Online Supplement for "A Bargaining Theory of Conflict with Evolutionary Preferences", *International Organization*.

Andrew T. Little* Thomas Zeitzoff†

March 24, 2017

In this document we present some formal analysis which is referenced in the published version of the paper.

Partially Observed Preferences

Suppose that when two players are matched, their preferences are observed with probability $q \in (0, 1)$, and with probability $1-q$ the actors only know the distribution of preferences. Since the game now includes incomplete information, our equilibrium requirement is now that the players use Perfect Bayesian Equilibrium strategies in the bargaining game given their preferences and consistent beliefs. (In this case, consistency only requires that proposers have a correct belief about the distribution of types)

We assume in this section that all actors' preferences are equal to the objective payoffs when in the proposer role, and in the responder role the conflict fitness is $v - k + \beta^r$. This is primarily to keep the uncertainty one-sided, and greatly reduces the number of cases to consider when solving for the equilibrium offer made and hence the optimal type given a preference distribution. Further, as shown in the online supplement, when the players preferences are allowed to vary based on role,

*Department of Government, Cornell University. andrew.little@cornell.edu.
†School of Public Affairs, American University. zeitzoff@american.edu.

there is no fitness benefit to having a $\beta > 0$ when in the proposer role, and so in any SRNE the average toughness in the proposer role is zero.

We also assume uniform noise, so if the type that gets the highest fitness is β^r_{max}, then the toughness parameters in the next generation are uniformly distributed on $[\beta^r_{\text{max}} - \epsilon^r, \beta^r_{\text{max}} + \epsilon^r]$.

When the types are observed, by standard logic the proposer (who again has preferences equal to her objective payoff, and hence gets $v - k$ for fighting) offers $v - k + \beta^r$, which is accepted if $\beta^r \leq 2k$, and makes an offer which is rejected otherwise.

When the type is unobserved with a population distributed on $[\beta^r_m - \epsilon^r, \beta^r_m + \epsilon^r]$, then proposer utility for making offer x is:

$$u^p(x; \beta^r_m) = \begin{cases} v - k & x < v - k + \beta^r_m - \epsilon^r \\ \frac{x-(v-k+\beta^r_m-\epsilon^r)}{2\epsilon^r}(2v - x) + \frac{v-k+\beta^r_m+\epsilon^r-x}{2\epsilon^r}(v - k) & x \in (v - k + \beta^r_m - \epsilon^r, v - k + \beta^r_m + \epsilon^r) \\ 2v - x & x \geq v - k + \beta^r_m + \epsilon^r \end{cases}$$

The middle segment (with respect to x) is a quadratic maximized at $v + \frac{\beta^r_m - \epsilon^r}{2}$. If this maximum lies below $v - k + \beta^r_m - \epsilon^r$, then the proposer makes an offer which is always rejected. If the maximum of the quadratic is above $v - k + \beta^r_m + \epsilon^r$, then the proposer makes this offer, which buys off all types. If the quadratic is maximized on the middle interval, the proposer makes that maximizing offer. So:

$$x^*_u = \begin{cases} v - k + \beta^r_m + \epsilon^r & \beta^r_m < 2k - 3\epsilon^r \\ v + \frac{\beta^r_m - \epsilon^r}{2} & \beta^r_m \in [2k - 3\epsilon^r, 2k + \epsilon^r] \\ v - k + \beta^r_m - \epsilon^r & \beta^r_m > 2k + \epsilon^r \end{cases}$$

If $\beta^r_m < 2k - 3\epsilon^r$, then the proposer buys off all types when the type is unobserved. This inequality also implies that the highest type is below $2k - 2\epsilon^r$, so a deal is also always reached

when the type is observed. Since a deal is always reached the highest type always gets the highest fitness, and hence the distribution is not stable. Conversely, if $\beta_m^r > 2k + \epsilon^r$, then all types fight regardless of whether the type is observed, which also violates the stability condition. So, in any stable equilibrium $\beta_m^r \in [2k - 3\epsilon^r, 2k + \epsilon^r]$ and an interior offer is made when the type is unobserved.

Next, we compute the fitness for a responder with toughness β_j^r when the average toughness is β_m^r (within the range of types in the distribution). When the type is observed, the resulting fitness is $v - k + \beta_j^r$ for $\beta_j^r \leq 2k$ and $v - k$ otherwise. When the type is unobserved, the responder accepts the offer made if and only if:

$$v + \frac{\beta_m^r - \epsilon^r}{2} \geq v - k + \beta_j^r \Longrightarrow \beta_j^r \leq k + \frac{\beta_m^r - \epsilon^r}{2}$$

So, the expected fitness for the responder is:

$$\Pi^r(\beta_j^r; \beta_m^r) = \begin{cases} q(v - k + \beta_j^r) + (1 - q)\left(v + \frac{\beta_m^r - \epsilon^r}{2}\right) & \beta_j^r \leq k + \frac{\beta_m^r - \epsilon^r}{2} \\ q(v - k + \beta_j^r) + (1 - q)(v - k) & \beta_j^r \in [k + \frac{\beta_m^r - \epsilon^r}{2}, 2k] \\ v - k & \beta_j^r > 2k \end{cases}$$

which is a piecewise linear function, increasing on the first two segments and flat on the third, and with two (downward) discontinuities. In words, the type attaining the highest fitness is either the toughest one that never fights (regardless of whether the type is observed) or the toughest type who does not fight when the type is observed (but does reject the offer when her type is unobserved). The first peak is strictly higher if and only of:

$$q\left(v - k + k + \frac{\beta_m^r - \epsilon^r}{2}\right) + (1 - q)\left(v + \frac{\beta_m^r - \epsilon^r}{2}\right) > q(v - k + 2k) + (1 - q)(v - k)$$

$$\beta_m^r > \epsilon^r + 2k(2q - 1)$$

So, for the range of β_m^r where $\beta_{\max}$ is well-defined, the next generation average toughness as a function of the current generation average is:

$$\beta_{\max}(\beta_m^r) = \begin{cases} 2k & \beta_m^r < 2k(2q-1) + \epsilon^r \\ k + \frac{\beta_m^r - \epsilon^r}{2} & \beta_m^r > 2k(2q-1) + \epsilon^r. \end{cases}$$

This function is flat in β_m^r on the first segment, then there is a downward discontinuity after which it is linearly increasing in β_m^r. So, there may be a fixed point on the first segment, a fixed point on the second segment, or no fixed point (as shown below, there is never more than one intersection). There is a fixed point where $\beta_{\max}(\beta^*) = \beta^*$ (and hence an equilibrium) at $\beta^* = 2k$ if and only if:

$$2k < 2k(2q-1) + \epsilon^r \implies q > 1 - \frac{\epsilon^r}{4k}$$

And there is a fixed point where

$$\beta^* = k + \frac{\beta^* - \epsilon^r}{2} \implies \beta^* = 2k - \epsilon^r$$

if and only if

$$2k - \epsilon^r > 2k(2q-1) + \epsilon^r \implies q < 1 - \frac{\epsilon^r}{2k}$$

So, unless $q \in [1 - \frac{\epsilon^r}{2k}, 1 - \frac{\epsilon^r}{4k}]$, there is a unique β^* such that $\beta_{\max}(\beta^*) = \beta^*$ and hence a unique SRNE.

If q lies within this range, then there is no stable population toughness. However, it is relatively straightforward to characterize a stable "cycle" of population toughness. To show this, we first extend our equilibrium definition:

Definition A strategy profile $\sigma^* = (\sigma_1^*, \sigma_2^*)$, finite sequence of types $(\beta^*(1), ..., \beta^*(l))$ such that

$\beta^*(1) > \beta^*(t)$ for $t = 2, ...l$, and noise distribution $G(\nu)$ comprise a *Cyclical Single Reproducer Noisy Equilibrium* (CSRNE) if:

(1) $(\sigma_1^*(\beta_1, \beta_2), \sigma_2^*(\beta_1, \beta_2))$ are PBE strategies of the bargaining game for all $(\beta_1, \beta_2) \in \mathbb{R}^2$.

(2) $\beta^*(t+1) = \beta_{\max}(\beta^*(t))$ for $i = 1, ..., l-1$ and $\beta^*(1) = \beta_{\max}(\beta^*(l))$ where

$$\beta_{\max}(\beta^*) = \underset{\beta_j \in \text{Supp}(G(\nu - \beta^*))}{\arg\max} \Pi(\beta_j; G(\nu - \beta^*), \sigma^*)$$

Note that a SRNE is a special case of a CSRNE where $l = 1$. The restriction that $\beta^*(1)$ is the highest type is to pin down a unique "starting place" for the cycle; without this requirement the existence of one stable cycle of length l would entail $l - 1$ other cycles with the same set of types but a different order. We can now state our main results for this extension:

Proposition 1. *In the model with incomplete informtion and uniform noise:*

(i) there exists a unique CSRNE for all but a countably infinite number of values of q (and no CSRNE for these values).

In the CSRNE:

(ii) The average toughness across the sequence of types is increasing in q,

(iii) The probability of conflict is continuous and decreasing in q almost everywhere, but:

(iv) the probability of conflict is non-monotone and strictly higher for any $q > 1 - \frac{\epsilon^r}{4k}$ than any $q < 1 - \frac{\epsilon^r}{2k}$

Proof We first construct an algorithm which generates a CSRNE with these properties here, and then demonstrate uniqueness.

For $q < 1 - \frac{\epsilon^r}{2k}$ and $q > 1 - \frac{\epsilon^r}{4k}$ we have already demonstrated the existence of a CSRNE with $l = 1$. So, suppose $q \in (1 - \frac{\epsilon^r}{2k}, 1 - \frac{\epsilon^r}{4k})$.

Let $\beta^*(1) = 2k$, and let the second generation average toughness be the best response to this

toughness level:

$$\beta^*(2) = \beta_{\max}(2k) = 2k - \epsilon^r/2$$

For the third generation, if:

$$2k - \epsilon^r/2 < \epsilon^r + 2k(2q-1) \Longrightarrow q > 1 - \frac{3\epsilon^r}{8k} \in (1 - \frac{\epsilon^r}{2k}, 1 - \frac{\epsilon^r}{4k})$$

then $\beta_{\max}(2k - \epsilon^r/2) = 2k$, and hence $(\beta^*(1), \beta^*(2)) = (2k, 2k - \epsilon^r/2)$ constitute the preference cycle for a CSRNE for this range of q.

If $q = 1 - \frac{3\epsilon^r}{8}$, then then $\beta_{\max}$ is not well defined, so our algorithm does not identify a CSRNE.

If $q < 1 - \frac{3\epsilon^r}{8}$, then let $\beta^*(3) = \beta_{\max}(2k - \epsilon^r/2) = k + \frac{2k - \epsilon^r/2 - \epsilon^r}{2} = 2k - 3\epsilon^r/4$.

Generally, let $\beta^*(t) = 2k - (1 - 2^{1-t})\epsilon^r$. If the current generation is centered at $\beta^*(t)$, then

$$\beta_{\max}(\beta^*(t)) = \begin{cases} 2k & q > 1 - \epsilon^r \frac{2+2^{1-t}}{4} \\ \beta^*(t+1) & q < 1 - \epsilon^r \frac{2+2^{1-t}}{4}, \end{cases}$$

and is undefined if $q = 1 - \epsilon^r \frac{2+2^{1-t}}{4k}$.

Rearranging the threshold determining whether the next generation resets to $2k$ gives:

$$2^{1-t} > \frac{2(\epsilon^r - 2k(1-q))}{\epsilon^r}$$

Since 2^{1-t} starts at 1 for $t = 1$ and converges to 0 as $t \to \infty$, there will be a smallest integer l where the inequality holds if and only if the right-hand side of this equation is between 0 and 1, which is true exactly when $q \in (1 - \frac{\epsilon^r}{2k}, 1 - \frac{\epsilon^r}{4k})$.

In particular so the cycle "resets" to $2k$ at the smallest integer where the inequality is met, or:

$$l = 1 - \lfloor \log_2 \left(\frac{2(\epsilon^r - 2k(1-q))}{\epsilon^r} \right) \rfloor \tag{1}$$

So, as long as:

$$q \notin \{q : 2^{1-t} = \frac{2(\epsilon^r - 2k(1-q))}{\epsilon^r}, t = 1, 2, ...\},$$

then $\beta^*(1), ..., \beta^*(l)$ for $t = 1, ...l$ constitutes the preferences for a CSRNE. If q is in this set, then the sequence starting at $2k$ will eventually lead to a generation where β_{max} is not well-defined. This set is countable, completing part i.

For part ii, the average toughness in a cycle of length l is:

$$l^{-1} \sum_{t=1}^{l} (2k - (1 - 2^{1-t})\epsilon^r) = 2k - l^{-1} \sum_{t=1}^{l} (1 - 2^{1-t})\epsilon^r$$

which is strictly greater than $2k - \epsilon^r$, less than $2k$, and decreasing in l. Over the range $(1 - \frac{\epsilon^r}{2k}, 1 - \frac{\epsilon^r}{4k})$, l is decreasing in q. So average toughness is increasing in q.

For parts iii-iv, first consider the probability of conflict for an arbitrary β_m^r. When the type is observed, then conflict never occurs if $\beta_m^r \leq 2k - \epsilon^r$, always occurs if $\beta_m^r \geq 2k + \epsilon^r$, and happens when $\beta^r > 2k$ for the intermediate range:

$$p_c(\beta_m^r; \text{type observed}) = \begin{cases} 0 & \beta_m^r < 2k - \epsilon^r \\ \frac{\beta_m^r + \epsilon^r - 2k}{2\epsilon^r} & \beta_{\in}^r [2k - \epsilon^r, 2k + \epsilon^r] \\ 1 & \beta_m^r > 2k + \epsilon^r \end{cases}$$

When the type is unobserved, the equilibrium offer in the relevant range is $v + \frac{\beta_m^r - \epsilon^r}{2}$ which is

accepted if $\beta_j^r \leq k + \frac{\beta_m^r - \epsilon^r}{2}$, which occurs with probability:

$$\frac{\beta_m^r + \epsilon^r - (k + \frac{\beta_m^r - \epsilon^r}{2})}{2\epsilon^r} = \frac{\beta_m + 3\epsilon^r - 2k}{4\epsilon^r}$$

Combining, the average probability of conflict is:

$$p_c(\beta_m^r) = \begin{cases} 0 & \beta_m^r < 2k - 3\epsilon^r \\ (1-q)\frac{\beta_m + 3\epsilon^r - 2k}{4\epsilon^r} & \beta_m^r \in [2k - 3\epsilon^r, 2k - \epsilon^r) \\ q\frac{\beta_m^r + \epsilon^r - 2k}{2\epsilon^r} + (1-q)\frac{\beta_m + 3\epsilon^r - 2k}{4\epsilon^r} & \beta_m^r \in [2k - \epsilon^r, 2k + \epsilon^r) \\ 1 & \beta_m^r \geq 2k + \epsilon^r \end{cases}$$

So, when $q < 1 - \frac{\epsilon^r}{2k}$ and $\beta^* = 2k - \epsilon^r$, the probability of conflict is:

$$(1-q)\frac{2k - \epsilon^r + 3\epsilon^r - 2k}{4\epsilon^r} = (1-q)/2$$

When $q > 1 - \frac{\epsilon^r}{4k}$ and $\beta^* = 2k$, the probability of conflict is:

$$q\frac{2k + \epsilon^r - 2k}{2\epsilon^r} + (1-q)\frac{2k + 3\epsilon^r - 2k}{4\epsilon^r} = 3/4 - q/4$$

Finally, in the CSNRE for the intermediate range, β_m^r is always between $2k - \epsilon^r$ and $2k$, and the probability of conflict is linear on this segment, so we can write the average probability of conflict across generations as:

$$q\frac{\mathbb{E}[\beta_m^r] + \epsilon^r - 2k}{2\epsilon^r} + (1-q)\frac{\mathbb{E}[\beta_m^r] + 3\epsilon^r - 2k}{4\epsilon^r} \tag{2}$$

where $\mathbb{E}[\beta_m^r]$ is the average of the center of the distribution over the cycle derived in part ii.

Summarizing, the (average) probability of conflict as a function of q is:

$$p_c(\beta^*) = \begin{cases} (1-q)/2 & q < 1 - \frac{\epsilon^r}{2k} \\ \frac{(3-q)\epsilon^r + (1-q)l^{-1}\sum_{t=1}^{l}(1-2^{1-t})}{4\epsilon^r} & q \in (1 - \frac{\epsilon^r}{2k} 1 - \frac{\epsilon^r}{4k}) \\ 3/4 - q/4 & q > 1 - \frac{\epsilon^r}{4k} \end{cases}$$

where l is given by equation **??**.

This is locally decreasing on every segment (part iii). However, it is strictly less than $1/2$ on the first segment and strictly greater than $1/2$ on the last segment, proving part iv.

To prove uniqueness, recall a CSRNE requires that iteratively applying the best response function gives $\beta_{\max}(\beta^*(t)) = \beta^*(t+1)$ for $t = 1, ..., l-1$ and $\beta_{\max}(\beta^*(l)) = \beta^*(1)$. So, to see if there is a cycle starting at $\beta^*(1)$, we check if iteratively applying the best response function "leads back" to $\beta^*(1)$ for some $\beta^*(t)$. There are three relevant cases:

Case 1: $q > 1 - \frac{\epsilon^r}{4k}$. In this range, the best response function has a unique fixed point at $2k$. The idea of the proof is that for any $\beta^*(t)$, either $\beta_{\max}(\beta^*(t)) = 2k$ or $\beta_{\max}(\beta^*(t)) < \beta^*(t)$. So, iteratively applying the best response function either leads to a perpetually decreasing sequence (which can not be a cycle) or it must eventually hit $2k$ at which point it becomes "stuck" (i.e., $\beta_{\max}(\beta^*(t')) = 2k$ for all $t' > t$). By the best response function, $\beta_{\max}(\beta^*(t)) = 2k$ if $\beta^*(t) < 2k(2q-1)+\epsilon^r$, so what remains to be shown is that if $\beta^*(t) > 2k(2q-1)+\epsilon^r$, then $\beta_{\max}(\beta^*(t+1)) < \beta^*(t)$. Over the relevant range, this is equivalent to:

$$\beta^*(t) > k + \frac{\beta^*(t) - \epsilon^r}{2} \Longrightarrow \beta^*(t) > 2k - \epsilon^r \tag{3}$$

And since $\beta^*(t) > 2k(2q-1) + \epsilon^r$ and $q > 1 - \frac{\epsilon^r}{4k}$, we know that:

$$\beta^*(t) > 2k(2(1 - \frac{\epsilon^r}{4k}) - 1) + \epsilon^r = 2k - \epsilon^r + \epsilon^r = 2k$$

and so equation 1 holds too. So, the only CSRNE is of length one with $\beta^*(1) = 2k$.

Case 2: $q < 1 - \frac{\epsilon^r}{2k}$. In this range, the best resposne function has a unique fixed point at $2k - \epsilon^r$. If $\beta^*(1) < 2k - \epsilon^r$, then $\beta_{\max}(\beta^*(1)) > \beta^*(1)$, as either $\beta_{\max}(\beta^*) = 2k$ or the converse of equation 1 holds. This violates the condition that $\beta^*(1)$ is the highest part of the sequence. If $\beta^*(1) > 2k - \epsilon^r$, then $\beta_{\max}(\beta^*(1)) = k + \frac{\beta^*(1) - \epsilon^r}{2} \in (2k - \epsilon^r, \beta^*(1))$. More generally, if $\beta^*(t) > 2k - \epsilon^r$, then $\beta_{\max}(\beta^*(t)) \in (2k - \epsilon^r, \beta^*(t))$. So, for any sequence that starts above $2k - \epsilon^r$, subsequent generations have decreasing toughness which approaches but never reaches $2k - \epsilon^r$, contradicting the length of the cycle being finite. So, the only CSNRE has $\beta^*(1) = 2k - \epsilon^r$, which is the fixed point of $\beta_{\max}$, and hence the cycle has length one.

Case 3: $q \in (1 - \frac{\epsilon^r}{2k}, 1 - \frac{\epsilon^r}{4k})$. For the intermediate range, there is no fixed point to the best response function. If $\beta^*(1) \leq 2k(2q-1) + \epsilon^r$, then $\beta_{\max}(\beta^*(1)) = 2k$, which is above $\beta^*(1)$ for this range. If not then the largest possible value of $\beta^*(1)$ where $\beta_{\max}$ is well defined is $\beta^*(1) = 2k + \epsilon^r$. Then:

$$\beta_{\max}(\beta^*(1)) = k + \frac{\beta_m^r - \epsilon^r}{2} \leq 2k$$

Further, as characterized above, the toughness level will continue to decrease in every generation until $\beta_{\max}(\beta^*(t)) = 2k$. If $\beta^*(1) < 2k$, then this violates $\beta^*(1)$ being the highest toughness level. If $\beta^*(1) > 2k$, then by the same argument as in the first case repeatedly applying the best response function either leads to decreasing values of $\beta^*(t)$ or jumping back to $2k$, in which case the cycle identified in the main text starts, and hence $\beta^*(t) > 2k$ is never reached again. So, the cycle identified in the main text is unique in this range as well. ∎

Comment on Values of q with no CSRNE

The values of q where the CSRNE is not defined reach problems because the best response function has two peaks, and we avoid defining what should happen in the next generation in these

cases. However, it is easy to see that if we modify our definition of a CSRNE to "select" either the first or second peak, there is a unique stable cycle. In particular, if we define:

$$\beta_{\max}(\beta^*) = \min_{\beta_j}\{\underset{\beta_j \in \text{Supp}(G(\nu-\beta^*))}{\arg\max} \Pi(\beta_j; G(\nu-\beta^*), \sigma^*)\}$$

then in the uniform case

$$\beta_{\max}(\beta^*(t)) = \begin{cases} 2k & \beta^*(t) < 2k(2q-1) + \epsilon^r \\ k + \frac{\beta^*(t)-\epsilon^r}{2} & \beta^*(t) \geq 2k(2q-1) + \epsilon^r. \end{cases}$$

where now the second case applies at the boundary case. The same algorithm now defines a sequence where in the last element of the cycle $\beta^*(l) = 2k(2q-1) + \epsilon^r$.

Static Solution Concept

The models in the paper use an evolutionary process where each generation's preferences which allows for noise in the evolutionary process. A more "standard" evolutionary concept uses a more static concept of preferences being stable (as in, e.g., Dekel, Ely and Yilankaya, 2007; Huck and Oechssler, 1999). The main result using the solution concept is indeterminate: there is a wide range of equilibria where any probability of conflict is possible. So, one way to think about the noisy solution concept used in the paper is as an equilibrium selection mechanism to break the indeterminacy found here (and allow for comparative static predictions).

Formal Definition of the Static Solution Concept

As in the paper, let $\pi(\beta_j; \beta_{-j}, \sigma)$ be the expected fitness for a player of toughness β_j when matched with a partner with toughness β_{-j} and they use strategies $\sigma = (\sigma_1, \sigma_2)$ when in the subscripted role in the bargaining game.

Let the β's follow a distribution F. Then the expected fitness for a player with toughness level β is:

$$\Pi(\beta; F, \sigma) = \int \pi(\beta; \beta_{-j}, \sigma) dF(\beta_{-j})$$

Our static equilibrium concept is:

Definition A strategy profile $\sigma^* = (\sigma_1^*, \sigma_2^*)$ and preference distribution F comprise a *Stable Preferences Subgame Perfect Equilibrium* (SP-SPE) if:
(1) $(\sigma_1^*(\beta_1, \beta_2), \sigma_2^*(\beta_1, \beta_2))$ is a SPNE of the bargaining game for all $(\beta_1, \beta_2) \in \mathbb{R}^2$.
(2) supp$(F) \in \arg\max_\beta \Pi(\beta; F, \sigma^*)$

The first part states that the outcome of the bargaining game is a SPNE given the preferences of the players. The second part states that all types in the support of the distribution (i.e., with positive probability or density) get the highest possible fitness when playing strategies meeting the first condition.

We first show that in an SP-SPE the preferences always make the actors more willing to fight than their preferences dictate (this is always true on average for the model in the paper, but given the noise in the evolutionary process it is possible to have some actors with $\beta_j < 0$):

Proposition 2. *In any SP-SPE,* $Pr(\beta_j \leq 0) = 0$

Proof The intuition behind the result is to divide the pool of types into those with a toughness less than or equal to $2k$ and those with a toughness strictly greater than $2k$. There is no incentive to have a toughness less than 0 against the first group because one can always become tougher and get better deals without fighting. There is also no incentive to have toughness less than zero against the latter group because any deal than can be struck with them is worse than fighting. So, it is always strictly better to be type $\beta' = 0$ than any $\beta' < 0$.

Formally, write the fitness for being type β' given type distribution F as:

$$\Pi(\beta', F) = \int_{-\infty}^{2k} \left(\mathbf{1}_{\beta'+\beta_{-j}\leq 2k} \left(v + \frac{\beta' - \beta_{-j}}{2} \right) + \mathbf{1}_{\beta'+\beta_{-j}>2k}(v-k) \right) dF(\beta_{-j}) + \int_{2k}^{\infty} \left(\mathbf{1}_{\beta'+\beta_{-j}\leq 2k} \left(v + \frac{\beta' - \beta_{-j}}{2} \right) + \mathbf{1}_{\beta'+\beta_{-j}>2k}(v-k) \right) dF(\beta_{-j}).$$

I.e., the first integral captures the fitness from being matched with a $\beta_{-j} \leq 2k$ type and the second being matched with a $\beta_{-j} > 2k$ type. There are two cases to consider:

i. If $F(2k) = 1$ (i.e., all types are less than $2k$), then the second integral drops out and $\beta' + \beta_{-j} \leq 2k$ for $\beta' < 0$, hence the fitness is strictly increasing for $\beta' < 0$.

ii. If $F(2k) < 1$, then $Pr(\beta_{-j} > 2k) > 0$. For the range $\beta' < 0$, $\beta' + \beta_{-j} \leq 2k$ for all β_{-j} corresponding to the first integral, so the fitness is strictly increasing in β' in this range (i.e., one gets better deals without fighting this group by getting tougher). If $\beta' < 0$ and $\beta_{-j} > 2k$, then $v + \frac{\beta' - \beta_{-j}}{2} < v - k$, so increasing β' can only lead to more fighting among this group, but fighting gives a higher fitness than striking a deal, so the fitness when matched against a $\beta_{-j} > 2k$ is weakly increasing for $\beta' < 0$.

So, in either case the fitness is strictly increasing for $\beta' < 0$ violating the condition for a SP-SPE which places positive probability on $\beta_j < 0$. ∎.

Before focusing on more interesting cases, we first note that there is a large class of equilibria where conflict always happens:

Proposition 3. *There is a SP-SPE with preferences given by any F such that $Pr(\beta_j > 2k) = 1$*

Proof As shown above, when faced with a type with $\beta_j > 2k$, the fitness from striking a deal is strictly less than $v - k$. So if all of the population is sufficiently tough, the highest possible fitness is from conflict, and all types fight in all interactions.

Recall the equilibria analogous to these in the noisy evolution model are ruled out by assumption that the maximizing type must be a global maximizer. So, hereafter we focus on equilibria

were $Pr(\beta_j > 2k) < 1$

Characterization of Two-Type Equilibrium

We next show that even when restricting to equilibria with only two types, there is always a continuum of SP-SPE where any probability of conflict is possible.

Consider a two type distribution with toughness levels β_l and $\beta_h > \beta_l$, where the probability of being a β_h type is p_h. By proposition 1, $0 < \beta_l$. Proposition **??** implies that there is always a SP-SPE where conflict always occurs if $2k < \beta_l$, so we restrict attention to the case where $\beta_l < 2k$.

The expected fitness to being type β given σ^* is:

$$\Pi(\beta; \beta_l, \beta_h, p_h, \sigma^*) \equiv p_h \pi(\beta; \beta_h, \sigma^*) + (1 - p_h)\pi(\beta; \beta_l, \sigma^*)$$

For the triple (β_l, β_h, p_h) to be a part of a SP-SPE, it must be the case that actors with toughness parameters β_l and β_h get same fitness as each other, and that no "invader" with a different toughness parameter would get a higher fitness. So, the equilibrium condition can be written:

$$\Pi(\beta_l; \beta_l, \beta_h, p_h, \sigma^*) = \Pi(\beta_h; \beta_l, \beta_h, p_h, \sigma^*) \geq \Pi(\beta'; \beta_l, \beta_h, p_h, \sigma^*) \tag{4}$$

for any $\beta' \in \mathbb{R}$.

Given the equilibrium strategies derived in lemma **??**, a type $\beta' \leq 2k - \beta_h$ strikes a deal with both high and low types, giving fitness $v + \frac{\beta' - \beta_{-j}}{2}$. Types with $\beta' \in (2k - \beta_h, 2k - \beta_l]$ strike a deal with the low types (fitness $v + \frac{\beta' - \beta_l}{2}$) and fight the high types (fitness $v - k$). Types with $\beta' > 2k - \beta_l$ fight everyone, giving fitness $v - k$. So the expected fitness for type β' given a

two-type distribution is:

$$\Pi(\beta';\beta_l,\beta_h,p_h,\sigma^*)=\begin{cases}(1-p_h)\left(v+\frac{\beta'-\beta_l}{2}\right)+p_h\left(v+\frac{\beta'-\beta_h}{2}\right) & \beta'\leq 2k-\beta_h\\ (1-p_h)\left(v+\frac{\beta'-\beta_l}{2}\right)+p_h\left(v-k\right) & \beta'\in(2k-\beta_h,2k-\beta_l]\\ v-k & \beta'>2k-\beta_l\end{cases}$$

This is a piecewise linear function with discontinuities at $\beta' = 2k - \beta_h$ and $\beta' = 2k - \beta_l$. Since $\beta_l < 2k$, both segments are strictly increasing,since tougher types get a higher fitness due to getting better deals when in the responder role without leading to more conflict. The third segment is flat. So, in order for β_l and β_h to be maximizers, these discontinuities must occur at β_l and β_h. For the first discontinuity to lie at β_l it must be the case that $\beta_l = 2k-\beta_h$, and the condition for the second discontinuity to be at β_h is $\beta_h = 2k - \beta_l$, so both hold if $\beta_l + \beta_h = 2k$. This also implies that $\beta_l \in [0, k]$ and $\beta_h \in [0, k]$.

The condition that fitness is equal at β_l and β when $\beta_l + \beta_j = 2k$ becomes:

$$\begin{aligned}&(1-p_h)\left(v+\frac{2k-\beta_h-\beta_l}{2}\right)+p_h\left(v+\frac{2k-\beta_h-\beta_h}{2}\right)\\ &=(1-p_h)\left(v+\frac{2k-\beta_l-\beta_l}{2}\right)+p_h\left(v-k\right)\end{aligned}$$

which simplifies to

$$(1-p_h)v+p_h(v-\delta)=(1-p_h)(v+\delta)+p_h(v-k)\Longrightarrow p_h=\delta/k$$

where $\delta = k - \beta_l$. Since δ ranges from 0 to k, p_h can take on any value between 0 and 1. The probability of conflict is p_h^2, so this can take any value between 0 and 1 as well.

Other Classes of Static Equilibria

The definition of an SP-SPE allows for any distribution for the β's. The analysis above characterizes all SP-SPE with two types. To demonstrate that our central results are not sensitive to this restriction, we demonstrate that (1) for any finite n, there is a class of SP-SPE with properties similar to the two-type equilibrium, and (2) there is no SP-SPE that admits a density where a deal is struck with positive probability.

Proposition 4. *For any finite integer m,*

i. there exists a class of SP-SPE with m distinct types, and

ii. in this class of equilibria conflict can occur with with any probability between zero and one.

As in the two type case, there is always an equilibrium where all types have toughness greater than $2k$ and always fight. So, what remains is to show that there are equilibria with an interior probability of conflict.

We prove the result for even numbers of types, the proof for an odd number of types is the same with one difference highlighted in the proof. so the number of types can be written $m = 2n$ for some integer n. Order the types such that $\beta_1 < \beta_2 < ...\beta_{2n}$, and let $Pr(\beta_j = \beta_i) = p_i$. The fitness for being type β' in such an equilibrium is:

$$Pr(\beta_{-j} \leq 2k - \beta') \left(v + \frac{\beta' - \mathbb{E}[\beta_{-j} | \beta_{-j} \leq 2k - \beta']}{2} \right) + (1 - Pr(\beta_{-j} \leq 2k - \beta'))(v - k)$$

As long as $Pr(\beta_j < 2k) < 1$, the fitness for being $\beta' = 2k - \beta_1$ is strictly higher than the fitness to being $\beta' > 2k$, so $\beta_{2n} < 2k$. By proposition 1, $\beta_j \geq 0$, so this is a piecewise linear function that is weakly increasing on each segment (and strictly increasing for $\beta' < 2k$, with discontinuities at $2k - \beta_j$ for all $j \in \{1, ..., 2n\}$. So for each type to be at a local maximum, a more general symmetry condition must hold:

Lemma 1. *In any finite even type distribution with $2n$ types, it must be the case that $0 < \beta_1 <$*

$...\beta_n < k$ *and* $\beta_i = 2k - \beta_{2n+1-i}$.

Proof If not, the fitness must be strictly increasing at some $\beta' = \beta_j$, violating the condition for a SP-SPE. If this condition holds, all β_j are at a local maxima.

The analogous symmetry condition with an odd number of types is that the median type is at exactly $\beta_j = k$ and the (even) remainder of types are symmetrically aligned as in the even case.

Next we need to derive a condition for when all of the types attain the same fitness, which will guarantee all are at a global maximizer of the fitness function. Given the symmetry restriction, type j strikes a deal with types $1, ...2n - j + 1$ and fights the rest, giving fitness:

$$\begin{aligned}\Pi(\beta_j) &= P_{2n-j+1}\left(v + \frac{\beta_j - \overline{\beta}_j}{2}\right) + (1 - P_{2n-j+1})(v - k) \\ &= v - k + P_{2n-j+1}\left(k + \frac{\beta_j - \overline{\beta}_j}{2}\right)\end{aligned}$$

where

$$\begin{aligned}P_j &= \sum_{i=1}^{j} p_i \\ \overline{\beta}_j &= \frac{\sum_{i=1}^{2n-j+1} p_i \beta_i}{\sum_{i=1}^{2n-j+1} p_i}\end{aligned}$$

Next we show that for any $\beta_1, ...\beta_n$ meeting the symmetry condition, there exists a SP-SPE with a probability distribution where $\beta_i = 2k - \beta_{2n+1-i}$. Let $Pr(\beta_i = \beta) = p_i$. Adjacent types

getting the same fitness requires:

$$P_{2n-j+1}\left(k+\frac{\beta_j-\overline{\beta}_j}{2}\right)=P_{2n-j}\left(k+\frac{\beta_{j+1}-\overline{\beta}_{j+1}}{2}\right)$$

$$p_{2n-j+1}2k=P_{2n-j}\beta_{j+1}-P_{2n-j+1}\beta_j+\left(\sum_{i=1}^{2n-j+1}p_i\beta_i-\sum_{i=1}^{2n-j}p_i\beta_i\right)$$

$$p_{2n-j+1}2k=(P_{2n-j+1}-p_{2n-j+1})\beta_{j+1}-P_{2n-j+1}\beta_j+(\beta_{2n-j+1}p_{2n-j+1})$$

$$p_{2n-j+1}2k=P_{2n-j+1}(\beta_{j+1}-\beta_j)+p_{2n-j+1}(\beta_{2n-j+1}-\beta_{j+1})$$

$$p_{2n-j+1}(2k-\beta_{2n-j+1}+\beta_{j+1})=P_{2n-j+1}(\beta_{j+1}-\beta_j)$$

$$p_{2n-j+1}=P_{2n-j+1}\frac{\beta_{j+1}-\beta_j}{\beta_{j+1}+\beta_j}$$

for $j=1,...,2n-1$. Setting $j=1$ gives:

$$p_{2n}=\frac{\beta_2-\beta_1}{\beta_1+\beta_2}$$

since $P_{2n}=1$. Given this, setting $j=2$ gives:

$$p_{2n-1}=P_{2n-1}\frac{\beta_{j+1}-\beta_j}{\beta_{j+1}+\beta_j}=(1-p_{2n})\frac{\beta_{j+1}-\beta_j}{\beta_{j+1}+\beta_j}$$

where p_{2n} is given above. More generally:

$$p_{2n-j+1}=\left(1-\sum_{i=2n-j}^{2n}p_i\right)\frac{\beta_{j+1}-\beta_j}{\beta_{j+1}+\beta_j} \tag{5}$$

This gives a recursive definition for the $p_2,...p_n$'s, which are all strictly positive. So, as long as $\sum_{j=2}^{2n}p_j\leq 1$, setting $p_1=1-\sum_{j=2}^{2n}p_j$ makes the p_i's a proper probability distribution. The

$j = 2n - 1$ adjacency condition gives:

$$p_2 = \left(1 - \sum_{j=3}^{2n} p_i\right) \frac{\beta_{2n} - \beta_{2n-1}}{\beta_{2n-1} + \beta_{2n}}$$

$$\frac{\beta_{2n} - \beta_{2n-1}}{\beta_{2n-1} + \beta_{2n}} p_2 \leq \left(1 - \sum_{j=3}^{2n} p_i\right) \frac{\beta_{2n} - \beta_{2n-1}}{\beta_{2n-1} + \beta_{2n}}$$

$$\sum_{i=2}^{2n} p_i \leq 1$$

So, there is a unique p_i which is a probability distribution that meets the adjacency conditions. This completes part i.

For part ii, the probability of conflict is given by the probability that the type indices sum to something strictly greater than $2n$:

$$p_c = \sum_{i=2}^{2n} \sum_{j=2n+i-j}^{2n} p_i p_j$$

Since $\beta_{n+1}, ...\beta_{2n}$ are uniquely determined by $\beta_1, ...\beta_n$ by the symmetry condition and the p_i's are recursively defined by the β_j's and continuous in each β_j, this can be written as $p_c(\beta_1, ..., \beta_n)$. Further, as $\beta_i \to k$ from below for $i = 1, ..., n$, which implies $\beta_i \to k$ from above for $i = n+1, ...2n$, then $p_i \to 0$ for all $i > 1$, hence $p_1 \to 1$ and $p_c \to 0$. Similarly, if $\beta_i \to 0$ for $i = 1, ..n$ and hence $\beta_i \to 2k$, for $i = n+1, ..., 2n$, then $p_{n+1} \to 1$ and $p_c \to 1$. So, by the continuity of p_c in the β's, for any $p > 0$ there exists a $\beta_1, ..., \beta_n$ and hence distribution of preferences such that $p_c = p$ by the Intermediate Value Theorem, proving the result for even m.

Again, the logic of the proof is similar for odd m, with the restriction that $\beta_{(m+1)/2} = k$ and the other $\beta_i = \beta_{2n+1-i}$. ∎

Finally, we show that there is no SP-SPE that admits a density.

Suppose the type distribution admits a density f. Let $\underline{\beta} \in -\infty \cup \mathbb{R}$ be the lowest β in the support of f and $\overline{\beta} \in \infty \cup \mathbb{R}$ the highest. There is a class of SP-SPE where $\underline{\beta} > 2k$, so suppose

this is not the case. The expected fitness for being type $\beta_j \in [\underline{\beta}, \overline{\beta}]$ is then:

$$\Pi(\beta_j; \underline{\beta}, \overline{\beta}, \sigma^*) = \begin{cases} \int_{\underline{\beta}}^{2k-\beta_j} \left(\frac{\beta_j+\beta_{-j}}{2}\right) f(\beta_{-j}) d\beta_{-j} + Pr(\beta_j + \beta_{-j} > 2k)\,(v-k) & \beta_j + \underline{\beta} < 2k \\ (v-k) & \text{otherwise} \end{cases}$$

Proposition 5. *There is no SP-SPE such that $\underline{\beta} < 2k$ where the type distribution admits a density.*

Proof Suppose not, let the density be f. This implies that $\Pi(\beta_j; f, \sigma^*)$ is continuous in β_j. We derive two contradictory inequalities:

i. It must be the case that $\overline{\beta} < 2k - \underline{\beta}$. If not, $\Pi(\overline{\beta}; f, \sigma^*) = v - k$. Consider a type $\beta_j = 2k - \underline{\beta} - \epsilon$. This type will strike a bargain with types $[\underline{\beta}, \underline{\beta} + \epsilon]$, and since $\underline{\beta} < 2k$ the expected fitness from these bargains will be greater than $v - k$. So for some $\epsilon > 0$, $\Pi(2k - \underline{\beta} - \epsilon; f, \sigma^*) > \Pi(\overline{\beta}; f, \sigma^*)$.

ii It must be the case that $\underline{\beta} > 2k - \overline{\beta}$. If not, types $\underline{\beta}$ never fight, and for small $\epsilon > 0$ a type $\beta_j = \underline{\beta} + \epsilon$ would never fight, and get a strictly higher fitness than $\underline{\beta}$.

So $\underline{\beta} + \overline{\beta} < 2k$ and $\underline{\beta} + \overline{\beta} > 2k$, a contradiction ∎

In words, if the aggregate toughness of the lowest and highest types is too high, a type slightly less tough than the toughest type would get a strictly higher fitness than the toughest type. If the aggregate toughness of the lowest and highest types is too low, a type slightly tougher than the least tough could always extract a better bargain than the lowest type.

More General Preferences

In principle, preferences could diverge from fitness for any outcome of the bargaining game. To explore how our results are sensitive to alternative specifications of how preferences evolve, we first allow the toughness to vary based on whether the actor is in the proposer or responder role.

Role-based Toughness

For the model where toughness can vary based on whether in the proposer or responder role, suppose that if a type $\beta_{\max} = (\beta^p_{\max}, \beta^r_{\max})$ gets the highest fitness, then the toughness parameters for each actor in the subsequent round are given by $(\beta^p_{\max} + \nu^p_i, \beta^r_{\max} + \nu^r_i)$ where ν^p_i and ν^r_i are independent and uniformly distributed on $[-\epsilon^p, \epsilon^p]$ and $[-\epsilon^r, \epsilon^r]$, respectively, where $\epsilon^p, \epsilon^r > 0$.

Following a similar analysis as the main uniform model, the fitness for being type β_j when matched with a population with proposer fitness uniform on $[\beta^p_m - \epsilon^p, \beta^p_m + \epsilon^p]$ and responder fitness uniform on $[\beta^r_m - \epsilon^r, \beta^p_m + \epsilon^r]$ is:

$$\Pi(\beta^p_j, \beta^r_j; F; \sigma) = \frac{1}{2}\Pi^p(\beta^p_j; F; \sigma) + \frac{1}{2}\Pi^r(\beta^r_j; F; \sigma)$$

where Π^j is the expected fitness in role j. When in the proposer role the only relevant part of the distribution is the responder toughness and when in the responder role the only relevant part of the distribution is the proposer toughness. So, finding the optimal type can be separated into finding the optimal type in each role.

For the proposer role, it is more straightforward to first consider the fitness for fixed types (when using SPNE strategies):

$$\pi^p(\beta^p_j, \beta^r_{-j}; \sigma^*) = \begin{cases} v + k - \beta^r_{-j} & \beta^p_j + \beta^r_{-j} \leq 2k \\ v - k & \beta^p_j + \beta^r_{-j} > 2k \end{cases}$$

The fitness for making a deal (the first segment) is higher if $v + k - \beta^r_{-j} < v - k$, or $\beta^r_{-j} < 2k$. So if $\beta^r_{-j} < 2k$ the optimal proposer toughness is any β^p_j such that $\beta^p_j < 2k - \beta^r_{-j}$, and if $\beta_{-j} > 2k$ then the optimal proposer toughness is any β^j_p such that $\beta^p_j > 2k - \beta^r_{-j}$. So, if all responder types are greater than $2k$ any type such that $\beta^p_j > 0$ gets the highest possible fitness, and hence there can be no stable distribution using an argument analogous to that in the main model. If all responders

have toughness less than $2k$ then any type such that $\beta_j^p < 0$ gets the highest possible expected fitness, so there can be no stable type distribution of this form either.

So in any stable distribution, there must be some responders with $\beta_{-j}^r > 2k$ and some with $\beta_{-j}^r < 2k$. The only proposer type that gets the highest fitness when matched with an individual with this distribution is $\beta_j^p = 0$. This is because types with $\beta_j^p < 0$ will strike a deal with some responders with $\beta_{-j}^r > 2k$, which gives a lower fitness than fighting, and types with $\beta_j^p > 0$ fight against some types with $\beta\_j^r < 2k$, which gives lower fitness than striking a deal.

So, this can only have a unique maximum at $\beta_j^p = 0$, and hence in any stable preference distribution $\beta^{p,*} = 0$. That is, there is never a benefit to having preferences that deviate from fitness *when in the proposer role*.

This also implies that the expected fitness for having toughness β_j^r in the responder role in any stable preference distribution is:

$$\Pi^r(\beta_j^r; F^p; \sigma^*) = Pr(\beta_{-j}^p \leq 2k - \beta_j^r)(v - k + \beta_j^r) + Pr(\beta_{-j}^p > 2k - \beta_j^r)(v - k)$$

$$= \begin{cases} v - k + \beta_j^r & \beta_j^r \leq 2k - \epsilon^p \\ \frac{2k+\epsilon^p-\beta_j^r}{2\epsilon^p}(v - k + \beta_j^r) + \frac{\beta_j^r-2k+\epsilon^p}{2\epsilon^p}(v - k) & \beta_j^r \in (2k - \epsilon^p, 2k + \epsilon^p] \\ v - k & \beta_j^r > 2k + \epsilon^p \end{cases}$$

Again, the first segment is linear and increasing, the second segment is quadratic, and the last segment is constant, though always at a lower level than the peak of the first segment. The quadratic is maximized at $\beta^r = k + \epsilon^p/2$, which is above $2k - \epsilon^p$ if and only if $k > 3\epsilon^p/2$, so the optimal toughness nevel in the stable preference distribution is:

$$\beta^{r,*} = \begin{cases} k + \frac{\epsilon^p}{2} & k < \frac{3}{2}\epsilon^p \\ 2k - \epsilon^p & k > \frac{3}{2}\epsilon^p \end{cases}$$

which is double the toughness of the equilibrium average toughness is the baseline model (if $\epsilon^p = \epsilon$). So, by allowing the toughness to be conditional on whether a player is the proposer or not, the *aggregate* toughness remains unchanged, though only responders are irrationally tough.

To compute the probability of conflict, write the type a actor j is in role i as $\beta^{i,*} + \nu_j$. So, the probability of conflict in the stable preference distribution is:

$$Pr(\beta^r + \beta^p > 2k) = Pr(\beta^{r,*} + \nu^r + \nu^p > 2k) = \begin{cases} Pr(\nu^r + \nu^p > k - \epsilon^p/2) & k < 3\epsilon^p/2 \\ Pr(\nu^r + \nu^p > \epsilon^p) & k > 3\epsilon^p/2 \end{cases} \tag{6}$$

Determining the probability of conflict for either case requires computing the distribution of $\nu^p + \nu^r$. It is useful to first state a general result about the sum of uniform random variables centered at zero but with different range:

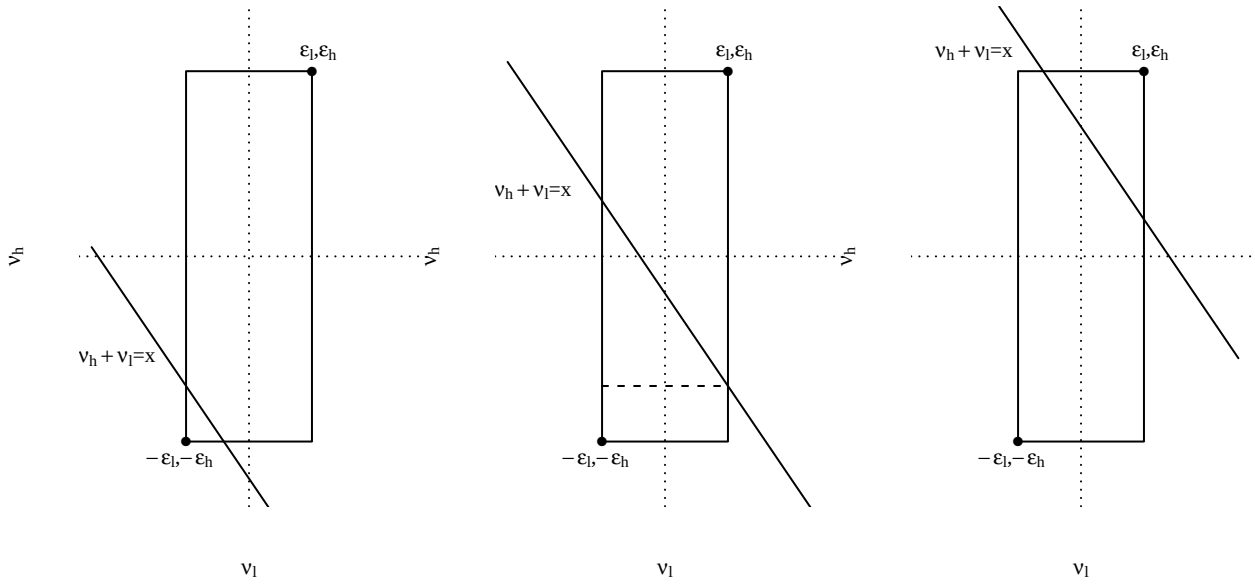
Lemma 2. *Let $\nu_h \sim U[-\epsilon_h, \epsilon_h]$ and $\nu_l \sim U[-\epsilon_l, \epsilon_l]$, where $\epsilon_l \leq \epsilon_h$. Then:*
i. the cumulative density function of $\nu_h + \nu_l$ is given by:

$$F^{\nu_h+\nu_l}(x) = \begin{cases} 0 & x < -\epsilon_h - \epsilon_l \\ \frac{(x+\epsilon_l+\epsilon_h)^2}{8\epsilon_l\epsilon_h} & x \in (-\epsilon_h - \epsilon_l, \epsilon_l - \epsilon_h) \\ x/(2\epsilon_h) + 1/2 & x \in (\epsilon_l - \epsilon_h, \epsilon_h - \epsilon_l) \\ 1 - \frac{(-x+\epsilon_l+\epsilon_h)^2}{8\epsilon_l\epsilon_h} & x \in (\epsilon_h - \epsilon_l, \epsilon_l + \epsilon_h) \\ 1 & x > \epsilon_l + \epsilon_h \end{cases}$$

ii. $F^{\nu_h+\nu_l}(-x) = 1 - F^{\nu_h+\nu_l}(x)$

Proof The result is easiest to prove visually. Figure 1 plots the joint density of ν_l and ν_h. For any x, the distribution function is area of the rectangle drawn by the bounds of the distribution below the line $\nu_l + \nu_h = x$ times the density over the rectangle, which is $\frac{1}{4\epsilon_l\epsilon_h}$ (i.e., the product of the individual densities). Clearly for $x < -\epsilon_l - \epsilon_h$ none of the rectangle is under the diagonal, so the

Figure 1: Illustration of CDF of Sum of Uniform Random Variables



distribution function is 0, and when $x > \epsilon_l + \epsilon_h$ the entire rectangle is below and the distribution function is 1.

The left panel shows that for $x \in (-\epsilon_l - \epsilon_h, \epsilon_l - \epsilon_h)$ the region below the diagonal is a right triangle with equal base and height. The diagonal intersects $\nu_h = -\epsilon_h$ at $\nu_h = x + \epsilon_h$, so the sides are length $x + \epsilon_h - (-\epsilon_l)$, and hence the area is $\frac{(x+\epsilon_l+\epsilon_h)^2}{2}$, and multiplying by the density gives $\frac{(x+\epsilon_l+\epsilon_h)^2}{8\epsilon_l\epsilon_h}$.

The middle panel shows that for $x \in (\epsilon_l - \epsilon_h, \epsilon_h - \epsilon_k)$, the area is a right triangle with area $\frac{(2\epsilon_l)^2}{2}$ plus a rectangle with area $2\epsilon_l(x - (\epsilon_l - \epsilon_h))$. Adding these and multiplying by the density gives:

$$\frac{2\epsilon_l(x - (\epsilon_l - \epsilon_h)) + 2\epsilon_l^2}{4\epsilon_l\epsilon_h} = \frac{x}{\epsilon_h} + \frac{1}{2}$$

The right panel shows that for $x \in (\epsilon_h - \epsilon_l, \epsilon_l + \epsilon_h)$, the area under the diagonal is the area of the entire rectangle $4\epsilon_l\epsilon_h$ minus the upper triangle with area $\frac{(\epsilon_l+\epsilon_h-x)^2}{2}$. Combining these pieces

gives part i.

Part ii follows from part i (or the symmetry of the densities of ν_l and ν_h around 0). ∎

So, there are many cases to consider for the probability of conflict, depending on the signs of $\epsilon^r - \epsilon^p$ and $k - \epsilon^p/2$, and then where $k - \epsilon^p/2$ and ϵ^p lie in the five segments of the CDF of $\nu^r + \nu^p$. Rather than enumerating all possible cases, we focus on comparative statics analogous to those in the main text:

Proposition 6. *In the unique stable preference distribution to the model with role-based toughness, the probability of conflict is:*

i. equal to the probability of conflict in the baseline if $\epsilon^p = \epsilon^r$,

ii. weakly decreasing in k, *and*

iii. for any $k > 3\epsilon^p/2$, *equal to*

$$\underline{p} = \begin{cases} \frac{\epsilon^r}{8\epsilon^p} & \epsilon^r < \epsilon^p \\ \frac{\epsilon^p}{8\epsilon^r} & \epsilon^r \in (\epsilon^p, 2\epsilon^p) \\ \frac{1}{2} - \frac{\epsilon^p}{2\epsilon^r} & \epsilon^r > 2\epsilon^p \end{cases}$$

Proof Part i follows from evaluating the fact that $\beta^{*,r} + \beta^{*,p} = 2\beta^*$ (where β^* is the average toughness for the main model) and if $\epsilon^r = \epsilon^p = \epsilon$ the distribution of $\epsilon^r + \epsilon^p$ is the triangle distribution with CDF given by lemma 2.

Part ii follows from the fact that the derivative of $\beta^{r,*}$ with respect to r is less than or equal to 2.

For part iii, for any $k > 3\epsilon^p/2$, the probability of conflict is $1 - F^{\nu_h+\nu_l}(\epsilon^p) = F^{\nu_h+\nu_l}(-\epsilon^p)$. When $\epsilon^p > \epsilon^r$, $-\epsilon^p$ must lie on the second segment of the CDF and and is hence the probability of

conflict is:

$$\frac{(-\epsilon^p + \epsilon^p + \epsilon^r)^2}{8\epsilon^p\epsilon^r} = \frac{\epsilon^r}{8\epsilon^p}.$$

When $\epsilon^p < \epsilon^r$, $-\epsilon^p$ lies on the second segment if $-\epsilon^p < \epsilon^p - \epsilon^r \implies \epsilon^r < 2\epsilon^p$ and on the third segment otherwise. Plugging $-\epsilon^p$ into the relevant segment of the PDF gives the desired result. ∎

Part i implies that allowing role-based toughness changes the equilibrium *preferences* but not the equilibrium probability of conflict if the amount of noise in the evolutionary process is the same for each role. Parts ii-iii examines what happens if the evolution of the toughness in different roles is more or less noisy. The only case where conflict approaches 0 is if k is large and $\epsilon^r \to 0$. This is because in this case $\beta^{r,*} \to 2k - \epsilon^p$. So, the probability of conflict approaches $Pr(\nu^p > 2k - \epsilon_p) = 0$. As noted in the main text, the case as $q \to 1$ in the incomplete information model in the paper is the same as the limiting case as $\epsilon^p \to 0$ here.

On the other hand, as $\epsilon^p \to 0$, the probability of conflict approaches $1/2$, four times the probability in the baseline model. This is because $\beta^{r,*} \to 2k$, and hence any responder with a positive draw of ν^r will fight every proposer.

Sacred Values

Next, we consider a very different type of deviation from the objective preferences, intended to capture the idea that issues subject to bargaining are "indivisible" or that actors may assign a "sacred value" to attaining a certain share of the prize.* For example, assigning a value to fairness can be captured by assuming an actor faces a large negative shock to their preferences when accepting a deal that gives them less than v (that is, $x < v$ for the responder and $x > v$ for the proposer). If both actors – perhaps ethnic or religious groups – are bargaining over territory

*Ginges et al. 2007; Ginges and Atran 2011.

they believe to be sacred, then they can assign an arbitrarily low subjective utility to any accepted deal that does not give them all of this land. Such attitudes are not limited to bargaining over land: Ryan finds that citizens punish leaders for compromising on issues that are "moralized," which certainly applies to many issues where disagreement leads to conflict.[†]

The formal definition of sacred value preferences is:

Definition A player has $(\overline{x}, \underline{x})$-*sacred value preferences* if her subjective utility function in role i is:

$$u(i,x,a) = \begin{cases} s & a = 1, i = p, x > \overline{x} \text{ or } a = 1, i = r, x < \underline{x} \\ 0 & a = 0 \\ g_i(x) & a = 1, i = p, x \leq \overline{x} \text{ or } a = 1, i = r, x \geq \underline{x} \end{cases}$$

for any $s < 0$, strictly positive and increasing $g_r(x)$, and strictly positive and decreasing $g_p(x)$.

When two actors i and j are matched with i in the proposer role and j in the responder role, by standard logic i will offer $\underline{x}_j$ and it will be accepted if $\underline{x}_j \leq \overline{x}_i$ and will make an offer which is rejected if $\underline{x}_j > \overline{x}_i$.

Proposition 7. *An actor with* $(\overline{x}, \underline{x})$*-sacred value preferences uses the same SPNE strategy and hence gets the same fitness as an actor with toughness* $(\beta^r, \beta^p) = (v + k - \overline{x}^p, v - k + \underline{x}^r)$

Now consider a noisy evolutionary process where in each generation the type that gets the highest fitness (call these $(\overline{x}_{\max}, \underline{x}_{\max})$) reproduces, and the next generation has sacred value preferences where $\overline{x}_i$ is uniformly distributed on $[\overline{x}_{\max} - \epsilon^p, \overline{x}_{\max} + \epsilon^p]$ and $\underline{x}_i$ is uniformly distributed on $[\underline{x}_{\max} - \epsilon^r, \underline{x}_{\max} + \epsilon^r]$. Then by an identical analysis, there is a unique stable distribution of

[†]Ryan 2016.

preferences centered around $\overline{x}^* = v + k - \beta^{p,*} = v + k$ and:

$$\underline{x}^* = v - k + \beta^{r,*} = \begin{cases} v + \epsilon^p/2 & k < 3\epsilon^p/2 \\ v + k - \epsilon^p & k \geq 3\epsilon^p/2 \end{cases}$$

and the probability of conflict is the same as in the role-based toughness case.

More General Preferences

This equivalence suggests that the same equilibrium behavior and chance of conflict can occur for a much wider class of preferences differing from the objective payoffs. A complete description of a players preferences is a $4-$tuple $u = (w^p, w^r, a^p(x), a^r(x))$, where $w^r \in \mathbb{R}$ and $w^p \in \mathbb{R}$ are the preferences over conflict when in the responder and proposer role, respectively, and $a^p(x)$ and $a^r(x)$ are the subjective utility when offer x is accepted in the respective roles. The only restrictions we place on the preferences are the following:

Assumption 1. *The preferences for the actors u are such that:*
i) a^p is weakly decreasing in x and a^r is weakly increasing in x
ii) there exists an $\overline{x} \in \mathbb{R}$ such that $\overline{x} = \min\{x : a^p(x) \geq w^p\}$ and an $\underline{x} \in \mathbb{R}$ such that $\underline{x} = \max\{x : a^r(x) \leq w^r\}$.

In words, i implies the proposer always prefers smaller accepted offers and the responder always prefers higher accepted offers. Part ii implies that there is a well defined "highest acceptable offer" for the proposer and a "lowest acceptable offer" for the responder. A somewhat more intuitive assumption which implies this property is if $a^p(x)$ is right-continuous and $w^p \in \text{Range}(a^p)$, and similarly $a^r(x)$ is left-continuous and $w^r \in \text{Range}(a^r)$. That is, the cases that need to be ruled out are when either the fighting fitness lies outside the range of possible fitness for acceptance or there is a discontinuity in the acceptance fitness which renders the minimum or maximum expressions undefined.

Suppose two players are matched to play the bargaining game, and call the player in the proposer role i and in the responder role j. Then the proposer role either offers $\underline{x}_j^r$ or an offer which is rejected, and prefers to offer $\underline{x}_j^r$ if and only if $\underline{x}_j^r \leq \overline{x}_i$. That is, if there is a division weakly preferred to war for both players, they strike a bargain at the minimal offer accepted by the responder. Otherwise, they fight. So, a more general statement of lemma **??** is:

Lemma 3. *Suppose players i and j have preferences meeting assumption 1, and i is placed in the proposer role with j in the responder role. Then in any SPNE:*

i. If $\underline{x}_j^r \leq \overline{x}_i^p$, then the proposer offers $\underline{x}_j^r$ and it is accepted

ii. If $\underline{x}_j^r > \overline{x}_i^p$, then the proposer makes an offer less than $\underline{x}_j^r$ which is rejected.

So, any preferences meeting equation 1 induce identical behavior as the $(\underline{x}, \overline{x})$-sacred value preferences. While explicitly modeling the evolution of preferences is more complex, as it requires specifying not just how a real-valued parameter changes but how the entire preference function evolves. However, as long as the resulting $\underline{x}$ and $\overline{x}$ behave in a similar manner defined above, identical results arise in this more general setting.

References

Dekel, Eddie, Jeffrey C. Ely and Okan Yilankaya. 2007. "Evolution of Preferences." *Review of Economic Studies* 74(3):685–704.

Ginges, Jeremy and Scott Atran. 2011. "War as a moral imperative (not just practical politics by other means)." *Proceedings of the Royal Society B: Biological Sciences* 278(1720):2930–2938.

Ginges, Jeremy, Scott Atran, Douglas Medin and Khalil Shikaki. 2007. "Sacred bounds on rational resolution of violent political conflict." *Proceedings of the National Academy of Sciences* 104(18):7357–7360.

Huck, Steffen and Jörg Oechssler. 1999. "The Indirect Evolutionary Approach to Explaining Fair Allocations." *Games and Economic Behavior* 28(1):13 – 24.

Ryan, Timothy J. 2016. "No Compromise: Political Consequences of Moralized Attitudes." *American Journal of Political Science* .