Mediation in reputational bargaining

Jack Fanning*

July 9, 2018

Abstract

This paper investigates the potential for mediation in a dynamic reputational bargaining model, where rational agents can imitate behavioral types. Agents are free to ignore the mediator, but she can affect behavior by eliciting information about their types and strategically releasing this over time. I first show that a simple mediation (communication) protocol in which the mediator immediately suggests a compromise when both parties accept its terms in private cannot improve on unmediated bargaining. Adding noise to this protocol, however, can improve payoffs if behavioral types are unlikely. My main result characterizes an essentially unique mediation protocol that maximizes rational agents' payoffs in symmetric problems. This always improves on unmediated outcomes if agents are risk averse, because the mediator reduces the dispersion of agreement terms. By reducing delay between pairs of rational agents more than between rational-behavioral agent pairs, the mediator can also improve risk neutral agents' payoffs, but only if behavioral types are unlikely or make large demands. By contrast, efficient outcomes can be obtained when a mechanism designer can impose outcomes if behavioral types are likely or make moderate demands. This efficiency shows that ensuring agents' obedience to instructions is a much bigger constraint on the mediator than incomplete information alone.

Keywords: Bargaining, reputation, behavioral types, mediation, delay

1 Introduction

In its broadest definition, mediation in bargaining refers to any instance in which a third party helps others reach a *voluntary agreement*. It is distinct from arbitration, which imposes an agreement. Mediation is widely used to help parties to resolve disputes ranging from international conflicts and industrial relations to divorce proceedings. For instance, Dixon (1996) finds that mediation occurred in 13% of dispute phases of international conflicts between 1947-1982.[1] In legal disputes, mediation is an increasingly popular form of Alternative Dispute Resolution (ADR). In a survey of general counsel for Fortune 1000 companies, Stipanowich and Lamare (2013) found that in each of corporate/commercial disputes, employment disputes and consumer disputes over 42% of companies "always" or "often" use mediation. By contrast binding arbitration was "always" or "often" used by less than 17% of companies in each category. Use of mediation increased in every category compared to a 1997 survey.

*Brown University. jack_fanning@brown.edu. Department of Economics, Robinson Hall, 64 Waterman Street, Brown University, Providence, RI 02912. See https://sites.google.com/a/brown.edu/jfanning for latest version.

[1]Dispute phases are distinguished by the level of conflict (e.g. threats of hostilities, open hostilities).

Attesting to mediations benefit's, Dixon found that mediated disputes were 47% less likely to escalate and 24% more likely to peacefully resolve compared to disputes with no conflict management.[2] More convincing evidence comes from Emery et al. (1991), who found that a treatment group, randomly selected to receive mediation services, settled 89% of contested custody cases out of court, compared to 28% of a control group. Mediation also halved the time spent reaching agreement and increased parties' satisfaction with the outcome.

Why might mediation help? Veteran mediator and Former Secretary of State for Labor, John Dunlop, describes the difficulties of "end-play" negotiations and the benefit of mediation as follows: "The critical problem is that each side would prefer the other to move to avoid a further concession itself, and that any move may create the impression of being willing to move all the way to the position of the other side... In these circumstances a third party may greatly facilitate agreement. The separate conditional acceptance to the mediator by one side of the proposal does not prejudice the position of that side if there is no agreement. It is not unusual for a mediator to secure the separate acceptance of each side of a "package" of the mediator's design and then to bring the parties together to announce that, even if they do not know it, they have an agreement."[3]

The claim is that mediators help in part by filtering information. Agents may resist proposing a compromise themselves for fear of being identified as a weak type, who is willing to concede entirely to her opponent's demand. By filtering the information that an agent is willing compromise (e.g. releasing it only when an opponent is also willing to compromise), the mediator can eliminate this fear, and so potentially encourage agreement.

This paper seeks to use economic theory to help understand why and when such mediation techniques can be effective. I do this in the context of the reputational bargaining model of Abreu and Gul (2000) (henceforth AG).

In AG's model, two agents must divide a dollar. They can make frequent offers over the course of an infinite horizon.[4] With positive probability an agent is a behavioral (commitment) type who always demands a fixed share of the dollar and accepts nothing less, otherwise the agent is rational. In the unique equilibrium, an agent identified as rational must concede immediately to a possibly behavioral opponent's demand. Given this, rational agents must imitate behavioral types and then slowly concede to their opponent in a war of attrition with inefficient delay. The model captures the difficulties of unmediated negotiations highlighted above, in particular, negotiators' fear that small concessions will necessitate larger ones is well justified. The unique equilibrium provides a clean benchmark against which to asses mediation's benefits.

[2] Wilkenfeld et al. (2003) and Beardsley et al. (2006) also present positive empirical assessments of mediation's effectiveness in resolving conflict.

[3] Dunlop (1984), p16-24.

[4] This might seem inconsistent with many mediation settings, where an impending trial implies a finite horizon (deadline). However, Fanning (2016) shows that infinite horizon and deadline reputational models are very similar if there is slight uncertainty about the deadline's timing (the last time at which agents can strike a deal).

My first result is negative: the simple mediation (communication) protocol outlined by Dunlop, in which the mediator immediately suggests a compromise agreement if both parties privately accept its terms, necessarily fails. Bargaining outcomes are identical to AG's unmediated equilibrium. The result highlights some of the subtle incentive problems of mediation. Even though the simple mediator protects the reputation of an agent who "confesses" rationality (a willingness to compromise) when her opponent does not, information still leaks out in this case. If both rational parties are expected to confess with positive probability, a confessing agent learns that her opponent is more likely to be behavioral if a deal is not immediately announced. This increases her incentive to subsequently concede to her opponent's demand. Because that opponent would prefer to receive his (behavioral) demand than a compromise, such concession destroys his incentive to confess rationality in the first place.

This negative result presents a puzzle given that "real world" mediators such as Dunlop claim to usefully employ the simple protocol. However, that puzzle is potentially explained by the paper's second (positive) result, which shows that adding noise to the simple protocol can improve on unmediated outcomes when behavioral types are unlikely.[5] This noise takes the form the mediator failing to announce a deal with (possibly small) positive probability when both parties confess rationality. This failure to announce a deal may arise because agents' messages sometimes go astray, or are misinterpreted by the mediator; it does not require the mediator to intentionally stay silent when she knows a compromise is feasible. The intuition for why adding noise helps is that it makes agents less pessimistic about an opponent's type if no deal is announced (because perhaps the mediator is at fault), and so reduces agents' incentives to subsequently concede. In conjunction with the first result, this may help explain why mediators are not explicitly incentivized to reach early agreements, but are instead typically paid by the hour (Velikonja (2009)), and is consistent with the advice of veteran mediators to not overly prioritize reaching an agreement (e.g. see Brazil (2007)). It also relates to Kydd (2001)'s finding that a neutral mediator who wants to minimize the possibility of war, will lack incentives to truthfully reveal her information to bargaining parties.

To more deeply understand how and when mediation can work, the paper's main analysis adopts a mechanism design approach.[6] I focus on equilibria with a mediator in which rational agents always have the option to concede to an opponent's behavioral demand. This restriction is with some loss of generality,[7] however, it represents all equilibria if agents initially make behavioral demands (mediators typically intervene only once parties are in conflict) and can recall an op-

[5]The previous literature on mediation in non-bargaining contexts (e.g. see Goltsman et al. (2009)) has also identified the importance of adding noise in addition to information filtration.

[6]Dynamic information design might seem to be a more accurate description of this exercise given that the mediator cannot affect the game structure, and affects behavior merely by acquiring information and managing its flow back to agents. However, the term information design is typically taken to refer to a situation where a "mediator" has access to her own information source about the state of the world.

[7]In subsection 4.2 I discuss other equilibria, and in the Supplementary Material (Appendix C) I show that these can improve on payoffs in the optimal equilibrium under the restriction, if behavioral types are sufficiently rare.

ponent's previous offer. The distribution of possible outcomes in such equilibria are completely characterized by two incentive constraints. The first is a *dynamic incentive constraint*, which says that each rational agent doesn't want to concede before the equilibrium specifies that she is supposed to reach an agreement. The second is a *type incentive constraint*, which says that each rational agent doesn't want pretend to be behavioral (even with the mediator) and then concede after receiving some amount of equilibrium concession from a rational opponent. A strong version of the revelation principle then applies, allowing me to restrict attention to direct mediation protocol in which the mediator asks (rational) agents to confess their type at the start of bargaining, and then only sends a single public message back with suggested agreement terms when that agreement should take place.

My main result characterizes an essentially unique mediation protocol that maximizes the sum of rational agents' payoffs for any symmetric bargaining problem.[8] This optimal protocol is symmetric: the mediator suggests an equal split of the dollar between two rational agents, and suggests rational agent i concede to behavioral agent j at the same rate regardless of whether $i = 1$ or $i = 2$. The distribution of agreements between two rational agents features a mass point at time zero, followed by continuous agreement at a rate which makes a confessing agent indifferent to conceding until an agreement is suggested. The distribution of agreements when a rational agent faces a behavioral opponent features no agreement on some initial interval, followed by a mass point, followed by continuous agreement at a rate which would make a non-confessing rational agent indifferent to conceding (although rational agents *do* confess).

One important way in which the optimal protocol improve on unmediated outcomes is to replace dispersed agreements terms between rational agents with a single average agreement (an equal split of the dollar). In an unmediated war of attrition equilibrium, agents sometimes concede and are sometimes conceded to, which is inefficient if agents are risk averse. This form of inefficiency has typically received much less attention in the theoretical literature than delayed agreement, but in practice it may be at least as important in explaining the benefit of mediation. For instance in contested custody cases, most parents would strictly prefer a 50/50 split of parenting time to a 50/50 chance of sole custody or none.

The second important way for a mediator to improve unmediated outcomes is to reduce delay, however, she must be careful about which agreements she prioritizes. In order to improve payoffs while preserving incentives for rational agents to confess, the mediator must reduce delay between pairs of rational agents by more than between rational-behavioral agent pairs.

By delaying rational-behavioral agreements more than rational-rational agreements, the mediator can improve on unmediated outcomes even for risk neutral agents, if the probability of behavioral types is small or behavioral demands are large. In either of these cases, delaying of rational-behavioral agreements imposes a smaller cost on an agent who confessed rational-

[8] Symmetric agents have equal impatience, behavioral probability, behavioral demands, and utility functions.

ity (such agreements aren't worth much in expected terms) but a larger cost on an agent who pretended to be behavioral, helping to preserve incentives while delay between rational agent pairs is reduced. If a behavioral type's probability is larger than her demand, however, then behavioral-rational delay is more costly to a confessing agent and so worsens the incentive to confess, meaning that even optimal mediation is no better than unmediated bargaining.

An interesting feature of the above characterization is that the further apart agents' demands are, the easier it is for the mediator to get them to compromise. In fact, when behavioral types demand almost the entire dollar, mediation is approximately efficient with payoffs arbitrarily close to those under complete information (this is also true when behavioral types are vanishingly rare). This is somewhat counterintuitive as one might think that parties have to give up too much for any compromise to work when demands are far apart.[9]

While unmediated bargaining is one benchmark with which to compare optimal mediation, another important benchmark is a standard mechanism design problem in which the designer can impose outcomes (e.g. there is no dynamic incentive constraint). Not surprisingly, this benchmark achieves strictly higher rational agent payoffs than optimal mediation. In fact, the difference is dramatic: when behavioral types are likely or make moderate demands (close to 1/2) it achieves full efficiency, with payoffs matching those under complete information.[10] Recall that for risk neutral agents, mediation cannot even improve on unmediated bargaining outcomes when behavioral types are likely and make moderate demands. For a smaller probability of behavioral types, or larger demands, the benchmark is less than fully efficient.

The possibility of full efficiency and very different comparative statics when the mechanism designer can impose outcomes shows that agents' freedom to ignore instructions constrains a mediator far more than the informational problem alone. In fact, it is the designer's ability to enforce (perpetual) disagreement between two reported behavioral types which is key to efficiency in this benchmark. This ability to impose disagreement would seem implausible for many bargaining settings, however, as parties would have incentives renegotiate and reach a deal on their own.

This second benchmark does not appear to be a close fit to arbitration (as a form of ADR), which always imposes an agreement (e.g. some dollar division) on parties rather than disagreement. Imposing an agreement between two behavioral agents, however, appears to be highly problematic, as such an agreement is by assumption worse than perpetual disagreement for at least one of those types. It may, therefore, be very hard to convince such types to volunteer for arbitration. Mediation, by contrast is inherently voluntary and does not require the active participation of commitment types. This may help explain the increased use of mediation compared

[9] On the other hand, unmediated bargaining is less efficient when demands are symmetric and far apart, so mediation would seem to offer larger potential gains.

[10] The possibility of efficiency might seem to conflict with the necessary inefficiency of Myerson and Satterthwaite (1983) given overlapping distributions of buyer and seller values. However, that result requires continuous distributions of values, while the reputational model is closer a setting with a discrete distributions of values.

to arbitration documented in Stipanowich and Lamare (2013).

In addition to the role for mediators I consider here, mediation has many other reputed benefits which I do not address (e.g. see Goldberg et al. (2012)). These include the acknowledgement of each side's grievances by a neutral party, the creation of a less confrontational atmosphere for negotiation, the mediator's ability to establish commonly accepted facts, her ability to offer an authoritative opinion on the legal merits of a case should it come to trial, and her ability to creatively integrate the multiple issues which may be at stake (convincing parties to sacrifice low value issues in return for higher value concessions elsewhere).[11][12]

There has been surprisingly little research on any of these benefits of mediation in dynamic bargaining. One notable exception is Jarque et al. (2003), which identifies (under certain conditions) an equilibrium in which a mediator adopting Dunlop's simple protocol can improve on an unmediated war attrition equilibrium, when agents have private information about their reservation values. It it unclear whether simple mediation "works" in this setting, however, because unmediated equilibria are not unique and need not resemble a war of attrition. I delay a full discussion of Jarque et al. (2003) and other related literature until Section 5.

The paper is arranged as follows. Section 2 outlines the model; Section 3.1 presents the unique equilibrium without mediation and considers simple mediation protocols inspired by Dunlop; Section 4 identifies the optimal mediation protocol for symmetric problems and a mechanism design benchmark; Section 5 discusses my results in relation to the existing literature.

2 The model

The model presented below encompass all the mediation protocols I consider in a consistent way. The setup adapts the the discrete-continuous time bargaining protocol advanced by Abreu and Pearce (2007), although for much of the analysis time can be treated as completely continuous. I discuss in Section 5 how results can be generalized to different bargaining protocols.

Two bargainers, $i = 1, 2$, must agree on how to divide a dollar and face an infinite horizon. Bargainers are either rational or behavioral types. If rational bargainer i obtains a share $x_i \in [0, 1]$ of the dollar at time t then her utility is $e^{-r_i t} u_i(x_i)$ where her discount rate is r_i and the twice continuously differentiable utility function u_i satisfies $u_i(0) = 0$, $u_i'(x) > 0$ and $u_i''(x) \leq 0$. This setup can capture any bargaining problem with a positive, concave, strictly decreasing utility possibility frontier. Behavioral types have no preferences, but mechanically implement an exogenously defined strategy. A third player is a mediator, $i = 3$, who is always a behavioral

[11] In labor disputes, these last two roles are often distinguished as "rights-based" and "interest-based" mediation, however, this distinction is not common elsewhere (Stipanowich and Lamare (2013)).

[12] There is an element of this final role for a mediator in my model: she collects and releases information to agents about the set out possible, mutually beneficial agreements, which are not known ex-ante.

type, with a fixed strategy.

Time is discrete-continuous to allow multiple events to occur at the same time in a sequential order. Each time $t \in [0, \infty)$ is divided into five different discrete times t^1, t^2, t^3, t^4, t^5. Time follows a lexicographic ordering so that $t^k < t^{k+1}$, and $t^k < s^l$ whenever $t < s$. The set of discrete continuous times is $DC = [0, \infty) \times \{1, 2, ..., 5\} \cup \infty$. There is no discounting of payoffs *within* each time t. The bargaining protocol is as follows: at time 0^1 each bargainer i simultaneously announces a demand $\alpha_i(0^1) \in [0, 1]$; at $t^1 > 0^1$ each bargainer can concede to her opponent's existing demand (accept the share $(1 - \alpha_j(t^1))$), ending the game; at any t^2 each bargainer can send a private message to the mediator, typically this will simply indicate that she is rational; at t^3 the mediator can send a public message to the agents, typically this will simply involve a suggested dollar division $(m_1(t^3), m_2(t^3))$ where $m_1(t^3) = 1 - m_2(t^3) \in [0, 1]$; at t^4 each bargainer can simultaneously change her demand to $\alpha_i(t^4)$; at t^5 each bargainer can concede to her opponent's (possibly new) existing demand. If both bargainers concede at the same time then each proposal is selected with probability $\frac{1}{2}$. It is possible to extend this setup without affecting the results to allow the mediator to also send private messages and the agents to send public messages, however, describing strategies and agents' information in this case is considerably more cumbersome, without affecting results

At every $t^k > 0^1$ each bargainer is associated with an existing demand. If bargainer i changes her demand at t^4 then she cannot change her demand again until time $(t + \Delta)^4$ for some $\Delta > 0$. That is, if $\alpha_i(t^3) \neq \alpha_i(t^4)$ then i's existing demand at s^k is $\alpha_i(s^k) = \alpha_i(t^4)$ for $t^4 \leq s^k < (t + \Delta)^4$. Similarly, if agent i sent a message at t^k, then she cannot send another message until $(t + \Delta)^k$. These restrictions mean that agents' bargaining environments are relatively stable, allowing strategies to be more easily defined.

Bargainer i is a behavioral type with probability $z_i \in (0, 1)$, and is otherwise rational. A behavioral type for bargainer i initially demands a share $\alpha_i(0^1) = \alpha_i \in (0, 1)$ and never changes this.[13] She concedes to her opponent's demand at $t^k \in \{t^1, t^5\}$ if and only if $(1 - \alpha_j(t^k)) \geq \alpha_i$. She never sends a message to the mediator, and so any message indicates rationality. Because of this, I say that an agent who sends a message to the mediator *confesses* rationality, and is a *confessing* agent, otherwise she is a *non-confessing* agent. The behavioral demands of the two bargainers are incompatible, $\alpha^1 + \alpha^2 > 1$.

The intuitive description of the game above does not define an explicit extensive form. I do that using stopping times. The idea is to define a new information set (private history) for an agent only when she observes a change in her bargaining environment. At each of her information sets, an agent chooses a planned future action and future action (stopping) time. Planned actions must maximize continuation payoffs at each information set, but not necessarily at later times. This allows for an effectively continuous time game, that does not create insurmountable

[13] In AG, agents can imitate multiple behavioral types and announce demands sequentially. I discuss this extension of the model in Section 5.

technical challenges. The description of bargainer strategies can then be simplified considerably after specifying the mediator's strategy, and by using well known reputational bargaining results.

At each non-terminal *private history* h_i for agent i (an information set), she chooses an *action plan* $a_i(h_i)$. An action plan $a_i(h_i) = (\tau_i(h_i), x_i(h_i), i)$ for agent i consists of three parts: a *future* time to take action, $\tau_i(h_i)$; an action to take at that time, $x_i(h_i)$; and a marker for agent i. She can plan to never take a future action by setting $\tau_i(h_i) = \infty$. Let any actions planned for t^1, t^5, ∞ (where $t^1 \neq 0^1$) be denoted $x_i(h_i) = X$: these are unambiguous (agents plan to concede, or do nothing). Actions at 0^1 or t^4 must specify a dollar division. Agents' actions at t^2 (private messages to the mediator) and the mediator's actions at t^3 (public messages to the agents) belong to some arbitrary message space $M \supseteq [0,1]^2$, implying that the space is rich enough to allow the mediator to suggest a division of the dollar. The set of i's possible action plans is then a subset of $A_i = DC \times \{X\} \cup M \cup [0,1] \times \{i\}$. A private history for agent i is then composed of a finite sequence of action plans which she has observed (of herself and others).

Which private histories are ultimately realized is determined inductively as follows. The initial *realized private history* is the null set, $h_i^1 = \emptyset$. Subsequent realized private histories, h_i^{k+1} (where $k \in \mathbb{N} \cup \infty$), are determined by the *joint realized history* $h^k = (h_1^k, h_2^k, h_3^k)$ and agents' action plans at realized private histories, $a_j(h_j^k)$. Let the time of the first action planned given h^k be $\tilde{\tau}(h^k) = \min\{\tau_1(h_i^k), \tau_2(h_i^k), \tau_3(h_i^k)\}$. Given h^k, let the set of players whose actions i observes at $\tilde{\tau}(h^k)$ be $J_i(h^k)$. If $J_i(h^k) = \emptyset$ then let $h_i^{k+1} = h_i^k$. If $J_i(h^k) = \{1,2\}$ then $h_i^{k+1} = (h_i^k, (a_i(h_i^k), a_1(h_1^k), a_2(h_2^k)))$. If $J_i(h^k) = \{j\}$ then $h_i^{k+1} = (h_i^k, (a_i(h_i^k), a_j(h_j^k)))$. The game ends if ever $\tilde{\tau}(h^k) \in \{t^1, t^5\}$ (one player concedes to a well defined existing demand) or else there is no agreement (players get a zero payoff). If $h_i^{k+1} \neq h_i^k$ then define the *reference time* of realized private history h_i^{k+1} as $\check{\tau}(h_i^{k+1}) = \tilde{\tau}(h^k)$ and let $\check{\tau}(\emptyset) = 0^1$. This is the effectively the time at which the history h_i^{k+1} occurs.

An example will help clarify this structure. At time 0^1, bargainers make initial behavioral demands, so bargainer i's action plan at $h_i^1 = \emptyset$ is $a_i(h_i^1) = (\tau_i(h_i^1), x_i(h_i^1), i)$ where $\tau_i(\emptyset) = 0^1$, $x_i(\emptyset) = \alpha_i$. The mediator intends to send no message until she receives messages from both agents, and so $a_3(\emptyset) = (\infty, X, 3)$. The minimum time $\tilde{\tau}(h^1)$ in the joint realized history $h^1 = (\emptyset, \emptyset, \emptyset)$ is therefore 0^1. As all players observe these demand announcements, the next realized private history for player i is $h_i^2 = (h_i^1, (a_1(h_1^1), a_2(h_2^1), a_i(h_i^1)))$. Given h_1^2, agent 1 plans to message the mediator at 0^2, $a_1(h_1^2) = (0^2, m, 1)$, agent 2 plans to concede at time t^5, $a_2(h_1^2) = (t^5, X, 2)$, and the mediator plans to say nothing, $a_3(h_1^2) = (\infty, X, 3)$. This means $\tilde{\tau}(h^2) = 0^2$. Agent 1's action at 0^2 is observed by agent 1 and the mediator but not player 2, so that $h_1^3 = (h_1^2, a_1(h_1^2))$, $h_3^3 = (h_3^2, (a_1(h_1^2), a_3(h_3^2))$, and $h_2^3 = h_2^2$. Given h_1^3, agent 1 plans to change her demand to the entire dollar at $s^4 > t^5$ so that $a_1(h_1^3) = (s^4, 1, 1)$, the mediator plans to say nothing $a_3(h_3^3) = (\infty, X, 3)$. And so, $\tilde{\tau}(h^3) = t^5$, at which point the game ends with player 2 conceding to player 1's existing (initial, behavioral) demand α_1.

Let the set of possible joint realized histories be H and the set of possible realized private histories be H_i. A behavior strategy for agent i randomizes over her possible action plans at each of her possible realized private histories, $\sigma_i : H_i \to \Delta(A_i)$. A belief for bargainer i is $\mu_i : H_i \to \Delta(\{Z, R\} \times H)$, this describes both her belief about her opponent's type and her belief about the joint history at each of her possible private histories.

A (weak) perfect Bayesian equilibrium requires that at each of bargainer i's possible realized private histories, h_i, her behavior strategy maximizes her reference time, $\check{\tau}(h_i)$, continuation payoff, given the strategies of others and her beliefs, where beliefs are determined by Bayes rule where possible.

3 Unmediated bargaining and simple mediation protocols

In this section, I first consider a Baseline version of the model without mediation, which AG show has a unique, inefficient, war of attrition equilibrium. I then consider three simple communication protocols for the mediator, motivated by Dunlop. In the first two protocols, the mediator seeks a specific compromise, and immediately suggests that agreement if and only if both agents provisionally accept it. The third protocol adds noise to the first protocol: the mediator sometimes fails to announce an agreement even when both agents confess. Neither of the first two protocols can improve payoffs above the unmediated Baseline equilibrium, however, the third mediator strategy can improve payoffs if behavioral types are unlikely.

3.1 A Baseline Without Mediation

In this subsection there is no mediator (so we can ignore times t^2 and t^3). In this setting, AG's results (Proposition 4) imply that if agent i is revealed to be rational at time t^k in equilibrium (i.e. just after i makes a non-behavioral demand), while agent j may be behavioral, then i must immediately concede. This relies only on continuation strategies being optimal at t^k. It mirrors the logic of the Coase conjecture (Coase (1972)) in that one-sided asymmetric information implies an immediate agreement favourable to the informed party.

Given AG's result, it is without loss of generality to assume that rational agents always imitate behavioral types and then simply choose when to concede. We can, therefore, move to fully continuous time and describe (the on equilibrium path part of) agent j's strategy with a cumulative distribution function $F^j : [0, \infty] \to [0, 1]$, where never conceding is described by a concession time of $t = \infty$. Let $F_j(t)$ be the *total* probability that agent j (who may be behavioral) has conceded before time t. This implies that agent j's reputation for being behavioral at

t is $\bar{z}_j(t) = \frac{z_j}{1-F_j(t)}$. Given agent j's strategy, agent i's expected payoff to conceding at t is:[14]

$$U_i(t) = \int_{s<t} e^{-r_i s} u_i(\alpha_i) dF_j(s) + (1-F_j(t)) e^{-r_i t} u_i(1-\alpha_j) + \left(F_j(t) - \sup_{s<t} F_j(s)\right) e^{-r_i t} \frac{1}{2}(u_i(\alpha_i) + u_i(1-\alpha_j))$$

Analysis

The unique equilibrium of this model is characterized by three properties: (i) at most one agent concedes with with positive probability at time zero; (ii) both agents reach a probability one reputation at the same time, $T^* < \infty$; and (iii) agents are indifferent to conceding at any time on $(0, T^*]$. This third indifference condition implies that agent j must concede on the interval $(0, T^*]$ at the constant rate:

$$\frac{f_j(t)}{1 - F_j(t)} = \lambda_j = \frac{r_i u_i(1 - \alpha_j)}{u_i(\alpha_i) - u_i(1 - \alpha_j)} \tag{1}$$

This implies that $1 - F_j(t) = (1 - F_j(0))e^{-\lambda_j t}$. Next define rational agent j's *exhaustion time*, $T_j = -\frac{1}{\lambda_j} ln(z_j)$, as the time by which she must have conceded even if she did not concede at time zero (so $1 = z_j e^{\lambda_j T_j}$). Condition (i) and (ii) then imply $T^* = \min\{T_1, T_2\}$, and finally:

$$1 - F_j(0) = z_j e^{\lambda_j T^*} = \min\left\{1, z_j z_i^{\frac{-\lambda_j}{\lambda_i}}\right\} \tag{2}$$

Proposition 1 (AG, Proposition 1)**.** *The Baseline model has a unique distribution of equilibrium outcomes, characterized by equations (1) and (2).*

The fact that $T^* > 0$ implies that rational agents sometimes inefficiently delay agreement. This offers scope for mediation to improve outcomes. Payoffs are:

$$U_i^B = u_i(\alpha_i) F_j(0) + u_i(1 - \alpha_j)(1 - F_j(0))$$

3.2 Immediate One-shot ($I1$) mediation

In this subsection, I let the mediator adopt the simplest possible version of the mediation protocol suggested by Dunlop: she suggests an agreement (m_1, m_2) immediately at 0^3 if both agents confess at 0^2 and otherwise remains silent. I call this Immediate One-shot ($I1$) mediation. I show that it cannot improve on unmediated bargaining outcomes.

If the mediator *does* make an announcement at 0^3, then both agents are revealed to be rational. In this case any dollar division or even perpetual delay is consistent with sequential rationality (e.g. agent i changes her demand to $\alpha_i(0^4) \in [0, 1]$ and subsequently doesn't concedes unless j

[14] Here and elsewhere, I suppress the explicit dependence of payoffs on strategies to minimize notation.

offers her more than that). In Section 5, I discuss how the mediator can get rational agents to agree to any dollar division between their behavioral demands in discrete time, by sequentially revealing their rationality. Given the eventual negative result of $I1$ mediation, it is without loss of generality to assume that agents do actually follow the mediator's suggestion .[15]

We can again simplify to a continuous time framework. Let agent i's (on equilibrium path) strategy be described as follows. Define $c_i \in [0,1]$ as the *total* probability that agent i (who might be behavioral) confesses at 0[2]. If both agents confess then agent i obtains the payoff $u_i(m_i)$, if not, she must choose when to concede. Agent i's concession choice is described by two cumulative distribution functions $F_i^c \in [0,1]^{[0,\infty]}$ and $F_i^n \in [0,1]^{[0,\infty]}$. Let $F_i^c(t)$ be the probability that agent i has conceded to her opponent before time t conditional on her confessing and no mediator suggestion. Similarly, let $F_i^n(t)$ be the probability that agent i has conceded before time t in the war of attrition, conditional on her not confessing. Finally, let $F_i(t) = c_i F_i^c(t) + (1-c_i)F_i^n(t)$ be the probability that agent i has conceded by time t conditional on no mediator suggestion.[16] Note that while I have not included additional subscripts or superscripts on F_i, it may be distinct from the function used to describe the Baseline model's equilibrium.

If agent j adopts a strategy σ_j then rational agent i's utility from confessing and then conceding at time t if the mediated makes no suggestion, is:

$$\begin{aligned} U_i^c(t) =& c_j u_i(m_i) + (1-c_j)\Bigg(\int_{s<t} e^{-r_i s} u_i(\alpha_i) dF_j^n(s) + (1-F_j^n(t)) e^{-r_i t} u_i(1-\alpha_j) \\ &+ \left(F_j^n(t) - \sup_{s<t} F_j^n(s) \right) e^{-r_i t} \frac{1}{2}(u_i(\alpha_i) + u_i(1-\alpha_j)) \Bigg) \end{aligned}$$

Alternatively, rational i's utility if she does not confess and concedes at time t is:

$$\begin{aligned} U_i^n(t) =& \int_{s<t} e^{-r_i s} u_i(\alpha_i) dF_j(s) + (1-F_j(t)) e^{-r_i t} u_i(1-\alpha_j) \\ &+ \left(F_j(t) - \sup_{s<t} F_j(s) \right) e^{-r_i t} \frac{1}{2}(u_i(\alpha_i) + u_i(1-\alpha_j)) \end{aligned}$$

Analysis

It is clear that the Baseline model's equilibrium can still be an equilibrium here, indeed this is the case in all mediation protocols considered. If agent j does not confess with positive

[15] Expected continuation payoffs must be weakly below those associated with some mediator proposal. I show that even the prospect of those higher payoffs cannot incentivize joint confession.

[16] There are potentially relevant, non-degenerate higher order beliefs in this game. If agent i confessed (but j didn't) then at time t, i believes j is behavioral with probability $\overline{z}_j^c(t) = \frac{z_j}{(1-c_j)(1-F_j^n(t))}$. If i did not confess, then at time t she believes j is behavioral with probability $\overline{z}_j^n(t) = \frac{z_j}{1-F_j(t)}$. If j did not confess, then she believes that i beliefs about her likelihood of being behavioral are $\overline{z}_j^c(t)$ with probability $\frac{c_i(1-F_i^c(t))}{1-F_i(t)}$ and $\overline{z}_j^n(t)$ otherwise.

probability then agent i has no incentive to do so either.

It is also clear that there can be no equilibrium with $m_i \in (1-\alpha_j, \alpha_i)$ in which rational agent i *always* confesses and j does so with positive probability. If there was, then a confessing agent j would learn for sure that i was behavioral if the mediator made no announcement, and so would subsequently concede immediately. Knowing this, a rational agent i would optimally choose not to confess because this would give her a larger payoff, $\alpha_i > m_i$. Any equilibrium with mediation and $m_i \in (1-\alpha_j, \alpha_i)$, therefore, must involve both rational agents mixing between confessing and not. The next proposition shows that there is no such equilibrium. Moreover, while there can sometimes be an equilibrium where both parties confess with positive probability if $m_i \in \{1-\alpha_j, \alpha_i\}$, outcomes in this case remain identical to those in the Baseline equilibrium.[17]

Proposition 2. *The distribution of outcomes in any equilibrium when the mediator adopt the $I1$ protocol is identical to that in the unique Baseline equilibrium without mediation.*

The explanation for this result is similar to why it is impossible for rational agents to always confess. I prove that if the mediator does not suggest an agreement (at 0^3), then at least one confessing agent, say j, must immediately concede with probability one ($F_j^c(0) = 1$). I loosely sketch the proof of that claim below. Such concession destroys the incentive for her opponent to confess in the first place.

Suppose that neither confessing agent immediately concedes with probability one ($F_j^c(0) < 1$). Standard arguments show that concession behavior after time zero must be continuous. I then show that if a confessing agent i concedes continuously on some interval (s,t), then so must a non-confessing agent i.[18] For this to be the case a non-confessing agent j must concede at rate $\frac{f_j^n(t)}{1-F_j^n(t)} = \lambda_j$ (to make confessing agent i indifferent regarding her concession) and the total concession for agent j must be at rate $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$ (to make non-confessing but rational i indifferent). The identity $F_i(t) = c_i F_i^c(t) + (1-c_i)F_i^n(t)$ then implies a confessing agent j must also concede at rate $\frac{f_j^c(t)}{1-F_j^c(t)} = \lambda_j$. But such a bounded concession rate would imply that a (rational) confessing agent j never concedes with probability one in finite time. Such behavior cannot be optimal for a rational agent, given the possibility of facing a behavioral opponent.

3.3 Immediate Infinite ($I\infty$) mediation

In this subsection I allow the mediator to respond to agents who confess rationality continuously over the infinite horizon, in what I call the *Immediate Infinite ($I\infty$)* mediation protocol. One concern about the negative result for the $I1$ protocol is that mediation is discontinuous, happening once and for all at time zero. It might be thought that allowing agents to confess

[17] This result also generalizes to a model in which behavioral type i "confesses" if $m_i = \alpha_i$.

[18] This claim is not immediate, but can be established by modifying standard war of attrition arguments.

continuously over time might allow for greater success. After all, in practice, mediators often hold multiple conferences with disputing parties before helping them reach a settlement (e.g. see Goldberg et al. (2012)). Nonetheless, $I\infty$ mediation still cannot improve on the unmediated outcomes.

In the $I\infty$ protocol, if agent i confesses at time t^2 and agent j confesses at $s^2 \geq t^2$, then at s^3 the mediator suggests the agreement (m_1, m_2). For tractability reasons, I focus on what I call $I\infty$ *equilibria* in which rational agents follow the mediator's suggestion by changing their demands to (m_1, m_2) at s^4. Focussing on such equilibria entails some loss of generality because this imposes constant continuation payoffs following an announcement by the mediator (we could imagine that such payoffs change over time and depend on which agent compromised first). It is, however, then without loss of generality to assume $m_i \in (1 - \alpha_j, \alpha_i)$, because if $m_i \geq \alpha_i$ then rational i would confess with probability one at 0^2 if this had any chance of affecting the outcome.

In an $I\infty$ equilibrium, agent i's strategy reduces to choosing a time to confess and a time to concede (to her opponent's behavioral demand). It is without loss of generality to assume that an agent never concedes at t^1 but only at t^5, and confesses before she concedes (because doing so strictly increases her payoff whenever it affects the game's outcome). We can again, therefore, analyze the game in continuous time. Agent i's strategy is described by two cumulative distribution functions, $F_i^c \in [0,1]^{[0,\infty]}$ and $F_i^d \in [0,1]^{[0,\infty]}$. Let $F_i^c(t)$ be the total probability that agent i has confessed before time t, and $F_i^d(t)$ be the total probability that agent i has conceded before time t (c=confessed, d=defeated) where $F_i^c(t) \geq F_i^d(t)$. Given j's equilibrium strategy, rational agent i's expected utility from confessing at time s and conceding at time $t \geq s$ is:

$$U_i(s,t) = \int_{v<s} e^{-r_i v} u_i(\alpha_i) dF_j^d(v) + \int_{v\in(s,t]} e^{-r_i v} u_i(m_i) dF_j^c(v)$$
$$+ (1 - F_j^c(t)) e^{-r_i t} u_i(1-\alpha_j) + (F_j^c(s) - \sup_{v<s} F_j^d(v)) e^{-r_i s} u_i(m_i)$$

Analysis

The Baseline equilibrium is still an $I\infty$ equilibrium, where $F_i^c(t) = F_i^d(t)$ for all t. The next proposition establishes that this is in fact the only $I\infty$ equilibrium.

Proposition 3. *The distribution of outcomes in any $I\infty$ equilibrium is identical to that in the unique Baseline equilibrium without mediation.*

The idea of the proof of is similar to that of Proposition 2 in that unless behavior matches the Baseline equilibrium with $F_i^c(t) = F_i^d(t)$, then indifference conditions for confessing and non-confessing agents imply a contradiction to the fact that rational agents must concede within finite time. However, it is somewhat more involved. I show that if ever $F_i^c(t') > F_i^d(t')$ for

some t', then there exists some $t'' > t'$ such that $F_i^c(t) > F_i^d(t)$ for all $t \in [t', t'')$ and $F_i^c(t'') = F_i^d(t'')$, where $t'' < \infty$ or else agents could not concede within finite time. I then argue that agents must confess and concede continuously on the interval $(t', t'']$. For that to be so, a confessing agent i (who has already confessed) and non-confessing agent i (who hasn't yet confessed) must be respectively indifferent between confession and concession times on the interval $(t', t'']$. These indifference conditions imply linear ODE that govern F_j^c and F_j^d, which imply that $F_j^c(t) - F_j^d(t) > 0$ for $t \in (t', t'']$. Applying similar arguments with the roles of i and j reversed, we must have $F_i^c(t) - F_i^d(t) > 0$ for $t \in (t', t'']$, a contradiction.

3.4 Noisy One-Shot ($N1$) mediation

In this subsection I consider a Noisy One-Shot ($N1$) mediation protocol which adds noise to the mediator's strategy in the $I1$ protocol. In both the $I1$ and $I\infty$ protocols, an agent who confesses receives an unambiguous signal that her opponent hasn't confessed, if the mediator doesn't make an immediate announcement. Adding noise, by having the mediator sometimes fail to suggest an agreement even when both parties confess, obfuscates that signal. I show that this can improve on unmediated outcomes if behavioral types are unlikely.

In the $N1$ protocol, if both agents confess at 0^2 the mediator suggests the agreement (m_1, m_2) at 0^3 with probability $b \in (0, 1)$, and otherwise remains silent. This noise can be interpreted as the mediator always suggesting agreement when she knows that both parties have confessed, but with probability $1 - \sqrt{b}$ each agent's message goes astray, or is misinterpreted. I focus attention on what I call $N1$ *equilibria*, which are equilibria where the mediator adopts the $N1$ protocol, while rational agents always confess at 0^2 and subsequently immediately implement any mediator suggestion.

If the mediator makes no suggestion at 0^3, then rational agents must decide when to concede. We can describe (on path) equilibrium strategies using the cumulative distribution function $H_i^c \in [0, 1]^{[0,\infty]}$, where $H_i^c(t)$ is the probability that a *rational* agent i has conceded before time t conditional on confessing and the mediator making no suggestion. In an $N1$ equilibrium, agent i's utility if she confesses and concedes at time t is then:

$$\begin{aligned} U_i^c(t) =& (1 - z_j)\left(bu_i(m_i) + (1 - b)\int_{s<t} e^{-r_i s} u_i(\alpha_i) dH_j^c(s)\right) \\ &+ \left((1 - z_j)(1 - b)(1 - H_j^c(t)) + z_j\right) e^{-r_i t} u_i(1 - \alpha_j) \\ &+ (1 - z_j)(1 - b)\left(H_j^c(t) - \sup_{s<t} H_j^c(s)\right) e^{-r_i t} \frac{1}{2}(u_i(\alpha_i) + u_i(1 - \alpha_j)) \end{aligned} \tag{3}$$

Agent i's utility if she does not confess and then concedes at time t is:

$$U_i^n(t) = (1-z_j)\int_{s<t} e^{-r_i s}u_i(\alpha_i)dH_j^c(s) + \Big((1-z_j)(1-H_j^c(t)) + z_j\Big)e^{-r_i t}u_i(1-\alpha_j)$$
$$+ (1-z_j)\left(H_j^c(t) - \sup_{s<t} H_j^c(s)\right)e^{-r_i t}\frac{1}{2}(u_i(\alpha_i) + u_i(1-\alpha_j))$$

Analysis

In an $N1$ equilibrium, if the mediator does not make a suggestion at 0[3], then rational agent j must believe that her opponent is behavioral with probability $\bar{z}_i = \frac{z_i}{1-(1-z_i)b}$. In this case, behavior in the continuation game must resemble that of the Baseline model but with initial reputations $\bar{z}_i$ instead of z_i. As noted previously, this equilibrium is characterized by three conditions: (i) at most one agent concedes with with positive probability at time zero; (ii) both agents reach a probability one reputation at the same time, $T^* < \infty$; and (iii) agents are indifferent to conceding at any time on $(0, T^*]$.

Let $F_j(t) = (1-\bar{z}_j)H_j^c(t)$ be the probability that a confessing agent i believes that j will concede before t conditional on no mediator announcement. We can then rewrite equation 3 as:

$$U_i^c(t) = (1-z_j)bu_i(m_i) + (1-b(1-z))\Bigg(\int_{s<t} e^{-r_i s}u_i(\alpha_i)dF_j(s)$$
$$+ (1-F_j(t))e^{-r_i t}u_i(1-\alpha_j) + \left(F_j^c(t) - \sup_{s<t} F_j^c(s)\right)e^{-r_i t}\frac{1}{2}(u_i(\alpha_i) + u_i(1-\alpha_j))\Bigg)$$

Condition (iii) then implies that agent i must expect j to concede at rate $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$ on $(0, T^*]$. Agent j's exhaustion time is now $T_j = -\frac{1}{\lambda_j}ln(\bar{z}_j)$. To ensure conditions (i) and (ii) are satisfied, we must have $T^* = min\{T^1, T^2\}$ and $1 - F_j(0) = \max\left\{1, \bar{z}_j\bar{z}_i^{-\frac{\lambda_j}{\lambda_i}}\right\}$. More generally for $t \leq T^*$ we must have $1 - F_j(t) = (1-F_j(0))e^{-\lambda_j t}$.

Given such behavior, a rational agent i who did not confess (not her equilibrium strategy) will subsequently find it in her interest to wait until T^* and then concede. This is because conditional on no mediator suggestion, the rate at which i expects j to concede on $(0, T^*]$ is larger if she did not confess than if she did. That is, $\frac{(1-z_j)h_j^c(t)}{(1-z_j)(1-H_j^c(t))+z_j} \geq \frac{(1-\bar{z}_j)h_j^c(t)}{(1-\bar{z}_j)(1-H_j^c(t))+\bar{z}_j}$, which implies $\frac{U_i^n(t)}{dt} > \frac{U_i^c(t)}{dt} = 0$ on $(0, T^*]$. Hence we must have:

$$U_i^{*n} = \max_t U_i^n(t) = (1-z_j)\int_{s<T*} e^{-r_i s}u_i(\alpha_i)dH_j^c(s) + z_j e^{-r_i T^*}u_i(1-\alpha_j)$$

When agent i confesses, she is subsequently indifferent to conceding at any $t \in (0, T^*]$ so that:

$$U_i^{*c} = \max_t U_i^c(t) = U_i^c(T^*) = (1-z_j)\left(bu_i(m_i) + (1-b)\int_{s<T^*} e^{-r_i s}u_i(\alpha_i)dH_j^c(s)\right) + z_j e^{-r_i T^*}u_i(1-\alpha_j)$$

A necessary and sufficient condition for an $N1$ equilibrium to exist therefore is:

$$Q_i = \frac{U_i^{*c} - U_i^{*n}}{(1 - z_j)b} = u_i(m_i) - \int_{s<T^*} e^{-r_i s} u_i(\alpha_i) dH_j^c(s) \geq 0 \tag{4}$$

That is, the share proposed by the mediator must be better than the stream of payoffs from a known rational agent's concession on $[0, T^*]$. The paper's first positive result shows that when agents' reputations are sufficiently small, an $N1$ equilibrium always exists which (strictly) Pareto dominates the equilibrium of the Baseline model.

Proposition 4. *For any given r_i, u_i, α_i for $i = 1, 2$, $b \in (0, 1)$ and fixed $K \geq 1$, there exists $\underline{z} > 0$ such that whenever $z_i \leq \underline{z}$ and $K \geq \frac{z_1}{z_2} \geq \frac{1}{K}$, an $N1$ equilibrium exists, which rational agents strictly prefer to the Baseline equilibrium.*

Some intuition for the result comes from examining equation (4). If agent i does not confess, she sacrifices an immediate payoff of $u_i(m_i)$ in return for the stream of payoffs $\int_{s<T^*} e^{-r_i s} u_i(\alpha_i) dH_j^c(s)$. That stream of payoffs comes relatively slowly when initial reputations are small. Notice, that for any $b \in (0, 1)$, if z_j is small then so is $\bar{z}_j$, and so the probability that agent j concedes before time t is very close to the probability that rational agent j concedes, $(1 - \bar{z}_j) H_j^c(t) \approx H_j^c(t)$. Hence the value of the stream of payoffs $\int_{s<T^*} e^{-r_i s} u_i(\alpha_i) dH_j^c(s)$, is approximately the same as her payoff in a Baseline equilibrium with reputations $\bar{z}_j$, i.e. $(1 - \bar{z}_j) \int_{s<T^*} e^{-r_i s} u_i(\alpha_i) dH_j^c(s) + \bar{z}_j e^{-r_i T^*} u_i(1 - \alpha_j)$. Because any Baseline equilibrium is inefficient, however, it is possible to choose an m_i to satisfy equation (4) for both agents.

It might seem strange that adding noise makes a difference to the success of mediation, after all agents could use mixed strategies in the $I1$ protocol (a particular form of noise). Indeed, if both agents confessed with probability b under the $I1$ protocol then conditional on agent i confessing and hearing no mediator announcement, she will believe that j is behavioral with probability $\bar{z}_j$. These situations are quite distinct, however. In particular, when agents mix under the $I1$ protocol, continuation play after time zero must provide incentives for both a confessing and non-confessing agent to concede continuously, whereas an $N1$ equilibrium only needs to provide dynamic incentives for a confessing agent.

Notice, that the result allows b to be chosen arbitrarily close to one. This shows that the mediator can guarantee an efficient outcome as the likelihood of behavioral types becomes vanishingly small. What about when behavioral types are not unlikely? The next Proposition shows that when the probability of at least one agent's behavioral type is close to one, no $N1$ equilibrium can exist. The reason for this, can be readily ascertained, by reexamining inequality (4), which must hold in any $N1$ equilibrium. By confessing, an agent effectively gains an immediate payoff of $u_i(m_i)$ but loses a delayed payoff of $u_i(\alpha_i)$, when she faces a rational opponent. When behavioral types are likely, however, there can very little delay (i.e. $T^* \leq \varepsilon$ for ε small), and so the inequality cannot hold (i.e. $\int_{s<T^*} e^{-r_i s} u_i(\alpha_i) dH_j^c(s) \geq e^{-r_i \varepsilon} u_i(\alpha_i) >$

$u_i(m_i)$ for some i).

Proposition 5. *For any given r_i, u_i, α_i for $i = 1, 2$ there exists $\underline{z} < 1$ such that if $z_1 \geq \underline{z}$, then no $N1$ equilibrium exists.*

4 Optimal mediation

In this section I adopt a mechanism design approach, in order to more better understand the constraints and possibilities of mediation, and ultimately to identify an optimal mediation protocol. The main results focus on symmetric bargaining problems. This limited scope allows me to make considerable progress, at the expense of some generality. The key insights gained seem likely to extend to asymmetric problems.

Consider any equilibrium of the reputational bargaining game (with a mediator) in which each agent always has the option to concede her opponent's behavioral demand prior to an agreement. That is, when rational i reaches agreement at some time t^k in the equilibrium, she must have demanded $\alpha_i(s^k) \leq \alpha_i$ for all $s < t$, (technically, this still allows for a time t^5 agreement with $\alpha_i(t^4) > \alpha_i$). I call such equilibria *α-optional*, in the sense that each agent j has the *option* of accepting a dollar share of at least $1 - \alpha_i$ prior to an agreement. For rest of this subsection I focus exclusively on such equilibria, but discuss justifications for this restriction, and equilibria which do not satisfy it in subsection 4.2.

We can describe the distribution of outcomes in any equilibrium as follows. Let $G^R \in [0,1]^{[0,\infty]}$ be the cumulative distribution of equilibrium agreement times conditional on two rational agents (R=rational), so that $G^R(t)$ is the probability of agreement before time t^5 conditional on both agents being rational (with $t = \infty$ again corresponding to no agreement). Likewise let $G^Z_j \in [0,1]^{[0,\infty]}$ be the cumulative distribution function of agreement times conditional on agent i being rational and j being behavioral behavioral (Z=behavioral). The terms of any such rational-behavioral equilibrium agreement must be $(\alpha_j, 1 - \alpha_j)$, because behavioral agent j always demands α_j allowing i to guarantee the dollar share $1 - \alpha_j$ (and j never accepts less than α_j). Let $M^t_i \in [0,1]^{[0,1]}$ be the cumulative distribution function of agent i's share conditional on an agreement between two rational agents at time t, so that $M^t_i(m)$ is the probability of agent i obtaining a share less than m, conditional on an agreement between two rational agents at time t. Feasibility implies $M^t_1(m) = 1 - M^t_2(1 - m)$ for all $m \in [0,1]$. The entire set of such distributions is described by the function $M_i : [0,\infty) \to [0,1]^{[0,1]}$ such that $M_i(t) = M^t_i$. Finally let $T^R = \min\{t : G^R(t) = 1\}$ and $T^Z_j = \min\{t : G^Z_j(t) = 1\}$.

We are interested in what constraints must hold in α-optional equilibria. Given any such equilibrium consider the following global deviation for rational agent i: act consistent with her equilibrium strategy up to time t^5 (this sometimes involves reaching agreement as $s \leq t^5$) but

always concede an instant after that time.[19] Agent i's expected payoff when she adopts this deviation assuming that she obtains exactly the share $1-\alpha_j$ when she concedes an instant after t^5 is:

$$\begin{aligned} U_i^c(t) =&(1-z_j)\int_{s\leq t} e^{-r_i s}\int u_i(m)dM_i^s(m)dG^R(s) + z_j\int_{s\leq t} e^{-r_i s}u_i(1-\alpha_j)dG_j^Z(s) \\ &+ e^{-r_i t}u_i(1-\alpha_j)\Big((1-z_j)(1-G^R(t)) + z_j(1-G_j^Z(t))\Big) \end{aligned}$$

In reality i may obtain a larger share than $1-\alpha_j$ when she concedes an instant after t^5, but not less, given that agent j always demands $\alpha_j(t)\leq\alpha_j$ prior to an agreement in equilibrium. Notice that agent i's actual expected equilibrium payoff is $U_i^c(\max\{T^R, T_j^Z\})$. For agent i's equilibrium strategy to be optimal, she must not want to make the global deviation, and so we must have:

$$U_i^c(\max\{T^R, T_j^Z\}) = \max_t U_i^c(t) \quad \text{(Dynamic IC)} \tag{5}$$

I call this the *dynamic incentive constraint*. An immediate observation is that for this constraint to be satisfied we must have $T_j^Z \leq T^R$. If this were not true, $T^R < T_j^Z$, then agent i would realize at T^R that she faced a behavioral opponent j and would profitably concede.

Rational agent i also has the option of making an alternative global "deviation" in which she acts consistent with a behavioral type prior to time t^5 (i.e. always demands α_i and never message the mediator) and concedes an instant after that time. Agent i's expected payoff from this deviation, again assuming that she obtains the share $1-\alpha_j$ when she concedes an instant after t^5 is:

$$U_i^n(t) = (1-z_j)\int_{s\leq t} e^{-r_i s}u_i(\alpha_i)dG_i^Z(s) + e^{-r_i t}u_i(1-\alpha_j)\Big((1-z_j)(1-G_i^Z(t)) + z_j\Big)$$

In this case, agent i makes exactly the same agreements as a behavioral type at $s\leq t^5$ (giving her a share α_i). For rational agent i not to want to make this deviation, we must have:

$$U_i^c(T^R) \geq \sup_t U_i^n(t) \quad \text{(Type IC)} \tag{6}$$

I call this the *type incentive constraint*.

I then define a *direct mediation protocol* as follows: If both agents initially demand $\alpha_i(0^4)=\alpha_i$ and message the mediator at 0^2, she sends a single message back to them at some time t^3 (if neither has changed demand prior to t^3) suggesting terms for an agreement. If both agents initially demand $\alpha_i(0^4)=\alpha_i$, but only agent i messages the mediator at 0^2, the mediator sends

[19] This deviation isn't well defined in the sense that there is no $s^* = \min\{s > t^5\}$, however, this not important. It is equivalent to the lack of a best possible deviation in a Betrand competition game when an opponent's price is greater than marginal cost.

a single message to the agents suggesting that i concede to j at some time t^3 (if neither has changed demand prior to t^3). If neither agent messages the mediator at 0^2, or some agent demands $\alpha_i(t^4) \neq \alpha_i$, the mediator is silent for the rest of the game. A *direct mediation equilibrium* is then an equilibrium in which the mediator adopts a direct mediation protocol, rational agents always message the mediator at 0^2, immediately implement any suggested agreement and demand α_i prior to the mediator's suggestion.

It is immediate that if a distribution of outcomes $(G^R, G_1^Z, G_2^Z, M_1, M_2)$ satisfies both the dynamic and type incentive constraints, then it arises in some direct mediation equilibrium (this is formalized in Observation 1, below). This implies that those two constraints completely characterize the set of possible equilibria, and that it is without loss of generality to restrict attention to direct mediation equilibria. The direct mediator suggests an agreement before time t^3 with probability $G^R(t)$ if both agents message her at 0^2, and with probability $G_i^Z(t)$ if only j messages her. In the former case, she suggests that i gets a share less than m with probability $M_i^t(m)$ in a time t agreement. As previously argued, following a suggestion from the mediator, it is always an equilibrium of the continuation game for agents to implement her suggestion. Additionally, if ever agent i changes her demand when not instructed by the mediator, she (optimally) immediately concedes to her opponent in the continuation equilibrium. The dynamic incentive constraint therefore ensures that a rational agent doesn't want to concede or change demand before the mediator's suggestion. The type incentive constraint ensures that she optimally confesses her rationality at 0^2.

Observation 1. *Any distribution of outcomes $(G^R, G_1^Z, G_2^Z, M_1, M_2)$ satisfying equations (5) and (6) can be obtained in some direct mediation equilibrium.*

The ability to restrict attention to direct mediation equilibria in which agents only send a single private message to the mediator and she only sends a single public message back, represents a strong version of the revelation principle. In general for multistage games, Myerson (1986) shows that in a communication equilibrium, a mediator must typically collect information and privately recommend actions to agents in each stage. While my formal game rules didn't actually allow the mediator to make private messages to the agents (to make the description of the game easier), it is clear that this had no bearing on the requirement that any equilibrium must satisfy the dynamic and type incentive constraints, and so in fact, restricting attention to public messages from the mediator is without loss of generality. Direct mediation equilibria are fully described by their agreement distributions. Given this, I henceforth treat all mediation protocols with the same distribution of equilibrium agreements as an equivalence class, and use the term mediation protocol interchangeably with the distribution of equilibrium outcomes which arise from that protocol.

The Baseline equilibrium is clearly α-optional, and so can be implemented as a direct mediation

equilibrium. The distribution of agreement times in that case is

$$1 - G_j^Z(t) = \frac{z}{1-z}(e^{\lambda_j(T^R - t)} - 1) \qquad \text{and} \qquad 1 - G^R(t) = (1 - G_i^Z(t))(1 - G_j^Z(t)),$$

for $t \leq T^R = T_1^Z = T_2^Z = \min\{-\frac{1}{\lambda_1}ln(z_1), -\frac{1}{\lambda_2}ln(z_2)\}$. The distribution of dollar shares conditional on an agreement at $t \leq T^R$ is:

$$M_i^t(m) = \begin{cases} 0 & \text{if } m < 1 - \alpha_j \\ \frac{\frac{g_j^Z(t)}{1-G_j^Z(t)}}{\frac{g_j^Z(t)}{1-G_j^Z(t)} + \frac{g_i^Z(t)}{1-G_i^Z(t)}} & \text{if } m \in [1-\alpha_j, \alpha_i) \text{ and } t \in (0, T^R] \\ \frac{G_j^Z(0)}{G_i^Z(0)+G_j^Z(0)} & \text{if } m \in [1-\alpha_j, \alpha_i),\ G_i^Z(0) + G_j^Z(0) > 0 \text{ and } t = 0 \\ 1 & \text{if } m \geq \alpha_i \end{cases}$$

These distributions cause incentive constraints to bind in the sense that $U_i^c(t) = U_i^n(t)$ for all $t \in [0, T^R]$.

Considering the Baseline equilibrium under a direct mediation protocol immediately suggests a way in which the mediator can improve outcomes for risk averse agents. The dispersed distribution of agreement shares between rational agents at $t > 0$ is clearly inefficient. Simply replacing this distribution by a single average agreement improves payoffs. That is, let $\hat{M}_i^t(m) = 0$ if $m < \hat{m}_i(t)$ and $\hat{M}_i^t(m) = 1$ otherwise where, $\hat{m}_i(t) = \int m dM_i^t(m)$. We then have $u_i(\hat{m}_i(t)) \geq \int u_i(m) dM_i^t$ for $t \leq T^R$, with a strict inequality for agent i at $t \in (0, T^R]$ if she is risk averse (i.e. $u_i'(1-\alpha_j) > u_i'(\alpha_j)$).

Leaving the distribution of agreement *times* unchanged then ensures that both incentive constraints are satisfied. Let $U_i^c(t)$ and $U_i^n(t)$ be the utilities under the Baseline equilibrium, and let $\hat{U}_i^c(t)$ and $\hat{U}_i^n(t)$ be utilities when the mediator suggests the average time t agreement between rational agents. Clearly we have $\hat{U}_i^c(t) \geq U_i^c(t) = U_i^n(t) = \hat{U}_i^n(t)$ for $t \leq T^R$, so that the type incentive constraint is satisfied. Moreover,

$$\left(\hat{U}_i^c(T^R) - \hat{U}_i^c(t)\right) - \left(U_i^c(T^R) - U_i^c(t)\right) = (1-z_j)\int_{t<s\leq T^R} e^{-r_i s}\left(u(\hat{m}_i(t)) - \int u_i(m) dM_i^s(m)\right) dG^R(s) \geq 0,$$

which means $\hat{U}_i^c(T^R) - \hat{U}_i^c(t) \geq U_i^c(T^R) - U_i^c(t) = 0$ so that the dynamic incentive constraint is satisfied. Moreover, the constraints are both slack for agent i when she is risk averse in the sense that $\hat{U}_i^c(T^R) > \hat{U}_i^c(t)$ for $t \neq T^R$ and $\hat{U}_i^c(T^R) > \sup_t U_i^n(t)$, so that even if j is not risk averse, it is possible for the mediator to strictly increase the payoff of *both* agents compared to the baseline equilibrium (i.e. by specifying $\check{m}_j(t) = 1 - \check{m}_i(t) = \hat{m}_j(t) + \varepsilon$ on $(0, \varepsilon]$ for some $\varepsilon > 0$ small). I formalize this observation below.

Observation 2. *Suppose that $u_1'(1-\alpha_2) > u_1'(\alpha_1)$, then there is a (direct mediation) equilibrium*

which delivers strictly higher payoffs for both rational agents than the Baseline equilibrium.

While this observation is technically trivial, it may be at least as important as reduced delay in explaining the benefits of mediation (the form of inefficiency focussed on in most of the theoretical bargaining literature). Moreover, the slackness of the dynamic and type incentive incentive constraints after eliminating dispersed outcomes, makes it is clear that the mediator can *also* then reduce delay compared to the distribution of agreements in the baseline equilibrium (i.e. choose some $\hat{G}^R(t) > G^R(t)$ and $\hat{G}^Z_i(t) > G^Z_i(t)$ for $t < T^R$).

We are interested in understanding more generally how a mediator *should* behave to benefit agents, given the above incentive constraints. In addition to the beneficial role for mediators highlighted in Observation 2, Proposition 4 showed that the mediator can improve even risk neutral agents payoffs when the probability of behavioral types is sufficiently small. However, identifying an optimal mediation protocol for general asymmetric bargaining problems is extremely challenging. It is clear that there are many moving parts and a great number of constraints (the type and dynamic incentive constraints really represent $[0, \infty)^4$ non-independent constraints). Moreover, it is not even obvious what objective function should be maximized to appropriately take account of any asymmetry between the agents.

The optimal mediation problem simplifies considerably, however, if agents fundamentals are symmetric in the sense that $u_i = u$, $r_i = r$, $z_i = z$, and $\alpha_i = \alpha$. For the rest of paper I will assume that this is so. Given symmetry and the fact that behavioral agents have no utility function and don't cooperate with the mediator, it seems natural to maximize the sum of rational agents payoffs, or perhaps their Nash product. I adopt the former objective, however, I will show that the optimal protocol identified necessarily implies symmetric payoffs, and so also maximizes the Nash product.

I formally, identify an optimal mediation protocol for symmetric bargaining problems as a solution to the following problem:

$$\arg\max_{G^R, G^Z_1, G^Z_2, M_1, M_2} U^c_1(T^R) + U^c_2(T^R) \text{ subject to equations (5) and (6)}$$

We can immediately start to simplify the arguments of this maximization and the agents' constraints. I say that a mediation protocol is *symmetric* if $G^Z_i = G^Z$ and $M^t_i = M^t$ for all t and $i = 1, 2$, and is *strongly symmetric* if it is symmetric and additionally $M^t(0.5) = 1$ (i.e. the mediator always suggest a 50/50 division between rational agents). Given an equilibrium of a game with symmetric fundamentals, it is simple to show that there is an equilibrium of a strongly symmetric mediation protocol which obtains a weakly higher objective (formalized in Observation 3, below).

Observation 3. *If $(G^R, G^Z_1, G^Z_2, M_1, M_2)$ describes an (α-optional) equilibrium (with symmetric fundamentals), then the strongly symmetric mediation protocol $(G^R, \hat{G}^Z)$ with $\hat{G}^Z = 0.5(G^Z_1 + G^Z_2)$*

describes an equilibrium that achieves a weakly higher objective, $U_1^c(T^R) + U_2^c(T^R)$.

This observation is almost immediate. First consider the associated symmetric protocol $(G^R, \check{G}^Z, \check{M})$ where $\check{G}^Z = \hat{G}^Z = 0.5(G_1^Z + G_2^Z)$ and $\check{M} = M_1 + M_2$. Let the utilities in the original equilibrium be $U_i^c(t)$ and $U_i^n(t)$ and the utilities in the symmetric protocol be $\check{U}^c(t)$, $\check{U}^n(t)$. Symmetry then implies:

$$\begin{aligned}\check{U}^c(t) =&(1-z)\int_{s\leq t} e^{-rs}\int u(m)0.5(dM_1^s(m)+dM_2^s(m))dG^R(s) + z\int_{s\leq t} e^{-rs}u(1-\alpha)0.5(dG_1^Z(s)+dG_2^Z(s))\\ &+e^{-rt}u(1-\alpha)\left((1-z)(1-G^R(t))+z(1-0.5(G_1^Z(t)+G_2^Z(t)))\right) = 0.5(U_1^c(t)+U_2^c(t)),\\ \check{U}^n(t) =&(1-z)\int_{s\leq t} e^{-rs}u(\alpha)0.5(dG_1^Z(s)+dG_1^Z(s)) + e^{-rt}u(1-\alpha)\left((1-z)(1-0.5(G_1^Z(t)+G_2^Z(t)))+z\right) = 0.5(U_1^n(t)+U_2^n(t)).\end{aligned}$$

This immediately ensures that the symmetric protocol obtains the same objective. Moreover, because the original equilibrium satisfied $U_i^c(T^R) = \max_t U_i^c(t) \geq \sup_t U_i^n(t)$, it is clear that $\check{U}^c(T^R) = \max_t \check{U}^c(t) \geq \sup_t \check{U}^n(t)$. The discussion around Observation 2 then illustrates that moving from this symmetric protocol to a strongly symmetric one preserves the incentive constraints and weakly increases the objective (because $u(0.5) \geq \int m d\check{M}^t(m)$).

Given Observation 3, I restrict attention to searching for an Optimal Strongly Symmetric Mediation Protocol (OSSMP). I shall proceed on that basis before ultimately showing that any optimal protocol is necessarily symmetric, and if agents are strictly risk averse, $u''(0.5) > 0$, then the optimal protocol is strongly symmetric.

Main Analysis

The main results of paper characterize the unique OSSMP. They are summarized in the following theorem.

Theorem 1. *A unique OSSMP exists. It delivers higher payoffs than the Baseline equilibrium if and only if either: i) agents are risk averse, in that* $u'(1-\alpha) > u'(\alpha)$*; or ii) the probability of a behavioral type is larger than its demand,* $z < \alpha$*. The optimal distribution* G^R *features an atom of agreement at time zero before increasing continuously on* $(0, T^R]$ *to keep a confessing agent indifferent to conceding, where* $T^R = T^Z$*. The optimal distribution* G^Z *implies no agreement strictly before some time* t^* *and an atom of agreement at* t^**, it then increases continuously on the non-degenerate interval* $(t^*, T^Z]$ *to keep a non-confessing agent indifferent to conceding. If behavioral types are sufficiently unlikely or make sufficiently large demands, in particular if* $z < \alpha$ *when agents are risk neutral, then* $t^* > 0$*.*

This characterization has many interesting features which help to illustrate how and when mediation works. I shall discuss these in turn as I establish the theorem's claims over the course of this section.

I first address the theorem's characterization of an optimal distribution of agreements between two rational agents. The claim is that the distribution is front-loaded as much as possible, with an atom at time zero, and continuous agreement thereafter, to keep a confessing agent indifferent to concession. If this were not true then the dynamic incentive constraint would be slack for some $t < T^R$ in the sense that $U^c(T^R) > U^c(t)$. But in which case keeping G^Z fixed, it is possible to move some rational-rational agreements forward in time while still satisfying the constraint for all t. This strictly increases rational agents' payoffs, $U^c(T^R)$, and so also relaxes the type incentive constraint because $\sup_t U^n(t)$ doesn't change. The claim is formally established in Lemma 1.

Lemma 1. *Given any distribution G^Z, then if the set of distributions G^R which satisfy both incentive constraints is non-empty, there is a unique such distribution, $G^{R*}_{G^Z}$, which maximizes rational agents' payoffs, characterized by $U^c(t) = U^c(T^R)$ for $t \leq T^R = T^Z$.*

The optimal distribution of rational-rational agreements given G^Z, which keeps rational agents indifferent to conceding on $[0, T^R]$, is denoted $G^{R*}_{G^Z}$. It can be identified in closed form by noting that the requirement of $\frac{U^c(t)}{dt} = 0$ on $[0, T^R]$ implies the linear ODE:

$$g^{R*}_{G^Z}(t) = \lambda^m \left((1 - G^{R*}_{G_Z}(t)) + \frac{z}{1-z}(1 - G^Z(t)) \right)$$

where

$$\lambda^m = \frac{ru(1-\alpha)}{u(0.5) - u(1-\alpha)}.$$

Solving this ODE using the boundary condition $G^{R*}_{G^Z}(T^Z) = 1$ gives:

$$1 - G^{R*}_{G^Z}(t) = e^{-\lambda^m t} \int_t^{T^Z} \lambda^m e^{\lambda^m s} \frac{z}{1-z}(1 - G^Z(s))ds \tag{7}$$

Having identified $G^{R*}_{G^Z}$, the OSSMP problem is reduced to finding the optimal G^Z. Because $G^{R*}_{G^Z}$ leaves agents indifferent to conceding on $[0, T^Z]$, the payoff from confessing is $U^c(0) = U^c(T^R) = G^{R*}_{G^Z}(0)\,(u(0.5) - u(1-\alpha))\,(1-z) + u(1-\alpha)$. Maximizing this payoff is clearly equivalent to maximizing $G^{R*}_{G^Z}(0)$.

Notice that a larger $G^Z(t)$ increases $G^{R*}_{G^Z}(0)$. More precisely, if $\tilde{G}^Z(t) \geq G^Z(t)$ for all t and $\tilde{G}^Z(s) > G^Z(s)$ for some s, then $G^{R*}_{\tilde{G}^Z}(0) > G^{R*}_{G^Z}(0)$. Other things equal, more behavioral-rational agreements before time t, relax the dynamic incentive constraint, allowing a slower rate of rational-rational agreements after time t (to keep the agent indifferent) and so ultimately more agreement at time 0. More insight into this relaxation can be gained by thinking about an agent's incentives from the perspective of time t, conditional on no suggested agreement thus far. A larger $G^Z(t)$ implies that the agent's belief that she faces a rational opponent is larger, $\frac{(1-z)(1-G^R(t))}{(1-z)(1-G^R(t))+z(1-G^Z(t))}$. Given that agreements with rational opponents give her a larger dollar

share (0.5 vs $1-\alpha$) this change implies a larger continuation payoff and less incentive to concede immediately.

Given how G^Z affects $U^c(T^R)$, we can now form a reduced *time t type incentive constraint* of $IC_{G^Z}(t) = U^c(0) - U^n(t) \geq 0$. Plugging in for $G^{R*}_{G^Z}(0)$, before integrating by parts gives:

$$\begin{aligned}
IC_{G^Z}(t) =& u(1-\alpha) + (u(0.5) - u(1-\alpha))(1-z)\left(1 - \int_0^{T^Z} \lambda^m e^{\lambda^m s}\frac{z}{1-z}(1-G^Z(s))ds\right) \\
&- (1-z)\int_{s\leq t} e^{-rs}u(\alpha)dG^Z(s) - e^{-rt}u(1-\alpha)\Big((1-z)(1-G^Z(t)) + z\Big) \\
=& u(1-\alpha)(1-e^{-rt}) + (u(0.5) - u(1-\alpha))(1-z)\left(1 - \int_0^{T^Z} \lambda^m e^{\lambda^m s}\frac{z}{1-z}ds\right) \\
&- e^{-rt}G^Z(t)(1-z)(u(\alpha) - u(1-\alpha)) + \int_0^{T^Z} G^Z(s)r\Big(u(1-\alpha)e^{\lambda^m s}z - \mathbb{1}_{[s\leq t]}u(\alpha)e^{-rs}(1-z)\Big)ds \qquad (8)
\end{aligned}$$

It is worth pausing to examine this constraint, because manipulating it is key to characterizing an optimal distribution of behavioral-rational agreement times, G^Z. In particular, consider the final integrand of $IC_{G^Z}(t)$, which is linear in $G^Z(s)$. This integrand captures the time t incentive costs and benefits of behavioral-rational agreements before time s (subject to not affecting $G(t)$). The cost is direct in that earlier agreements increase a non-confessing agent's payoff, $-u(\alpha)e^{-rs}(1-z)$ for $s < t$. That cost is decreasing in s because later payoffs are discounted. The benefit, meanwhile is indirect through relaxing the dynamic incentive constraint and so allowing earlier rational-rational agreements, $u(1-\alpha)e^{\lambda^m s}z$. To intuitively understand why this is increasing in s, notice that a change in $G^Z(s)$ has a larger effect on the agent's belief that she faces a behavioral opponent, $\bar{z}(s) = \frac{z(1-G^Z(s))}{z(1-G^Z(s))+(1-z)(1-G^R(s))}$ when $(1-G^R(s))$ is larger, and so on her incentive to concede immediately.

The discussion in the previous paragraph already provides the basic logic behind the characterization of the optimal distribution G^Z highlighted in Theorem 1. Because the incentive benefits minus costs of earlier behavioral-rational agreements, $u(1-\alpha)e^{\lambda^m s}z - u(\alpha)e^{-rs}(1-z)$, are increasing in s, it is best for incentives to have a larger $G^Z(s)$ later on in bargaining. The optimal distribution features $G^Z(t) = 0$ for $t \leq t^*$ and a binding type incentive constraint, $IC_{G^Z}(t) = 0$, for $t \in [t^*, T^Z]$ (i.e. $G^Z(t)$ is as large as possible later on).

A first step towards formalizing this idea is Lemma 2, which consists of two claims. The first states that the type incentive constraint must bind at T^Z, $IC_{G^Z}(T^Z) = 0$. This is fairly intuitive. If it did not hold, then it is possible to slightly increase $G^Z(t)$ for $t \in [T^Z - \varepsilon, T^Z]$ while maintaining $IC_{G^Z}(t) \geq 0$ for such t. This change would also increase $G^{R*}_{G^Z}(0)$, and by extension improves incentives for $t < T^Z - \varepsilon$ (as $U^n(t)$ does not change for such t).

The second claim of Lemma 2 is that the time $t \leq T^Z$ type incentive constraint must also bind, $IC_{G^Z}(t) = 0$, if $t \geq \hat{t} = \frac{1}{\lambda^m + r} ln\left(\frac{(1-z)u(\alpha)}{zu(1-\alpha)}\right)$. To see this, notice that the final integrand of the time t incentive constraint (equation (8)) is exactly zero at $s = \hat{t} \leq t$, and positive for $s \in (\hat{t}, T^Z]$. For such s, therefore, having $G^Z(s)$ as large as possible subject to not changing $G^Z(t)$, relaxes the

time t type constraint. If ever $IC_{G^Z}(t) > 0$ for $t \in [\hat{t}, T^Z)$, therefore, we can increase $G^Z(s)$ for $s \in [t, t+\varepsilon)$ and so ultimately $G^{R*}_{G^Z}(0)$, while preserving $IC_{G^Z}(t) \geq 0$ for all t.

Lemma 2. *Consider any distribution G^Z such that $\inf_t IC_{G^Z}(t) \geq 0$. If $IC_{G^Z}(t) > 0$ for some $t \in \left[\min\{\hat{t}, T^Z\}, T^Z\right]$ then there is an alternative distribution $\tilde{G}^Z$ such that $\min_s IC_{\tilde{G}^Z}(s) \geq 0 = IC_{\tilde{G}^Z}(t)$ for $t \in \left[\min\{\hat{t}, T^Z\}, T^Z\right]$, and $G^{R*}_{\tilde{G}^Z}(0) > G^{R*}_{G^Z}(0)$.*

Now suppose that behavioral types are fairly likely relative to their demands, in the sense that $z \geq \frac{u(\alpha)}{u(\alpha)+u(1-\alpha)}$ ($z \geq \alpha$ if agents are risk neutral). This implies $\hat{t} \leq 0$ and so by Lemma 2 we may restrict attention to distributions with $IC_{G^Z}(t) = 0$ for $t \leq T^Z$ and so $\frac{dU^n(t)}{dt} = 0$. Such distributions must therefore satisfy the linear ODE:

$$g^Z(t) = \lambda\left(1 - G^Z(t) + \frac{z}{1-z}\right)$$

which combined with the boundary condition $G^Z(T^Z) = 1$ gives $G^Z(t) = \frac{1-ze^{\lambda(T^Z-t)}}{1-z}$. Theorem 1 stated that generally the optimal behavioral-rational agreement distribution featured no agreement on some initial interval $[0, t^*)$, an atom of agreement at t^* and continuous agreement on $(t^*, T^Z]$ so as to leave a non-confessing rational agent indifferent to concession. Generally, therefore, the optimal distribution has the form:

$$G^Z_{\check{t},T^Z}(t) = \begin{cases} 0 \text{ for } t < \check{t} \\ \frac{1-ze^{\lambda(T^Z-t)}}{1-z} \text{ for } t \in [\check{t}, T^Z] \end{cases} \tag{9}$$

for some $\check{t}$. In the special case of $\hat{t} \leq 0$, however, Lemma 2 implies that $\check{t} = t^* = 0$. Clearly $G^Z_{0,T^Z}(t)$ is strictly decreasing in T^Z, and so the OSSMP problem can be reduced to finding the minimum T^Z such that $IC_{G^Z_{0,T^Z}}(T^Z) \geq 0$ (this maximizes $G^{R*}_{G^Z_{0,T^Z}}(0)$). To that end define:

$$\begin{aligned} IC_{G^Z_{\check{t},T^Z}}(T^Z) =& (1-z)\left(u(0.5) - u(1-\alpha)\right)\left(1 - \frac{z\lambda^m}{1-z}\int_0^{\check{t}} e^{\lambda^m s}ds + \frac{z^2\lambda^m}{(1-z)^2}\int_{\check{t}}^{T^Z} e^{\lambda T^Z + (\lambda^m-\lambda)s} - e^{\lambda^m s}ds\right) \\ &+ u(1-\alpha)(1-e^{r\check{t}}) - e^{-r\check{t}}\left(u(\alpha) - u(1-\alpha)\right)\left(1 - e^{\lambda(T^Z-\check{t})}z\right) \end{aligned} \tag{10}$$

Notice that $IC_{G^Z_{0,T^Z}}(T^Z)$ is continuous in T^Z. Moreover, if $T^Z = -\frac{1}{\lambda}ln(z)$, then G^Z corresponds to the Baseline equilibrium distribution, where $IC_{G^Z_{0,T^Z}}(T^Z) = 0$ if agents are risk-neutral and $IC_{G^Z_{0,T^Z}}(T^Z) > 0$ otherwise.[20] An OSSMP, therefore, must not only exist but be unique in this case. I denote such an optimal distribution as G^{Z*}.

[20] Evaluating at $T^Z = -\frac{1}{\lambda}ln(z)(0)$ we get $IC_{G^Z_{0,T^Z}}(T^Z) = (1-z)\left(1 - \left(\frac{\lambda z^{-\frac{\lambda^m}{\lambda}} - \lambda^m z^{-1}}{\lambda^m - \lambda} + 1\right)\left(\frac{z}{1-z}\right)^2\right)$. It is not hard to verify that this expression is decreasing in the ratio $\frac{\lambda^m}{\lambda}$. When $u(x) = x$, we have $\lambda^m = 2\lambda$ so that $IC_{G^Z_{0,T^Z}}(T^Z) = 0$, and otherwise $\frac{\lambda^m}{\lambda} \in (1,2)$ so that $IC_{G^Z_{0,T^Z}}(T^Z) > 0$.

Theorem 1 makes an additional claim about OSSPM in this case. When agents are risk neutral the optimal distribution G^{Z*} and hence $G^{R*}_{G^{Z*}}$ and payoffs are identical to the Baseline equilibrium. This is established in Lemma 3. The proof is just a few lines of algebra, which show that $\frac{dIC_{G^Z_{0,T^Z}}(T^Z)}{dT^Z} < 0$ for $T^Z < -\frac{1}{\lambda}ln(z)$.

Lemma 3. *If agents are risk neutral and $z \geq \alpha$, then the distribution of agreement times (and payoffs) in the unique OSSMP are identical to those in the Baseline equilibrium.*

This is a fairly strong negative result regarding the potential for mediation. Proposition 5 showed a similar negative result for the $N1$ protocol, however, there was no associated claim that the protocol was optimal.

Theorem 1 also makes a converse positive claim for risk neutral agents. If the probability of behavioral types is less than their demand, $z < \alpha$, then mediation can always improve on unmediated outcomes. This is established in Lemma 4, below. Combined with the preliminary analysis showing that mediation always benefits risk averse agents, this completes Theorem 1's characterization of when mediation is beneficial.

Lemma 4. *If $z < \alpha$, then an OSSMP delivers higher payoffs than the Baseline equilibrium.*

A somewhat counterintuitive implication of Lemmas 3 and 4 is that there is more chance that the mediator can benefit agents whose demands are initially further apart (although if 50% of agents are rational, beneficial mediation is possible regardless of demands). A basic intuition for the result is that Baseline equilibrium is less efficient for more extreme demands (with payoffs of $1-\alpha$), and so it is easier for the mediator to improve outcomes.

A more complete intuition for these results comes from considering how a mediator can reduce delay (the only way she can benefit risk neutral agents), while preserving incentives to confess rationality. Clearly, in order to preserve the incentive to confess, the mediator must delay rational-behavioral agreements by more than rational-rational agreements (we need $G^{R*}_{G^Z}(0) > G^Z(0)$ given $\alpha > 0.5$). Rational-behavioral agreements give a confessing agent a dollar share $1-\alpha$ with probability z, and a non-confessing agent α with probability $1-z$. If $z(1-\alpha) < (1-z)\alpha$ (or equivalently $z < \alpha$), therefore, delaying these agreements hurts a non-confessing agent more than a confessing agent and so improves the type incentive constraint compared to the Baseline equilibrium. On the other hand, if $z \geq \alpha$, then this delay worsens the type incentive constraint (which was already binding in the Baseline equilibrium) leading to the negative mediation result of Lemma 3.[21]

The proof of Lemma 4 is by construction, using a distribution which strictly delays some rational-behavioral agreements compared to the Baseline equilibrium (a feature of the optimal

[21] It might initially seem surprising that the cutoff for beneficial mediation doesn't depend on agents' impatience, after all the mediator improves outcomes by front loading rational-rational agreements while delaying behavioral-rational agreements. However, with continuous time, rescaling the discount rate (e.g. let $r' = 2r$) is no different from rescaling time, which clearly has no effect (e.g. let bargaining occur on $\{t' = t/2 : t \in [0, \infty)\} = [0, \infty)$).

distribution). Specifically, it uses the distribution $G^Z_{\check{t},T^Z(\check{t},1-\alpha)}$, where

$$T^Z(\check{t}, M) = \check{t} + \frac{1}{\lambda} ln\left(\frac{u(\alpha) - Me^{r\check{t}}}{z(u(\alpha) - u(1-\alpha))}\right) \tag{11}$$

is defined to ensure $U^n(\check{t}) = M$ (so $U^n(\check{t}) = M = 1 - \alpha$ in our particular case).[22] I then show that $G^R_{G^Z_{\check{t},T^Z(\check{t},1-\alpha)}}(0)$ is strictly increasing in $\check{t}$ for $\check{t} \approx 0$ if and only if $z > \alpha$.[23]

Of course, even the non-optimal N1 protocol offered improvements on the Baseline equilibrium when the probability of behavioral types was small. Indeed, that protocol also reduced delay of rational-rational agreements by less than rational-behavioral agreements. The symmetric version of the N1 protocol features a positive mass of rational-rational agreements at time zero compared to a continuous distribution of rational-behavioral agreements. The optimal distribution of rational-behavioral agreements is nonetheless importantly different from the N1 protocol, in featuring a positive interval with no agreement for risk neutral agents when $z < \alpha$ ($G^Z_{\check{t},T^Z}$ with $\check{t} > 0$). Illustrating this sub-optimatlity the $N1$ protocol can improve on Baseline payoffs for risk neutral agents if and only if $z < \hat{z}$ for some $\hat{z} < \alpha$. For example, when $\alpha = 0.6$ we have $\hat{z} = 0.41$.[24]

Let us now return to the task of identifying the optimal distribution of behavioral-rational agreement times more generally. Examining equation (10), it is clear that $IC_{G^Z_{\check{t},T^Z}}(T^Z)$ is continuous in $\check{t}$ and T^Z and strictly increasing in the former for $\check{t} < \hat{t}$, because the final integrand is negative for $s < \hat{t}$. Indeed, the negative final integrand of equation (8) for $s < t$, implies that $G^Z_{\min\{T^Z,\hat{t}\},T^Z}$ maximizes $IC_{G^Z}(T^Z)$ among all distributions G^Z with the same T^Z. This means that T^Z is consistent with the incentive constraints if and only if $IC_{G^Z_{\min\{T^Z,\hat{t}\},T^Z}}(T^Z) \geq 0$. In this case, define $t^*(T^Z) = \min\{\check{t} : IC_{G^Z_{\check{t},T^Z}}(T^Z) \geq 0\}$. Notice that if $IC_{G^Z_{0,T^Z}}(T^Z) \leq 0$, then $IC_{G^Z_{t^*(T^Z),T^Z}}(t) = 0$ for $t \in [t^*(T^Z), T^Z]$.

Lemma 5, below, establishes that an optimal distribution G^{Z*} must be of the form $G^Z_{t^*(T^Z),T^Z}$. The proof considers any distribution (not been ruled out by Lemma 2), where $G^Z(t) > 0$ for some $t < T^Z(t^*)$. It shows that reducing $G^Z(s)$ for some $s < T^Z(t^*)$, and increasing $G^Z(v)$ for some $v > t^*(T^Z)$ while leaving the time T^Z incentive constraint unchanged, increases $G^{R*}_{G^Z}(0)$.

---

[22] We need $U^n(t) = M = e^{-r\check{t}}[(1-z)G^Z_{\check{t},T^Z}(\check{t})(u(\alpha) - u(1-\alpha)) + u(1-\alpha)]$ and $G^Z_{\check{t},T^Z}(\check{t}) = \frac{1-ze^{\lambda(T^Z-\check{t})}}{1-z}$. The unique T^Z which allows both these equations to hold is $T^Z(\check{t}, M)$. To ensure $T^Z(\check{t}, M) \geq \check{t}$ we need $(1-z)u(\alpha)+zu(1-\alpha) \geq Me^{r\check{t}}$.

[23] To see how initial delay of behavioral-rational agreements, $\check{t} > 0$, allows for reduced delay of rational-rational agreements, notice that $U^n(\check{t}) = 1 - \alpha$ implies more rational-behavioral agreements by time $\check{t}$ (i.e. $G^Z_{\check{t},T^Z(\check{t},1-\alpha)}(\check{t}) > G^Z_{0,T^Z(0,1-\alpha)}(\check{t})$.)

[24] The symmetric $N1$ protocol agreement time distributions are $1 - G^Z(t) = \frac{e^{-\lambda t} - \bar{z}}{1-\bar{z}}$ and $1 - G^R(t) = (1-b)e^{-2\lambda t} - \frac{ze^{-2\lambda t}\left(2(e^{\lambda t}-1) - \bar{z}(e^{2\lambda t}-1)\right)}{(1-z)(1-\bar{z})}$, for $t \leq T^R = T^Z = -\frac{1}{\lambda} ln(\bar{z})$ where $\bar{z} = \frac{z}{1-(1-z)b}$. An $N1$ equilibrium with $b > 0$ exists if and only if $Q \geq 0$ (defined in equation (4)). It is readily verified that this condition is harder to satisfy for larger b and z. Such mediation, therefore, offers improvements over the Baseline equilibrium for risk neutral agents if and only if $\lim_{b\to 0} Q = \alpha - 0.5 - \frac{z}{1-z}(1-\alpha)(1-z^{\frac{r}{\lambda}}) > 0$. This defines an implicit cutoff $\hat{z}$ such that payoff improvements are possible if and only if $z < \hat{z}$. To see that $\hat{z} < \alpha$ notice that $\lim_{b\to 0} Q|_{z=\alpha} = -0.5 + \alpha^{\frac{\alpha}{1-\alpha}}$, which is positive whenever, $Q' = \frac{\alpha}{1-\alpha} ln(\alpha) - ln(0.5) \geq 0$. Because Q' is decreasing in α and $Q'|_{\alpha=0.5} = 0$ we must have $Q' < 0$ for $\alpha > 0.5$.

The logic behind this result is exactly the one highlighted in the discussion of the time t type incentive constraint (equation (8)). Recall that the final integrand records the time t incentive benefit less costs of increasing $G^Z(s)$, $u(1-\alpha)e^{\lambda^m s}z - u(\alpha)e^{-rs}(1-z)$, subject to not changing $G^Z(t)$. This is increasing in s, and so one by decreasing $G^Z(s)$ slightly early on in bargaining, you can increase $G^Z(v)$ by a lot later on in bargaining, while maintaining incentives, ultimately increasing $G^{R*}_{G^Z}(0)$. The distribution $G^Z_{t^*(T^Z),T^Z}$ is exactly the one which prioritizes making $G^Z(t)$ large as large as possible at the end of bargaining, while still satisfying $IC_{G^Z}(t) \geq 0$.

Lemma 5. *For any distribution* $G^Z \neq G^Z_{t^*(T^Z),T^Z}$ *with* $\min_s IC_{G^Z}(s) = 0 = IC_{G^Z}(t)$ *for* $t \in [\min\{\hat{t}, T^Z\}, T^Z]$*, we have* $G^{R*}_{G^Z_{t^*(T^Z),T^Z}}(0) > G^{R*}_{G^Z}(0)$.

The ability to restrict attention to distributions of the form $G^Z_{t^*(T^Z),T^Z}$ allows us to rapidly establish existence of an OSSMP. Because $IC_{G^Z_{\check{t},T^Z}}(T^Z)$ is continuous in $\check{t}$ and T^Z, and strictly increasing in the former, the set of T^Z for which $t^*(T^Z)$ is defined is closed. The fact that incentive constraints are satisfied under the Baseline distribution $G^Z_{0,-\frac{1}{\lambda}ln(z)}$ implies that this set is also non-empty. Moreover, $t^*(T^Z)$ is continuous on that closed set. Hence, the OSSMP problem can be reduced to maximizing a continuous function of T^Z, $G^{R*}_{G^Z_{t^*(T^Z),T^Z}}(0)$ on a compact set $\{T^Z : IC_{G^Z_{\min\{T^Z,\hat{t}\},T^Z}}(T^Z) \geq 0\} \cap \left[0, -\frac{1}{\lambda}ln(z)\right]$.[25]

Lemma 6, below, establishes three further facts about an OSSMP. Firstly, it is unique. Secondly, it features $t^*(T^Z) < T^Z$, so that the optimal distribution G^{Z*} is not degenerate. Thirdly, it features $t^*(T^Z) > 0$ when the probability of behavioral types is small or behavioral demands are large (in particular, $z < \alpha$ for risk neutral agents), so that that there is a non-degenerate interval with no behavioral-rational agreements.

The non-degeneracy of the distribution ($t^*(T^Z) < T^Z$), is interesting because it implies that the mediator is needed to help rational agents back down against behavioral opponents at the right time, and not just broker compromises. If $t^*(T^Z) = T^Z$ in the optimal distribution then a confessing agent could simply concede at T^Z, without the need for direction. It may not seem immediately obvious why this should hold. One might expect that (at least sometimes) one should delay behavioral-rational agreements as much as possible, with a probability one atom at some distant date. After all, for small behavioral probabilities these agreements seem to always give a non-confessing agent a larger payoff $(1-z)u(\alpha)$ than they could ever give to a confessing agent, $u(1-\alpha)$. To understand why a degenerate distribution is not optimal, suppose that it was, so that $t^*(T^Z) = T^Z$. In that case, the posterior probability of facing a behavioral type $\bar{z}(t) = \frac{z}{z+(1-z)(1-G^R(t))}$ would be close to one at $t \approx T^Z$, creating a strong incentive for a confession agent to concede, and making the dynamic incentive constraint impossible to satisfy without rapid rational-rational agreements. Moving to $t^*(T^Z) < T^Z$ reduces the dynamic

[25] Any distribution $G^Z_{\check{t},T^Z}$ with $T^Z > -\frac{1}{\lambda}ln(z)$ is worse than the Baseline equilibrium. That is, $G^Z_{\check{t},T^Z}(t) \leq G^Z_{0,T^{Z'}}(t)$ where $T^{Z'} = -\frac{1}{\lambda}ln(z)$, so that $G^{R*}_{G^Z_{0,T^{Z'}}}(0) > G^{R*}_{G^Z_{\check{t},T^Z}}(0)$, while $IC_{G^Z_{0,T^{Z'}}}(t) \geq 0$.

incentive constraint considerably and so allows much slower rational-rational agreements on $(t^*(T^Z), T^Z]$ and so ultimately more agreement at time zero. Moreover, the reduced delay of rational-behavioral agreements only increases a non-confessing agent's payoff very slightly if $t^*(T^Z) \approx T^Z$.

The claim that there is always a non-degenerate interval without rational-behavioral agreements, $t^*(T^Z) > 0$, when the probability of behavioral types is small or behavioral demands are large, illustrates that delaying those agreements (not just compared to rational-rational agreements but even with respect to the Baseline distribution) is a key feature of (optimal) mediation. This is consistent with the discussion following Lemma 4.

Lemma 6. *A unique OSSMP exists. The optimal behavioral rational distribution* $G^{Z*} = G^Z_{t^*(T^Z),T^Z}$ *satisfies* $t^*(T^Z) < T^Z$. *Moreover, there exists* $\underline{z}(\alpha, u) > 0$ *and* $\underline{\alpha}(z, u) < 1$ *such that if* $z < \underline{z}(\alpha, u)$ *or* $\alpha > \underline{\alpha}(z, u)$ *then* $t^*(T^Z) > 0$ *where* $\underline{z}(\alpha, u) = \alpha$ *and* $\underline{\alpha}(z, u) = z$ *if agents are risk neutral.*

Lemma 6's proof of uniqueness of the OSSMP is somewhat indirect. It takes as given the (already established) existence of a maximum utility $\overline{u} = U^c(0)$ achieved in an OSSMP. Because $IC_{G^{Z*}}(t) = 0$ for $t \in [t^*(T^Z), T^Z]$, we must also have $U^n(t) = \overline{u}$ for such t. Knowing this, we can consider a reduced problem of maximizing $G^{R*}_{G^Z_{\check{t},T^Z(\check{t},\underline{u})}}(0)$, with respect to $\check{t}$ where recall that $T^Z(\check{t}, \underline{u})$ is defined to ensure $U^n(T^Z) = \overline{u}$. This reduced problem has the same maximizers as the original problem, however, the objective function is more easily shown to be strictly quasi-concave. Moreover, the objective is strictly decreasing when $\check{t} \approx T^Z(\check{t}, \underline{u})$ and is strictly increasing for $\check{t} = 0$ when $z \approx 0$ or $\alpha \approx 1$, which establishes the Lemma's claims regarding $t^*(T^Z)$.

This Lemma completes the proof of Theorem 1, however, there is still more to be said about optimal mediation. In particular, Proposition 6, below, establishes that (given symmetric fundamentals) any optimal protocol is necessarily symmetric with the same distribution of agreement times as in the unique OSSMP. Given this, the OSSMP is clearly the unique optimal protocol if agents are risk averse in the sense that $u''(0.5) < 0$.[26] The proof supposes a optimal non-symmetric protocol, before showing that the weakly superior strongly symmetric protocol derived from this (i.e. with $\hat{G}^Z = 0.5(G^Z_1 + G^Z_2)$, $\hat{M}(0.5) = 1$) cannot be the OSSMP. This extension of Observation 3 (which said that a strongly symmetric protocol could always do at least as well as any other protocol) is somewhat interesting in that it suggests that a mediator should treat agents fairly, and not pick favorites if she wants to maximize total payoffs.[27]

Proposition 6. *Any optimal mediation protocol is symmetric with the same distribution of agreement times as in the unique OSSMP. If* $u''(0.5) < 0$ *then the unique optimal protocol*

[26] Or more precisely, any optimal protocol has the same distribution of equilibrium outcomes as the OSSMP.

[27] This contrasts with Kydd (2001)'s finding that biased mediator's do better than impartial ones, although the settings are quite different.

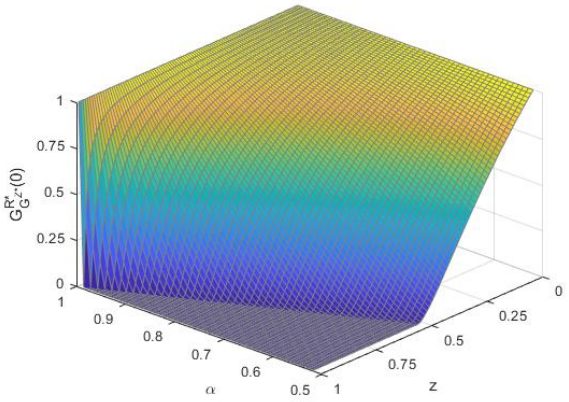

Figure 1. Numerical calculations of $G^{R*}_{G^{Z*}}(0)$ for risk neutral agents

is strongly symmetric.

The final result of this subsection, Proposition 7, provides some limiting comparative statics on behavioral demands, behavioral probabilities, and utility functions. These extend the findings of Theorem 1 that mediation was more feasible for risk neutral agents when behavioral demands were larger and behavioral probabilities were smaller, and was always feasible for risk averse agents. The proposition shows that if the probability of behavioral types is arbitrarily small, behavioral demands are arbitrarily large, or agents are arbitrarily risk averse, then mediation is approximately efficient in the sense that $G^{R*}_{G^{Z*}}(0) \approx 1$ (so that payoffs are approximately match those possible under complete information $(1-z)u(0.5) + zu(1-\alpha)$). These findings are fairly intuitive given the previous analysis, and indeed they can be established by using a lower bound on $G^{R*}_{G^{Z*}}(0)$ provided by the (non-optimal) $N1$ mediation protocol.

Proposition 7. *Consider sequences of symmetric bargaining games,* $B^n = (\alpha, z^n, u, r)$ *with* $\lim_n z^n = 0$, $B^n = (\alpha^n, z, u, r)$ *with* $\lim_n \alpha^n = 1$, *and* $B^n = (\alpha, z, u^n, r)$ *with* $\lim_n u^n(\alpha) = \lim_n u^n(0.5) > \lim_n u^n(1-\alpha)$. *In the associated sequences of OSSMPs* $\lim_n G^{R*}_{G^{Z*}}(0) = 1$.

Establishing comparative statics more generally is made difficult be the fact that the optimal mediation protocol does not have a closed form solution. However, Figure 1 presents numerical calculations of $G^{R*}_{G^{Z*}}(0)$ for risk neutral agents, and shows that this probability is always increasing in α and $-z$. The calculations also illustrate how large $G^{R*}_{G^{Z*}}(0)$ can be, even for quite large behavioral probabilities. For instance, if $z = 0.5$ and $\alpha = 0.9$ then $G^{R*}_{G^{Z*}}(0) = 0.85$ so that rational payoffs (of 0.27) are very close to those under complete information (of $0.3 = (1-z)u(0.5) + zu(1-\alpha)$) and almost three times larger than those in unmediated bargaining (of 0.1).

4.1 A mechanism design benchmark

In this subsection, I compare optimal mediation to a mechanism design benchmark when the mechanism designer can impose agreement and (perpetual) disagreement between agents. That is, there is no need to ensure that agents follow through on the designers' suggestions. Agents report their types to the designer, who chooses an agreement time and terms, or perpetual disagreement, based on the reported types. Behavioral agents always report their true type to the designer, but rational agents may lie. I require that the designer imposes perpetual disagreement between two reported behavioral types and only ever imposes agreements of $(\alpha, 1-\alpha)$ between rational-behavioral pairs.[28] Again, I restrict attention to symmetric bargaining problems and have the designer maximize the sum of rational agents payoffs.

For the same reasons as in Observation 3 we can restrict attention to strongly symmetric mechanisms with $G_i^Z = G^Z$ and $M_i^t(0.5) = 1$ (transforming an arbitrary mechanism into a strongly symmetric one weakly improves incentives and payoffs). I call an optimal mechanism of this kind an an Optimal Strongly Symmetric Delegation Mechanism (OSSDM), because agents delegate their subsequent decision making power. Individual rationality constraints can also be added, to ensure that agents want to delegate their decision making power, however, these will typically not bind (see discussion below). Formally, the OSSDM problem is as follows:

$$\begin{aligned}\max_{G^R,G^Z} U^c =&(1-z)\int e^{-rs}u(0.5)dG^R(s) + z\int e^{-rs}u(1-\alpha)dG^Z(s)\\ \text{s.t. } U^c \geq U^n =& (1-z)\int e^{-rs}u(\alpha)dG^Z(s) \qquad \text{(Type IC}^*\text{)}\end{aligned}$$

This problem is much simpler than the optimal mediation problem (OSSMP). I characterize its solution in the following proposition.

Proposition 8. *In any OSSDM, pairs of rational agents agree immediately,* $G^R(0) = 1$. *Furthermore, if* $z \geq \frac{u(\alpha)-u(0.5)}{u(1-\alpha)+u(\alpha)-u(0.5)}$ *then* $G^Z(0) = 1$, *and otherwise* $\int e^{-rs}dG^Z(s) = \frac{(1-z)u(0.5)}{(1-z)u(\alpha)-zu(1-\alpha)}$.

The proof of this characterization is trivial. Increasing $G^R(0)$ strictly improves the objective function and the type incentive constraint, so that we must have $G^R(0) = 1$. On the other hand, increasing $\int e^{-rs}dG^Z(s)$ strictly improves the objective function, but worsens the incentive constraint. If $z \geq \frac{u(\alpha)-u(0.5)}{u(1-\alpha)+u(\alpha)-u(0.5)}$ then $(1-z)u(0.5) + zu(1-\alpha) \geq (1-z)u(\alpha)$, so that the incentive constraint is satisfied even with $G^Z(0) = 1$, and otherwise the constraint must bind in an OSSDM, so that $\int e^{-rs}dG^Z(s) = \frac{(1-z)u(0.5)}{(1-z)u(\alpha)-zu(1-\alpha)}$.

There are many distributions which make the constraint bind when $z < \frac{u(\alpha)-u(0.5)}{u(1-\alpha)+u(\alpha)-u(0.5)}$. One such distribution implies agents either agree immediately with probability $G^Z(0) = G^Z(\infty) =$

[28] Allowing the designer to give a behavioral type more than α would not affect things, because the designer would never choose to do so.

$\frac{(1-z)u(0.5)}{(1-z)u(\alpha)-zu(1-\alpha)}$, or never agree. Others, however, have eventual agreement, for instance $G^Z(t) = 0$ if $t < -\frac{1}{r}ln\left(\frac{(1-z)u(0.5)}{(1-z)u(\alpha)-zu(1-\alpha)}\right)$ and $G^Z(t) = 1$ otherwise. Interesting, as the probability of behavioral types vanishes, this latter distribution converges to a point mass at $-\frac{1}{r}ln\left(\frac{u(0.5)}{u(\alpha)}\right)$, which is exactly the limiting distribution of behavioral-rational agreements in the OSSMP (this is shown in the proof of Lemma 6). Proposition 7 shows also $G^{R*}_{G^{Z*}}(0)$ converges to one in the OSSMP in that case. In other words, the OSSMP and OSSDM are identical in the limit.

More generally, however, this characterization shows how much the agents' freedom to ignore the mediator's suggestions constrained mediation. Certainly, an OSSDM always achieves a strictly higher payoff than the OSSMP.[29] Not only do rational agent pairs reach immediate agreement, $G^R(0) = 1$, but because the type incentive constraint is less strict than under mediation (agents can't pretend to be behavioral and then subsequently concede) we can have an efficient outcome without any delay, $G^Z(0) = 1$, when behavioral types are likely or make moderate demands (close to 1/2), $z \geq \frac{u(\alpha)-u(0.5)}{u(1-\alpha)+u(\alpha)-u(0.5)}$. For risk neutral agents this reduces to $z \geq 2\alpha - 1$. By contrast for risk neutral agents and parameters $z \geq \alpha > 2\alpha - 1$, mediation is unable to improve on Baseline equilibrium payoffs at all!

More generally, comparative statics appear to be quite different to those established under mediation. Payoffs in an OSSMP can be written as $G^{R*}_{G^{Z*}}(0)(1-z)(u(0.5)-u(1-\alpha))+u(1-\alpha)$ where $G^{R*}_{G^{Z*}}(0)$ is increasing in α and $-z$ for risk neutral agents (see Figure 1). In an OSSDM, payoffs are effectively of the form $(1-z)u(0.5) + G^Z(0)zu(1-\alpha)$ where $G^Z(0)$ is decreasing in α and $-z$. Perhaps a more consistent way to compare these problems is using a measure of efficiency given by the difference between these payoffs (U) and the Baseline equilibrium ($U^B = u(1-\alpha)$), divided by the difference between complete information payoffs ($U^{CI} = (1-z)u(0.5)+zu(1-\alpha)$) and the Baseline equilibrium payoffs (i.e. $e = \frac{U-U^B}{U^{CI}-U^B}$). Efficiency is then 100% in an OSSDM when behavioral types are likely or their demands are moderate, but strictly smaller otherwsise. For risk neutral agents, efficiency is 0% in an OSSMP when behavioral types are likely and their demands are moderate, but strictly larger otherwise. In all cases, however, efficiency approaches 100% when the probability of behavioral types is very large, $z \approx 0$, or their demands are very small very unlikely, $\alpha \approx 1$.

As mentioned, the OSSDM problem as written above, lacks individual rationality constraints, which might be thought to constrain its efficiency. Why should agents delegate their decision making power to the mediator? A natural constraint is that agents do better than in the Baseline equilibrium. This is certainly true for rational agents, as $U^c > u(1-\alpha)$. For behavioral types, a plausible assumption is that they have the same discount rate and same utility as rational agents for dollar shares greater than α but obtain $-D$ for any smaller share, for D large. A weak improvement on the Baseline equilibrium then translates into a constraint, $U^n \geq (1-z)u(\alpha)\int_0^{T^Z} e^{-rs}dG^Z = u(1-\alpha)(1-e^{rT^Z}z)$, where $T^Z = -\frac{ln(z)}{\lambda}$ (behavioral types must expect a

[29] The OSSMP distributions $G^{R*}_{G^{Z*}}$ and G^{Z*} satisfy the OSSDM type incentive constraint strictly. Setting $G^R(0) = 1$ instead, strictly increases payoffs while preserving incentives.

payoff $e^{-rt}u(1-\alpha)z$ less than a rational agent who concedes at T^Z). Clearly this constraint is also satisfied as $U^n = U^c > u(1-\alpha)$ if $z < \frac{u(\alpha)-u(0.5)}{u(1-\alpha)+u(\alpha)-u(0.5)}$ and $U^n = (1-z)u(\alpha)$ otherwise.

We can now imagine an extended bargaining game, where at time zero agents can agree to participate in an OSSDM, with reputational bargaining continuing if at least one agent refuses to participate. Because all agents (strictly) prefer the OSSDM to their Baseline equilibrium payoffs, beliefs can remain unchanged if one agent did in fact refuse to participate.

While an OSSDM achieves a strictly higher objective than the OSSMP, the credibility of this mechanism seems slightly dubious. It requires agents to fully delegate future decision making to the designer, who can then maintain perpetual disagreement between two (reportedly) behavioral types. While contracts constraining future agreements may sometimes drawn up, courts typically do not enforce contracts in the absence of a harmed party (i.e. there is no party with standing to enforce the contract). Seemingly, therefore, a rational agent should always have the option to pretend to be behavioral and then change her mind and accept her opponent's demand.

While the designer could impose agreements in an OSSDM, I have not referred to it as arbitration because it sometimes imposes perpetual disagreement, which conflicts with the typical practice of arbitration as a form of Alternative Dispute Resolution. An arbitrator who always immediately imposes some dollar division even between behavioral types would seem likely to face fierce opposition. Using the assumptions about utility outlined above, this would necessarily give at least one behavioral agent a payoff of less than $u(1) - D\frac{z}{2}$, which is worse than perpetual disagreement for large D. In this case, therefore, the designer would seem unable to satisfy any reasonable individual rationality constraint for behavioral types. This illustrates an important point. The knowledge that a mediator will never impose an agreement that an agent strongly dislikes, may be an important selling point of mediation compared to arbitration. This can help explain its greater popularity in Stipanowich and Lamare (2013)'s survey.

4.2 Non α-optional equilibria

In this subsection, I first discuss how the restriction to α-optional equilibria (under which I identified an optimal mediation protocol) is substantive, before arguing that there are good reasons to make that restriction.

The α-optional restriction is substantive not only because there are other equilibria, but because these can deliver strictly higher payoffs than the OSSMP identified in Theorem 1 for risk neutral agents when behavioral types are unlikely. In the Supplementary Material (Appendix B), I consider an equilibrium which does this, with the following form. Rational agents always confess at time 0^2. If both agents confess, at time 0^3 the mediator selects one of the two agents with equal probability, and suggests that she demand $\alpha_i(0^4) = 1$. When the mediator (later)

suggests an agreement, she always suggests that the selected agent, i, gets the whole dollar. If only agent i messages the mediator, however, then the mediator also always tells her to demand $\alpha_i(0^4) = 1$. In this case, clearly agent i obtains the share $(1 - \alpha)$ in any subsequent agreement with her behavioral opponent. If neither agent confesses, then the mediator says nothing. If an agent fails to follow the mediator's suggestion at some time, then the mediator subsequently says nothing and the continuation equilibrium specifies that she should concede to her opponent immediately.

The reason such an equilibrium can outperform an OSSMP is that by taking away the option to concede from one of the rational agents, the mediator may face less demanding incentive constraints. There is only a single dynamic incentive constraint for the lucky rational agent who is told to demand $\alpha_i(0^4) = 1$, and to satisfy this constraint, the agreement rate can be extremely slow, because she is being promised a large payoff, the entire dollar in any future agreement with a rational opponent. As a result, the probability of agreement at time zero can be made larger than in an OSSMP. For large probabilities of behavioral types, however, low option equilibria cannot exist (rational agents would never confess), because the information that the mediator immediately gives to a non-confessing rational agent (that she she faces a behavioral opponent and so can concede) is too valuable.

While there is an important insight to be gained from low option equilibria, that the mediator may benefit by limiting agents' opportunities to strike deals without her, there are several reasons to believe that such mediation strategies will not be used in practice. Most notably, the entire set of equilibria are α-optional if agents initially make behavioral demands and can always recall past offers. The first assumption seems reasonable, since mediators typically only get involved after parties begin a dispute (recall, if one agent initially reveals rational by making a non-behavioral demand then there is immediate agreement even in unmediated bargaining). The second assumption also seems reasonable for many bargaining situations. It would hold endogenously if a rational agent's preferences change, so that she becomes unwilling to accept an agreement less than her opponent's most generous previous offer, the assumption of Fershtman and Seidmann (1993).

There are several further reasons to doubt the extreme/dispersed agreements of low option equilibria. First, quite simply, they are nonintuitive. In practice, mediators typically try get both parties to compromise on their initial demands, rather than encouraging one of them to demand even more. This may be explained by risk aversion, which is in clear tension with the dispersed agreements (I only show low option equilibria can benefit risk neutral agents).

Another (out of model) reason to doubt such agreements is that they may be impossible to implement if bargaining actually takes place in discrete time (rather than discrete-continuous time). In an alternating offer model, for instance, after agent i reveals her rationality (by demanding the entire dollar) it would seem that her rational opponent j could always subsequently

reveal her own rationality to secure her positive alternating offer payoff, instead of 0. In Section 5 I explain how in discrete time, an informed mediator can implement any agreement for rational agents between behavioral types' demands (i.e. any α-optional equilibrium agreement) by revealing agents' private information sequentially, but that explanation does not extend to more extreme demands.

Finally, there are practical reasons to limit attention to α-optional equilibria. Such equilibria represent a tractable class, with their possible outcomes exactly characterized as those satisfying the dynamic and type incentive constraints. This allowed me to restrict attention to direct mediation equilibria, where the mediator only sends a single message to agents suggesting the terms of an agreement.

5 Discussion and literature review

Although the potential for mediators to expand payoff sets is well known in economics (for instance, the set of correlated equilibria is typically larger than the set of Nash equilibria), the role for mediators in dynamic bargaining settings, where real world mediators practice, has received relatively little attention.[30] As mentioned in the Introduction, an important exception to this is Jarque et al. (2003).

Jarque et al. (2003) consider a model with incomplete information about reservation values. Time is continuous. Agent i gets utility $e^{-rt}(x_i - s_i)$ when she gets a share x_i of the dollar at time t and has reservation value $s_i \in [s_i^L, s_i^H]$. A war of attrition equilibrium always exists in this setting, where only on two possible agreements are used (agent i demands $\overline{x}_i > s_i^H$). A mediator adopts Dunlop's simple mediation protocol (effectively my $I\infty$ protocol), immediately announcing an agreement $m_i \in (1 - \overline{x}_j, \overline{x}_i)$, whenever both parties accept this in private. The main result is that when fundamentals are symmetric, there is an equilibrium with mediation if and only if the fraction of types willing to concede in the war of attrition is sufficiently small ($s_i^L \approx 1 - \overline{x}_j$). This is ex-ante more efficient than the war of attrition. Čopič and Ponsatí (2008) extends this model to allow for a continuum of possible agreement terms, and illustrates the existence of a mediator supported ex-post efficient equilibrium.

Why should Dunlop's simple mediation protocol appear to be effective when agents have incomplete information about reservation values, but not in the reputational model? The reason is that with reservation values, the mediator facilitates agreement between types who would never agree in the war of attrition. For example, if $\overline{x}_1 = \overline{x}_2 = 0.7$, then types $s_1 = s_2 = 0.4$ will never concede in a war of attrition, but by introducing the alternative $m_i = 0.5$, these types can

[30] There is a game theoretic literature on mediation of international conflicts in political science, which considers conflict as a bargaining impasse caused by asymmetric information. However, these models typically abstract from the dynamic nature of bargaining, instead focussing on simple take-it-or-leave-it protocols (see Powell (2002)).

reach agreement. The extra payoffs created by such agreements add enough grease to the system to overcome the inherent difficulties of mediation. In the reputational model, by contrast, introducing a compromise agreement does not expand the set of types who ultimately agree.

One "problem" of the reservation value model of Jarque et al. (2003) (and Čopič and Ponsatí (2008)) is that it is not clear whether mediation offers Pareto improvements over unmediated bargaining, because of the existence of multiple equilibria absent mediation. Ponsati (1997) shows in the reservation value setting, that if the game rules allow only three alternatives agreements and strategies are Markov, then at least one alternative will not be used when there is no mediator (i.e. there is a war of attrition). However, she also shows that it is possible to construct non-Markov equilibria, in which all three alternatives are used (i.e. there are compromise agreements) and these provide ex-ante Pareto improvements on the war of attrition equilibria. The uniqueness of the Baseline equilibrium in the reputational model, allows me to say more clearly when mediation is beneficial and when it is not.

Proposition 4's finding that noise (added to the simple mediation protocol) can allow for more effective communication is in line with the existing literature. Myerson (1991) highlights how noiseless communication may be completely uninformative (absent any filtration) in very simple sender-receiver games, while noisy communication is informative. Goltsman et al. (2009) characterize the full extent to which mediation, arbitration and negotiation (finitely many rounds of communication with no discounting) can improve receiver payoffs in a cheap talk game. They show that both mediators and arbitrators should both filter information, but mediators should also add noise. They further show that arbitration is (generically) more effective than mediation, while mediation is only sometimes more effective than communication. Hörner et al. (2015) by contrast show that arbitration and mediation are equally effective at deterring conflict in a simple game in which parties choose whether to go to war.

Although my mediator has no source of her own information about the state of the world, there are some strong similarities to the literature on dynamic information design. For instance, Ely (2017) considers a model in which the information designer knows whether an agent has received an email but would like to keep that agent working for as long as possible. The agent always has the option of checking her email but would only like to do so if she has actually received an email. The designer's optimal solution features the gradual release of information about the existence of an email to prevent the agent checking on her own. Despite my mediator wanting to minimize agents' delay of agreement, she also releases her information only gradually, because not doing so would destroy agents' incentives to reveal their information to her in the first place.

Basak (2016) considers a model very similar to reputational bargaining with a form of information design. A third-party has access to an informative signal about the likelihood that an agent is committed to her bargaining demand. If the signal is perfectly informative then its

release makes bargaining efficient (eliminating delay when at least one party is not committed), however, if the signal is only moderately informative, then its release may lower payoffs and increase the expected delay.

While the paper has focussed on the potential of mediation to improve on unmediated outcomes, it is also of interest to consider whether mediation can make players worse off. Because players always have the option of conceding, at least one player can guarantee her Baseline equilibrium payoff, $u_i(1-\alpha_j)$. I show in the Supplementary Material (Appendix C), however, that when behavioral types are sufficiently unlikely, there are mediation protocols which hold both players down to this outside option, generically implying Pareto losses.[31]

One obvious direction for future work is to extend the characterization of optimal mediation to non-symmetric problems. While a worthy goal, this also appears quite challenging. Many of the arguments that dramatically simplified the symmetric problem do not immediately extend, in particular the argument restricting attention to strongly symmetric protocols, which reduced four (infinite dimensional) objects G^R, G_1^Z and G_2^Z and M_1 with $[0,\infty)^4$ (non-independent) constraints into two objects G^R and G^Z with just $[0,\infty)^2$ constraints. Another complicating factor is that when one agent is more patient than her opponent, other things equal, it is more efficient to give her a smaller share in earlier agreements and a larger share in later agreements.

Another direction to extend the analysis is to allow players to imitate multiple behavioral types, a feature of AG. Clearly, this makes the design of an optimal mediation protocol more complex still because we must consider that mediation may affect rational players' demand choices. For instance, mediating bargaining when players make extreme demands (typically implying low payoffs for both agents) can cause players to make those demands more frequently. In fact, in the Supplementary Material (Appendix C), I show that as a result of this demand distorting effect, mediation can strictly lower the payoffs of *both* rational agents compared to the Baseline equilibrium when behavioral types are unlikely.

Another difference from AG is my use of discrete continuous time, instead of starting with discrete time and taking a continuous time limit. It may be thought that this involves an important loss of generality, because when the mediator reveals that both agents are rational any surplus division is consistent with equilibrium continuation play, but that typically isn't the case in discrete time. For instance in a complete information alternating offer model with period length Δ, agents must agree to a unique division $(\alpha_1^R, 1-\alpha_1^R)$ immediately (e.g. $\alpha_1^R = \frac{1-e^{-r_2\Delta}}{1-e^{-(r_1+r_2)\Delta}}$ with risk neutrality). However, by revealing information sequentially, an informed mediator can still implement any compromise for rational agents between behavioral types' demands when Δ is small.

To illustrate the idea, suppose that the informed mediator wants to immediately implement a

[31] The potential for a third party to increase delay in bargaining is explored by Manzini and Ponsati (2006), who show that bargainers may delay agreement in a complete information alternating offer model until a third party arrives. Agents do this in order to extract additional resources from the third party, who has a stake in the outcome.

division (m_1, m_2) between rational agents with $m_1 \in (1-\alpha_2, \min\{\alpha_1^R, \alpha_1\})$ and wants rational agents to concede to behavioral types. If agent 1 is behavioral, the mediator announces this immediately. If (rational) agent 1 then immediately demands m_1, the mediator reveals whether agent 2 is rational in period 2 if agent 2 doesn't accept, but otherwise remains silent. Accepting m_1 is clearly optimal for a rational agent 2 given that $m_1 \leq \alpha_1^R$. Demanding m_1 is optimal for rational agent 1 for small Δ because otherwise the resulting game has one-sided private information and so can only give her a payoff of at most marginally more than $u_1(1-\alpha_2)$ by the logic of the Coase conjecture (see AG). This argument does not extend to implementing rational-rational agreements with $m_1 > \max\{\alpha_1^R, \alpha_1\}$, suggests that more extreme agreements, present in the low option of subsection 4.2, will be impossible to implement in many discrete time games. If rational agent 1 demanded $m_1 > \max\{\alpha_1^R, \alpha_1\}$, a rational opponent could subsequently also reveal rationality to guarantee $e^{-r_2\Delta}u_2(\alpha_2^R) = u_2(1-\alpha_1^R)$.

6 Appendix A: Proofs

Proof of Proposition 2. Suppose there is an equilibrium $\sigma = (\sigma_1, \sigma_2)$ with $c_i c_j > 0$. Let $A_i^c = \{t : U_i^c(t) = \max_s U_i^c(s)\}$ and $A_i^n = \{t : U_i^n(t) = \max_s U_i^n(s)\}$. Since σ is an equilibrium, $A_i^n \neq \emptyset \neq A_i^c$. Define $T_i^c = \inf\{t : F_i^c(t) = 1\}$, as the final time by which a confessing agent i concedes to her opponent. Similarly, define $T_i^n = \inf\{t : (1-c_i)(1-F_i^n(t)) = z_i\}$ as the final time a rational, non-confessing agent i concedes to her opponent. Finally, define $T^* = \max\{T_j^c, T_j^n, T_i^c, T_i^n\}$ and $T^c = \min\{T_i^c, T_j^c\}$. I next prove a series of claims, which help establish the result.

(a) *We must have* $T_j^c \leq T_i^n < \infty$. To establish $T_i^n \geq T_j^c$ suppose instead that $T_i^n < T_j^c$ then after time T_i^n a confessing agent j knows that she faces a behavioral opponent, and so would prefer to concede immediately rather than wait until T_j^c.
To establish $T_i^n < \infty$, let π_j^t be the conditional probability that agent j continues to act consistent with a behavioral type on the interval $[s, s+t)$ for arbitrary s. For agent i not to concede at s it must be that:

$$\begin{aligned} u_i(1-\alpha_j) \leq & (1-\pi_j^t)u_i(1) + \pi_j^t e^{-r_i t}u_i(1) \\ \pi_j^t \leq & \frac{u_i(1) - u_i(1-\alpha_j)}{(1-e^{-r_i t})u_i(1)} \end{aligned}$$

where the second line simply rearranges the first. Fix $\delta \in \left(\frac{u_i(1)-u_i(1-\alpha_j)}{u_i(1)}, 1\right)$, and consider K such that $\delta^K < z_i$ and t' such that $\delta = \frac{u_i(1)-u_i(1-\alpha_j)}{(1-e^{-r_i t'})u_i(1)}$. Suppose agent i did not to concede on the interval $[0, t'K)$ then it must be that the probability j acts consistent with a behavioral type on that interval is less than $(\pi_j^{t'})^K \leq \delta^K < z_i$, but this contradicts the fact that a behavioral type acts like itself. And so rational agent i will always concede by $T_i^n \leq t'K$

(b) *We must have* $\max\{T_j^c, T_j^n\} = T^* < \infty$. I first claim that $T_i^n \leq \max\{T_j^c, T_j^n\}$. Suppose not, so that $T_i^n > \max\{T_j^c, T_j^n\}$. Then after time $\max\{T_j^c, T_j^n\}$ a non-confessing rational agent i knows that she faces a behavioral opponent, and so would prefer to concede immediately rather than wait until T_i^n. By claim (a) we already know that $T_i^c \leq T_j^n \leq \max\{T_j^c, T_j^n\}$, hence $\max\{T_i^c, T_i^n\} \leq \max\{T_j^c, T_j^n\}$. Reversing the labelling we also have $\max\{T_i^c, T_i^n\} \geq \max\{T_j^c, T_j^n\}$, which establishes $\max\{T_j^c, T_j^n\} = T^*$. Claim (a) implies $\max\{T_i^c, T_j^c\} \leq \max\{T_j^n, T_i^n\} < \infty$, so that $T^* \leq \infty$.

(c) *There is no jump in F_i^c at $t \in (0, T^*]$. Furthermore, if $F_j^n(0) > 0$ then $F_i^c(0) = 0$.* Suppose F_i^c jumped at $t \in (0, T^*]$, then F_j^n must be constant on $[t - \varepsilon, t]$ for some $\varepsilon > 0$, as non-confessing agent j would prefer instead to concede an instant after t rather than on the interval $[t - \varepsilon, t]$. But in which case, a confessing agent i would prefer to concede at $t - \varepsilon$ rather than wait until t. Finally, if $F_j^n(0) > 0$ then a confessing agent i would strictly prefer to concede an instant after zero rather than at zero, so that $F_i^c(0) = 0$.

(d) *There is no jump in F_i^n at $t \in (0, T^*]$. Furthermore, if $F_j(0) > 0$ then $F_i^n(0) = 0$.* Suppose that F_i^n did jump at $t \in (0, T^*]$, then F_j is constant on $[t - \varepsilon, t]$ for some $\varepsilon > 0$, as a rational agent j would prefer instead to concede an instant after t rather than on the interval $[t - \varepsilon, t]$. But in which case, a non-confessing agent i would prefer to concede at $t - \varepsilon$ rather than wait until t. Finally, if $F_j(0) > 0$ then a non-confessing agent i would strictly prefer to concede an instant after zero rather than at zero, so that $F_i^n(0) = 0$.

(e) *If F_i^n is continuous at t then so is U_j^c. If F_i is continuous at t then so is $U_j^n(s)$.* This follows from the definitions.

(f) *If $T^* \geq t'' > t'$ then $F_i(t'') > F_i(t')$.* Suppose not, then let $t_i^* = \sup\{t : F_i(t) = F_i(t')\} \in [t'', T^*]$. First, notice that no rational agent j can concede at $s \in (t', t_i^*)$ because this is strictly worse than conceding slightly earlier (e.g. at $\frac{s+t'}{2}$). Combining this with the continuity of F_j, U_i^c and U_i^n on $(0, T^*]$, established in claims (c), (d) and (e), implies that rational agent i (whether she confessed or not) would strictly prefer to concede at some early point in (t', t_i^*), such as $\frac{t_i^*+t'}{2}$, than wait to concede at or just after t_i^*. This, however, contradicts the definition $t_i^* \leq T^* < \infty$.

(g) *If $T_j^c \geq t'' > t'$ then $F_i^n(t'') > F_i^n(t')$.* Suppose not, then let $t_i^{*n} = \sup\{t : F_i^n(t) = F_i^n(t')\} \in [t'', T^*]$. First, notice that a confessing agent j will not concede at $s \in (t', t_i^{*n})$ because this is strictly worse than conceding slightly earlier (e.g. at $\frac{t'+s}{2}$). This ensures that $T_j^c \geq t_i^{*n}$. When combined with claim (f), we must have that F_n^j and F_i^c are strictly increasing on the interval (t', t_i^{*n}), i.e. $F_j^n(t'') > F_j^n(t')$ and $F_i^c(t'') > F_i^c(t')$. Because F_i^c is strictly increasing on (t', t_i^{*n}), we must have that A_i^c is dense on that interval. By claims (d) and (e) U_i^c is continuous and hence constant on (t', t_i^{*n}). In turn that ensures that U_i^c is differentiable on (t', t_i^{*n}) with $\frac{dU_i^c(t)}{dt} = 0$, so that a non-confessing agent j must be conceding at rate $\frac{f_j^n(t)}{1-F_j^n(t)} = \lambda_j$. Notice, however, that because $c_j(1 - F_j^c(t)) > 0$ for $t < T_j^c$ where $T_j^c \geq t_i^{*n}$ for $t \in (t', t_i^{*n})$ we must have:

$$\frac{f_j(t)}{1 - F_j(t)} = \frac{(1 - c_j) f_j^n(t)}{(1 - c_j)(1 - F_j^n(t)) + c_j(1 - F_j^c(t))} < \frac{f_j^n(t)}{1 - F_j^n(t)} = \lambda_j$$

A concession rate of exactly $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$ would make a non-confessing agent i indifferent between conceding at any $t \in (t', t_i^{*n})$ and so a smaller concession rate, $\frac{f_j(t)}{1-F_j(t)} < \lambda_j$, means that she would strictly prefer to concede earlier rather than later on the interval (t', t_i^{*n}). The continuity of F_j and hence U_i^n on $(0, T^*]$, established in (c) and (d), then means that a non-confessing agent i must get a strictly lower payoff when conceding at or just after t_i^{*n} than if she conceded earlier (e.g. at $\frac{t'+t_i^{*n}}{2}$). This means that t_i^{*n} cannot be the supremum, a contradiction.

(h) *If $T_j^c > 0$, then agent j must concede at rate $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$ rate on $(0, T_j^c]$.* If $T_j^c > 0$, then claim (g) implies that F_i^n is strictly increasing on $[0, T_j^c]$, and so A_i^n is dense in $[0, T_j^c]$. From claims (c), (d), and (e) it follows that U_i^n is continuous. Hence, U_i^n is constant on this interval, and so differentiable with $\frac{dU_i^n(t)}{dt} = 0$, which implies that agent j concedes at rate $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$.

(i) *If $T_j^c < T^*$, then $\frac{f_j(t)}{1-F_j(t)} = \frac{f_j^n(t)}{1-F_j^n(t)} = \lambda_j$ on $(T_j^c, T^*]$.* First, suppose that $T_j^c \geq T_i^c$, then by claim (f) we must have that F_i^n is strictly increasing on $[T_i^c, T^*]$, and so A_i^n is dense in $[T_i^c, T^*]$. From claims (c), (d), and (e) it follows that U_i^n is continuous and hence is also constant on $(T_i^c, T^*]$. In turn that implies that U_i^n is differentiable on $(T_i^c, T^*]$ with $\frac{dU_i^n(t)}{dt} = 0$, and so agent j must concede at rate $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$. Next, suppose that $T_j^c < T_i^c$, so that at $t > T_j^c$, a confessing and non-confessing agent i have the same beliefs (in particular, both are certain that they face a non-confessing opponent j). This implies $A_i^n \cap (T_j^c, T^*] = A_i^c \cap (T_j^c, T^*]$. By claim (f) we know that

F_i is strictly increasing on $(T_j^c, T^*]$, which implies that $A_i^n \cup A_i^c$ is dense in $(T_j^c, T^*]$, and so A_i^n is also dense on that interval. From claims (c), (d), and (e) it follows that U_i^n is continuous and hence constant on $(T_j^c, T^*]$. In turn that implies that U_i^n is differentiable on $(T_j^c, T^*]$ with $\frac{dU_i^n(t)}{dt} = 0$, and so agent j must concede at rate $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$. Finally, notice that for $t \geq T_j^c$ we have $1 - F_i(t) = (1 - c_i)(1 - F_i^n(t))$ and so $\frac{f_j(t)}{1-F_j(t)} = \frac{f_j^n(t)}{1-F_j^n(t)}$.

(j) *If $T^c \geq t'' > t'$, and $F_j^c(t'') = F_j^c(t')$ then $F_j^c(t'') = F_i^c(t'') = 0$.* I first claim that $F_i^c(t'') = F_i^c(t')$. To see this, notice that if $F_j^c(t'') = F_j^c(t')$, then to ensure that agent j on average concedes at rate $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$ on (t', t'') as required by claim (h), a non-confessing agent j must concede at rate:

$$\frac{f_j^n(t)}{1 - F_j^n(t)} = \lambda_j \left(1 + \frac{c_j(1 - F_j^c(t))}{(1 - c_j)(1 - F_j^n(t))} \right) \tag{12}$$

For $t < T^c$, however, $c_j(1 - F_j^c(t)) > 0$ and so this rate is strictly greater than λ_j, which implies that a confessing agent i would strictly prefer to concede at t'' rather than on the interval (t', t''). Next, define $t_i^{**} = \inf\{s : F_i^c(s) = F_i^c(t')\} \leq t'$. The previous argument implies $t_i^{**} = t_j^{**}$. Anticipating that a non-confessing agent j will concede at a rate larger than λ_j on (t_i^{**}, t'') (as specified in equation (12)), a confessing agent i cannot concede on $[t_i^{**} - \varepsilon, t'')$ for some $\varepsilon > 0$, because she would prefer to concede at t'' instead. This implies $t_i^{**} = 0$ and $F_i^c(t') = 0$. The continuity of F_i^c on $(0, T^*]$, established in claim (c), then implies $F_i^c(t'') = F_i^c(t') = 0$.

(k) *Suppose $T^c > 0$, let $t^{*c} = \inf\{t : F_1^c(t) > 0$ or $F_2^c(t) > 0\}$, and suppose $t^{*c} \leq t' < t'' \leq T^c$, then $F_i^c(t'') > F_i^c(t')$.* Suppose not, so that $F_i^c(t'') = F_i^c(t')$. Claim (j) then implies that $F_1^c(t'') = F_2^c(t'') = 0$. Because $t^{*c} < t''$, however, we must have either $F_1^c(t'') > 0$ or $F_2^c(t'') > 0$ given $t^{*c} < t''$, a contradiction.

(l) *If $T^c > 0$ then $\frac{f_j^n(t)}{1-F_j^n(t)} = \frac{f_j^c(t)}{1-F_j^c(t)} = \lambda_j$ on $(t^{*c}, T^c]$, where t^{*c} is defined in claim (k).* First notice that by (k) A_i^c must be dense in $[t^{*c}, T^c]$. From claims (d) and (e) U_i^c is continuous on $(0, T^c]$ and hence constant. In turn this implies that U_i^c is differentiable on $(t^{*c}, T^c]$ with $\frac{dU_i^c(t)}{dt} = 0$, which implies that a non-confessing agent j concedes at rate $\frac{f_j^n(t)}{1-F_j^n(t)} = \lambda_j$. By claim (h) we must also have a total concession rate $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$ on $(t^{*c}, T^c]$. If both these concession rates hold, then:

$$\lambda_j = \frac{f_j(t)}{1 - F_j(t)} = \frac{c_j f_j^c(t) + (1 - c_j) f_j^n(t)}{c_j(1 - F_j^c(t)) + (1 - c_j)(1 - F_j^n(t))} = \frac{c_j f_j^c(t) + (1 - c_j)(1 - F_j^n(t))\lambda_j}{c_j(1 - F_j^c(t)) + (1 - c_j)(1 - F_j^n(t))},$$

which rearranges to give $\frac{f_j^c(t)}{1-F_j^c(t)} = \lambda_j$.

(m) *We must have $T^c = 0$.* Suppose not, and so $T^c = T_j^c > 0$ for some agent j. This clearly implies $F_j^c(0) < 1$. Notice that the continuity of F_i^c on $(0, T^*]$ implies that either $t^{*c} = 0$ or $F_i^c(t^{*c}) = F_i^c(0) = 0$ (where t^{*c} is defined in claim (k)). Claim (l) then implies that if $t \in [t^{*c}, T^c]$, then $F_j^c(t) = 1 - (1 - F_j^c(0))e^{-\lambda_j t} < 1$, however, this contradicts $T_j^c < \infty$.

We are almost done. Suppose that $T_j^c = 0$ so that $F_j^c(0) = 1$ (and so $F_i^n(0) = 0$ by claim (d)). Claim (i) then implies that $\frac{f_j(t)}{1-F_j(t)} = \frac{f_j^n(t)}{1-F_j^n(t)} = \lambda_j$ on $(0, T^*]$. This implies that $(0, T^*] \subseteq A_i^c = A_i^n$, and so a confessing agent i who concedes at $t \in A_i^c$ must get the payoff:

$$U_i^c(t) = c_j u_i(m_i) + (1 - c_j)\left(F_j^n(0) u_i(\alpha_i) + (1 - F_j^n(0)) u_i(1 - \alpha_j) \right)$$

Whereas a non-confessing agent i's who concedes at $t \in A_i^n$ must get the payoff:

$$U_i^n(t) = c_j u_i(\alpha_i) + (1 - c_j)\left(F_j^n(0) u_i(\alpha_i) + (1 - F_j^n(0)) u_i(1 - \alpha_j) \right)$$

Therefore, if $c_j > 0$ we must have $m_i \geq \alpha_i$, or agent i would not find it optimal to confess. Clearly we cannot have

$m_j < 1-\alpha_i$, or confessing would deliver agent j a payoff of $c_i u_j(m_i) + (1-c_i)u_j(1-\alpha_j)$, which is strictly less than the payoff $u_j(1-\alpha_i)$ which she could guarantee by not confessing and then conceding (recall that $F_j^c(0) = 1$ and $c_i > 0$).

Suppose finally that $m_j = 1-\alpha_i$. In this case, we must have $U_j^c(t) \leq u_i(1-\alpha)$ for all t (or we could not have $F_j^c(0) = 1$) and so we must similarly have $U_j^n(t) \leq u_i(1-\alpha)$ for all t, which in particular implies $F_i(0) = 0$. Given $T_j^c = 0$, claim (b) implies that $T_j^n = T^*$.

Analogous to the requirement that both agents reach a probability one reputation at the same time in the Baseline model, we must have $T_j^n = T^* = \max\{T_i^c, T_i^n\}$. If $T_j^n < \max\{T_i^c, T_i^n\}$ then because $T_j^c = 0$, any rational agent i would know she faced a behavioral type at T_j^n and would concede at most an instant after. Similarly if $T_j^n > \max\{T_i^c, T_i^n\}$, then a non-confessing agent j would know she faced a behavioral type at $\max\{T_i^c, T_i^n\}$ and would concede at most an instant after.

Claims (h) and (i) then imply that agents must concede at rates $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$ and $\frac{f_i(t)}{1-F_i(t)} = \lambda_i$ on $(0, T^*]$. Combined with the fact that $F_i(0) = 0$, this implies $1-F_i(t) = e^{-\lambda_i t}$ for $t \leq T^*$. The boundary conditions $(1-c_i)(1-F_i^n(T^*)) = z_i$ and $(1-F_i^c(T^*)) = 0$, therefore imply $1-F_i(T^*) = e^{-\lambda_i T^*} = z_i$ or $T^* = -\frac{1}{\lambda_i} ln(z_i)$. For agent j, these concession rates as well as $F_j^c(0) = 1$ imply that $(1-c_j)(1-F_j^n(t)) = (1-F_j(t)) = (1-F_j(0))e^{-\lambda_j t}$. The boundary condition $(1-c_j)(1-F_j^n(T^*)) = z_j$ then implies $(1-F_j(0))e^{-\lambda_j T^*} = z_j$. Clearly if $T_j^* = -\frac{ln(z_j)}{\lambda^j} < -\frac{ln(z_i)}{\lambda^i} = T_i^* = T^*$, we have an immediate contradiction. Otherwise, $(1-F_j(0)) = z_i e^{\lambda_j T^*} = z_i z_j^{-\frac{\lambda_j}{\lambda_i}}$. But in which case, any such an equilibrium has exactly the same distribution of outcomes as the Baseline equilibrium! Such an equilibrium "involving" the mediator exists whenever $T_i^* \neq T_j^*$ (e.g. let $c_i = 1-z_i$ and $c_j \in \left(0, 1-z_i z_j^{-\frac{\lambda_j}{\lambda_i}}\right]$ when $T_i^* < T_j^*$). □

Proof of Proposition 3. Suppose there is an equilibrium $\sigma = (\sigma_1, \sigma_2)$. In this setup, I refer to agent j who has confessed but not yet conceded, as a confessing agent. Let $A_i = \{(s,t) : U_i(s,t) = \max_{v,w} U_i(v,w)\}$. Since σ is an equilibrium, $A_i \neq \emptyset$. Finally, define $T_i^d = \inf\{t : F_i^d(t) = 1-z_i\}$ and $T^* = \max\{T_1^d, T_2^d\}$.

(a) *We must have* $T_i^d = T^* < \infty$. This follows for the reasons as outlined in the proof of Proposition 2, claim (a). We must have $T_i^d = T_j^d$, because if a rational agent knows she faces a behavioral opponent she will concede immediately. We must have $T_j^d < \infty$ because if a rational agent j does not concede at some t to get $u_j(1-\alpha_i) > 0$, she must expect her opponent to stop acting like a behavioral type soon and therefore must eventually become convinced that her opponent is behavioral.

(b) *If* F_i^d *jumps at* $t \in [0, T^*]$ *then* F_j^c *is constant on* $[t-\varepsilon, t]$ *for some* $\varepsilon > 0$. This follows because if agent j has not confessed before $t-\varepsilon$, she would strictly increase her payoff by confessing an instant after t compared to slightly before as this would give her $u_j(\alpha_j)$ rather than $u_j(m_j)$ with positive probability (at least $F_i^d(t) - \sup_s F_i^d(s) > 0$).

(c) *If* F_i^c *jumps at* $t \in (0, T^*]$ *then* F_j^d *is constant on* $[t-\varepsilon, t)$ *for some* $\varepsilon > 0$. This follows because agent j would prefer to concede an instant after t rather than slightly before as this would give her $u_j(m_j)$ rather than $u_j(1-\alpha_i)$ with positive probability (at least $F_i^c(t) - \sup_{s<t} F_i^c(t) > 0$).

(d) *Let* $t' \leq t'' < t''' \leq T^*$. *If* $F_i^c(t''') = F_i^c(t')$ *and* $F_j^c(t'') > F_j^d(t'')$ *then* $F_j^d(t'') = F_j^d(t''')$. If this is not true then there must exist some $s \leq t''$ and some $t \in (t'', t''']$ such that $(s,t) \in A_i$. However, given that $F_i^c(t'') = F_i^c(t''')$ the alternative strategy of conceding slightly earlier (e.g. at $\frac{1}{2}(t''+t)$) while still confessing at s is strictly more profitable as it moves the concession payoff $u_j(1-\alpha_i)$ forward in time (with probability greater than $z_i > 0$).

(e) *Let* $t' < t''' \leq T^*$. *If* $F_i^c(t''') = F_i^c(t')$ *then either* $F_j^d(t') = F_j^d(t''')$ *or* $F_j^c(t) = F_j^d(t)$ *for all* $t \in [t', t''')$. Suppose not, then for some $t'' \in [t', t''')$ we have $F_j^d(t'') < F_j^c(t'')$ and $F_j^d(t') < F_j^d(t''')$. Define $\check{t}_i = \sup\{t : F_i^c(t) = F_i^c(t')\}$. By claim (d) we have $F_j^d(t'') = \sup_{s<\hat{t}_i} F_j^d(s)$ and $F_j^c(t'') > F_j^d(t'')$. This implies that F_i^c must be continuous at $\check{t}_i$, i.e. $F_i^c(\check{t}_i) = \sup_{s<\hat{t}_i} F_i^c(s)$. To see this, notice that confessing at $\check{t}_i$ and conceding at some later

date t must give i a strictly lower payoff than confessing slightly earlier (e.g. at $\frac{1}{2}(\check{t}_i + t'')$ and still conceding at t (with probability $F_j^c(t'') - F_j^d(t'') > 0$ she receives the payoff $u_i(m_i)$ earlier). By claim (d), therefore, we must have $F_j^d(t''') = F_j^d(\check{t}_i)$. But in which case any strategy in which agent i confesses an instant after $\check{t}_i$ cannot be optimal either, contradicting the definition of the supremum $\check{t}_i$.

(f) *Let $T^* \geq t'' > t'$. If $F_i^d(t'') = F_i^d(t')$ and $F_i^c(t') > F_i^d(t')$ then $F_j^c(t') = F_j^c(t'')$* Suppose not so that $F_j^c(t') < F_j^c(t'')$. Then there exists $(s,t) \in A_j$ such that $s \in (t', t'']$. However, given $F_i^d(t'') = F_i^d(t')$, the alternative plan of confessing slightly earlier (e.g. at $\hat{s} = \frac{1}{2}(t' + s)$) while still conceding at t would be strictly better for j as this gives her the payoff $u_i(m_i)$ with positive probability at an earlier time (at least $(F_i^c(t') - F_i^d(t')) > 0$).

(g) *There is no jump in F_i^d at $t \in (0, T^*]$.* Suppose not, then by claim (b) F_j^c is constant on $[t-\varepsilon, t]$ for some $\varepsilon > 0$. Hence, by claim (e) either $F_i^d(t) = F_i^d(t-\varepsilon)$ (a direct contradiction) or $F_i^c(s) = F_i^d(s)$ for $s \in [t-\varepsilon, t)$. It must then be that F_i^c also jumps at t, because we must have $\sup_{s<t} F_i^c(s) = \sup_{s<t} F_i^d(s) < F_i^d(t) \leq F_i^c(t)$. Hence by claim (c), F_j^d is constant on $[t-\varepsilon, t)$ for some $\varepsilon > 0$ (assume the same ε without loss of generality). Given that F_i^c and F_i^d jump at t, we must have $(t,t) \in A_i$. However, the alternative strategy for i of both confessing and immediately conceding slightly earlier (e.g. at $t - \frac{\varepsilon}{2}$) delivers strictly higher expected profits as she gets the payoffs $(F_j^c(t-\varepsilon) - F_j^d(t-\varepsilon))u_i(m_i)$ and $(1 - F_j^c(t-\varepsilon))u_i(1-\alpha_j) > 0$ at an earlier date, without affecting other payoffs.

(h) *If F_i^d is continuous at $s \leq t$ then $U_i(s,t)$ is continuous at s, and if F_i^c is continuous at t then $U_i(s,t)$ is continuous at t.* This follows from how $U_i(s,t)$ is defined.

For claims (i)-(m) suppose that $F_1^c(t') > F_1^d(t')$ for some $t' \in [0, \infty)$ (symmetric arguments apply if $F_2^c(t') > F_2^d(t')$). Define $\bar{t}_1 = \inf\{t \geq t' : F_1^c(t) = F_1^d(t)\}$ and $\underline{t}_1 = \inf\{t : F_1^c(s) < F_1^d(s)\ \forall s \in [t, t']\}$. Notice that by claim (g), the continuity of F_1^d, we have $F_1^c(\bar{t}_1) = F_1^d(\bar{t}_1)$. Also note that $\bar{t}_1 > t' \geq \underline{t}_1$ and $F_1^c(t) > F_1^d(t)$ for all $t \in (\underline{t}_1, \bar{t}_1)$. Let $\bar{t}_1 \geq t''' > t'' > \underline{t}_1$.

(i) *We must have $F_2^c(t''') > F_2^c(t'')$.* Suppose not, and so let $\check{t}_2 = \sup\{t : F_2^c(t) = F_2^c(t'')\} \geq t'''$. I first establish the subclaim (i') that this must imply either $F_1^d(t'') = F_1^d(\check{t}_2)$ or $F_1^c(t) = F_1^d(t)$ for $t \in [t'', \check{t}_2)$. Suppose not (again), then $F_1^d(t'') < F_1^d(\check{t}_2)$ and there is some $t \in [t'', \check{t}_2)$ such that $F_1^d(t) < F_1^c(t)$. By claim (g), the continuity of F_1^d, we must have $F_1^d(t'') < F_1^d(\check{t}_2 - \varepsilon)$ for all $\varepsilon > 0$ sufficiently small. Choose such an appropriately small $\varepsilon < \check{t}_2 - t$, then we have $F_2^c(t'') = F_2^c(\check{t}_2 - \varepsilon)$, $F_1^d(t) < F_1^c(t)$ for some $t \in [t'', \check{t}_2 - \varepsilon)$ and $F_1^d(t'') < F_1^d(\check{t}_2 - \varepsilon)$, which contradicts claim (e).
By assumption we have $F_1^c(t) > F_1^d(t)$ for all $t \in [t'', \bar{t}_1)$ so that subclaim (i') in fact implies $F_1^d(t'') = F_1^d(\check{t}_2)$. This in turn ensures $\bar{t}_1 > \check{t}_2$ because $F_1^d(\check{t}_2) = F_1^d(t'') < F_1^c(t'') \leq F_1^c(\check{t}_2)$ and $F_1^c(\bar{t}_1) = F_1^d(\bar{t}_1)$. I next claim that it can't be optimal for agent 2 to confess at $\check{t}_2$ while conceding at some $t \geq \check{t}_2$. To see this, notice that agent 2 would do strictly better confessing slightly earlier (e.g. at $\frac{1}{2}(\check{t}_2 + t'')$) while still conceding at t as this would bring forward the payoff $u_2(m^2)$ with positive probability (at least $F_1^c(t'') - F_1^d(t'') > 0$), without affecting other payoffs. Given claim (g), the continuity of F_1^d, this argument similarly also implies that confessing an instant after $\check{t}_2$ is strictly worse than confessing at $\frac{1}{2}(\check{t}_2 + t'')$. This contradicts the definition of the supremum $\check{t}_2$.

(j) *We must have $F_1^d(t''') > F_1^d(t'')$.* Suppose not, then let $\check{t}_1 = \sup\{t : F_1^d(t) = F_1^d(t'')\} \geq t'''$. Given claim (g), the continuity of $F - i^d$, we have $F_1^d(\check{t}_1) = F_1^d(t'')$. Given $F_1^d(\check{t}_1) = F_1^d(t'') < F_1^c(t'') \leq F_1^c(\check{t}_1)$ we must have $\check{t}_1 < \bar{t}_1$. By claim (f) we must then have $F_2^c(\check{t}_1) = F_2^c(t'')$ which contradicts claim (i), that F_2^c is increasing on $(\underline{t}_1, \bar{t}_1]$.

(k) *We must have $F_2^d(t''') > F_2^d(t'')$.* Suppose not so that $F_2^d(t''') = F_2^d(t'')$. Given that F_2^c is increasing on the interval $[t'', t''']$ by claim (i), we must have $F_2^c(t) > F_2^d(t)$ for $t \in (t'', t''']$. Define $\bar{t}_2 = \inf\{t \geq t''' : F_2^c(t) = F_2^d(t)\}$ and $\underline{t}_2 = \inf\{t : F_2^c(s) < F_2^d(s)\ \forall s \in [t, t''']\}$, then switching the labelling for 1 and 2, claim (i) implies $F_1^c(t''') > F_1^c(t'')$ and claim (j) implies $F_2^d(t''') > F_2^d(t'')$, a contradiction.

(l) *We must have $F_1^c(t''') > F_1^c(t'')$.* Suppose not, and so $F_1^c(t''') = F_1^c(t'')$. Let $\check{t}_1 = \inf\{t : F_1^c(t) = F_1^c(t'')\}$. The right continuity of F_1^c ensures that $F_1^c(\check{t}_1) = F_1^c(t'')$. Clearly, we have $\check{t}_1 \geq \underline{t}_1$ (if $\check{t}_1 < \underline{t}_1$ then certainly at

some $t \in (\check{t}_1, t'']$ we must have $F_1^c(t) = F_1^d(t) = F_1^c(t'') \geq F_1^d(t'') \geq F_1^d(t)$, which contradicts $F_1^c(t'') > F_1^d(t'')$). By claim (e), we then have either $F_2^d(t''') = F_2^d(\check{t}_1)$, which contradicts claim (k), or $F_2^c(t) = F_2^d(t)$ for all $t \in [\check{t}_1, t''')$. Notice that because F_1^d is strictly increasing on $[\check{t}_1, t''')$ by claim (j) while F_1^c is by assumption constant, for some $s \leq \check{t}_1$ and some $t \in (\check{t}_1, t''')$ we must have $(s,t) \in A_1$. Furthermore, if $(s',t') \in A_1$ where $s' \in [s, \check{t}_1]$ then $(s',t) \in A_1$. This is simply because at time s' an agent who confessed at s and another who previously confessed at s' have the same incentives to concede thereafter. I claim, however, that $(\check{t}_1, t) \notin A_1$. To see this, notice that such a strategy is strictly worse than both confessing and conceding at t, which gives agent 1 the higher payoff of $u_1(\alpha_1)$ instead of $u_1(m_1)$ from the positive concession of agent 2 on the interval $[\check{t}_1, t)$. That is:

$$\begin{aligned} U_1(t,t) - U_1(\check{t}_1, t) &\geq \int_{\check{t}_1 \leq v \leq t} (u_1(\alpha_1) - u_1(m_1)) e^{-r_1 v} dF_2^c(v) \\ &\geq e^{-r_1 t}(u_1(\alpha_1) - u_1(m_1))(\sup_{v<t} F_2^c(v) - F_2^c(\check{t}_1)) > 0 \end{aligned}$$

where the first inequality follows from $F_2^c(t) = F_2^d(t)$ on $[\check{t}_1, t''')$, the second from $t \geq v \in [\check{t}_1, t]$ and the third from claim (i). For the same reason, confessing an instant before $\check{t}_1$ and conceding at t cannot be optimal either. This either contradicts the definition of $\check{t}_1$ as an infimum or implies $\check{t}_1 = 0$ and $F_2^c(0) = 0$. The latter possibility, however, clearly contradicts $F_1^c(v) > F_1^d(v)$ for all $v \in (\check{t}_1, t''')$.

(m) *F_i^c is continuous on $(\underline{t}_1, \bar{t}_1]$.* If F_i^c did jump at $t \in (\underline{t}_1, \bar{t}_1]$ then by (c), F_j^d is constant on $(t-\varepsilon, t)$ for some $\varepsilon > 0$, contradicting either claim (j) or (k).

We are almost done. Because F_1^c, F_1^d are increasing on $(\underline{t}_1, \bar{t}_1)$, established in claims (j) and (l), while by assumption $F_1^d(t) < F_1^c(t)$ on this interval, it follows that there is some $s' \in (\underline{t}_1, \bar{t}_1)$ such that A_1 is dense in the set $\{(s',t) : t \in [s', \bar{t}_1]\}$. Notice that regardless of whether agent 1 confesses at s' or $s \in (s', \bar{t}_1)$, she faces the same incentives to concede after s if she has not already done so. Notice also, that there is always a positive probability that agent 1 has confessed before s but has not conceded. From the continuity of F_2^c on $(\underline{t}_1, \bar{t}_1]$ it follows that $U_1(s',t)$ is constant on $[s', \bar{t}_1]$, and hence differentiable with respect to t with zero partial derivative, $\frac{\partial U_1(s',t)}{\partial t} = 0$. This implies:

$$\frac{f_2^c(t)}{1 - F_2^c(t)} = \lambda_2^c = \frac{r_1 u_1(1-\alpha_2)}{u_1(m_1) - u_1(1-\alpha_2)}$$

for $t \in [\underline{t}_1, \bar{t}_1]$. Solving this linear ODE gives $(1 - F_2^c(s)) = (1 - F_j^c(\underline{t}_1))e^{-\lambda_2^c(s-\underline{t}_1)}$.

By the same reasoning there must be some $s'' \in (\underline{t}_1, \bar{t}_1)$ such that A_1 is dense in the set $\{(s,s'') : s, \in [\underline{t}_1, s'']\}$. The continuity of F_2^d on $(\underline{t}_1, \bar{t}_1]$ then implies that $U_1(s,s'')$ is constant on $(\underline{t}_1, s'']$, and hence differentiable with respect to s with zero partial derivative, $\frac{\partial U_1(s,s'')}{\partial s} = 0$. Rearranging this zero derivative condition gives:

$$\frac{f_2^d(s)}{F_2^c(s) - F_2^d(s)} = \lambda_2^d = \frac{r_1 u_1(m_1)}{u_1(\alpha_1) - u_1(m_1)}$$

This should already suggest a problem. When $F_2^c(s) - F_2^d(s)$ becomes arbitrarily small $f_2^d(s)$ must be similarly small. However, $f_2^c(t) \geq \lambda_2^c(1 - F_2^c(t)) \geq \lambda_2^c z_2$ is bounded above zero, implying $F_2^c(t) - F_2^d(t) > 0$ on $(\underline{t}_1, \bar{t}_1]$. To be more precise, the above linear ODE is solved to give:

$$(1 - F_2^d(s)) = \begin{cases} \phi_2^d e^{-\lambda_2^d(s-\underline{t}_1)} + \phi_2^c \theta_2 (e^{-\lambda_2^c(s-\underline{t}_1)} - e^{-\lambda_2^d(s-\underline{t}_1)}) \text{ if } \lambda_2^d \neq \lambda_2^c \\ (\phi_2^d + \lambda_2^d \phi_2^c (s - \underline{t}_1)) e^{-\lambda_2^d(s-\underline{t}_1)} \text{ if } \lambda_2^d = \lambda_2^c \end{cases}$$

where, $\theta_2 = \frac{\lambda_2^d}{\lambda_2^d - \lambda_2^c}$ and $\phi_2^d = (1 - F_2^d(\underline{t}_1)) \geq (1 - F_2^c(\underline{t}_1)) = \phi_2^c$. Define the gap between F_2^c and F_2^d as $d_2(s) =$

$F_2^c(s) - F_2^d(s)$, and consider the following transformations of this gap:

$$\begin{aligned} d_2(s)\frac{e^{\lambda_2^d(s-\underline{t}_1)}}{\theta_2 - 1} =& \frac{\phi_2^d - \theta_2\phi_2^c}{\theta_2 - 1} + e^{(\lambda_2^d-\lambda_2^c)(s-\underline{t}_1)} \quad \text{if} \quad \lambda_2^d > \lambda_2^c \\ d_2(s)\frac{e^{\lambda_2^c(s-\underline{t}_1)}}{\phi_2^d - \theta_2\phi_2^c} =& e^{(\lambda_2^c-\lambda_2^d)(s-\underline{t}_1)} + \frac{\theta_2 - 1}{\phi_2^d - \theta_2\phi_2^c} \quad \text{if} \quad \lambda_2^d < \lambda_2^c \\ d_2(s)e^{\lambda_2^d(s-\underline{t}_1)} =& \phi_2^d + \lambda_2^d\phi_2^c(s - \underline{t}_1) - \phi_2^c \quad \text{if} \quad \lambda_2^d = \lambda_2^c \end{aligned}$$

I claim that each of these transformations is positive. Notice that $\theta_2 - 1 = \frac{\lambda_2^c}{\lambda_2^d - \lambda_2^c} > 0$ when $\lambda_2^d > \lambda_2^c$. Similarly $\phi_2^d - \theta_2\phi_2^c \geq -\phi_2^c\frac{\lambda_2^c}{\lambda_2^d-\lambda_2^c} > 0$ when $\lambda_2^d < \lambda_2^c$, where the first inequality follows from $\phi_2^d \geq \phi_2^c$. Each to the transformed gaps is strictly increasing in s, implying that $d_2(s) > 0$ for $s \in (\underline{t}_1, \bar{t}_1]$. Recall that we must have $\bar{t}_1 \leq T^* < \infty$, and $F_1^c(\bar{t}_1) = F_1^d(\bar{t}_1)$. Now define $\bar{t}_2 = \inf\{t > \underline{t}_1 : F_2^c(t) = F_2^d(t)\} \leq T^* < \infty$, where this is consistent with the definition of $\bar{t}_2$ in the proof of claim (k). We can now repeat the above arguments with the roles of agent 1 and 2 reverse to find that $d_1(s) > 0$ for $s \in (\underline{t}_1, \bar{t}_2]$ and $F_2^c(\bar{t}_2) = F_2^d(\bar{t}_2)$. Let $\bar{t} = \min\{\bar{t}_1, \bar{t}_2\}$. For some i we must have $\bar{t} = \bar{t}_i$, but that implies both $F_i^c(\bar{t}_i) = F_i^d(\bar{t}_i)$ and $d_i(\bar{t}_i) = F_i^c(\bar{t}_i) - F_i^d(\bar{t}_i) > 0$, a contradiction. We must, therefore, have $F_i^c(t) = F_i^d(t)$ for $t \in [0, \infty)$. Given this, the unique equilibrium must match that of the Baseline model by standard arguments (see AG).

□

Proof of Proposition 4. Agent i is indifferent to conceding at any $t \in (0, T^*]$, in particularly therefore because it is optimal to concede an instant after time zero we must have:

$$U_i^{*c} = \max_t U_i^c(t) = (1 - z_j)\Big(bu_i(m_i) + (1 - b)H_j^c(0)u_i(\alpha_i)\Big) + \Big(z_j + (1 - z_j)(1 - b)(1 - H_j^c(0))\Big)u_i(1 - \alpha_j)$$

Setting this expression equal to the expression in the main text for $U_i^{*c} = U_i^c(T^*)$ and rearranging gives:

$$\int_{s<T^*} e^{-r_i s}u_i(\alpha_i)dH_j^c(s) = u_i(\alpha_i) - (1 - H_j^c(0))(u_i(\alpha_i) - u_i(1 - \alpha_j)) + \frac{z_j(1 - e^{-r_i T^*})u_i(1 - \alpha_j)}{(1 - z_j)(1 - b)}. \tag{13}$$

And so, Q_i reduces to:

$$Q_i = u_i(m_i) - \left(u_i(\alpha_j) - (1 - H_j^c(0))(u_i(\alpha_i) - u_i(1 - \alpha_j) + \frac{z_j(1 - e^{-r_i T^*})u_i(1 - \alpha_j)}{(1 - z_j)(1 - b)} +)\right)$$

Suppose that $T^* = T_j = -\frac{1}{\lambda_j}ln(\bar{z}_j) \leq T_i$, so that $1 - H_j^c(0) = 1$ and $1 - H_i^c(0) = \frac{\bar{z}_i}{1-\bar{z}_i}\left(\bar{z}_j^{-\frac{\lambda_i}{\lambda_j}} - 1\right)$. Substituting in for these equalities gives:

$$\begin{aligned} Q_i =& u_i(m_i) - u_i(1 - \alpha_j) - \frac{z_j\left(1 - \left(\frac{z_j}{1-(1-z_j)b}\right)^{\frac{r_i}{\lambda_j}}\right)u_i(1 - \alpha_j)}{(1 - z_j)(1 - b)} \\ Q_j =& u_j(m_j) - u_j(\alpha_j) - \frac{z_i\left(1 - \left(\frac{z_j}{1-(1-z_j)b}\right)^{\frac{r_j}{\lambda_j}}\right)u_j(1 - \alpha_i)}{(1 - z_i)(1 - b)} + (u_j(\alpha_j) - u_j(1 - \alpha_i))\frac{z_i\left(\left(\frac{z_j}{1-(1-z_j)b}\right)^{-\frac{\lambda_i}{\lambda_j}} - 1\right)}{(1 - z_i)(1 - b)} \end{aligned}$$

Define $\underline{m}_i < \alpha_i$ as the mediation share that causes $Q_i = 0$, that is:

$$\underline{m}_i = u_i^{-1}\left(u_i(1-\alpha_j) + \frac{z_j\left(1-\left(\frac{z_j}{1-(1-z_j)b}\right)^{\frac{r_i}{\lambda_j}}\right)u_i(1-\alpha_j)}{(1-z_j)(1-b)}\right)$$

Notice that $\underline{m}_i \to 1-\alpha_j$ as $z_j \to 0$. Setting $m_i = 1-m_j = 1 = \underline{m}_i$ we then have:

$$\frac{Q_j}{z_i} = \frac{u_j(1-\underline{m}_i)-u_j(\alpha_j)}{z_i} - \frac{\left(1-\left(\frac{z_j}{1-(1-z_j)b}\right)^{\frac{r_j}{\lambda_j}}\right)u_j(1-\alpha_i)}{(1-z_j)(1-b)} + (u_i(\alpha_i)-u_i(1-\alpha_j))\frac{\left(\left(\frac{z_j}{1-(1-z_j)b}\right)^{-\frac{\lambda_i}{\lambda_j}}-1\right)}{(1-z_i)(1-b)} \tag{14}$$

We are interested in taking the limit of $\frac{Q_j}{z_i}$ as $z_j \to 0$. It is clear that $\lim_{z_j\to 0}\frac{1-\left(\frac{z_j}{1-(1-z_j)b}\right)^{\frac{r_j}{\lambda_j}}}{(1-z_j)} = 1$ while $\frac{\left(\frac{z_j}{1-(1-z_j)b}\right)^{-\frac{\lambda_i}{\lambda_j}}-1}{(1-z_i)} = \infty$. By assumption we have $\frac{1}{z_i} \geq \frac{K}{z_j}$. We can then use l'Hopital's rule and the inverse function to show that:

$$\lim_{z_j\to 0}\frac{u_j(1-\underline{m}_i)-u_j(\alpha_j)}{z_j} = -\frac{u_j'(\alpha_j)u_i(1-\alpha_j)}{u_i'(1-\alpha_j)(1-b)} > -\infty$$

where this uses the fact that $\left.\frac{\partial\left(1-\left(\frac{z_j}{1-(1-z_j)b}\right)^{\frac{r_i}{\lambda_j}}\right)z_j(1-z_j)^{-1}}{\partial z_j}\right|_{z_j=0} = 1$ And so we must have $\lim_{z_j\to 0}\frac{Q_j}{z_i} = \infty$. This ensures the existence of some $\underline{z}' > 0$ such that if $z_j \leq \underline{z}'$ we must have $Q_j \geq 0$ and $Q_i \geq 0$ and so there is an $N1$ equilibrium.

It remains to show that such equilibrium can strictly improve the payoff of both agents. If $\lambda_j \geq \lambda_i$ and $z_j \geq z_i$ then clearly we must $T_i \geq T_j$ in any $N1$ equilibrium and in the Baseline equilibrium. Alternatively, suppose that $\lambda_j > \lambda_i$ (and possibly $z_j < z_i$). In this case let $\overline{z}'' > 0$ be such that for $z_j \leq \overline{z}''$ we have $\left(\frac{z_j}{1-(1-z_j)b}\right)^{\frac{\lambda_i}{\lambda_j}-1} \geq K$. This implies:

$$\left(\frac{z_j}{1-(1-z_j)b}\right)^{\frac{\lambda_i}{\lambda_j}} \geq \frac{Kz_j}{1-(1-z_j)b} \geq \frac{Kz_j}{1-(1-Kz_j)b} \geq \frac{z_i}{1-(1-z_i)b}$$
$$T_j = -\frac{1}{\lambda_j}ln\left(\frac{z_j}{1-(1-z_j)b}\right) \leq -\frac{1}{\lambda_i}ln\left(\frac{z_i}{1-(1-z_i)b}\right) = T_i$$

The first inequality on the first line is directly implied, the second follows because $K \geq 1$, the third because $Kz_j \geq z_i$. The second line is then simply a rearrangement of the inequality of the first and final term on line one, and implies that $T_j \leq T_i$ in an $N1$ equilibrium. The bound also ensures that $z_j^{\frac{\lambda_i}{\lambda_j}} \geq Kz_j \geq z_i$ so that in the Baseline equilibrium we must also have $T_j \leq T_i$

Let $z_j \leq \min\{\overline{z}', \overline{z}'', \frac{1}{2}\}$, then the payoff to player i in the Baseline equilibrium is $u_i(1-\alpha_j)$. Given that $U_i^{*n} > u_i(1-\alpha_j)$ in any $N1$ equilibrium it is clear that we must also have $U_i^{*c} > u_i(1-\alpha_j)$. In the Baseline equilibrium agent j's payoff is $U_j^B = u_j(\alpha_i) - (u_j(\alpha_j) - u_j(1-\alpha_i))z_i z_j^{-\frac{\lambda_i}{\lambda_j}}$. We need to compare this to her payoff in an $N1$ equilibrium, which can be expressed as:

$$U_j^{*c} = (1-z_i)bu_j(m_j) + (1-b(1-z_i))u_i(\alpha_j) - (u_i(\alpha_i)-u_i(1-\alpha_j))z_i\left(\frac{z_j}{1-(1-z_j)b}\right)^{-\frac{\lambda_i}{\lambda_j}}$$

The best possible $N1$ equilibrium for agent j (consistent with a fixed b) has $m_j = 1-\underline{m}_i$. For that equilibrium we

have:

$$\begin{aligned}\frac{U_j^{*c}-U_j^B}{z_i} =&(1-z_i)b\frac{u_j(1-\underline{m_i})-u_i(\alpha_j)}{z_i}+(u_i(\alpha_i)-u_i(1-\alpha_j))\left(z_j^{-\frac{\lambda_i}{\lambda_j}}-\left(\frac{z_j}{1-(1-z_j)b}\right)^{-\frac{\lambda_i}{\lambda_j}}\right)\\ \geq& b(K-z_j)\frac{u_j(1-\underline{m_i})-u_i(\alpha_j)}{z_j}+(u_i(\alpha_i)-u_i(1-\alpha_j))z_j^{-\frac{\lambda_i}{\lambda_j}}\left(1-\left(\frac{2}{2-b}\right)^{-\frac{\lambda_i}{\lambda_j}}\right)\end{aligned}$$

where the second line follows from the assumption $K \geq \frac{z_i}{z_j} \geq \frac{1}{K}$ and $\frac{1}{1-(1-z_j)b} \geq \frac{2}{2-b}$ when $z_j \leq \frac{1}{2}$ (this is equivalent to $2-b \geq 2(1-(1-z_j)b)$).

We previously established that the $\lim_{z_j\to 0}\frac{u_j(1-\underline{m_i})-u_i(\alpha_j)}{z_j} > -\infty$. Additionally noticing that $\left(\frac{2}{2-b}\right)^{-\frac{\lambda_i}{\lambda_j}} < 1$ and that $\lim_{z_j\to 0} z_j^{-\frac{\lambda_i}{\lambda_j}} = \infty$ it is clear that $\lim_{z_j\to 0}\frac{U_j^{*c}-U_j^B}{z_i} = \infty$. This implies that there exists $\bar{z} > 0$ such for $z_j \leq \bar{z}$ we have an $N1$ equilibrium with $m_i = \underline{m_i}$ where both rational player's expected payoffs exceed their payoff in the Baseline equilibrium. This completes the proof. □

Proof of Proposition 5. Suppose this were not true, then there must exist some sequence of games $(r_i, u_i, \alpha_i, z_i^n, m^n, b^n)$ with $z_1^n \to 1$ and a sequence of $N1$ equilibria in each. I first claim that any (sub)sequence of these $N1$ equilibria must satisfy $\lim_n T^* = 0$. This follows immediately from the fact that $T^* \leq T_1 = -\frac{1}{\lambda^1} ln\left(\frac{z_1^n}{(1-z_1^n)(1-b)+z_1^n}\right)$.

Notice that $\int_{s<T^*} e^{-r_i s}u_i(\alpha_i)dH_j^c(s) \geq e^{-r_i T^*}u_i(\alpha_i)$, hence for any $\varepsilon > 0$, for all sufficiently large n we need $m_i > \alpha_i - \varepsilon$ for $i = 1, 2$, in order to have $Q_i \geq 0$. Choosing $\varepsilon = \frac{\alpha_1+\alpha_2-1}{2}$ we have $m_1 + m_2 > \alpha_1 + \alpha_2 - 2\varepsilon = 1$, a contradiction. This completes the proof. □

Proof of Lemma 1. Throughout this proof I hold G^Z fixed and consider G^R such that both incentive constraints are satisfied. Suppose that such a distribution G^R, implies $T^R > T^Z$. In that case, the alternative distribution $\tilde{G}^R$ with $\tilde{G}^R(t) = G^R(t)$ for $t < T^Z$ and $\tilde{G}^R(T^Z) = 1$ so that $T^R = T^Z$ strictly increases $U^c(T^R)$, while relaxing both incentive constraints. It is therefore, without loss of generality to focus on G^R that imply $T^R = T^Z$.

Suppose that G^R implies that the dynamic incentive does not bind, in the sense that $U^c(T^Z) - U^c(\vec{t}^1) = \delta > 0$ for some real valued time $\vec{t}^1 < T^Z$. We next want to show that in this case there must exist some alternative distribution $\check{G}^R$ with $\check{G}^R(t) \geq G^R(t)$ satisfying both constraints, which increases $U^c(T^Z)$. To that end define $\vec{t}^2 = T^Z$ if $T^Z < \infty$ and $\vec{t}^2 = \min\{t : e^{-rt}u(1) \leq \frac{\delta}{2}\}$ otherwise. Notice that we must have $U^c(T^Z) - U^c(\vec{t}^2) \leq e^{-r\vec{t}^2}u(1) \leq \frac{\delta}{2}$ and so $U^c(t^2) - U^c(\vec{t}^1) \geq \frac{\delta}{2} > 0$.

Next define $\vec{t}^3 = \min\{t \in [\vec{t}^1, \vec{t}^2] : U^c(\vec{t}^2) - U^c(t) \leq (\vec{t}^2 - t)\frac{\delta}{4(\vec{t}^2-\vec{t}^1)}\}$. This is well defined because the the right continuity of G^z and G^R ensure that $U^c(t)$ is right continuous also. By construction $\vec{t}^3 > \vec{t}^1$, and $U(\vec{t}^3) - U(t) = [U^c(\vec{t}^2) - U^c(t)] - [U^c(\vec{t}^2) - U^c(\vec{t}^3) > \frac{\delta(\vec{t}^3-t)}{4(\vec{t}^2-\vec{t}^1)}$ for all $t \in [\vec{t}^1, \vec{t}^3)$. For such t we have:

$$\begin{aligned}U^c(\vec{t}^3) - U^c(t) =&(1-z)\int_{s\in(t,\vec{t}^3]} e^{-rs}u(0.5)dG^R(s) + z\int_{s\in(t,\vec{t}^3]} e^{-rs}u(1-\alpha)dG^Z(s)\\ &+ e^{-r\vec{t}^3}u(1-\alpha)\left((1-z)(1-G^R(\vec{t}^3)) + z(1-G^Z(\vec{t}^3))\right) - e^{-rt}u(1-\alpha)\left((1-z)(1-G^R(t)) + z(1-G^Z(t))\right)\\ &\leq (1-z)e^{-rt}(G^R(\vec{t}^3) - G^R(t))\left(u(0.5) - u(1-\alpha)\right) + (e^{-r\vec{t}^3} - e^{-rt})\left((1-z)(1-G^R(\vec{t}^3)) + z(1-G^Z(\vec{t}^3))\right)u(1-\alpha)\end{aligned}$$

Where the inequality follows from the fact that the integrals in the first line are respectively smaller than $(1-z)e^{-rt}u(0.5)(G^R(\vec{t}^3) - G^R(t))$ and $ze^{-rt}u(1-\alpha)(G^Z(\vec{t}^3) - G^Z(t))$, and some rearrangement. Let $\varepsilon = \frac{\delta 4}{\vec{t}^2-\vec{t}^1}$ so that $U(\vec{t}^3) - U(t) > 2\varepsilon(\vec{t}^3 - t)$ for $t \in [\vec{t}^1, \vec{t}^3)$. By dividing the right hand side of the above inequality by $(\vec{t}^3 - t)$ and

taking its limit as $t \to \vec{t}^3$ gives:

$$e^{-r\vec{t}^3}\left(u(0.5)-u(1-\alpha^j)\right)(1-z)lim_{t\to\vec{t}^3}\frac{G^R(\vec{t}^3)-G^R(t)}{\vec{t}^3-t}-ru(1-\alpha)e^{-r\vec{t}^3}\left((1-z)(1-G^R(\vec{t}^3))+z(1-G^Z(\vec{t}^3))\right)\geq\varepsilon$$

This in turn implies that there exists $\varepsilon'>0$ and $\vec{t}^4<\vec{t}^3$ such that for all $t\in[\vec{t}^4,\vec{t}^3]$,

$$G^R(\vec{t}^3)-G^R(t)\geq\left((1-G^R(\vec{t}^3))+\frac{z}{1-z}(1-G^Z(\vec{t}^3))\right)\lambda^m(\vec{t}^3-t)+\varepsilon'(\vec{t}^3-t).$$

Consider then an alternative distribution, $\hat{G}^R$. This is defined by $\hat{G}^R(t)=G^R(t)$ for $t\geq\vec{t}^3$, and satisfies the indifference condition $U^c_{\hat{G}^R}(t)=U^c_{\hat{G}^R}(\vec{t}^3)$ for $t\leq\vec{t}^3$ (where $U^c_{\hat{G}^R}(t)$ is the utility of conceding at t given $\tilde{G}^R$). This indifference condition implies that $\hat{G}^R(t)$ is differentiable on $[0,\vec{t}^3]$ with $\hat{g}^R(t)=\left((1-\hat{G}^R(t))+\frac{z}{1-z}(1-G^Z(t))\right)\lambda^m$. It is clear that this implies the existence of some $\vec{t}^5<\vec{t}^3$ such that for all $t\in[\vec{t}^5,\vec{t}^3]$,

$$\hat{G}^R(\vec{t}^3)-\hat{G}^R(t)\leq\left((1-z)(1-\hat{G}^R(\vec{t}^3))+z(1-G^Z(\vec{t}^3))\right)\lambda^m(\vec{t}^3-t)+\frac{\varepsilon'}{2}(\vec{t}^3-t).$$

Letting $\vec{t}^6=\max\{\vec{t}^4,\vec{t}^5\}$ we must then have $\hat{G}^R(t)>G^R(t)$ for all $t\in[\vec{t}^6,\vec{t}^3)$. We can now define $\check{G}^R(t)=\hat{G}^R(t)$ for $t\geq\vec{t}^6$ and $\check{G}^R(t)=G^R(t)$ elsewhere. This distribution implies $\check{G}^R(t)\geq G^R(t)$ for all t, and $\check{G}^R(t)>G^R(t)$ for $t\in[\vec{t}^6,\vec{t}^3)$. This ensures that $U^c_{\check{G}^R}(t)>U^c_{G^R}(t)$ for all $t\in[\vec{t}^6,T^Z]$.

I claim that $\check{G}^R$ must satisfy the Dynamic IC constraint. For $t\geq\vec{t}^3$ we have $U^c_{\check{G}^R}(T^Z)-U^c_{\check{G}^R}(t)=U^c(T^Z)_{G^R}-U^c_{G^R}(t)\geq 0$. For $t\in[\vec{t}^6,\vec{t}^3]$ we have $U^c_{\check{G}^R}(T^Z)-U^c_{\check{G}^R}(t)=U^c_{\check{G}^R}(T^Z)-U^c_{\check{G}^R}(\vec{t}^3)\geq 0$ (recall that $U^c_{\check{G}^R}(t)=U^c_{\check{G}^R}(\vec{t}^3)$). Finally for $t<\vec{t}^6$ we have $U^c_{\check{G}^R}(t)=U^c_{G^R}(t)$ and so $U^c_{\check{G}^R}(T^Z)>U^c_{\check{G}^R}(t)$. The new distribution $\check{G}^R$ must certainly also satisfy the type IC constraint, because G^Z is unchanged and therefore so is $U^n(t)$.

For arbitrary cumulative distribution function G^R on $[0,T^Z]$ with $T^Z=T^R$ let $T=T^Z$, as well as $v(T,G^R)=U^c_{G^R}(T)$, $w(t,G^R)=U^c_{G^R}(T)-U^c_{G^R}(t)$. The proof above establishes that for any G^R satisfying both constraints, if the dynamic incentive constraint doesn't bind ($w(t,G^R)>0$ for some $t<T$), then there is some alternative distribution $\check{G}^R$ on $[0,T]$ with $\check{G}^R(t)\geq G^R(t)$ which improves rational payoffs but still satisfies both incentive constraint ($v(T,\check{G}^R)>v(T,G^R)\geq U^n(t)$ and $w(t,\check{G}^R)\geq 0$ for $t\in[0,T]$). We can then apply Lemma 7, see below, which in this case establishes the existence of some $\overline{G}^R$ which satisfies both incentive constraints and delivers a higher time T payoff than G^R (i.e. $v(T,\overline{G}^R)>v(T,G^R)\geq U^n(t)$ and $w(t,\overline{G}^R)\geq 0$ for $t\in[0,T]$), for which there is no other distribution on $[0,T]$ with $\tilde{G}^R(t)\geq\overline{G}^R(t)$. This implies that the dynamic incentive constraint must bind for $\overline{G}$ (i.e. $w(t,\check{G}^R)=0$ for $t\in[0,T]$ for $t\in[0,T^Z]$).

□

Lemma 7. *For fixed $T\leq\infty$, consider functions of the form:*

$$v(T,G)=\int_{s\leq T}e^{-rs}A_1(s)dG(s)+A_2(T)$$
$$w(t,G)=\int_{s\leq t}A_3(s)dG(s)+\int_{s\leq T}e^{-rs}A_4(s)dG(s)+A_5(t)+e^{-rt}A_6(t)G(t)$$

where each $A_k(s)$ is a continuous bounded function on $[0,T]$ and G is some cumulative distribution function on $[0,T]$. Let Δ be the set of cumulative distribution functions on $[0,T^Z]$, and define $X=\{G\in\Delta:w(t,G)\geq 0,\forall t\in[0,T]\}$. Define the partial order $\gtrsim$ on X by $G\sim G$, and $\tilde{G}\succ G$, if $v(T,\tilde{G})>v(T,G)$, $\tilde{G}(t)\geq G(t)$ for all t. Given some $\tilde{t}<\infty$ and $\underline{G}\in X$, define $\bar{t}_G=\inf\{t:w(s,G)=0 \text{ for } s\in[t,T]\}$ and the partial order $\gtrsim_{\tilde{t}}$ on X as follows: Let $G\sim_{\tilde{t}}G$, and $\tilde{G}\succ_{\tilde{t}}G$, if $v(T,\tilde{G})>v(T,G)$, $\bar{t}_{\tilde{G}}\leq\bar{t}_G$, $\tilde{G}(t)\geq G(t)$ for $t\geq\tilde{t}$, and $\tilde{G}(s)\leq G(s)$ for $s<\tilde{t}$. Suppose there exists some $\tilde{G}\succ G$, then there exists some $\overline{G}\succ G$ for which there is no $\hat{G}$ such that $\hat{G}\succ\overline{G}$ (i.e. $\overline{G}$ is $\gtrsim$ maximal). Suppose there exists some $\tilde{G}\succ_{\tilde{t}}G$, then there exists some $\overline{G}\succ_{\tilde{t}}G$ for which there is no $\hat{G}$ such that $\hat{G}\succ_{\tilde{t}}\overline{G}$ (i.e. $\overline{G}$

is $\gtrsim_{\tilde{t}}$ maximal).

Proof. Define $\overline{u}(\hat{G}) = \sup_{\tilde{G}\in X}\{v(T,\tilde{G}) : \tilde{G} \gtrsim \hat{G}\}$ (respectively $\overline{u}_{\tilde{t}}(\hat{G}) = \sup_{\tilde{G}\in X}\{v(T,\tilde{G}) : \tilde{G} \gtrsim_{\tilde{t}} \hat{G}\}$). Let $G^0 = G$ and choose $G^{k+1} \gtrsim G^k$ (respectively $G^{k+1} \gtrsim_{\tilde{t}} G^k$) such that $v(T,G^{k+1}) \geq \frac{\overline{u}(G)+v(T,G^k)}{2}$ (respectively $v(T,G^{k+1}) \geq \frac{\overline{u}_{\tilde{t}}(G)+v(T,G^k)}{2}$). Let $\underline{G}(t) = \lim_k G^k(t)$ and then define the cumulative distribution function $\overline{G}$ by $\overline{G}(t) = \inf\{\underline{G}(s) : s > t\}$. Clearly we have $G^k \overset{}{\underset{w}{\rightarrow}} \overline{G}$.

Given $G^k(T) = \overline{G}(T) = 1$ and the weak convergence of G^k, we clearly have $\lim_k \int_{s\leq T} A_k(s)dG^k(s) = \int_{s\leq T} A_k(s)d\overline{G}(s)$ and so ultimately $\lim_k v(T,G^k) = v(T,\overline{G})$ and $\lim_k w(T,G^k) = w(T,\overline{G}) \geq 0$. Noticing that $\overline{G}$ is continuous almost everywhere, let $Y = \{t : \overline{G} \text{ is continuous at } t\}$. For $t \in Y$ we have $\lim_k G^k(t) = \overline{G}(t)$. Define the cumulative distribution functions $G^{k,t}$ on $[0,t]$ by $G^{k,t}(s) = \frac{G^k(s)}{G^k(t)}$ and $\overline{G}^t(s) = \frac{\overline{G}(s)}{\overline{G}(t)}$, then $G^{k,t} \overset{}{\underset{w}{\rightarrow}} \overline{G}^t$. This ensures $\lim_k \int_{s\leq t} A_k(s)dG^k(s) = \int_{s\leq t} A_k(s)d\overline{G}(s)$ and so ultimately for $t \in Y$ we have $\lim_k w(t,G^k) = w(t,\overline{G}) \geq 0$. For $t \notin Y$, the right continuity of $\overline{G}$ implies that $w(t,\overline{G}) \geq \sup_{v>t} \inf_{s\in(t,v]\cap Y} w(s,\overline{G}) \geq 0$. This establishes $\overline{G} \in X$.

If $G^{k+1} \gtrsim G^k$ then $\overline{G}(t) \geq G^k(t)$ and $v(T,\overline{G}) \geq v(T,G^k) > v(T,G^0)$ so that $\overline{G} \gtrsim G^k \succ G^0$. If $G^{k+1} \gtrsim_{\tilde{t}} G^k$ we have $\bar{t}_{\overline{G}} \leq \bar{t}_{G^k}$ (indeed, $G^k(t) = \overline{G}(t)$ for $t \geq \bar{t}_{G^k}$), $\overline{G}(t) \geq G^k(t)$ for $t \geq \tilde{t}$ and $\overline{G}(t) \leq G^k(t)$ for $t \leq \tilde{t}$ and $v(T,\overline{G}) \geq v(T,G^k) > v(T,G^0)$ so that $\overline{G} \gtrsim_{\tilde{t}} G^k \succ_{\tilde{t}} G^0$. Suppose then that there exists $\hat{G} \in X$ such that $\hat{G} \succ \overline{G}$ (respectively $\hat{G} \succ_{\tilde{t}} \overline{G}$), then define $\varepsilon = v(T,\hat{G}) - v(T,\overline{G}) > 0$. Clearly, $\hat{G} \succ G^k$ so that $\overline{u}(G^k) \geq v(T,G^k) + \varepsilon$ (respectively $\hat{G} \succ \overline{G}$ so that $\overline{u}_{\tilde{t}}(G^k) \geq v(T,G^k)+\varepsilon$. That in turn implies $v(T,G^k) \geq v(T,G^0)+k\frac{\varepsilon}{2}$ and so $v(T,\overline{G}) = \infty$, which contradicts the fact that $v(T,\overline{G})$ must be bounded. □

Proof of Lemma 2. We are given that $IC_{G^Z}(t) > 0$ for some $t \in [\min\{\hat{t}, T^Z\}, T^Z]$. We first want to find an alternative distribution $\hat{G}^Z$ with $IC_{\hat{G}^Z}(t) \geq 0$ and $\hat{G}^Z(t) \geq G^Z(t)$ for all t such that $G^{R*}_{\hat{G}^Z}(0) > G^{R*}_{G^Z}(0)$. Initially suppose that $IC_{G^Z}(T^Z) = \delta > 0$. If $T^Z < \infty$ then let $t' = T^Z$. If $T^Z = \infty$ then let $e^{-rt'}u(1-\alpha) = \frac{\delta}{3}$ so that $IC_{G^Z}(t) \geq \frac{2\delta}{3}$ for $t \geq t'$. Given the right continuity of G^Z, we must have $IC_{G^Z}(t) \geq \frac{\delta}{3}$ for $t \geq t' - \varepsilon$ for some $\varepsilon > 0$. Consider the alternative distribution $\hat{G}^Z$, such that $\hat{G}^Z(t) = G^Z(t)$ for $t < T^Z - \varepsilon$ and $\hat{G}^Z(t) = \min\{G^Z(t)+\varepsilon', 1\}$ for $t \geq t' - \varepsilon$ and some $\varepsilon' > 0$. Notice that for $t < t' - \varepsilon$ we must have $IC_{\hat{G}^Z}(t) \geq IC_{G^Z}(t)$. For all $t \geq t' - \varepsilon$ we have $IC_{\hat{G}^Z}(t) \geq IC_{G^Z}(t) - \varepsilon' u(\alpha)$. Given $IC_{G^Z}(t) \geq \frac{\delta}{3}$, by selecting $\varepsilon' > 0$ sufficiently small, we must have $IC_{\hat{G}^Z}(t) \geq 0$ for all $t \geq t' - \varepsilon$. Given $\hat{G}^Z(t) \geq G^Z(t)$ for all t, this alternative distribution implies $G^{R*}_{\hat{G}^Z}(0) > G^{R*}_{G^Z}(0)$.

Next, suppose that $IC_{G^Z}(T^Z) = 0$ but $IC_{G^Z}(t') = \delta > 0$ for some $t' \in [\hat{t}, T^Z)$. Let $\hat{G}^Z(t) = G^Z(t)$ if $t < t'$ and $\hat{G}^Z(t) = \max\{G^Z(t')+\varepsilon, G^Z(t)\}$ otherwise, for some $\varepsilon > 0$. Clearly, $IC_{\hat{G}^Z}(t') \geq IC_{G^Z}(t') - \varepsilon u(\alpha)$ so that for $\varepsilon \leq \frac{\delta}{2u(\alpha)}$ we have $IC_{\hat{G}^Z}(t') \geq \frac{\delta}{2}$. If $\hat{G}^Z(t) = G^Z(t)$ then because the final the integrand in equation (8) is always positive for $s \geq t'$ and $\hat{G}^Z(s) \geq G^Z(s)$, we must have $IC_{\hat{G}^Z}(t) \geq IC_{G^Z}(t) \geq 0$. If $\hat{G}^Z(t) = \hat{G}^Z(t') > G^Z(t)$, however, then $U^n_{\hat{G}^Z}(t) \leq U^n_{\hat{G}^Z}(t')$ and so $IC_{\hat{G}^Z}(t) \geq IC_{\hat{G}^Z}(t') > 0$ (a larger t simply delays a non-confessing agent's payoff from concession).

For arbitrary $\tilde{G}^Z$ define $v(T,\tilde{G}^Z) = G^{R*}_{G^Z}(0)$ and $w(t,\tilde{G}^Z) = IC_{\tilde{G}^Z}(t)$, as well as $T = \infty$. The proof above establishes that if the time t type incentive constraint doesn't bind for some $t \in [\min\{\hat{t}, T^Z\}, T^Z]$ (i.e. $w(t,\tilde{G}^Z) > 0$) for some then there is some alternative incentive compatible $\hat{G}^Z$ ($w(t,\hat{G}^Z) \geq 0$ for all $t \leq T$) delivering higher payoffs ($v(T,\hat{G}^Z) > v(T,\tilde{G}^Z)$) with $\hat{G}^Z(s) \geq \tilde{G}^Z(s)$. Invoking Lemma 7, this implies the existence of some $\overline{G}^Z$ with $\overline{G}^Z(t) \geq \tilde{G}^R(t)$, $v(T,\overline{G}^Z) > v(T,\tilde{G}^Z)$ and $w(t,\overline{G}^Z) \geq 0$ for all $t \leq T$ such that there is no alternative $\check{G}^Z$ with $\check{G}^Z(t) \geq \overline{G}^Z(t)$, $v(T,\check{G}^Z) > v(T,\check{G}^Z)$ and $w(t,\check{G}^Z) \geq 0$. And so, this incentive compatible distribution delivers higher payoffs and must satisfy $IC_{\overline{G}^Z}(t) = 0$ for $t \in [\min\{\hat{t}, T^Z\}, T^Z]$, completing the proof. □

Proof of Lemma 3. As argued in the text, we can restrict attention to the reduced problem of finding the minimum T^Z such that $IC_{G^Z_{0,T^Z}}(T^Z) \geq 0$. Recall that for risk neutral agents we have $IC_{G^Z_{0,T^Z}}(T^Z) = 0$ for $T^Z = -\frac{1}{\lambda}ln(z)$. Taking

the derivative of $IC_{G^Z_{0,T^Z}}(T^Z)$ we get

$$\begin{aligned}\frac{dIC_{G^Z_{0,T^Z}}(T^Z)}{dT^Z} &= -\frac{z^2}{1-z}ru(1-\alpha)\int_0^{T^Z}\lambda e^{\lambda T^Z+(\lambda^m-\lambda)s}ds + zru(1-\alpha)e^{\lambda T^Z}\\ &= -\frac{z^2}{1-z}ru(1-\alpha)\frac{\lambda}{\lambda^m-\lambda}(e^{\lambda^m T^Z}-e^{\lambda T^Z}) + zru(1-\alpha)e^{\lambda T^Z}\end{aligned}$$

Notice that $u(x) = x$ implies $\lambda^m = 2\lambda$ and so $\frac{\lambda}{\lambda^m-\lambda} = 1$. In turn, this implies $\frac{dIC(0)}{dT^Z}\big|_{T^Z=-\frac{1}{\lambda}ln(z)} = 0$. Finally, notice that $\frac{dIC_{G^Z_{0,T^Z}}(T^Z)}{dT^Z}e^{-\lambda T^Z}$ is strictly decreasing in T^Z and so $\frac{dIC_{G^Z_{0,T^Z}}(T^Z)}{dT^Z} > 0$ for $T^Z < -\frac{1}{\lambda}ln(z)$. Hence, we must have $IC_{G^Z_{0,T^Z}}(T^Z) < 0$ whenever $T^Z < -\frac{1}{\lambda}ln(z)$. In the OSSMP, therefore, we must have $T^Z = -\frac{1}{\lambda}ln(z)$, so that the optimal distributions G^{Z*} and $G^{R*}_{G^{Z*}}$ correspond exactly to the Baseline equilibrium. □

Proof of Lemma 4. Given Observation 2, we only need to contend with the risk neutral case. This proof derives some expressions in greater detail than is needed, but which are used in later proofs. Consider a distribution of the form $G^Z_{T^Z(\check{t},M),T^Z}$ where $G^Z_{\check{t},T^Z}$ is defined in the equation (9) and $T^Z(\check{t},M)$ is defined in equation (11) to ensure that $U^n(t) = M$ for $t \in [\check{t}, T^Z]$. Such a distribution implies that:

$$G^{R*}_{G^Z_{\check{t},T^Z(\check{t},M)}}(0) = 1 - \frac{z}{1-z}\int_0^{\check{t}}\lambda^m e^{\lambda^m s}ds - \left(\frac{z}{1-z}\right)^2\lambda^m\int_{\check{t}}^{T^Z}e^{\lambda T^Z+(\lambda^m-\lambda)s} - e^{\lambda^m s}ds$$

and so

$$\frac{dG^{R*}_{G^Z_{\check{t},T^Z(\check{t},M)}}(0)}{d\check{t}} = \frac{z\lambda^m}{(1-z)^2}\left(ze^{\lambda T^Z+(\lambda^m-\lambda)\check{t}} - e^{\lambda^m\check{t}}\right) - \frac{dT^Z}{d\check{t}}\frac{z^2}{(1-z)^2}\lambda\lambda^m\int_{\check{t}}^{T^Z}e^{\lambda T^Z+(\lambda^m-\lambda)s}ds \tag{15}$$

which has the same sign as

$$\begin{aligned}Y(w) =& \frac{dG^{R*}_{G^Z_{\check{t},T^Z(\check{t},M)}}(0)}{d\check{t}}\frac{(1-z)^2}{z\lambda^m e^{\lambda^m\check{t}}} = -1 + ze^{\lambda(T^Z-\check{t})} - \frac{dT^Z}{d\check{t}}\frac{z\lambda}{\lambda^m-\lambda}e^{\lambda(T^Z-\check{t})}(e^{(\lambda^m-\lambda)(T^Z-\check{t})}-1) \qquad (16)\\ =& -1 + \left(\frac{u(\alpha)-w}{u(\alpha)-u(1-\alpha)}\right) - \frac{\lambda u(\alpha)}{(\lambda^m-\lambda)u(1-\alpha)}\left(\frac{u(1-\alpha)-w}{u(\alpha)-u(1-\alpha)}\right)\left(\left(\frac{u(\alpha)-w}{z(u(\alpha)-u(1-\alpha))}\right)^{\frac{\lambda^m-\lambda}{\lambda}} - 1\right)\end{aligned}$$

where $w = Me^{r\check{t}}$. The first line evaluates the integral and the second imposes

$$\frac{dT^Z(\check{t},M)}{d\check{t}} = \frac{u(\alpha)\lambda - (r+\lambda)Me^{r\check{t}}}{\lambda(u(\alpha)-Me^{r\check{t}})} = \frac{u(\alpha)}{u(1-\alpha)}\frac{u(1-\alpha)-Me^{r\check{t}}}{u(\alpha)-Me^{r\check{t}}}$$

where this uses $\frac{r+\lambda}{\lambda} = \frac{u(\alpha)}{u(1-\alpha)}$. Clearly $Y(u(1-\alpha)) = 0$, as occurs when $\check{t} = 0$ and $M = u(1-\alpha)$. Taking the derivative of Y we get

$$\frac{dY(w)}{dw}(u(\alpha)-u(1-\alpha)) = -1+\frac{u(\alpha)\lambda}{u(1-\alpha)(\lambda^m-\lambda)}\left(\left(\frac{u(\alpha)-w}{z(u(\alpha)-u(1-\alpha))}\right)^{\frac{\lambda^m-\lambda}{\lambda}} - 1 + (u(1-\alpha)-w)\frac{\lambda^m-\lambda}{\lambda}\frac{(u(\alpha)-w)^{\frac{\lambda^m-2\lambda}{\lambda}}}{(z(u(\alpha)-u(1-\alpha)))^{\frac{\lambda^m-\lambda}{\lambda}}}\right) \tag{17}$$

Imposing risk neutrality, $u(x) = x$, so that $\lambda^m = 2\lambda$, and evaluating this at $w = (1-\alpha)$ gives:

$$\frac{dY(w)}{dw}\Big|_{w=1-\alpha}(2\alpha-1) = -1 + \frac{\alpha}{1-\alpha}\left(\frac{1}{z}-1\right) = \frac{\alpha-z}{z(1-\alpha)}$$

This is clearly positive whenever $\alpha > z$, which implies $Y(w) > 0$ when w is slightly greater than $(1-\alpha)$, and so $G^{R*}_{G^Z_{\check{t},T^Z(\check{t},M)}}(0) > 0$ when $M = (1-\alpha)$ and $\check{t}$ is slightly greater than 0. Recalling that $U^n(t) = M$ for $t \in [\check{t}, T^Z(\check{t},M)]$ implies $IC_{G^Z_{\check{t},T^Z(\check{t},M)}}(t) > 0$ for all t. □

$t \in [\min\{\hat{t}, T^Z\}, T^Z]$

Proof of Lemma 5. We are given the existence of some $\check{G}^Z$ with $\min_s IC_{\check{G}^Z}(s) = 0 = IC_{\check{G}^Z}(t)$ for $t \in [\min\{\hat{t}, T^Z\}, T^Z]$ and $\check{G}^Z \neq G^Z_{t^*(T^Z),T^Z}$. I consider distributions G^Z with the same fixed T^Z throughout the proof and $\min_s IC_{G^Z}(s) = 0 = IC_{G^Z}(t)$ for $t \in [\min\{\hat{t}, T^Z\}, T^Z]$. For such G^Z, let $\underline{t}_{G^Z} = \inf\{t : G^Z(t) > 0\}$ and $\bar{t}_{G^Z} = \min\{t : IC_{G^Z}(s) = 0 \text{ for } s \in [t, T^Z]\}$. Clearly, we have, $\underline{t}_{\check{G}^Z} \leq \bar{t}_{\check{G}^Z} \leq \min\{\hat{t}, T^Z\}$. Moreover, notice that we must have $G^Z(t) = G^Z_{t^*(T^Z),T^Z}(t)$ for $t \geq \bar{t}_{\check{G}^Z}$.

Given $\check{G}^Z \neq G^Z_{t^*(T^Z),T^Z}$, I claim that $\bar{t}_{\check{G}^Z} > t^*(T^Z)$. Suppose not, then $\check{G}^Z(t) \geq G^Z_{t^*(T^Z),T^Z}(t)$ for all t and $IC_{G^Z_{t^*(T^Z),T^Z}}(T^Z) = 0$. We must then have $IC_{\check{G}^Z}(T^Z) - IC_{G^Z_{t^*(T^Z),T^Z}}(T^Z) < 0$, presenting a contradiction. I further claim that $\underline{t}_{\check{G}^Z} < \bar{t}_{\check{G}^Z}$. Suppose not, so that $\underline{t}_{\check{G}^Z} = \bar{t}_{\check{G}^Z} > t^*(T^Z)$. In this case $\check{G}^Z(t) \leq G^Z_{t^*(T^Z),T^Z}(t)$ for all t, which implies $IC_{\check{G}^Z}(T^Z) - IC_{G^Z_{t^*(T^Z),T^Z}}(T^Z) > 0$. Given that by assumption $IC_{\check{G}^Z}(T^Z) = 0$ we have a contradiction.

Given some fixed $\tilde{t}$, define the relation $\gtrsim^*_{\tilde{t}}$ over distributions with $\min_s IC_{G^Z}(s) = 0 = IC_{G^Z}(t)$ for $t \in [\min\{\hat{t}, T^Z\}, T^Z]$ as follows. Let $G^Z \sim^*_{\tilde{t}} G^Z$, and let $G^Z \succ^*_{\tilde{t}} \tilde{G}^Z$ if $\underline{t}_{\tilde{G}^Z} \leq \underline{t}_{G^Z}$, $\bar{t}_{\tilde{G}^Z} \geq \bar{t}_{G^Z}$, $G^{R*}_{G^Z}(0) > G^{R*}_{\tilde{G}^Z}(0)$ and either $\underline{t}_{G^Z} = \tilde{t}$ and $G^Z(t) \geq \tilde{G}^Z(t)$ for $t \geq \tilde{t}$, or $\bar{t}_{G^Z} = \tilde{t}$ and $G^Z(t) \leq \tilde{G}^Z(t)$ for $t < \tilde{t}$.

Given $\check{G}^Z \neq G^Z_{t^*(T^Z),T^Z}$, I claim that for any $\tilde{t} \in (\underline{t}_{\check{G}^Z}, \bar{t}_{\check{G}^Z})$, there exists a distribution $\bar{G}^Z_{\tilde{t}}$ such that $\bar{G}^Z_{\tilde{t}} \succ^*_{\tilde{t}} \check{G}^Z$, and there is no distribution G^Z for which $G^Z \succ^*_{\tilde{t}} G^Z_{\tilde{t}}$.

I first simply look to find an alternative distribution $\hat{G}^Z$, which generates higher payoffs. For some $\varepsilon' \in (0, \tilde{t} - \underline{t}_{\check{G}^Z})$ and $\varepsilon^1 \geq 0$, let $\hat{G}^Z(t) = \min\{\check{G}^Z(t) - \varepsilon^1, 0\}$ for $t \leq \underline{t}_{\check{G}^Z} + \varepsilon'$. Suppose that $\check{G}^Z$ is discontinuous at $t' \in (\tilde{t}, \bar{t}]$ so that $\sup_{s<t'} \check{G}(s) < \check{G}(t')$. In this case let $\hat{G}^Z(t) = \sup_{s<t'} \check{G}(s) + \varepsilon^2$ for $t \in [t' - \varepsilon', t')$, where $\varepsilon^2 \geq 0$ is still to be defined and $\varepsilon' \in (0, t' - \tilde{t}]$. If on the other hand $\check{G}(t)$ is continuous on $(\tilde{t}, \bar{t}]$ then there must exists some $t' \in (\tilde{t}, \bar{t}]$ such that $IC_{G^Z}(t') > 0$ and $G^Z(t') > G^Z(t)$ for $t < t'$. In this case define $\hat{G}^Z(t) = \max\{\check{G}(t) + \varepsilon^2, \check{G}(t')\}$ for $t \in [t' - \varepsilon', t')$. Let $\hat{G}^Z(t) = \check{G}^Z(t)$ elsewhere.

For $t \geq t'$ we have:

$$IC_{\hat{G}^Z}(t) - IC_{\check{G}^Z}(t) = \int_{\underline{t}_{\hat{G}^Z}}^{\underline{t}_{\hat{G}^Z}+\varepsilon} (\hat{G}^Z - \check{G}^Z(s)) r\left(u(1-\alpha)e^{\lambda^m s} z - u(\alpha)e^{-rs}(1-z)\right) ds$$
$$- \int_{t'-\varepsilon}^{t'} (\hat{G}^Z - \check{G}^Z(s)) r\left(u(1-\alpha)e^{\lambda^m s} z - u(\alpha)e^{-rs}(1-z)\right) ds$$

This difference is continuous and strictly increasing in ε^1 and $-\varepsilon^2$, is positive for $\varepsilon^2 = 0$ and negative for $\varepsilon^1 = 0$. For all sufficiently small ε^1 therefore, there is a uniquely defined ε^2 such that $IC_{\hat{G}^Z}(t) - IC_{\check{G}^Z}(t) = 0$ for $t \geq t'$. This leaves $\hat{G}^Z$ as a function of ε^1 and ε'. For $IC_{\hat{G}^Z}(t) - IC_{\check{G}^Z}(t) = 0$ as $\varepsilon^1 \to 0$ and then $\varepsilon' \to 0$ we must have:

$$\lim_{\varepsilon' \to 0} \lim_{\varepsilon^1 \to 0} \frac{IC_{\hat{G}^Z}(t) - IC_{\check{G}^Z}(t)}{\varepsilon^1 \varepsilon'} = -r\left(u(1-\alpha)e^{\lambda^m \underline{t}_{\hat{G}^Z}} z - u(\alpha)e^{-r\underline{t}_{\hat{G}^Z}}(1-z)\right) + \lim_{\varepsilon' \to 0} \lim_{\varepsilon^1 \to 0} \frac{\varepsilon^2}{\varepsilon^1} r\left(u(1-\alpha)e^{\lambda^m t'} z - u(\alpha)e^{-rt'}(1-z)\right) = 0$$

Notice that for sufficiently small ε', we have $IC_{\check{G}^Z}(s) \geq \delta$ for $s \in [t' - \varepsilon', t')$ and some $\delta > 0$ and so for such s, $IC_{\hat{G}^Z}(s) > 0$ for all sufficiently small ε^2. For $s < t' - \varepsilon$ we have $U^n_{\hat{G}^Z}(s) \leq U^n_{\hat{G}^Z}(s)$, hence, so long as we can show

$G^{R*}_{\hat{G}^Z}(0) > G^{R*}_{\check{G}^Z}(0)$ then all time t type incentive constraints will be satisfied. To that end, notice that:

$$G^{R*}_{\hat{G}^Z}(0) - G^{R*}_{\check{G}^Z}(0) = \int_0^{T^Z} \lambda^m e^{\lambda^m s} \frac{z}{1-z} (\hat{G}^Z(s) - \check{G}^Z(s)) ds$$

$$\begin{aligned}
\lim_{\varepsilon' \to 0} \lim_{\varepsilon^1 \to 0} \frac{G^{R*}_{\hat{G}^Z}(0) - G^{R*}_{\check{G}^Z}(0)}{\varepsilon' \varepsilon^1} \frac{1-z}{z\lambda^m} &= e^{\lambda^m t'} \lim_{\varepsilon' \to 0} \lim_{\varepsilon^1 \to 0} \frac{\varepsilon^2}{\varepsilon^1} - e^{\lambda^m \underline{t}_{\hat{G}^Z}} \\
&= \frac{e^{\lambda^m t'} \left(u(1-\alpha) e^{\lambda^m \underline{t}_{\hat{G}^Z}} z - u(\alpha) e^{-r \underline{t}_{\hat{G}^Z}} (1-z) \right) - e^{\lambda^m \underline{t}_{\hat{G}^Z}} \left(u(1-\alpha) e^{\lambda^m t'} z - u(\alpha) e^{-rt'} (1-z) \right)}{u(1-\alpha) e^{\lambda^m t'} z - u(\alpha) e^{-rt'} (1-z)} \\
&= \frac{u(\alpha)(1-z) e^{(\lambda^m - r) \underline{t}_{\hat{G}^Z}} (e^{-r(t' - \underline{t}_{\hat{G}^Z})} - e^{\lambda^m (t' - \underline{t}_{\hat{G}^Z})})}{u(1-\alpha) e^{\lambda^m t'} z - u(\alpha) e^{-rt'} (1-z)} > 0
\end{aligned}$$

Where the final two lines hold for $t' < \hat{t}$, and the inequality in the final line follows because the denominator is negative for $t' < \hat{t}$. If $t' = \hat{t}$, on the other hand then we must have $\lim_{\varepsilon' \to 0} \lim_{\varepsilon^1 \to 0} \frac{\varepsilon^2}{\varepsilon^1} = \infty$ and so the second line must certainly be strictly positive. The implication of this is that for sufficiently small ε^1 and ε', $G^{R*}_{\hat{G}^Z}(0) > G^{R*}_{\check{G}^Z}(0)$.

Fix $\tilde{t} \in (\underline{t}_{\check{G}^Z}, \bar{t}_{\check{G}^Z})$, $T = T^Z$. For arbitrary G^Z define $v(T, G^Z) = G^{R*}_{G^Z}(0)$ and $w(t, G^Z) = IC_{G^Z}(t)$. The proof above establishes that there exists some $\hat{G}^Z$ such that $\bar{t}_{\hat{G}^Z} \leq \bar{t}_{\check{G}^Z}$, $\hat{G}^Z(t) \geq \check{G}^Z(t)$ for $t \geq \tilde{t}$, $\hat{G}^Z(s) \leq \check{G}^Z(s)$ for $s < \tilde{t}$, $v(T, \hat{G}^Z) > v(T, \check{G}^Z)$ and $w(t, \hat{G}^Z) \geq 0$ for all $t \leq T$. Invoking Lemma 7, therefore, there exists $\overline{G}^Z_{\tilde{t}}$ such that $IC_{\overline{G}^Z_{\tilde{t}}}(t) \geq 0$, $G^{R*}_{\overline{G}^Z_{\tilde{t}}}(0) > G^{R*}_{\check{G}^Z}(0)$, $\bar{t}_{\bar{G}^Z_{\tilde{t}}} \leq \bar{t}_{\check{G}^Z}$, $\bar{G}^Z_{\tilde{t}}(t) \geq \check{G}^Z(t)$ for $t \geq \tilde{t}$, $\bar{G}^Z_{\tilde{t}}(s) \leq \check{G}^Z(s)$ for $s < \tilde{t}$, and there is no alternative incentive compatible G^Z with $G^Z(t) \geq \overline{G}^Z_{\tilde{t}}(t)$ for $t \geq \tilde{t}$ and $G^Z(s) \leq \overline{G}^Z_{\tilde{t}}(s)$ for $s < \tilde{t}$, $\bar{t}_{\bar{G}^Z_{\tilde{t}}} \geq \bar{t}_{G^Z}$ such that $G^{R*}_{G^Z}(0) > G^{R*}_{\overline{G}^Z_{\tilde{t}}}(0)$.

I claim that we have $\bar{G}^Z_{\tilde{t}} \succ^* \check{G}^Z$. Clearly we have $\underline{t}_{\check{G}^Z} \leq \underline{t}_{\overline{G}^Z_{\tilde{t}}}$, and so for this not to be true requires that both $\bar{t}_{G^Z} < \tilde{t}$ and $\bar{t}_{G^Z} > \tilde{t}$. But in this case, we know from the previous construction that there must exist an incentive compatible distribution $\tilde{G}^Z$ with $G^Z(t) \geq \overline{G}^Z_{\tilde{t}}(t)$ for $t \geq \tilde{t}$ and $G^Z(t) \leq \overline{G}^Z_{\tilde{t}}(t)$ for $t < \tilde{t}$, $\bar{t}_{\bar{G}^Z_{\tilde{t}}} \geq \bar{t}_{G^Z}$ such that $G^{R*}_{G^Z}(0) > G^{R*}_{\overline{G}^Z_{\tilde{t}}}(0)$, a contradiction.

I next define the relation $\gtrsim^*$ over distributions with $\min_s IC_{G^Z}(s) = 0 = IC_{G^Z}(t)$ for $t \in [\min\{\hat{t}, T^Z\}, T^Z]$ as follows. Let $G^Z \sim^* G^Z$, and let $G^Z \succ^* G^Z$ if there exists some $\tilde{t}$ such that $G^Z \succ^*_{\tilde{t}} G^Z$. Let $\overline{u}^*(G^Z) = \sup\{G^{R*}_{\tilde{G}^Z}(0) : \tilde{G}^Z \gtrsim^* G^Z\}$. Define a sequence of distribution functions by $G^0 = \check{G}^Z \neq G^Z_{t^*(T^Z), T^Z}$, and $G^{k+1} \gtrsim^* G^k$ such that $G^{R*}_{G^{k+1}}(0) \geq \frac{\overline{u}^*(G^k) + G^{R*}_{G^k}(0)}{2}$. Clearly, if $G^k = G^Z_{t^*(T^Z), T^Z}$ for some k then the proof is complete (as then $G^{R*}_{G^Z_{t^*(T^Z), T^Z}}(0) \geq G^{R*}_{\check{G}^Z}(0)$ and $\check{G}^Z$ was arbitrary), so suppose not and $G^{k+1} \succ^* G^k$ for all k.

We want to establish that $G^k \overset{w}{\to} \overline{G}^Z_{t^*(T^Z), T^Z}$, which if shown completes the proof (as then $G^{R*}_{G^Z_{t^*(T^Z), T^Z}}(0) > G^{R*}_{\check{G}^Z}(0)$). Suppose not. Let $\tilde{t}^k \in [\underline{t}_{G^0}, \bar{t}_{G^0}]$ be such that $G^{k+1} \succ^*_{\tilde{t}^k} G^k$, and (taking a subsequence if necessary) let $\tilde{t}^k \to \tilde{t}^*$.

I claim that $\underline{G}(t) = \lim_k G^k(t)$ is well defined except possibly at $t = \tilde{t}^*$, where we can let it be defined some arbitrary subsequence if necessary. This is clearly the case if $\underline{t}_{G^k} \to \tilde{t}^*$ and $\bar{t}_{G^k} \to \tilde{t}^*$. If $\underline{t}_{G^k} \not\to \tilde{t}^*$, so that for some $t < \tilde{t}^*$ we have $\underline{t}_{G^k} < t$ for all sufficiently large k, (i.e. $G^k(t) > 0$). For all sufficiently large k we must have $\tilde{t}^{k-1} > t$, and so $G^k(t) > 0$ implies $IC_{G^k}(s) = 0$ for $s \in [\tilde{t}^{k-1}, T^Z]$ and $G^k(v) \leq G^{k-1}(v)$ for $v \leq \tilde{t}^{k-1}$. But in which case $\underline{G}(t)$ is defined for $t \neq \tilde{t}^*$. Suppose instead that $\bar{t}_{G^k} \not\to \tilde{t}^*$, and so for some $t > \tilde{t}^*$ we have $\bar{t}_{G^k} \geq t$ for all sufficiently large k. For all sufficiently large k we must have $\tilde{t}^{k-1} < t$, and so $\bar{t}_{G^k} \geq t$ implies $G^k(s) = 0$ for $s \leq \tilde{t}^{k-1}$ and $G^k(s) \geq G^{k-1}(s)$ for $s \geq \tilde{t}^{k-1}$. But in which case $\underline{G}(t)$ is again defined for $t \neq \tilde{t}^*$.

We can now define the cumulative distribution function $\overline{G}$ by $\overline{G}(t) = \inf\{\underline{G}(s) : s > t\}$, so that $G^k \overset{w}{\to} \overline{G}$. Replicating the arguments from Lemma 7, we have $\min_s IC_{\overline{G}}(s) = 0 = IC_{\overline{G}}(t)$ for $t \in [\min\{\hat{t}, T^Z\}, T^Z]$. Given the supposition that $\overline{G} \neq G^Z_{t^*(T^Z), T^Z}$, then we know there exists some $\tilde{G}^Z \succ \overline{G}$. But in which case there exists some $\tilde{G}^Z \succ G^k$ and so $G^{R*}_{G^{k+1}}(0) - G^{R*}_{G^{k+1}}(0) \geq \frac{\overline{G}^{R*}_{\tilde{G}^Z} - G^{R*}_{\overline{G}}(0)}{2} > 0$, contradicting the fact that $G^{R*}_{G^k}(0) \leq 1$ for sufficiently large k. This completes the proof.

□

Proof of Lemma 6. We can restrict attention to parameters $z < \frac{u(\alpha)}{u(\alpha)+u(1-\alpha)}$, because otherwise we have $\hat{t} \leq 0$, a case which has already been dealt with. The argument in the main text establishes that an OSSMP exists.

By Lemma 5 we know that $G^{Z*} = G^{Z}_{t^*(T^Z),T^Z}$ for some T^Z. Following the sketched argument for uniqueness outlined in the main text, let the maximized objective be $U^{c*}_{G^{Z*}}(0) = \underline{u} > u(1-\alpha)$. Because $IC_{G^{Z*}}(t) = 0$ for $t \in [t^*(T^Z), T^Z]$, we must also have $U^n(t) = \overline{u}$ for such t. Knowing this, we can consider a reduced problem of maximizing $G^{R*}_{G^{Z}_{\check{t},T^Z(\check{t},\underline{u})}}(0)$, with respect to $\check{t}$ where $T^Z(\check{t}, M)$ is defined in equation (11), to ensure that $U^n(T^Z) = M$. This reduced problem must have the same maximizers (i.e. implied distribution function) as the original problem. Equation (16) in the proof of Lemma 4 defines the variable $Y(w)$ which has the same sign as $\frac{dG^{R*}_{G^{Z}_{\check{t},T^Z(\check{t},\underline{u})}}(0)}{d\check{t}}$ where $w = \underline{u}e^{r\check{t}}$. Equation 17 evaluates $\frac{dY(w)}{dw}$. The second derivative is:

$$\frac{d^2Y(w)}{dw^2} = -\frac{u(\alpha)}{u(1-\alpha)(u(\alpha)-u(1-\alpha))}\left(\frac{(u(\alpha)-w)^{\frac{\lambda^m-2\lambda}{\lambda}}}{(z(u(\alpha)-u(1-\alpha)))^{\frac{\lambda^m-\lambda}{\lambda}}} + \frac{(u(\alpha)-w)^{\frac{\lambda^m-2\lambda}{\lambda}}}{(z(u(\alpha)-u(1-\alpha)))^{\frac{\lambda^m-\lambda}{\lambda}}} + (u(1-\alpha)-w)\frac{\lambda^m-2\lambda}{\lambda}\frac{(u(\alpha)-w)^{\frac{\lambda^m-3\lambda}{\lambda}}}{(z(u(\alpha)-u(1-\alpha)))^{\frac{\lambda^m-\lambda}{\lambda}}}\right)$$

Evaluating at $w = \overline{u}e^{r\check{t}} \in (u(1-\alpha), u(\alpha))$ and remembering that $\lambda^m \in (\lambda, 2\lambda]$, it is clear that this expression is strictly negative as each of the bracketed terms is positive, the first two strictly. The implication is that $G^{R*}_{G^{Z}_{\check{t},T^Z(\check{t},\overline{u})}}$ is strictly quasiconcave in $\check{t}$, and so has a unique maximizer $\check{t}^*$. This completes the proof of uniqueness.

Next, notice that $\check{t} = T^Z(\check{t}, \underline{u})$ if and only if $w = \overline{u}e^{r\check{t}} = u(\alpha) - z(u(\alpha) - u(1-\alpha))$, but in this case $Y(w) = -(1-z) < 0$ and so decreasing $\check{t}$ slightly (which is certainly possible given $T^Z > 0$) would strictly increase $G^{R*}_{G^{Z}_{\check{t},T^Z(\check{t},M)}}(0)$.

As already noted in the text, the proof of Lemma 3 establishes that the distribution G^{Z}_{0,T^Z} can only satisfy both incentive constraints for risk neutral agents agents when it matches the Baseline equilibrium distribution (i.e. $T^Z = -\frac{1}{\lambda}ln(z)$), and so we must have $t^*(T^Z) > 0$ in the optimal distribution when $z < \alpha$. More generally, notice that Proposition 7 establishes that $\lim_n G^{R*}_{G^{Z*}}(0) = 1$ if $z^n \to 0$ or $\alpha^n \to 1$, and so in either case $\lim_n u \to u(0.5)$. This implies that

$$\lim_{z\to 0} z^{\frac{\lambda^m-\lambda}{\lambda}} Y(w) = -\frac{\lambda u(\alpha)}{(\lambda^m-\lambda)u(1-\alpha)}\frac{u(1-\alpha)-\lim_n w}{(u(\alpha)-u(1-\alpha))}\left(\frac{u(\alpha)-\lim_n w}{u(\alpha)-u(1-\alpha)}\right)^{\frac{\lambda^m-\lambda}{\lambda}}$$

is strictly positive whenever $\lim_n w = u(0.5)e^{r\lim_n \check{t}} < u(\alpha)$ (taking a subsequence if necessary to ensure convergence, noting that $e^{-r\check{t}}u(\alpha) \geq u(1-\alpha)$). Clearly, therefore, the associated sequence of optimal distributions must satisfy $\lim_n \check{t} = \frac{1}{r}ln\left(\frac{u(\alpha)}{u(0.5)}\right) > 0$ or else $\frac{dG^{R*}_{G^{Z}_{\check{t},T^Z(\check{t},\underline{u})}}(0)}{d\check{t}} > 0$ for all sufficiently small z or all sufficiently large α, a contradiction. This establishes the existence of some $\underline{z}(u,\alpha) > 0$ and $\underline{\alpha}(u,z) < 1$ such that if $z < \underline{z}(u,\alpha)$ or $\alpha > \underline{\alpha}(u,z)$ then $\check{t} > 0$ in the optimal distribution.

□

Proof of Proposition 7. Notice that payoffs in an OSSMP and under a strongly symmetric $N1$ protocol are of the form $G^R(0)(1-z)(u(0.5)-u(1-\alpha)) + u(1-\alpha)$ where $G^R(0) = b$ in the latter. Hence, in an OSSMP we must have $G^{R*}_{G^{Z*}}(0) \geq b$, for any b which is part of an $N1$ equilibrium. Let $b \in (0,1)$ arbitrary, then the condition for an $N1$ equilibrium to exists as highlighted by the proof of Proposition 4 is that

$$Q = u(0.5) - u(1-\alpha) - \frac{z\left(1-\left(\frac{z}{1-(1-z)b}\right)^{\frac{r}{\lambda}}\right)u(1-\alpha)}{(1-z)(1-b)} \geq 0$$

Suppose first that $B^n = (\alpha, z^n, u, r)$ with $\lim_n z^n = 0$. It is clear that the final expression vanishes so that $\lim_n Q = u(0.5) - u(1-\alpha) > 0$. Suppose next that $B^n = (\alpha^n, z, u, r)$ with $\lim \alpha^n = 1$. In this case $\lim u(1-\alpha^n) = 0$ and $\lim \frac{r}{\bar{\lambda}} = \infty$ so that $\lim_n Q = u(0.5) > 0$.

Finally, suppose that $B^n = (\alpha, z, u^n, r)$ with $\lim_n u^n(\alpha) = \lim_n u^n(0.5) > \lim_n u^n(1-\alpha)$, and without loss of generality normalize $u^n(1-\alpha) = \underline{u}(1-\alpha) > 0$ for all n. Evaluating the limit of Q and rescaling gives

$$\hat{Q}(b) = \lim_n \frac{Q(1-b)(1-z)}{\underline{u}(1-\alpha)} = \frac{r}{\bar{\lambda}}(1-b)(1-z) - z\left(1 - \left(\frac{z}{1-(1-z)b}\right)^{\frac{r}{\bar{\lambda}}}\right)$$

where $\frac{r}{\bar{\lambda}} = \frac{\lim_n u^n(\alpha) - \underline{u}(1-\alpha)}{\underline{u}(1-\alpha)}$. Moreover,

$$\begin{aligned}
\frac{d\hat{Q}(b)}{db} &= -\frac{r}{\bar{\lambda}}(1-z) + \frac{r}{\bar{\lambda}}(1-z)z^{\frac{r+\bar{\lambda}}{\bar{\lambda}}}(1-(1-z)b)^{-\frac{r+\bar{\lambda}}{\bar{\lambda}}} \\
\frac{d^2\hat{Q}(b)}{db^2} &= \frac{r}{\bar{\lambda}}\frac{r+\bar{\lambda}}{\bar{\lambda}}(1-z)^2 z^{\frac{r+\bar{\lambda}}{\bar{\lambda}}}(1-(1-z)b)^{-\frac{r+2\bar{\lambda}}{\bar{\lambda}}} > 0
\end{aligned}$$

Evaluating at $b = 1$ we get $\hat{Q}(1) = 0$ and $\frac{d\hat{Q}(b)}{db}\big|_{b=1} = 0$, so that $\hat{Q}(b) > 0$ for any $b < 1$. □

Proof of Proposition 6. Suppose there is some optimal mediation protocol which is not symmetric, $(G^R, G_1^Z, G_2^Z, M_1, M_2)$. The discussion of Observation 3 highlighted that the strongly symmetric mediation protocol $(G^R, \check{G}^Z)$ with $\check{G}^Z = 0.5(G_1^Z + G_2^Z)$ implied an equilibrium with utilities $\hat{U}^c(t) \geq 0.5(U_1^c(t) + U_2^c(t))$ and $\hat{U}^n(t) = 0.5(U_1^n(t) + U_2^n(t))$. Clearly, if this is not the OSSMP, then the non-symmetric protocol is not optimal. The OSSMP implies $\hat{U}^c(T^R) = \hat{U}^c(t)$ for $t \leq T^R = T^Z$, $\hat{U}^c(T^R) = \hat{U}^n(t)$ for $t \in [t^*, T^R]$, and $\hat{G}^Z(t) = 0$ for $t < t^*$. This immediately implies $G_i^Z(t) = 0$ for $t < t^*$.

Because the non-symmetric protocol is an equilibrium, we must have $U_i^c(T^R) \geq U_i^c(t)$ and $U_i^c(T^R) \geq U_i^n(t)$. Suppose that $U_i^c(T^R) > U_i^n(t)$ for some $i \in \{1,2\}$ and some $t \in [t^*, T^R]$, then clearly $\hat{U}^c(T^R) > U_i^n(t)$, a contradiction. However, if we have $U_i^c(T^R) = U_i^n(t)$ for $t \in [t^*, T^R]$ then on this interval we must have $G_i^Z(t) = \frac{1-ze^{\lambda(T^R-t)}}{1-z} = \hat{G}^Z$ and so the original protocol must in fact be symmetric. Finally, notice that when $u''(0.5) < 0$, then $u(0.5) < \int u(m)d\check{M}^t(m)$ for any symmetric $\check{M}^t$ where $\check{M}^t(0.5) < 1$. Hence, we must have $\check{M}^0(0.5) = 1$ and $\int \mathbb{1}_{[t:\check{M}^t(0.5)=1]}dG^R(t) = 1$. □

References

Abreu, D. and F. Gul (2000). Bargaining and Reputation. *Econometrica 68*(1), pp. 85–117. 2

Abreu, D. and D. Pearce (2007). Bargaining, reputation, and equilibrium selection in repeated games with contracts. *Econometrica 75*(3), 653–710. 6

Basak, D. (2016). Transparency and delay in bargaining. *Unpublished manuscript*. 36

Beardsley, K. C., D. M. Quinn, B. Biswas, and J. Wilkenfeld (2006). Mediation style and crisis outcomes. *Journal of conflict resolution 50*(1), 58–86. 2

Brazil, W. D. (2007). Hosting mediations as a representative of the system of civil justice. *Ohio State Journal on Dispute Resolution 22*(2), 227–276. 3

Coase, R. H. (1972). Durability and Monopoly. *Journal of Law and Economics 15*(1), pp. 143–149. 9

Čopič, J. and C. Ponsatí (2008). Robust bilateral trade and mediated bargaining. *Journal of the European Economic Association 6*(2-3), 570–580. 35, 36

Dixon, W. J. (1996). Third-party techniques for preventing conflict escalation and promoting peaceful settlement. *International Organization 50*(04), 653–681. 1

Dunlop, J. T. (1984). *Dispute resolution: Negotiation and consensus building*. Greenwood Publishing Group. 2

Ely, J. C. (2017, January). Beeps. *American Economic Review 107*(1), 31–53. 36

Emery, R. E., S. G. Matthews, and M. M. Wyer (1991). Child custody mediation and litigation: Further evidence on the differing views of mothers and fathers. *Journal of Consulting and Clinical Psychology 59*(3), 410. 2

Fanning, J. (2016). Reputational bargaining and deadlines. *Econometrica*. 2

Fershtman, C. and D. Seidmann (1993). Deadline Effects and Inefficient Delay in Bargaining with Endogenous Commitment. *Journal of Economic Theory 60*, 306–321. 34

Goldberg, S. B., F. E. Sander, N. H. Rogers, and S. Rudolph Cole (2012). *Dispute Resolution: Noegotitation, Mediation, Arbitration, and Other Processes* (6 ed.). Wolters Kluwer. 6, 13

Goltsman, M., J. Hörner, G. Pavlov, and F. Squintani (2009). Mediation, arbitration and negotiation. *Journal of Economic Theory 144*(4), 1397–1420. 3, 36

Hörner, J., M. Morelli, and F. Squintani (2015). Mediation and peace. *The Review of Economic Studies 82*(4), 1483–1501. 36

Jarque, X., C. Ponsatı, and J. Sákovics (2003). Mediation: incomplete information bargaining with filtered communication. *Journal of Mathematical Economics 39*(7), 803–830. 6, 35, 36

Kydd, A. (2001). Which side are you on? mediation as cheap talk. *American Journal of Political Science 47*(3), 596–611. 3, 29

Manzini, P. and C. Ponsati (2006). Stakeholder bargaining games. *International Journal of Game Theory 34*(1), 67–77. 37

Myerson, R. (1991). *Game Theory: Analysis of Conflict*. Cambridge, Massachusetts: Harvard University Press. 36

Myerson, R. B. (1986). Multistage games with communication. *Econometrica: Journal of the Econometric Society*, 323–358. 19

Myerson, R. B. and M. A. Satterthwaite (1983, April). Efficient mechanisms for bilateral trading. *Journal of Economic Theory 29*(2), 265281. 5

Ponsati, C. (1997). Compromise vs. capitulation in bargaining with incomplete information. *Annales d'Economie et de Statistique*, 191–210. 36

Powell, R. (2002). Bargaining theory and international conflict. *Annual Review of Political Science 5*(1), 1–30. 35

Rubinstein, A. (1982). Perfect Equilibrium in a Bargaining Model. *Econometrica 50*(1), pp. 97–109. 58

Stipanowich, T. and J. R. Lamare (2013). Living with'adr': Evolving perceptions and use of mediation, arbitration and conflict management in fortune 1,000 corporations. *Arbitration and Conflict Management in Fortune 1*. 1, 6, 33

Velikonja, U. (2009). Making peace and making money: economic analysis of the market for mediators in private practice. *Alb. L. Rev. 72*, 257. 3

Wilkenfeld, J., K. Young, V. Asal, and D. Quinn (2003). Mediating international crises: Cross-national and experimental perspectives. *Journal of Conflict Resolution 47*(3), 279–301. 2

Supplementary Material (for online publication)

Appendix B

In this appendix, I show that when the probability of behavioral types is small a low option equilibrium exists which delivers higher payoffs to rational agents than the OSSMP. Recall that a low option takes the following form: Rational agents confess rationality to the mediator at time 0^2. If both agents confess rationality then at time 0^3, with probability 0.5 the mediator (publicly) tells agent i to demand $\alpha_i = 1$ (revealing rationality), and otherwise tells agent j to demand $\alpha_j = 1$. In the former case, the mediator suggests that agent i the whole dollar in any subsequent agreement. If only agent i confesses rationality, then the mediator always tells her to demand $\alpha_i = 1$. In this case agent i obtains the share $(1-\alpha)$ in all subsequent agreements. If neither agent confesses, then the mediator says nothing. If an agent fails to follow the mediator's suggestion at some time, then the mediator subsequently says nothing and the continuation equilibrium specifies that she should concede to her opponent immediately.

Conditional on agent i (upon instruction) demanding $\alpha_i = 1$ at 0^4, let the cumulative distributions of her agreement times with a rational and behavioral opponent be $G^{R,o}$ and $G^{Z,o}$ respectively. Let $T^{R,o} = \sup_t\{t : G^{R,0}(t) < 1\}$ and $T^{Z,o} = \sup_t\{t : G^{Z,0}(t) < 1\}$. The conditional probability that agent i faces a behavioral opponent after her instruction is $\bar{z} = \frac{2z}{1+z}$. If she subsequently concedes at t^5 (without being instructed to by the mediator), the she obtains the expected utility:

$$\begin{aligned} U^{c,o}(t) =&(1-\bar{z})\int_{s\leq t} e^{-rs}u(1)dG^{R,o}(s) + \bar{z}\int_{s\leq t} e^{-rs}u(1-\alpha)dG^{Z,o}(s) \\ &+ e^{-rt}u(1-\alpha)\Big((1-z)(1-G^{R,o}(t)) + z(1-G^{Z,o}(t))\Big) \end{aligned}$$

which leads to a new dynamic incentive constraint

$$U^{c,o}(T^{R,o}) = \max_t U^{c,o}(t).$$

An agent j who confessed rationality but is not told to change her demand, can obtain $u(0) = 0$ from conceding and also gets a continuation payoff of zero, and so is indifferent to subsequently following the mediator's instructions.

If rational agent i does not confess and the mediator announces nothing at 0^3, then the agent realizes she must face a behavioral opponent and subsequently immediately concedes. If the mediator instead tells her opponent at 0^3 to demand α_j, then because i obtains $u_i(0) = 0$ from conceding (or otherwise revealing rationality), her continuation payoff is $\int_{t\leq T^R} e^{-rs}u(\alpha)dG^{Z,o}(s)$. In sum, her expected payoff to not confessing is:

$$U^{n,o} = zu(1-\alpha) + (1-z)\int_{t\leq T^R} e^{-rs}u(\alpha)dG^{Z,o}(s)$$

The new type incentive constraint is then simply:

$$\frac{(1+z)U^{c,o}(T^R)}{2} \geq U^{n,o}.$$

Consider the distributions $G^{Z,o}_*(t) = 0$ for $t < T^{R,o} = T^{Z,o}$ and $1 - G^{R,o}_*(t) = \frac{\bar{z}}{1-\bar{z}}(e^{\bar{\lambda}(T^{R,o}-t)} - 1)$ where $\bar{\lambda} = \frac{ru(1-\alpha)}{u(1)-u(1-\alpha)}$. These ensure that the dynamic incentive constraint binds in the sense that $U^{c,0}(T^{R,o}) = U^{c,o}(t)$ for $t \leq T^{R,o}$. We

then attempt to select the minimum $T^{R,o}$ such that the type incentive constraint binds. To that end define:

$$W(T^{R,o}) = \frac{(1+z)U^{c,o}(T^{R,o})}{2} - U^n = \frac{1-z}{2}\left(1 - \frac{2z}{1-z}(e^{\bar{\lambda}(T^{R,o}-t)} - 1)\right)(u(1) - u(1-\alpha)) + \frac{1+z}{2}u(1-\alpha) - zu(1-\alpha) - (1-z)e^{-rT^{R,o}}u(\alpha)$$

and let $T_*^{R,o} = \min\{T^{R,o} : W(T^{R,o}) \geq 0\}$. Notice that $W(T^{R,o})$ is strictly concave in $T^{R,o}$. For all sufficiently large z, $T_*^{R,o}$ is not well defined (this is certainly the case for $z \leq \alpha$ for risk neutral agents), however, it is well defined for sufficiently small z because as $z \to 0$ we have $W(T^{R,o}) \to 0.5u(1) - e^{-rT^{R,o}}u(\alpha)$, which also implies that $T_*^{R,o} \to \frac{1}{r}ln\left(\frac{2u(\alpha)}{u(1)}\right)$. For all sufficiently small z, therefore, there is an equilibrium of the low option mediation protocol. We are interested in payoffs under this protocol as compared to payoffs in the OSSMP as $z \to 0$ when agents are risk neutral. The difference between these payoffs is:

$$\begin{aligned} U^{c,o}(T^{R,o}) - U^c(T^R) =& \frac{1-z}{2}G_*^{R,o}(0)(u(1) - u(1-\alpha)) + \frac{1+z}{2}u(1-\alpha) - (1-z)G_{G^{Z*}}^{R*}(0)(u(0.5) - u(1-\alpha)) - u(1-\alpha) \\ =& \frac{1-z}{2}\left(\alpha(G_*^{R,o}(0) - G_{G^{Z*}}^{R*}(0)) - (1 - G^R*(0))(1-\alpha)\right) \end{aligned}$$

where the second equality imposes $u(x) = x$. By rearranging it is clear that $U^{c,o}(T^{R,o}) - U^c(T^R) \geq 0$ if and only if

$$\frac{2\alpha - 1}{\alpha} \geq \frac{1 - G_*^{R,o}(0)}{1 - G_{G^{Z*}}^{R*}(0)}.$$

The optimal distribution of rational-behavioral agreement times in the OSSMP satisfies $G^{Z*} = 0$ for $t < t^*$. This implies that $1 - G_{G^{Z*}}^{R*}(0) \geq \frac{z}{1-z}(e^{\lambda^m t^*} - 1)$ whereas $1 - G_*^{R,o}(0) = \frac{2z}{1-z}(e^{\bar{\lambda}T_*^{R,o}} - 1)$. Therefore, if we can show that

$$\frac{2\alpha - 1}{\alpha} \geq \frac{2(e^{\bar{\lambda}T_*^{R,o}} - 1)}{e^{\lambda^m t^*} - 1}$$

then the low option equilibrium delivers higher higher payoffs than the OSSMP. Above we noted that as $z \to 0$ we have $T_*^{R,o} \to \frac{1}{r}ln(2\alpha)$. The proof of Lemma 6 likewise showed that $t^* \to \frac{1}{r}(2\alpha)$ in this case. And so the low option equilibrium gives higher payoffs than the OSSMP for all sufficiently small z if:

$$\frac{2\alpha - 1}{\alpha} > \lim_{z \to 0} \frac{2(e^{\bar{\lambda}T_*^{R,o}} - 1)}{e^{\lambda^m t^*} - 1} = \frac{2\left((2\alpha)^{\frac{1-\alpha}{\alpha}} - 1\right)}{(2\alpha)^{\frac{2-2\alpha}{2\alpha-1}} - 1}.$$

Showing this inequality holds analytically is surprisingly tricky, however, it can be easily verified numerically that it holds for all $\alpha \in (0.5, 1)$.

Appendix C

I this Appendix I investigate the possibility of mediation protocols that lower agents payoffs compared to unmediated outcomes. The first result (Proposition 9) shows that when the probability of behavioral types is small, there always exists a mediation protocol that gives each agent i a payoff $u_i(1-\alpha_j)$, her payoff from conceding immediately. In the Baseline equilibrium, agent i's payoff were $F_j(0)u_i(\alpha_i) + (1 - F_j(0))u_i(1-\alpha_j)$ where $F_j(0) > 0$ if and only if $T_i = -\frac{1}{\lambda_i}ln(z_i) < T_j$. Hence, generically (whenever $T_i \neq T_j$) such mediation is Pareto inferior to the Baseline equilibrium. Clearly, this only lowers payoffs for at most one of the two agents.

The second result (Proposition 10) highlights some of the difficulties for mediation when agents can imitate multiple behavioral types. I extend the model slightly to allow for multiple types, and so mediation can affect rational

agents' initial demand choices. I show that when the probability of behavioral types is sufficiently small each agent obtains a payoff approximately equal to that which she would receive facing her most aggressive possible behavioral opponent for sure. This is strictly lower than the Baseline equilibrium payoff for both agents when the set of behavioral types is even moderately rich.

Proposition 9. *For any given r_i, u_i, α_i for $i = 1, 2$ and fixed $K \geq 1$, there exists $\underline{z} > 0$ such that whenever $z_i \leq \underline{z}$ and $K \geq \frac{z_1}{z_2} \geq \frac{1}{K}$, there is an equilibrium with mediation where each player i's payoff is exactly $u_i(1 - \alpha_j)$.*

Proof. I prove this by construction. Following the notation of Section 4 in the main text, consider the distributions $G_j^Z(t) = 0$ for $t < T^{Z_1} = T^{Z_2} = T^R$. We would like to find some T^R, some mediation proposals $m_i : [0, T^R] \to [0, 1]$ and a distribution G^R such that *both* agents are indifferent to conceding on $[0, T^R]$ and $G^R(0) = 0$. For arbitrary m_i and T^R the fraction of remaining j agents at $t < T^R$ is $1 - F_j(t) = (1 - G^R(t))(1 - z_j) + z_j$. For confessing agent i to be indifferent to concession on $(0, T^R]$ we must have

$$\frac{f_j(t)}{1 - F_j(t)} = \lambda_j^m(t) = \frac{r_i u_i(1 - \alpha_j)}{u_i(m_i(t)) - u_i(1 - \alpha_j)}$$

Imposing the boundary condition $1 - F_j(T^R) = z_j$ and solving the linear ODE, we get

$$1 - G^R(t) = 1 - \frac{F_j(t)}{1 - z_j} = \frac{z_j}{1 - z_j}\left(exp\left(\int_t^{T^R} \lambda_j^m(s)ds\right) - 1\right) \tag{18}$$

We want this equation for hold for both agent i and j for all t and so we get:

$$P(t) = \frac{g^R(t) - g^R(t)}{\lambda_j^m(t)\lambda_i^m(t)} = \frac{z_j}{(1 - z_j)\lambda_i^m(t)} exp\left(\int_t^{T^R} \lambda_j^m(s)ds\right) - \frac{z_i}{(1 - z_i)\lambda_j^m(t)} exp\left(\int_t^{T^R} \lambda_i^m(s)ds\right) = 0$$

It is immediate that we can identify $m_j(T^R)$ as the unique value which solves this at T^R:

$$\frac{z_j(1 - z_i)}{z_i(1 - z_j)} = \frac{\lambda_i^m(T^R)}{\lambda_j^m(T^R)} = \frac{u_j(m_j(T^R)) - u_j(1 - \alpha_i)}{u_i(1 - m_j(T^R)) - u_i(1 - \alpha_j)} \frac{r_i u_i(1 - \alpha_j)}{r_j u_j(1 - \alpha_i)}. \tag{19}$$

More generally, imposing $m_i(t) = 1 - m_j(t)$ and differentiating gives:

$$\frac{dP(t)}{dt} = \frac{z_j}{1 - z_j}\left(\frac{u_j'(m_j(t))m_j'(t)}{r_j u_j(1 - \alpha_i)} - \frac{\lambda_j^m(t)}{\lambda_i^m(t)}\right) exp\left(\int_t^{T^R} \lambda_j^m(s)ds\right) + \frac{z_i}{1 - z_i}\left(\frac{u_i'(1 - m_j(t))m_j'(t)}{r_i u_i(1 - \alpha_j)} + \frac{\lambda_i^m(t)}{\lambda_j^m(t)}\right) exp\left(\int_t^{T^R} \lambda_i^m(s)ds\right) = 0$$

Combining the two above equations and solving for $m_j'(t)$ gives:

$$m_j'(t) = \frac{\lambda_j^m(t) - \lambda_i^m(t)}{\frac{\lambda_i^m(t)u_j'(m_j(t))}{r_j u_j(1-\alpha_j)} + \frac{\lambda_j^m(t)u_i'(1-m_j(t))}{r_i u_i(1-\alpha_i)}} = \frac{r_j u_j(1 - \alpha_i)(u_j(m_j(t)) - u_j(1 - \alpha_i)) - r_i u_i(1 - \alpha_j)(u_i(1 - m_j(t)) - u_i(1 - \alpha_j))}{(u_i(1 - m_j(t)) - u_i(1 - \alpha_j))u_j'(m_j(t)) + (u_j(m_j(t)) - u_j(1 - \alpha_i))u_i'(1 - m_i(t))}$$

This is uniformly Lipschitz continuous in $m_j(t)$ for $t \in [0, T^R]$ (its derivative is continuous) and is continuous in t, hence by Picard's Theorem it has a unique solution. Define $\bar{m}_j$ as the unique value which solves the equality:

$$1 = \frac{u_j(\bar{m}_j) - u_j(1 - \alpha_i)}{u_i(1 - \bar{m}_j) - u_i(1 - \alpha_j)} \frac{r_i u_i(1 - \alpha_j)}{r_j u_j(1 - \alpha_i)}$$

and $\bar{\lambda}^m = \frac{r_j u_j(1-\alpha_i)}{u_j(\bar{m}_j)-u_j(1-\alpha_i)}$. When $z_i = z_j$, it is clear that we must have $m_j(t) = \bar{m}_j$ and $\lambda_j^m(t) = \lambda_i^m(t) = \bar{\lambda}^m$ for all t. More generally, it is clear that $m_j(t)$ (respectively $\lambda_j(t)$) is a convex combination of $\bar{m}_j$ and $m_j(T^R)$ (respectively $\bar{\lambda}^m$ and $\lambda_j^m(T^R)$) given that $m_j'(t) > 0$ when $\lambda_j^m(t) > \lambda_i^m(t)$ (which decreases $\frac{\lambda_j^m(t)}{\lambda_i^m(t)}$).

Let the solution be indexed by T^R, $m_j^{T^R}$, with associated agreement time distribution $G^R_{T^R}$. Clearly $m_j^{T^R}(T^R - t)$ is independent of T_R and so $G^R_{T^R}(0)$ is continuous and strictly decreasing in T^R (see equation (18)). This ensures that there is a unique value of T^R such that $G^R_{T^R}(0) = 0$, call this $T^{R,0}$. Clearly the distribution $G^R_{T^{R,0}}$ ensures that a confessing agent i obtains the payoff $u_i(1-\alpha_j)$.

Now consider the payoff of an agent i who doesn't confess. Given that $G_i^Z(t) = 0$ for $t < T^R$, conceding at $t \in (0, T^R)$ is strictly worse than conceding at 0 for a payoff of $u_i(1-\alpha_j)$. Conceding after T^R gives at most $e^{-r_i T^R} u_i(\alpha_i)$. If $T^R \geq \bar{T}^R = \max_{i\in 1,2}\left\{-\frac{1}{r_i} ln\left(\frac{u_i(1-\alpha_j)}{u_i(\alpha_i)}\right)\right\}$, therefore, agent i who doesn't confess obtains a payoff of exactly $u_i(1-\alpha_j)$.

Assume $z_i, z_j \leq 1-\varepsilon$ for any $\varepsilon > 0$ and fix u_i, r_i and α_i. Examining equation (19) it is clear that the bound $\frac{z_i}{z_j} \in \left[\frac{1}{K}, K\right]$ implies that we can uniformly bound $\lambda_j(T^R)$ and hence $\lambda_j(t)$, so that $\lambda_j(t) \in \left[\frac{1}{L}, L\right]$ for some $L \geq 1$. Given this bound, equation (18) shows that in order to have $G^R_{T^{R,0}}(0) = 0$ as $z_j \to 0$, we must have $T^{R,0} \to \infty$. And so, there exists $\bar{z} > 0$ such that if ever $z_i \leq \bar{z}$ then $T^{R,0} \geq \bar{T}^R$. This completes the proof. □

Next I consider a generalization of the model, following AG, in which agents make their demand announcements sequentially and agents can imitate multiple different behavioral types. To do this I introduce a new time 0^0 at which agent 1 makes her initial demand announcement. Agent 2 can then either immediately concede at 0^1 or announce a counterdemand. For each agent i there is a finite set of behavioral type demands C_i, where the conditional probability of a behavioral agent i being of type α_i is $\pi_i(\alpha_i)$. If a behavioral type has $\alpha_2 < 1-\alpha_1$ then she immediately concedes at 0^1. Assume that $\max C_i > 1 - \min C_j$. Let rational agent 1's demand choice be described by probability distribution μ_1 on C_1, and rational 2's choice after observing α_1, be $\mu_1^{\alpha_1}$ on $C_2 \cup Q$ where Q indicates immediate concession. Reputations after demand choices are:

$$\bar{z}_1(\alpha_1) = \frac{z_1\pi_1(\alpha_1)}{z_1\pi_1(\alpha_1) + (1-z_1)\mu_1(\alpha_1)} \qquad \bar{z}_2^{\alpha_1}(\alpha_2) = \frac{z_2\pi_2(\alpha_2)}{z_2\pi_2(\alpha_2) + (1-z_2)\mu_2^{\alpha_1}(\alpha_2)}$$

AG establish a unique equilibrium of this game with without mediation, characterized by the condition that each type a rational agent imitates must give her the same expected continuation payoff. After time zero, behavior matches the Baseline equilibrium described in the main text but with z_i replaced by $\bar{z}_i$. Let $\lambda_j^{\alpha_j,\alpha_i} = \frac{r_i u_i(1-\alpha_i)}{u_i(\alpha_i)-u_i(1-\alpha_j)}$ be the concession rate by j that would keep i indifferent to conceding after time zero without a mediator and demands α_i, α_j. AG show that as the fraction of behavioral types becomes small ($z_i \to 0$ and $\frac{z_1}{z_2} \in \left[\frac{1}{K}, K\right]$ for some $K \geq 1$) then bargaining becomes arbitrarily efficient so long as $\lambda_j^{\alpha_j,\alpha_i} \neq \lambda_i^{\alpha_i,\alpha_j}$ for each pair of incompatible demands α_i, α_j. Moreover, let $\alpha_i^R = \arg\max_{\alpha_i} u_i(\alpha_i)^{\frac{r_j}{r_i+r_j}} u_j(1-\alpha_i)^{\frac{r_i}{r_i+r_j}}$. This is the complete information alternating offers demand (Rubinstein (1982)) when the time between offers converges to zero. If agent i can imitate some type $\alpha_i' \leq \alpha_i^R$, then AG show she must obtain an equilibrium payoff greater than $u_i(\alpha_i')$. This holds because $\alpha_i' \leq \alpha_i^R$ implies $\lambda_j^{\alpha_j,\alpha_i'} < \lambda_i^{\alpha_i',\alpha_j}$ for any $\alpha_j > \alpha_i'$ so that agent i builds reputation exponentially more quickly than j. If agents imitate the demands (α_i', α_j) with positive limit probability as $z_i \to 0$, therefore, agent j must concede with probability approaching one at 0^4 to ensure that both agents reach a probability one reputation at the same time.

The next result, by contrast, shows that as the fraction of behavioral types become small, there exist equilibria with mediation which give rational agents payoffs arbitrarily close to $u_i(1-\max C_j)$. The mediation protocol used for each incompatible demand pair is the same as in Proposition 9. Demand choices can then be distorted so that both agents almost exclusively start imitating their maximum demand type. Clearly, whenever the type space is rich enough that agent i can imitate a type $\alpha_i' \in (1-\max C_j, \alpha_i^R]$, then this implies strictly lower payoff under mediation than without.

Proposition 10. *Consider a sequence of bargaining games $B^n = \{u_i, r_i, C_i, \pi_i, z_i^n\}$ such that $\lim_n z_i^n = 0$ and $\frac{z_1^n}{z_2^n} \in \left[\frac{1}{K}, K\right]$ for some constant $K \geq 1$. Then there is a sequence of equilibria with mediation such that the limit of agent i's equilibrium payoffs is $\lim_n U_i = u_i(1-\max C_j)$ for $i = 1, 2$, $i \neq j$.*

Proof. I first construct an equilibrium which will hold for all arbitrarily large n. Given any demand α_1 suppose that whenever agent 2 makes counterdemand $\alpha_2 > 1 - \alpha_1$ then agent 1's continuation payoff is $u_1(1 - \alpha_2)$. In this case agent 1's expected payoff from demanding α_1 is:

$$\underline{U}_1(\alpha_1) = u_1(\alpha_1)\left((1 - z_2)\mu^{\alpha_1}(Q) + \sum_{\alpha_2 \leq 1-\alpha_1} z_2\pi_2(\alpha_2)\right) + \sum_{\alpha_2 > 1-\alpha_1} u_1(1 - \alpha_2)\left(z_2\pi_2(\alpha_2) + (1 - z_2)\mu_2^{\alpha_1}(\alpha_2)\right)$$

Whenever $|C_1| = 1$ then let $\mu_1(\max C_1) = 1$. Otherwise, consider any $\varepsilon \in (0, 1)$ and define $\mu_1(\max C_1) = 1 - \varepsilon$ and $\mu_1(\alpha_1) = \frac{\varepsilon}{|C_1|-1}$. If $|C_2| = 1$ let $\mu_2^{\alpha_1}(\max C_2) = 1$, in this case because $\max C_2 > 1 - \min C_1$, we have that agent 1's continuation payoff is $u_1(1 - \max C_2)$ for all α_1. Suppose then $|C_2| > 1$. Let $\mu_2^{\max C_1}(\max C_2) = 1 - \varepsilon$ and $\mu_2^{\max C_1}(\alpha_2) = \frac{\varepsilon}{|C_2|-1}$ if $\alpha_2 < \max C_2$. If $\alpha_1 < \max C_1$ then $\mu_2^{\alpha_1}(\max C_2) = 1 - \varepsilon^{\alpha_1}$ and $\mu_2^{\max C_1}(\alpha_2) = \frac{\varepsilon^{\alpha_1}}{|C_2|-1}$ if $\alpha_2 \in D_2(\alpha_1) = \{\alpha_2 \in (1 - \alpha_1, \max C_2)\}$ and $\mu_2^{\max C_1}(Q) = \frac{\varepsilon^{\alpha_1}(|C_2|-1-|D_2|}{|C_2|-1}$, where ε^{α_1} is defined to ensure that $\underline{U}_1(\alpha_1) = \underline{U}_1(\max C_1)$. To check that ε^{α_1} is well defined, consider:

$$\begin{aligned}\underline{U}_1(\max C_1) - \underline{U}_1(\alpha_1) = &\sum_{\alpha_2 \leq 1-\alpha_1}\left((u_1(1 - \alpha_1) - u(\alpha_1))\, z_2\pi_2(\alpha_2) + \frac{(1 - z_2)}{|C_2| - 1}\left(\varepsilon u_1(1 - \alpha_2) - \varepsilon^{\alpha_1}u_1(\alpha_1)\right)\right)\\ &+ (\varepsilon^{\alpha_1} - \varepsilon)(1 - z_2)\left(u_1(1 - \max C_2) - \sum_{\alpha_2 > 1-\alpha_1}\frac{u_1(1 - \alpha_2)}{|C_2| - 1}\right)\end{aligned}$$

Given $1 - \max C_2 > \min C_2$, this expression is continuous and strictly increasing in ε and decreasing in ε^{α_1}. Choose $\bar{\varepsilon} > 0$ such that

$$\sum_{\alpha_2 \leq 1-\alpha_1}\frac{\bar{\varepsilon}u_1(1 - \alpha_2) - u_1(\alpha_1)}{|C_2| - 1} + (1 - \bar{\varepsilon})\left(u_1(1 - \max C_2) - \sum_{\alpha_2 > 1-\alpha_1}\frac{u_1(1 - \alpha_2)}{|C_2| - 1}\right) < 0.$$

for $\alpha_1 = \min C_2$. It is clear that for all $\varepsilon \leq \bar{\varepsilon}$, that for all $\alpha_1 \in C_1$ there must exist some $\underline{z}_2$ such that if $z_2 \leq \underline{z}_2$, then $\underline{U}_1(\max C_1) - \underline{U}_1(\alpha_1) < 0$ if $\varepsilon^{\alpha_1} = 1$, hence $\varepsilon^{\alpha_1} \geq \varepsilon$ is well defined and indeed, bounded away from 1.

Now consider the sequence of bargaining games B^n and choose N sufficiently large that $z_2^n \leq \underline{z}_2$ for all $n \geq N$. Suppose that agents play the demand choice strategies above for all such n. Given that agent 1 imitates all her types with probability bounded away from 0, and agent 2 likewise imitates all possible incompatible counterdemands with probability bounded away from 0, it is clear that $z_i^n \to 0$ implies $\bar{z}_1(\alpha_1) \to 0$ and moreover there exists $L \geq 1$ such that $\frac{\bar{z}_1(\alpha_1)}{\bar{z}_1^{\alpha_1}(\alpha_2)} \in \left[\frac{1}{L}, L\right]$ for all incompatible behavioral demand pairs. By Proposition 9, therefore, there exists some $N'_{\alpha_1,\alpha_2} \geq N$ such that if $n \geq N'_{\alpha_1,\alpha_2}$, we can find an equilibrium with mediation for the continuation game with incompatible demand pair (α_1, α_2) such that continuation payoffs for each agent i are exactly $u_i(1 - \alpha_j)$. For all $n \geq N' = \max_{\alpha,\alpha_2}\{N'_{\alpha_1,\alpha_2}\}$, by construction agent 1 is indifferent between all her demand choices, agent 2 is indifferent between all her incompatible counterdemands and conceding immediately.

The above equilibrium with mediation gives players an expected utility of at most $\lim_n U_i \leq (1-\varepsilon)u_i(1-\max C_j) + \varepsilon u_i(\max C_i)$. Given that $\varepsilon \in (0, \bar{\varepsilon}]$ was arbitrary it is clear that we can choose a sequence $\varepsilon^n \to 0$ such there is an equilibrium with mediation of the above form for all $n \geq N'$, and so i's payoff converges to $u_i(1 - \max C_j)$. □