Appendices for “Interethnic Complementarities and Decentralized Intergroup Cooperation” (2014)∗

Sangick Jeon†
Groundtruth.io

A Equilibrium Conditions

A.1 Re-Analyzing the In-Group Policing Equilibrium

First, define a “cooperator” as any player that is not in punishment phase, and define a “defector” as any player being punished. Then, the in-group policing equilibrium is a subgame perfect Nash equilibrium if: (1a) cooperators have no incentives to defect against out-group members, (1b) cooperators have no incentives to defect against in-group cooperators, (1c) cooperators have no incentives to cooperate with in-group defectors, (2a) defectors have no incentives to defect against out-group members, (2b) defectors have no incentives to defect against in-group cooperators, and (2c) defectors have no incentives to cooperate with in-group defectors (see Fearon and Laitin 1996). Conditions (1c) and (2c) are satisfied for any state of the world since a deviation would strictly lower the player’s payoff. So just need to check conditions (1a), (1b), (2a), and (2b).

Condition (1a) holds if and only if an individual is willing to forego gains from immediate defection for future beneficial cooperation:

\[ \alpha \Delta^{out} + \sum_{i=1}^{T^{igp}} \delta^i [p \Delta^{out} + (1-p)q(-\beta)] \leq \Delta^{out} + \sum_{i=1}^{T^{igp}} \delta^i [p \Delta^{out} + (1-p)(q\Delta^{in} + \alpha \Delta^{in}(1-q))] \]

\[ \Leftrightarrow \alpha \Delta^{out} - \Delta^{out} \leq (\delta \frac{1 - \delta^{T^{igp}}}{1 - \delta})(1-p)[q\Delta^{in} + \alpha \Delta^{in}(1-q) + q\beta]. \]

Similarly, condition (2a) holds if and only if a defector is not willing to stay in punish-
ment phase for an additional $t$ rounds:

$$\alpha \Delta^\text{out} + \sum_{i=1}^{T^\text{igp}+t} \delta^i[p\Delta^\text{out} + (1-p)q(-\beta)]$$

$$\leq \Delta^\text{out} + \sum_{i=1}^{T^\text{igp}} \delta^i[p\Delta^\text{out} + (1-p)q(-\beta)] + \sum_{j=1}^{T^\text{igp}+t} \delta^j[p\Delta^\text{out} + (1-p)(q\Delta^\text{in} + \alpha\Delta^\text{in}(1-q))]$$

$$\Leftrightarrow \alpha \Delta^\text{out} - \Delta^\text{out} \leq (\delta \frac{1-\delta^\text{igp}}{1-\delta})(1-p)[q\Delta^\text{in} + \alpha\Delta^\text{in}(1-q) + q\beta].$$

Both of these inequalities suggest that in-group policing becomes harder to support under conditions of low $\Delta^\text{in}$ and high $\Delta^\text{out}$. This is consistent with the logic developed throughout the paper: higher levels of $\Delta^\text{out}$ increase the payoff for defecting in cross-group interactions and lower levels of $\Delta^\text{in}$ decrease the effectiveness of punishment mechanisms (since punishment is the removal of future beneficial cooperation).

Furthermore, these inequalities are minimized by (i) $q = 1$ if $\alpha\Delta^\text{in} > \Delta^\text{in} + \beta$, of (ii) $q = 0$ if $\alpha\Delta^\text{in} \leq \Delta^\text{in} + \beta$.

If (i) $\alpha\Delta^\text{in} > \Delta^\text{in} + \beta$, then subgame perfection requires:

$$\frac{\Delta^\text{out}(\alpha - 1)}{(\Delta^\text{in} + \beta)(1-p)} \leq (\delta \frac{1-\delta^\text{igp}}{1-\delta})$$

$$\Leftrightarrow \delta^\text{igp} \leq 1 - \frac{\Delta^\text{out}(\alpha - 1)}{(\Delta^\text{in} + \beta)(1-p)} \frac{1-\delta}{\delta} \quad (A-1)$$

If (ii) $\alpha\Delta^\text{in} \leq \Delta^\text{in} + \beta$, then subgame perfection requires:

$$\frac{\Delta^\text{out}(\alpha - 1)}{\alpha(1-p)} \leq (\delta \frac{1-\delta^\text{igp}}{1-\delta})$$

$$\Leftrightarrow \delta^\text{igp} \leq 1 - \frac{\Delta^\text{out}(\alpha - 1)}{\Delta^\text{in}\alpha(1-p)} \frac{1-\delta}{\delta} \quad (A-2)$$

But since (ii) $\alpha\Delta^\text{in} \leq \Delta^\text{in} + \beta$, it is the case that $\frac{\Delta^\text{out}(\alpha - 1)}{(\Delta^\text{in} + \beta)(1-p)} \leq \frac{\Delta^\text{out}(\alpha - 1)}{\Delta^\text{in}\alpha(1-p)}$. Thus, expressions A-1 and A-2 together suggest that the minimum necessary to satisfy conditions (1a) and (2a) is:

$$\tilde{\delta}_1 = \delta^\text{igp} = \frac{\Delta^\text{out}(\alpha - 1)}{(\Delta^\text{in} + \beta)(1-p)}. \quad (A-3)$$

However, we need to make sure that $\tilde{\delta}_1 \in (0, 1)$ in order to choose $T^\text{igp} = 1$. For purposes of simplification, let $\Delta^\text{in} = 1$ and normalize $\Delta^\text{out}/\Delta^\text{in} = \Delta^\text{out}_N = \Delta^\text{out}_N$. Then, $\Delta^\text{out}_N$ is the maximum $\Delta^\text{out}/\Delta^\text{in}$ ratio that ensures $\tilde{\delta}_1 < 1$:

$$\Delta^\text{out}_N < \frac{(1-p)(1+\beta)}{(\alpha - 1)}. \quad (A-4)$$
If this inequality holds, then it is always possible to choose $T^{igp} = 1$, since $\delta T^{igp}$ is decreasing in $T^{igp}$. Notice that high out-group diversity must be balanced by sufficient in-group diversity (since $\Delta^*_N = \Delta^*_N / \Delta^*_{in}$). Otherwise, must enforce stronger in-group punishments by choosing $T^{igp} > 1$. More specifically, when $\Delta^*_N < \Delta^*_N$, the minimum $T_{min}^{igp}$ necessary to satisfy conditions (1a) and (2a) is given by:

$$\delta T^{igp} \leq 1 - \frac{\Delta^*_N(\alpha - 1)}{(\Delta^*_{in} + \beta)(1 - p)} \frac{1 - \delta}{\delta}$$

$$\iff T_{min}^{igp} \geq \ln \left(1 - \frac{\Delta^*_N(\alpha - 1)}{(\Delta^*_{in} + \beta)(1 - p)} \frac{1 - \delta}{\delta}\right) / \ln(\delta). \quad (A-5)$$

Notice that this still requires $0 < 1 - \frac{\Delta^*_N(\alpha - 1)}{(\Delta^*_{in} + \beta)(1 - p)} \frac{1 - \delta}{\delta} \iff \delta > \frac{\Delta^*_N(\alpha - 1)}{(\Delta^*_{in} + \beta)(1 - p)}$. Thus, when $\Delta^*_N < \Delta^*_N$, the minimum $\bar{\delta}_{1d}$ necessary to satisfy conditions (1a) and (2a) is:

$$\bar{\delta}_{1d} = \frac{\Delta^*_N(\alpha - 1)}{\Delta^*_N(\alpha - 1) + (1 - p)(\Delta^*_{in} + \beta)}. \quad (A-6)$$

Next, consider condition (1b), which holds if and only if an individual is willing to forego gains from immediate defection for future beneficial cooperation:

$$\alpha \Delta^*_{in} + \sum_{i=1}^{T^{igp}} \delta^i[p\Delta^*_N + (1 - p)q(-\beta)] \leq \Delta^*_{in} + \sum_{i=1}^{T^{igp}} \delta^i[p\Delta^*_N + (1 - p)(q\Delta^*_{in} + \alpha \Delta^*_{in}(1 - q))]$$

$$\iff \alpha \Delta^*_{in} - \Delta^*_{in} \leq (\delta \frac{1 - \delta T^{igp}}{1 - \delta})(1 - p)[q\Delta^*_{in} + \alpha \Delta^*_{in}(1 - q) + q\beta].$$

$$\iff \alpha - 1 \leq (\delta \frac{1 - \delta T^{igp}}{1 - \delta})(1 - p)[q + \alpha(1 - q) + \frac{q\beta}{\Delta^*_{in}}].$$

This expression is minimized by (i) $q = 1$ if $1 + \beta / \Delta^*_{in} \leq \alpha$, or (ii) $q = 0$ if $1 + \beta / \Delta^*_{in} > \alpha$.

(i) If $1 + \beta / \Delta^*_{in} \leq \alpha$, then subgame perfection requires:

$$\frac{\alpha - 1}{(1 - p)(1 + \beta / \Delta^*_{in})} \leq (\delta \frac{1 - \delta T^{igp}}{1 - \delta})$$

$$\iff \delta T^{igp} \leq 1 - \frac{\alpha - 1}{(1 - p)(1 + \beta / \Delta^*_{in})} \frac{1 - \delta}{\delta}. \quad (A-7)$$

(ii) If $1 + \beta / \Delta^*_{in} > \alpha$, then subgame perfection requires:

$$\frac{\alpha - 1}{\alpha(1 - p)} \leq (\delta \frac{1 - \delta T^{igp}}{1 - \delta})$$

$$\iff \delta T^{igp} \leq 1 - \frac{\alpha - 1}{\alpha(1 - p)} \frac{1 - \delta}{\delta}. \quad (A-8)$$

However, since (ii) $1 + \beta / \Delta^*_{in} > \alpha$, $\frac{\alpha - 1}{(1 - p)(1 + \beta / \Delta^*_{in})} < \frac{\alpha - 1}{\alpha(1 - p)}$, expressions A-7 and A-8 together imply that the minimum $\bar{\delta}_{2}$ necessary to satisfy conditions (1b):
\[ \delta_2 = \delta^{T_{igp}} = \frac{\alpha - 1}{(1 - p)(1 + \beta/\Delta^m)}. \] (A-9)

Now consider condition (2b), which holds if and only if a defector is not willing to stay in punishment phase for an additional \( t \) rounds:

\[
0 + \sum_{i=1}^{T_{igp}+t} \delta^i[p\Delta^\text{out} + (1 - p)q(-\beta)] \\
\leq -\beta + \sum_{i=1}^{T_{igp}} \delta^i[p\Delta^\text{out} + (1 - p)q(-\beta)] + \sum_{i=1}^{T_{igp}+t} \delta^i[p\Delta^\text{out} + (1 - p)(q\Delta^\text{in} + \alpha\Delta^\text{in}(1 - q))] \\
\Leftrightarrow \beta \leq (\delta^{1 - \delta^{T_{igp}}} - 1)(1 - p)[q\Delta^\text{in} + \alpha\Delta^\text{in}(1 - q) + q\beta].
\]

This expression is minimized by (i) \( q = 1 \) if \( \Delta^\text{in} + \beta \leq \alpha\Delta^\text{in} \), or (ii) \( q = 0 \) if \( \Delta^\text{in} + \beta > \alpha\Delta^\text{in} \).

(i) If \( \Delta^\text{in} + \beta \leq \alpha\Delta^\text{in} \), then subgame perfection requires:

\[
\frac{\beta}{(1 - p)(\Delta^\text{in} + \beta)} \leq (\delta^{1 - \delta^{T_{igp}}} - 1) \\
\Leftrightarrow \delta^{T_{igp}} \leq 1 - \frac{\beta}{(1 - p)(\Delta^\text{in} + \beta)} \frac{1 - \delta}{\delta}. \tag{A-10}
\]

(ii) If \( \Delta^\text{in} + \beta > \alpha\Delta^\text{in} \), then subgame perfection requires:

\[
\frac{\beta}{\alpha\Delta^\text{in}(1 - p)} \leq (\delta^{1 - \delta^{T_{igp}}} - 1) \\
\Leftrightarrow \delta^{T_{igp}} \leq 1 - \frac{\beta}{\alpha\Delta^\text{in}(1 - p)} \frac{1 - \delta}{\delta}. \tag{A-11}
\]

But since (ii) \( \Delta^\text{in} + \beta > \alpha\Delta^\text{in} \), \( \frac{\beta}{(1 - p)(\Delta^\text{in} + \beta)} > \frac{\beta}{\alpha\Delta^\text{in}(1 - p)} \), expressions A-10 and A-11 together imply that the minimum \( \bar{\delta}_3 \) necessary to satisfy condition (2b) is:

\[ \bar{\delta}_3 = \delta^{T_{igp}} = \frac{\beta}{(1 - p)(\Delta^\text{in} + \beta)}. \tag{A-12} \]

Finally, consider expressions A-3, A-9, and A-12 all together, which give the minimum \( \bar{\delta}_{igp} \) necessary to sustain in-group policing as a subgame Nash equilibrium when \( \Delta_N^{out*} < \frac{1 - \Delta_N^\text{out}(\alpha - 1)}{(1 - p)(1 + \beta)/\alpha - 1} \):

\[ \bar{\delta}_{igp} = \max\{\bar{\delta}_1, \bar{\delta}_2, \bar{\delta}_3\} = \max\{\frac{\Delta_N^{out}(\alpha - 1)}{(1 - p)(1 + \beta)}, \frac{\alpha - 1}{(1 - p)(1 + \beta)}, \frac{\beta}{(1 - p)(1 + \beta)}\}. \tag{A-13} \]

If this equality holds, it is always possible to choose \( T_{igp} = 1 \). In the case of \( \Delta_N^{out*} \geq \frac{1 - \Delta_N^\text{out}(\alpha - 1)}{(1 - p)(1 + \beta)/\alpha - 1} \), subgame perfection requires the additional restriction given by A-6:
\[ \delta_{igp} = \max\left\{ \frac{\Delta_{N}^{\text{out}}(\alpha - 1)}{\Delta_{N}^{\text{out}}(\alpha - 1) + (1 - p)(1 + \beta)} \cdot \delta_{2}, \delta_{3} \right\}, \quad (A-14) \]

and must choose \( T_{\text{min}}^{\text{igp}} = \ln \left( 1 - \frac{\Delta_{N}^{\text{out}}(\alpha - 1)}{(1 + \beta)(1 - p)} \right) \cdot \delta \) / \( \ln(\delta) \).

### A.2 Re-Analyzing the Spiral Equilibrium

The spiral equilibrium is a subgame perfect Nash equilibrium if (1a) cooperators have no incentives to defect against out-group members when neither group is in punishment phase, (1b) cooperators have no incentives to cooperate in an out-group interaction when groups are in punishment phase, (1c) cooperators have no incentives to defect against in-group cooperators, (1d) cooperators have no incentives to cooperate with in-group defectors, (2a) defectors have no incentives to defect against out-group members when neither group is in punishment phase, (2b) defectors have no incentives to cooperate in an out-group pairing when groups are in punishment phase, (2c) defectors have no incentives to defect against in-group cooperators, and (2d) defectors have no incentives to cooperate with in-group defectors (see Fearon and Laitin 1996). Conditions (1b), (1d), (2b), and (2d) are satisfied for any state of the world since a deviation would strictly lower the player’s payoff. So just need to check conditions (1a), (1c), (2a), and (2c).

Condition (1a) holds if and only if an individual is willing to forego gains from immediate defection for future beneficial cooperation:

\[
\alpha \Delta_{\text{out}}^{\text{out}} + \sum_{i=1}^{T_{\text{out}}} \delta^{i} (1 - p)(q \Delta_{\text{in}} + \alpha \Delta_{\text{in}}^{\text{in}}(1 - q)) \leq \Delta_{\text{out}}^{\text{out}} + \sum_{i=1}^{T_{\text{out}}} \delta^{i} (p \Delta_{\text{out}} + (1 - p)(q \Delta_{\text{in}} + \alpha \Delta_{\text{in}}^{\text{in}}(1 - q))
\]

\[ \equiv \alpha \Delta_{\text{out}}^{\text{out}} - \Delta_{\text{out}}^{\text{out}} \leq p \Delta_{\text{out}}(\delta \frac{1 - \delta^{T_{\text{out}}}}{1 - \delta}) \]

\[ \equiv \alpha - 1 \leq p(\delta \frac{1 - \delta^{T_{\text{out}}}}{1 - \delta}). \]

Similar, condition (2a) holds if and only if an individual is willing to forego gains from immediate defection for future beneficial cooperation:

\[
\alpha \Delta_{\text{out}}^{\text{out}} + \sum_{i=1}^{T_{\text{in}}} \delta^{i} (1 - p)(- \beta) + \sum_{j=1}^{T_{\text{out}} - T_{\text{in}}} \delta^{j} (1 - p)(q \Delta_{\text{in}} + \alpha \Delta_{\text{in}}^{\text{in}}(1 - q))
\]

\[ \leq \Delta_{\text{out}}^{\text{out}} + \sum_{i=1}^{T_{\text{in}}} \delta^{i} (p \Delta_{\text{out}} + (1 - p)(- \beta)) + \sum_{j=1}^{T_{\text{out}} - T_{\text{in}}} \delta^{j} (p \Delta_{\text{out}} + (1 - p)(q \Delta_{\text{in}} + \alpha \Delta_{\text{in}}^{\text{in}}(1 - q))
\]

\[ \equiv \alpha \Delta_{\text{out}}^{\text{out}} - \Delta_{\text{out}}^{\text{out}} \leq p \Delta_{\text{out}}(\delta \frac{1 - \delta^{T_{\text{in}} + T_{\text{out}}}}{1 - \delta}) \]
\[ \Leftrightarrow \alpha - 1 \leq p(\delta \frac{1 - \delta^{T_{\text{in}+T_{\text{out}}}}}{1 - \delta}). \]

Notice that both within-group and between-group diversity has canceled out of these equations. They also suggest that condition (1a) and (2a) require the following inequality to hold:

\[ \frac{\alpha - 1}{p} \leq \delta \frac{1 - \delta^{T_{\text{out}}}}{1 - \delta}, \]

implying that \( T_{\text{out}} \) is given by:

\[ T_{\text{out}} \geq \ln(1 - \frac{\alpha - 1}{p} \cdot \frac{1 - \delta}{\delta}) / \ln(\delta). \] (A-15)

And, in order to deter defection in out-group pairings under the strongest possible punishment \( (T_{\text{out}} = \infty) \), condition (1a) and (2a) further require:

\[ \delta \geq \frac{\alpha - 1}{\alpha - 1 + p}. \] (A-16)

Next, consider condition (1c), which holds if and only if an individual is willing to forego gains from immediate defection for future beneficial cooperation:

\[ \alpha \Delta^{\text{in}} + \sum_{i=1}^{T_{\text{in}}} \delta^i[p \Delta^{\text{out}} + (1 - p)q(-\beta)] \leq \Delta^{\text{in}} + \sum_{i=1}^{T_{\text{out}}} \delta^i[p \Delta^{\text{out}} + (1 - p)(q \Delta^{\text{in}} + \alpha \Delta^{\text{in}}(1 - q))] \]

\[ \Leftrightarrow \alpha \Delta^{\text{in}} - \Delta^{\text{in}} \leq (\delta \frac{1 - \delta^{T_{\text{in}}}}{1 - \delta})(1 - p)[q \Delta^{\text{in}} + \alpha \Delta^{\text{in}}(1 - q) + q \beta]. \]

\[ \Leftrightarrow \alpha - 1 \leq (\delta \frac{1 - \delta^{T_{\text{in}}}}{1 - \delta})(1 - p)[q + \alpha(1 - q) + q \beta] \Delta^{\text{in}}. \]

This inequality suggests that condition (1c) begins to fail under conditions of high \( \Delta^{\text{in}} \), that is, as the temptation to defect increases. It is also the same expression that is broken down in A-7 and A-8, which means that \( \bar{\delta}_2 \) is the minimum \( \delta \) necessary to satisfy this condition:

\[ \bar{\delta}_2 = \delta^{T_{\text{in}}} = \frac{\alpha - 1}{(1 - p)(1 + \beta / \Delta^{\text{in}})}. \] (A-17)

Now consider condition (2c) holds if and only if a defector is not willing to stay in punishment phase for an additional \( t \) rounds:

\[ 0 + \sum_{i=1}^{T_{\text{in}+t}} \delta^i[p \Delta^{\text{out}} + (1 - p)q(-\beta)] \]

\[ \leq -\beta + \sum_{i=1}^{T_{\text{in}}} \delta^i[p \Delta^{\text{out}} + (1 - p)q(-\beta)] + \sum_{i=1}^{T_{\text{in}+t}} \delta^i[p \Delta^{\text{out}} + (1 - p)(q \Delta^{\text{in}} + \alpha \Delta^{\text{in}}(1 - q))] \]
\[\iffalse \beta \leq (\delta^{1 - \delta_{in} + t} - 1)(1 - p)[q\Delta^{in} + \alpha \Delta^{in}(1 - q) + q\beta]. \fi\]

This inequality fails to hold under conditions of low \(\Delta^{in}\), and is analyzed in A-10 and A-11. This means that \(\bar{\delta}_3\) is the minimum \(\delta\) necessary to satisfy condition (2c):

\[
\bar{\delta}_3 = \delta^{T_{in}} = \frac{\beta}{(1 - p)(\Delta^{in} + \beta)}. \tag{A-18}
\]

Then, finally, expressions A-16, A-17, and A-18 altogether suggest that the minimum \(\bar{\delta}_s\) necessary to sustain the spiral regime as a subgame Nash equilibrium is:

\[
\bar{\delta}_s = \max\{\frac{\alpha - 1}{\alpha - 1 + p}, \frac{\alpha - 1}{(1 - p)(1 + \beta/\Delta^{in})}, \frac{\beta}{(1 - p)(\Delta^{in} + \beta)}\}, \tag{A-19}
\]

where \(T_{min}^{out} = \ln\left(1 - \frac{\alpha - 1}{p} \cdot \frac{1 - \delta}{\delta}\right)/\ln(\delta)\).

A.3 The Mediator System

**Proposition:** Let \(\bar{\delta}_{ms} = \max\{1 - \frac{p(\Delta^{out} - Q)}{\gamma}, \frac{\alpha - 1}{(1 - p)(1 + \beta/\Delta^{in})}, \frac{\beta}{(1 - p)(\Delta^{in} + \beta)}\}\). Then, the Mediator System is a subgame perfect Nash equilibrium if and only if \(\Delta^{out} \geq Q^*\), \(\Delta^{out}(\alpha - 1) \leq \gamma^* \leq \frac{p}{1 - \delta}(\Delta^{out} - Q)\), and \((1 - \delta)^\gamma + Q \leq \Delta^{out}\). These conditions imply both that \(\bar{\delta}_{ms}\) is the smallest discount factor that can support the Mediator System as a subgame perfect equilibrium, and that if the Mediator System is a subgame perfect equilibrium for given parameters, then it is always possible to choose \(T^{ms} = 1\).

**Proof:** Consider \(i\), a member of group A, who is paired with a partner \(j\) (the same arguments apply to members of group B as well). Let \(s_t = (t_1, t_2, ..., t_n)\) represent the state of the system at the beginning of period \(t\), where \(t_j > 0\) indicates that player \(j\) is in out-group punishment phase, and \(t_j = 0\) means that player \(j\) is not in out-group punishment phase. Individuals who are not in in-group punishment phase will be called “Cooperators” and individuals who are in in-group punishment phase will be called “Defectors.” Then, according to the optimality principle of dynamic programming, the Mediator System is a subgame perfect Nash equilibrium if for any state of the world:

(1) A Cooperator \(i\) who is not in out-group punishment phase has no incentive (a) not to query for cost \(Q > 0\) when paired out-group, (b) to defect in an out-group pairing when \(t_i = 0\), (c) to cooperate in an out-group pairing if \(t_i > 0\), (d) to defect against an in-group Cooperator, (e) to cooperate with an in-group Defector, (f) not to appeal for cost \(c > 0\) if unilaterally defected upon, and (g) to refuse paying damage award \(\gamma > 0\) if judged against.

(2) A Defector \(i\) who is not in out-group punishment phase has no incentive (a) not to query for cost \(Q > 0\) when paired out-group, (b) to defect in an out-group pairing when \(t_i = 0\), (c) to cooperate in an out-group pairing if \(t_i > 0\), (d) to defect against an in-group Cooperator, (e) to cooperate with an in-group Defector, (f) not to appeal for cost \(c > 0\) if unilaterally defected upon, and (g) to refuse paying damage award \(\gamma > 0\) if judged against.
And (3) individuals in out-group punishment phase must have no incentives (a) to query for cost $Q > 0$ when paired out-group, (b) to cooperate in out-group pairings, and (c) pay damage award $\gamma > 0$ if judged against.

It is clear that conditions (1c), (1e), (1f), (2c), (2e), (2f), (3a), (3b), and (3c) are satisfied for all parameter values, since they strictly lower an individual’s payoff. Notice also that for in-group pairings, the Mediator System is identical to the in-group policing regime. This means that the equilibrium conditions for (1d) and (2d) are given by A-9 and A-12, respectively, and that together the minimum discount factor necessary to satisfy conditions (1d) and (2d) is given by:

$$\max\{\tilde{\delta}_2, \tilde{\delta}_3\} = \max\{\frac{\alpha - 1}{(1 - p)(1 + \beta/\Delta^m)}, \frac{\beta}{(1 - p)(\Delta^m + \beta)}\}. \quad (A-20)$$

And since out-group strategies are implemented independent of in-group interactions, conditions (1a) and (2a), conditions (1b) and (2b), and conditions (1g) and (2g) may be considered together. The following analysis identifies the equilibrium conditions for (1a/2a), (1b/2b), and (1g/2g):

For condition (1a/2a) to hold, it must be the case that individuals are better off querying at cost $Q > 0$ than not. Suppose an individual is paired out-group and does not query. Then the player receives utility zero, when by querying she would have received utility $\Delta^\text{out} - Q$. This means that the cost of query must be less than the value of cooperation:

$$Q^* = \Delta^\text{out}. \quad (A-21)$$

Notice that groups can afford to pay higher query costs when gains from intergroup exchange are high.

For condition (1b/2b) to hold, individuals must be willing to forego gains from immediate defection for future beneficial cooperation. Consider player $i$ (who is not in out-group punishment phase). By defecting, player $i$ receives an immediate payoff of $\alpha \Delta^\text{out}$, but must then pay for a damage award $\gamma > 0$. Hence, condition (1b/2b) requires $\alpha \Delta^\text{out} - Q - \gamma \leq \Delta^\text{out} - Q$, which implies that the damage award must be sufficiently large to deter defection:

$$\Delta^\text{out}(\alpha - 1) \leq \gamma^*. \quad (A-22)$$

Notice that the damage award increases in the value of out-group cheating.

For condition (1g/2g) to hold, individuals must be willing to pay damage award $\gamma > 0$ if judged against. In particular there must be credible punishments for refusing to pay. By not paying $\gamma > 0$, a player will gain $\alpha \Delta^\text{out}$, but will enter the out-group punishment phase indefinitely. This means that condition (1g/2g) requires the following inequality to hold:

$$\alpha \Delta^\text{out} - Q \leq (\alpha \Delta^\text{out} - Q - \gamma) + \delta p(\Delta^\text{out} - Q) + \delta^2 p(\Delta^\text{out} - Q) + ...$$

$$\Leftrightarrow (1 - \delta)\frac{\gamma}{p} + Q \leq \Delta^\text{outs}, \quad (A-23)$$

where $\Delta^\text{outs}$ gives the minimum level of between-group diversity necessary to support this equilibrium. When between-group diversity is low, the Mediator System is not a credible deterrent to opportunistic behavior. A-23 can be further rearranged to give the minimum
discount factor required to satisfy conditions (1g/2g):

\[
(1 - \delta)\frac{\gamma}{p} + Q \leq \Delta_{out}^* \Leftrightarrow 1 - \frac{p(\Delta_{out}^*)}{\gamma} \leq \bar{\delta}_d. \tag{A-24}
\]

Finally, consider A-22 and A-23 together, which give the lower and upper bounds on \(\gamma > 0\). This damage award must be large enough to deter defection, but also small enough so that a defector that is judged against has incentive to pay voluntarily. This means that the restrictions on \(\gamma\) are given by:

\[
\Delta_{out}^*(\alpha - 1) \leq \gamma^* \leq \frac{p}{1 - \delta}(\Delta_{out}^* - Q) \tag{A-25}
\]

and that \(\gamma^* > 0\) always exists since A-25 further solves to:

\[
\frac{1 - \delta}{p} \Delta_{out}^*(\alpha - 1) \leq (\Delta_{out}^* - Q) \Leftrightarrow \frac{1 - \delta}{p} \Delta_{out}^*(\alpha - 1) \leq \frac{Q}{\Delta_{out}^*} \leq 1 - \frac{1 - \delta}{p}(\alpha - 1)
\]

\[
\Leftrightarrow \frac{pQ}{p - (1 - \delta)(\alpha - 1)} \leq \Delta_{out}^* \tag{A-26}
\]

which is always possible.

Altogether then, the Mediator System is a subgame perfect Nash equilibrium if: \(\bar{\delta}_ms = \max\{1 - \frac{p(\Delta_{out}^* - Q)}{\gamma}, \frac{\alpha - 1}{(1 - p)(1 + \beta/\Delta_{in}^*)}, \frac{\beta}{(1 - p)(\Delta_{in}^* + \beta)}\}\), and \(\Delta_{out}^* \geq Q^*, \Delta_{out}^*(\alpha - 1) \leq \gamma^* \leq \frac{p}{1 - \delta}(\Delta_{out}^* - Q), \) and \((1 - \delta)\frac{\gamma}{p} + Q \leq \Delta_{out}^*\).

### B Expected Payoffs with “Noise” in Interethnic Interactions

Because of language barriers and cultural differences, amongst others, players intending to cooperate in interethnic interactions may accidentally “defect.” Such accidents and misinterpretations can be modeled as an \(\epsilon > 0\) probability that a player accidentally defects in a cross-group pairing. Suppose also that \(\epsilon = 0\) for all in-group pairings (since within groups individuals share common language and culture). Then, a player \(i\)’s ex ante expected equilibrium payoff in a spiral regime is given by the following recursive equation:

\[
V^*_s = pB + (1 - p)\Delta_{in}^* + \delta \left( (1 - \epsilon)^{2k}V^*_s + (1 - (1 - \epsilon)^{2k}) \cdot \left[ (1 - p)\Delta_{in}^* \frac{1 - \delta\Delta_{out}^*}{1 - \delta} + \delta\Delta_{out}^* V^*_s \right] \right),
\]

where \(B = \Delta_{out}^*(1 - \epsilon)^2 + (\alpha\Delta_{out}^* - \beta)\epsilon(1 - \epsilon)\) is the expected payoff if paired out-group in period \(t\), \((1 - \epsilon)^{2k}\) is the probability that on one defected in an out-group pairing, and the expression in brackets is the expected payoff if cross-group cooperation breaks down.
Fearon and Laitin (1996) show that this solves to:

\[V^*_s = \frac{pB + (1 - p)\Delta^{in} + \delta(1 - p)\Delta^{in}\frac{1 - \delta^{T_{out}}}{1 - \delta}(1 - (1 - \epsilon)^{2k})}{1 - \delta\left(\delta^{T_{out}} + (1 - \delta^{T_{out}})(1 - \epsilon)^{2k}\right)}.\] (A-27)

On the other hand, if player \(i\) belongs to an in-group policing regime, the ex ante expected equilibrium payoff is:

\[V^*_igp = pB + (1 - p)\Delta^{in} + \frac{\delta}{1 - \delta}\hat{V}^*_igp,\] (A-28)

where \(B\) is defined as before, and \(\hat{V}^*_igp\) is the expected payoff for all future rounds:

\[
\hat{V}^*_igp = pB + (1 - p)\left[pe\left(K_{-1}\epsilon 0 + (-\beta)(1 - K_{-1}\epsilon)\right) + p(1 - \epsilon)\left(K_{-1}\epsilon\alpha\Delta^{in} + \Delta^{in}(1 - K_{-1}\epsilon)\right)\right] + (1 - p)\left(K\epsilon\alpha\Delta^{in} + \Delta^{in}(1 - K\epsilon)\right).
\]

where \(K = \frac{k}{n - 1}\) is the fraction of in-group players other than \(i\) who were paired out-group in the last period if \(i\) was paired in-group that period, and \(K_{-1} = \frac{k - 1}{n - 1}\) is the fraction of in-group players other than \(i\) who were paired out-group in the last period if \(i\) was paired out-group that period.\(^1\)

Finally, if player \(i\) is part of the Mediator System, the ex ante expected equilibrium payoff is given by:

\[V^*_ms = pM + (1 - p)\Delta^{in} + \frac{\delta}{1 - \delta}\hat{V}^*_ms,\] (A-29)

where \(M = (\Delta^{out} - Q)(1 - \epsilon)^2 + (\alpha\Delta^{out} - Q - \gamma)\epsilon(1 - \epsilon) + (-\beta - Q + \gamma)\epsilon(1 - \epsilon)\) is the expected payoff for an out-group interaction, and \(\hat{V}^*_ms\) is the expected payoff for all future rounds:

\[
\hat{V}^*_ms = pM + (1 - p)\left[pe\left(K_{-1}\epsilon 0 + (-\beta)(1 - K_{-1}\epsilon)\right) + p(1 - \epsilon)\left(K_{-1}\epsilon\alpha\Delta^{in} + \Delta^{in}(1 - K_{-1}\epsilon)\right)\right] + (1 - p)\left(K\epsilon\alpha\Delta^{in} + \Delta^{in}(1 - K\epsilon)\right).
\]

\(^{1}\)Suppose provisionally that \(T_{igp} = 1\).