^*(\mathbf{B}|h^t)(V(n, \mathbf{B}) - V(c, \mathbf{B}) + s_c) \leq \mu_1^*(\mathbf{D}|h^t)(-s_c + \rho V(c, \mathbf{D}) - V(n, \mathbf{D}))$$
$$\mu_1'(\mathbf{B}|(h')^t)(V(n, \mathbf{B}) - V(c, \mathbf{B}) + s_c) > \mu_1'(\mathbf{D}|(h')^t)(-s_c + \rho V(c, \mathbf{D}) - V(n, \mathbf{D}))$$

By Lemma 4.13 and the inductive hypothesis, $V(n, \mathbf{B}) - V(c, \mathbf{B}) + s_c \geq 0$. If $V(n, \mathbf{B}) - V(c, \mathbf{B}) + s_c = 0$, an immediate contradiction arises:

$$-s_c + \rho V(c, \mathbf{D}) - V(n, \mathbf{D})) < 0 \leq -s_c + \rho V(c, \mathbf{D}) - V(n, \mathbf{D})$$

Thus, assuming that $V(n, \mathbf{B}) - V(c, \mathbf{B}) + s_c > 0$, we have

$$\frac{\mu_1^*(\mathbf{B}|h^t)}{\mu_1^*(\mathbf{D}|h^t)} \leq \frac{-s_c + \rho V(c, \mathbf{D}) - V(n, \mathbf{D})}{V(n, \mathbf{B}) - V(c, \mathbf{B}) + s_c} < \frac{\mu_1'(\mathbf{B}|(h')^t)}{\mu_1'(\mathbf{D}|(h')^t)}$$

However, $\mu_1'(\mathbf{D}|(h')^t) > \mu_1^*(\mathbf{D}|h^t)$ and

$$\mu_1^*(\mathbf{B}|h^t) = 1 - \mu_1^*(\mathbf{D}|h^t) > 1 - \mu_1'(\mathbf{D}|(h')^t) = \mu_1'(\mathbf{B}|(h')^t) + \mu_1'(\neg\mathbf{D}|(h')^t) \geq \mu_1'(\mathbf{B}|(h')^t)$$

Contradiction. □

5 Working the numbers

To gain a better understanding for the implications of Section 4 on the design of MAD cooperative services, we explore, through simulation, the parameter space for which its claims hold. We are especially interested in answering four questions: 1) What percentage of altruistic peers

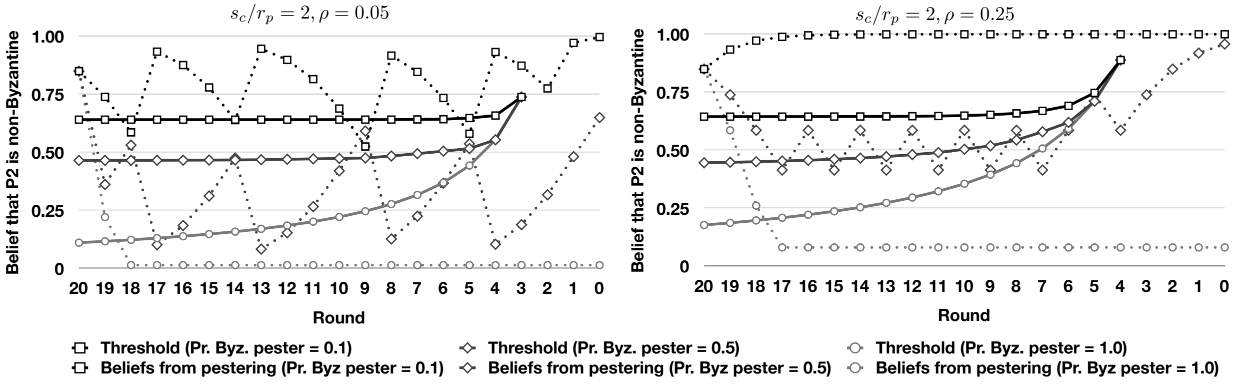

Figure 1: The belief thresholds and behavior of $\mathcal{P}_1$'s beliefs. Solid lines represent the threshold; dotted lines represent how an initial belief of $1 - \mu_1(\mathbf{B}) = 0.85$ evolves if $\mathcal{P}_1$ is continually pestered.

is sufficient to motivate a rational peer to pester—and what percentage of Byzantine peers does it take to discourage it? 2) If all non-Byzantine peers are motivated to pester, what percentage of Byzantine peers is required to discourage a rational peer from contributing? 3) How sensitive is cooperation to the generosity of altruistic nodes? and 4) How sensitive is cooperation to the number of rounds in the last exchange game?

Motivating $\mathcal{P}_2$. Figure 2 shows the conditions under which a rational $\mathcal{P}_2$ is willing to pester for the entire game (save the last round). We assume that an altruistic $\mathcal{P}_1$ contributes in round T and then with some fixed probability in the following rounds; a Byzantine $\mathcal{P}_1$ never contributes. Finally, we assume that a rational $\mathcal{P}_1$'s beliefs are such that $\mathcal{P}_1$ is willing to contribute in round T. To simplify our simulation, for tHe remainder of the game we assume that $\mathcal{P}_1$ never contributes. This is of course a worst case and leads to conservative estimates for the altruism necessary to motivate $\mathcal{P}_2$; in reality, we expect the threshold of altruism required to be lower than we report.

Not surprisingly, a larger ratio between the benefit of a contribution and the cost of pestering lowers the initial belief in the altruism of $\mathcal{P}_1$ sufficient to motivate $\mathcal{P}_2$ and increases $\mathcal{P}_2$'s resolve to pester in the face of higher odds that $\mathcal{P}_1$ may be Byzantine.

Motivating $\mathcal{P}_1$. We numerically compute the belief threshold for which $\mathcal{P}_1$ is motivated to contribute. We assign $\mathcal{P}_1$ some initial belief on the probability that $\mathcal{P}_2$ is not Byzantine and construct the entire game tree to determine whether that initial belief is sufficient to motivate $\mathcal{P}_1$ to contribute at the beginning of the game. Through a binary search, we can identify the threshold with an accuracy of 10^{-9}.

The solid lines in Figure 1 show how the threshold changes as a function of the number of rounds and different system parameters. As expected, the solid lines stop before the end of the game when the remaining cost of pestering is not enough to overcome the cost of contributing. The dotted lines in the figure represent how beliefs evolve if $\mathcal{P}_1$, starting with $1 - \mu_1(\mathbf{B}) = 0.85$, is continually pestered throughout the game. $\mathcal{P}_1$ contributes whenever the dotted line exceeds the corresponding threshold.

The maximum number of contributions, which varied from 2 to 17 times, increases when ρ increases, s_c/r_p decreases, and if Byzantine $\mathcal{P}_2$ are less likely to pester. Whenever $\mathcal{P}_1$ contributes and is subsequently pestered, then $\mathcal{P}_1$'s belief in $\mathcal{P}_2$ being non-Byzantine almost always drops. If $\mathcal{P}_1$ does nothing and is pestered in subsequent rounds, then the effect on $\mathcal{P}_1$'s beliefs depends on

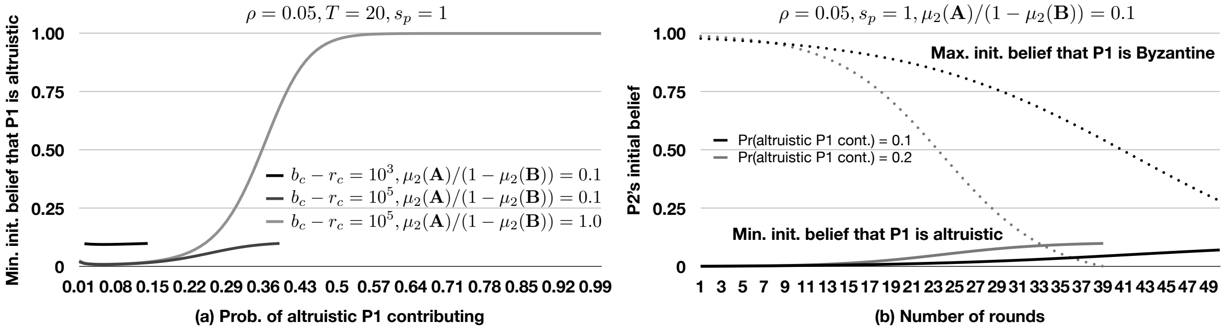

Figure 2: Sufficient bounds on the initial beliefs to incentivize $\mathcal{P}_2$ to pester for varying amounts of altruistic generosity (left) and numbers of rounds (right).

$\mathcal{P}_1$'s expectation of what Byzantine $\mathcal{P}_2$ would do. If a Byzantine $\mathcal{P}_2$ always pesters, then observing a pester gives no clue as to $\mathcal{P}_2$'s type. If a Byzantine $\mathcal{P}_2$ rarely pesters, however, then observing a pester increases $\mathcal{P}_1$'s belief that $\mathcal{P}_2$ non-Byzantine. Increasing network loss also makes $\mathcal{P}_1$ more likely to contribute again when pestered, as it is more likely that the contribution was simply lost on the network. Eventually, if $\mathcal{P}_1$'s beliefs rise beyond the threshold, $\mathcal{P}_1$ contributes again.

Too much of a good thing. Perhaps the most intriguing conclusion to come from Figure 2(a), however, is that altruistic prodigality can make it far more difficult to motivate $\mathcal{P}_2$ to pester. The reason is that the more generous altruistic peers are, the easier it is for a rational $\mathcal{P}_2$ to determine, from observed signals, whether $\mathcal{P}_1$ is altruistic or not, which in turn affects whether $\mathcal{P}_2$ continues to pester. Figure 2 shows that, with network loss at 5%, if there is a more than one-in-three chance that an altruistic $\mathcal{P}_1$ will contribute when pestered, motivating a rational $\mathcal{P}_2$ to pester in every round essentially requires $\mathcal{P}_2$ to believe that all non-Byzantine nodes are altruistic. Altruistic generosity becomes a more obvious discriminant if a Byzantine $\mathcal{P}_1$ never contributes, while it becomes less conspicuous with higher rates of network loss, for the same reason we saw when discussing Figure 1. Still, our results (not shown) show that spurring $\mathcal{P}_2$ to pester when there is two-third chance than an altruistic peer will contribute requires an implausible loss rate of over 25%.

Short is beautiful. Figure 2(b) illustrates how the bounds that determine whether $\mathcal{P}_2$ will pester evolve as a function of the number of rounds. Although a large number of rounds allow $\mathcal{P}_1$ to contribute more times and thereby may increase the chance that $\mathcal{P}_2$ may receive the contribution, we find that $\mathcal{P}_2$'s resolve to continue pestering wanes as the number of rounds increases. In particular, we see that $\mathcal{P}_2$ can be motivated to continue pestering only if Byzantine participants are increasingly unlikely. Indeed, more generous altruistic nodes, who are more likely to contribute when pestered, discourage $\mathcal{P}_2$ more quickly in the absence of a contribution.

6 Related work

Incentive-compatible systems. There has been a lot of work in incentive-compatible systems (*e.g.*, [3, 4, 11, 12, 14, 17]). None of these systems assume the existence of altruistic players, and only a few [3, 14, 17] consider the possibility of Byzantine peers. Our techniques can be applied to many of these systems. For example, BAR Gossip [14], FOX [12], and PropShare [11] can use altruism to incentivize key exchange.

Irrationality in incentive-compatible protocols. Eliaz [7] proposed the generalization of

Nash equilibrium to scenarios where some number of peers may be Byzantine. Aiyer et. al [3] later generalized this to the BAR model, which introduced the possibility of altruistic peers and on which our model is based. Abraham et. al [1] describe (k, t)-robust equilibrium, a strategy which is in a rational player's best interest despite the possibility of collusion by groups up to size k and up to t "irrational" agents that may play any strategy. Similarly, Martin [16] introduces his own equilibrium concept in which rational players do not deviate regardless of what Byzantine or altruistic participants do. Our work differs from previous work by showing the need for altruism to address a key problem in cooperative services and considering real-world issues such as network costs and lossy links.

Finally, Vassilakis et. al [19] study how altruism affects content sharing in P2P services at the application level. Their approach and our own are complementary; we focus on network-level incentives that motivate participants to actually send the content they share at the application level. Moreover, they do not consider the possibility of Byzantine participants or lossy links.

Game-theory. There has been extensive work that has covered imperfect knowledge, private signaling, and the use of altruism in game theory. The use of altruism to achieve cooperation in the finitely-repeated prisoner's dilemma game was first proposed by Kreps et. al [10]. It was shown that reputations could be maintained even when there was imperfect observation of actions [8]. Cripps et. al later showed that, under certain conditions, reputations cannot be maintained forever unless the action played by the irrational player was part of a rational player's equilibrium strategy [5, 6]. None of the previous work consider both the possibility of Byzantine and altruistic players. Many of them also assume that actions or their corresponding signals can either be observed at least publicly [5, 8], if not perfectly [10]. More importantly, the focus of this work is the existence (or nonexistence) of equilibrium under general conditions and thus use models which differ from ours. We focus on the application of theory to a specific problem and a realistic model that we believe to be applicable to many distributed protocols and show how to construct such an cooperative equilibrium that works under realistic conditions.

7 Conclusion

Despite the presence of altruistic peers in real-world MAD systems, little attention has been given to their role in establishing rational cooperation. In this paper, we take the first step in understanding their function by showing that altruism is necessary and sufficient to motivate rational cooperation in the crucial last exchange between MAD peers. Our results suggest that, while a small fraction of altruistic peers is sufficient to spur rational peers into action even in systems with a large fraction of Byzantine peers, overly generous altruistic peers can irreparably harm rational cooperation.

References

[1] I. Abraham, D. Dolev, R. Gonen, and J. Halpern. Distributed computing meets game theory: robust mechanisms for rational secret sharing and multiparty computation. In *Proc. 25th PODC*, pages 53–62, July 2006.

[2] E. Adar and B. A. Huberman. Free riding on Gnutella. *First Monday*, 5(10):2–13, Oct. 2000.

[3] A. S. Aiyer, L. Alvisi, A. Clement, M. Dahlin, J.-P. Martin, and C. Porth. BAR fault tolerance for cooperative services. In *Proc. 20th SOSP*, pages 45–58, Oct. 2005.

[4] B. Cohen. Incentives build robustness in BitTorrent. In *First Workshop on the Economics of Peer-to-Peer Systems*, June 2003.

[5] M. W. Cripps, G. J. Mailath, and L. Samuelson. Imperfect monitoring and impermanent reputations. *Econometrica*, 72(2):407–432, 2004.

[6] M. W. Cripps, G. J. Mailath, and L. Samuelson. Disappearing private reputations in long-run relationships. *Journal of Economic Theory*, 127(1):287–316, May 2007.

[7] K. Eliaz. Fault tolerant implementation. *Review of Economic Studies*, 69:589–610, Aug 2002.

[8] D. Fudenberg and D. K. Levine. Maintaining a reputation when strategies are imperfectly observed. *Review of Economic Studies*, 59(3):561–79, July 1992.

[9] A. Haeberlen, P. Kouznetsov, and P. Druschel. PeerReview: Practical accountability for distributed systems. In *Proc. 21th SOSP*, Oct. 2007.

[10] D. Kreps, P. Milgrom, J. Roberts, and R. Wilson. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27(2):245–52, Aug. 1982.

[11] D. Levin, K. LaCurts, N. Spring, and B. Bhattacharjee. BitTorrent is an auction: analyzing and improving BitTorrent's incentives. *SIGCOMM Comput. Commun. Rev.*, 38(4):243–254, 2008.

[12] D. Levin, R. Sherwood, and B. Bhattacharjee. Fair file swarming with FOX. In *Proc. 5th IPTPS*, Feb 2006.

[13] H. Li, A. Clement, M. Marchetti, M. Kapritsos, L. Robinson, L. Alvisi, and M. Dahlin. FlightPath: Obedience vs choice in cooperative services. In *OSDI 2008*, Dec 2008.

[14] H. C. Li, A. Clement, E. Wong, J. Napper, I. Roy, L. Alvisi, and M. Dahlin. BAR Gossip. In *Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI '06)*, pages 191–204, Nov. 2006.

[15] T. Locher, P. Moor, S. Schmid, and R. Wattenhofer. Free riding in bittorrent is cheap. In *Proceedings of the 5th Workshop on Hot Topics in Networks (HotNets '06)*, Nov. 2006.

[16] J.-P. Martin. Leveraging altruism in cooperative services. Technical Report MSR-TR-2007-76, Microsoft Research, June 2007.

[17] R. S. Peterson and E. G. Sirer. Antfarm: efficient content distribution with managed swarms. In *NSDI'09: Proceedings of the 6th USENIX symposium on Networked systems design and implementation*, pages 107–122, Berkeley, CA, USA, 2009. USENIX Association.

[18] M. Piatek, T. Isdal, T. Anderson, A. Krishnamurthy, and A. Venkataramani. Do incentives build robustness in BitTorrent? In *Proc. 4th NSDI*, Apr. 2007.

[19] D. K. Vassilakis and V. Vassalos. An analysis of peer-to-peer networks with altruistic peers. *Peer-to-Peer Networking and Applications*, 2(2):109–127, 2009.