MATHEMATICAL APPENDIX:

The paper focuses on core results that are non-intuitive, important, and robust. The paper has provided a clear exposition of the model's behavior and has demonstrated that the core conclusions are broadly robust to relaxing the various specifying assumptions of the simulation. This supplement will provide an exhaustive analysis of the model's behavior, giving an intricate account of the simulation results. Here we will prove some general propositions about the importance of network structure for the robustness of cooperation in the model using a regular grid with Moore neighborhoods as the network topology. We will define a set of payoff boundaries that determine the stability (or spreading) of cooperation in various structures, and apply those analytic boundaries to the behavioral surface observed in the experiments.

This analysis will offer a complete explanation of the underlying micro-foundations of the simulation dynamics and the specific shape of the experimental results. It will demonstrate that every aspect of the model's behavior (including every bump or ridge on the behavioral surface) is mathematically intelligible. This should assure readers that the results reported in the article are not random flukes, programming bugs, or artifacts of some auxiliary assumption. However, the exhaustive analysis here comes at a great cost in complexity and accessibility of exposition, and the main conclusions are the same as those presented in the paper. We thus focus in the paper on the most prominent and general qualitative conclusions, rather than the intricate details described here. Poring over these pages is not necessary to derive the full value of the paper, but the details are available here for interested readers.

DEFINITIONS

A concise exposition of the mathematical analysis requires definition of a few specialized terms. On the last page of this document, we provide a glossary of important terms that we employ throughout this section.

We begin with a review of the formal definition of the game. We know (by the definition of the Prisoner's Dilemma) that DC > CC > DD > CD. Let us define three parameters to represent the differences between adjacent payoffs:

Fear (F) = DD – CD
Value of Exchange (V) = CC – DD
Greed = DC – CC

Thus, parameters {F, V, G} represent the incremental benefit to an agent as a result of getting the next higher payoff in the game. The absolute scale of the payoffs is arbitrary so without loss of generality we may assume CD=0 and express the remaining payoffs in terms of these three parameters:

CD = 0
DD = F
CC = F + V
DC = F + V + G

Recall that these three parameters (F, V, G) are all positive by definition, in order for the PD inequalities to hold:

F > 0 **V > 0** **G > 0**

Further, by the elaborated definition of the Prisoner's dilemma (2CC > DC + CD; DC + CD > 2DD), we can say the following:

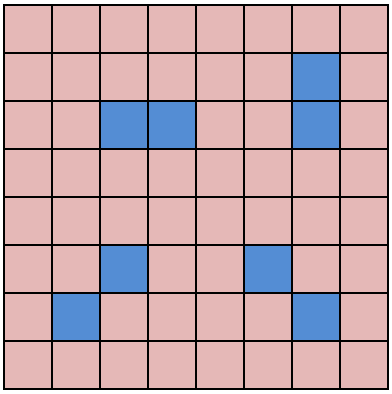
2CC > DC + CD implies 2(F + V) > (F + V + G) which simplifies to **V > G - F**

DC + CD > 2DD implies 2F < (F + V + G) which simplifies to **V > F – G**

For the purpose of replicating strategies, let us define the **neighborhood** as the 9 cells including the focal cell itself and its surrounding 8 adjacent cells (N, NE, E, SE, S, SW, W, NW).

Let us define the **C-paragon** as the highest scoring C neighbor in that neighborhood, and the **D-paragon** as the highest scoring D neighbor in that neighborhood.

It is obvious that an individual C cannot 'survive' (maintain cooperating) in this model when surrounded by Ds, because the Ds will outscore it in every dyadic interaction and the C will immediately imitate the better-performing D neighbors. We know that some clustering of Cs is necessary. Let us first consider the simplest clusters, just two adjacent Cs. Given an isolated cluster of two Cooperators ('pair'), we can identify the prospects for the cluster surviving into the next round. In this case, the two Cs in the cluster have equivalent neighborhoods so we can focus on just one of the agents, *i*.



In a pair of cooperators, *i*'s C-paragon (which could be either C in the cluster) has one C neighbor and seven D neighbors (yielding a payoff of CC + 7CD). Agent *i* compares this score to the D-paragon, which has two C neighbors and six D neighbors (yielding a payoff of 2DC + 6DD). Thus, for a cluster of two Cs to survive, the following would need to be true:

$$CC + 7CD \geq 6DD + 2DC$$

Using the parameters above, we may re-express this condition as

$(F + V) \geq 6F + 2(F + V + G)$ which simplifies to $7F + V + 2G \leq 0$

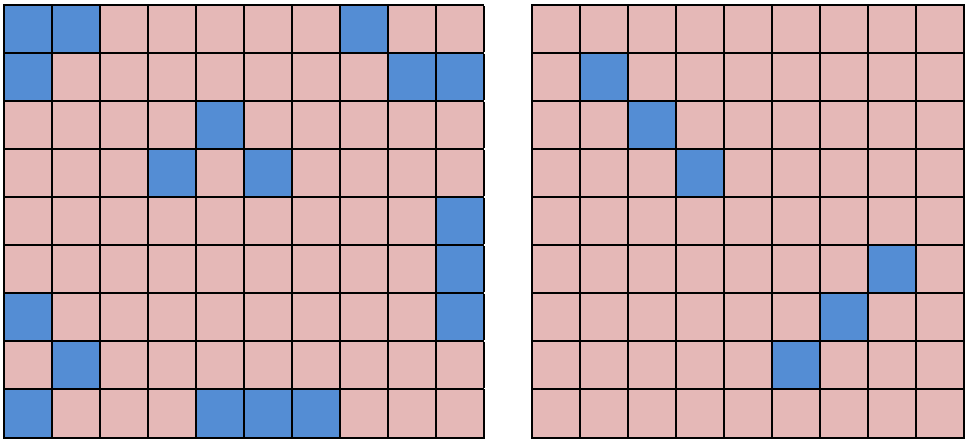
Which can never be true because F, V, G are all positive by definition of the Prisoner's Dilemma; therefore a pair of Cs surrounded by Ds will always convert to defection in the PD, regardless of the particular value of Fear, Greed, or the Value of exchange. This yields proposition 1a:

PROPOSITION 1a: Cooperation can never survive in a pair

Now let us consider the condition for cooperation to spread from this configuration. Looking at the perspective of the neighboring D that is exposed to the strongest force to cooperate (the greatest payoff advantage for its C paragon vs. its D paragon), note that the same comparison of payoffs above would govern a neighboring D's choice to switch to cooperation: $7F + V + 2G < 0$. Because the inequality will never be satisfied (in the PD), cooperation also cannot spread from C pairs. This yields proposition 1b:

PROPOSITION 1b: Cooperation can never spread from a pair

Now let us consider a trio of 3 adjacent cooperators. In both case (1) and case (2) below, the cluster's C-paragon has two C neighbors and six D neighbors. In case (1), the D-paragon will have three C neighbors and five D neighbors, whereas in case (2) the D-paragon will have two C neighbors and six D neighbors.



(1) (2)

All of the Cs compare the C-paragon's score (2CC + 6CD) to the D-paragon's score. Which for case (1) is 3DC + 5DD and for case (2) is 2DC + 6DD. Thus, for a C-trio to survive, these would need to be true:

Case (1) $2CC + 6CD \geq 3DC + 5DD$
Case (2) $2CC + 6CD \geq 6DD + 2DC$

Using the parameters above, we may re-express this condition as

Case (1)	$2(F + V) \geq 3(F + V + G) + 5F$	which simplifies to	$6F + V + 3G \leq 0$
Case (2)	$2(F + V) \geq 2(F + V + G) + 6F$	which simplifies to	$6F + V + 2G \leq 0$

Neither of which can ever be true because F, V, G are all positive by definition; therefore a trio of cooperators will always convert to defection. This yields proposition 2a:

PROPOSITION 2a: Cooperation can never survive in a trio

Examining the conditions for cooperation spreading from a trio, we restrict our condition that is most favorable for cooperation spreading: the end of a diagonal cluster of Cs as in case (2). Here the weakest D neighbor has only one C neighbor (seven Ds) and compares to a C paragon with two C neighbors.

$2CC + 6CD \geq DC + 7DD$ $2(F + V) \geq F + V + G + 7F$ which simplifies to $V \geq 6F + 3G$

Thus, there is a narrow range of payoffs (negligible Fear and low Greed) where a D neighbor at the end of a diagonal line will convert to C. However, even in the most favorable configuration and payoffs this outcome is ephemeral, because the cluster itself will evaporate (see proposition 2a), so the neighboring cooperator will also promptly disappear. Thus, proposition 2b states that cooperation cannot diffuse from a trio.

PROPOSITION 2b: Cooperation can never spread from a trio.

For clusters with three or fewer agents, all members were directly connected to the same C-paragon. With four or more agents, some members may not have direct access to the highest scoring agent in the cluster. Moreover, some may be connected to D-paragons with different scores. To assess the stability of a C cluster, then, we need to first identify the *weakest link* in the cluster – that is, the individual C agent most apt to convert to D. The weakest link *i* will be the C whose D-paragon scores highest relative to its C-paragon. If this weakest link will continue cooperating, the entire cluster will survive.

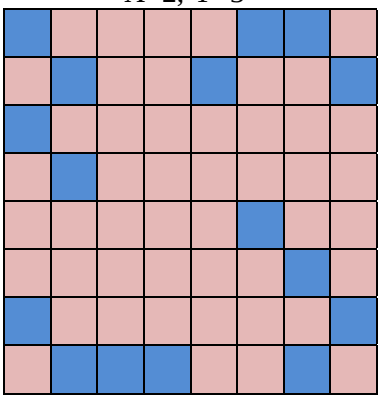
The weakest link in an open cluster of four Cs compares its C-paragon (including itself) with its D-paragon. Here we introduce new notation for brevity.

Let X = the number of Cs in the neighborhood of the weakest link's C-paragon

Let Y = the number of Cs in the neighborhood of the weakest link's D-paragon

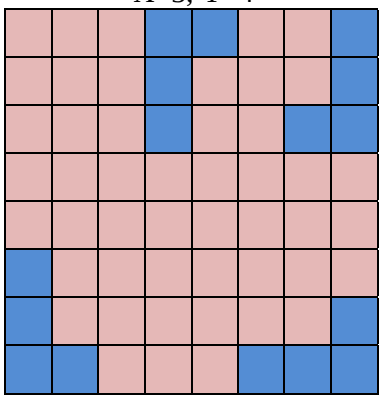
See the values of X and Y for the following three cases:

X=2; Y=3



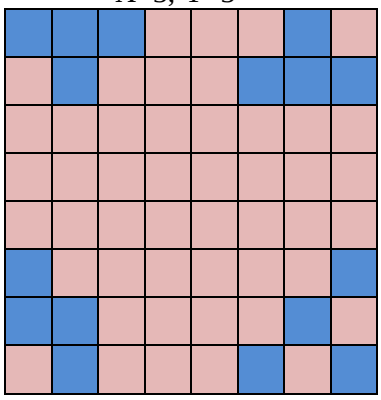
(1)

X=3; Y=4



(2)

X=3; Y=3



(3)

See that X and Y determine the payoff coefficients on the criteria for cluster survival.

$$(X)CC + (8\text{-}X)CD \geq (Y)DC + (8\text{-}Y)DD$$

For an open cluster of four Cs to survive, the following would need to be true:

Case (1) $2CC + 6CD \geq 3DC + 5DD$

Case (2) $3CC + 5CD \geq 4DC + 4DD$

Case (3) $3CC + 5CD \geq 3DC + 5DD$

Using the parameters above, we may re-express these conditions as:

Case (1)	$2(F + V) \geq 3(F + V + G) + 5F$	which simplifies to	$6F + 3G \leq 0$
Case (2)	$3(F + V) \geq 4(F + V + G) + 4F$	which simplifies to	$5F + 4G \leq 0$
Case (3)	$3(F + V) \geq 3(F + V + G) + 5F$	which simplifies to	$5F + 3G \leq 0$

None of which can be true because F and G are positive by definition; therefore an open cluster of four cooperators will always convert to defection.

PROPOSITION 3a. Cooperation can never survive in an open (non-square) 4-cluster

Examining the conditions for cooperation spreading from an open 4 cluster, we restrict our condition that is most favorable for cooperation spreading: the end of a diagonal cluster of Cs as in case (3). Here the weakest D neighbor has only one C neighbor (seven Ds) and compares to a C paragon with three C neighbors.

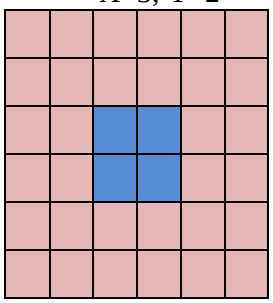
$3CC + 5CD \geq DC + 7DD$ $\quad 3(F + V) \geq F + V + G + 7F$ which simplifies to $\quad V \geq 5/2F + 1/2G$

Thus, there is a narrow range of payoffs (particularly sensitive to Fear) where a defector neighbor off the corner of an open 4 cluster will convert to cooperation. However, even in the most favorable configuration and payoffs this outcome is ephemeral, because the cluster itself will evaporate (see proposition 3a), so the neighboring cooperator will also disappear. Thus, cooperation cannot diffuse from an open 4 cluster.

PROPOSITION 3b: Cooperation can never spread from an open (non-square) 4-cluster.

Given a 2x2 square cluster of Cooperators, the C-paragon will have three C neighbors and five D neighbors (yielding a payoff of 3CC + 5CD) and the D-paragon will have two C neighbors and six D neighbors (yielding a payoff of 2DC + 6DD).

X=3; Y=2



Thus, for a C to survive in a square cluster of four Cs, the following would need to be true:

$3CC + 5CD \geq 2DC + 6DD$

Using the parameters above, we may re-express this condition as:

$3(F + V) \geq 6F + 2(F + V + G)$ which simplifies to: $\quad V \geq 5F + 2G$

For Cooperation to be stable in weak square clusters, the value of exchange (V) must be great relative to the premium for unilateral defection (G) or the penalty for unilateral cooperation (F). *This boundary is especially sensitive to Fear. Even if Greed is negligible, Fear must be less than 20% of the value of exchange for cooperation to survive in such small clusters.*

Comparing this configuration to the figures on the previous page, see that a 2x2 square is an *optimal* 4-C cluster. That is, no other configuration of 4 Cs is more stable. Unless otherwise specified, we will always mean a 2x2 square when we refer to a 4-C cluster. We focus on optimal clusters because this allows us to identify the analytical boundaries (payoff values) where *no* clusters of this size are able to survive.

PROPOSITION 4a: Cooperation can survive in *minimal square* clusters (2x2) if and only if: $V \geq 5F + 2G$ Thus, the stability of minimal square clusters depends strongly on *Fear*.

Examining the conditions for cooperation spreading from a minimal square cluster, we consider the case of spreading at the corners and at the sides. At the corners, the weakest D neighbor has only one C neighbor (seven Ds) and compares to a C paragon with three C neighbors. At the sides, the weakest D neighbor has two C neighbors (six Ds) and compares to a C paragon with three C neighbors.

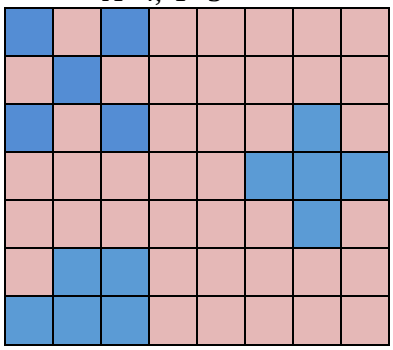
3CC + 5CD > DC + 7DD	3(F + V) > F + V + G + 7F	which simplifies to	V > 5/2F + 1/2G
3CC + 5CD > 2DC + 6DD	3(F + V) > 2(F + V + G) + 6F	which simplifies to	V > 5F + 2G

Note that the condition for cooperation to spread at the sides of minimal square clusters is the same as the stability condition for these clusters (and also subsumes spreading at the corners). Thus, the payoff conditions necessary for minimal square clusters to survive also enable cooperation to spread into surrounding fields of Ds from minimal square clusters of Cs, so we never see minimal square clusters of cooperation at equilibrium. Cooperation will always disappear if no clusters of at least 2x2 appear, but if at least one such cluster appears and Fear and Greed are within the region described above, then cooperation will diffuse broadly. *Thus minimal square clusters may have a crucial role in getting cooperation started in conditions of low Fear.*

PROPOSITION 4b: Cooperation can diffuse from *minimal square* clusters (2x2) if and only if: $V > 5F + 2G$ Thus, cooperation spreads from minimal square clusters in the same conditions that minimal square clusters become stable, and depends largely on Fear.

Given a 'star' of Cooperators, the weakest link's C-paragon will have four C neighbors and four D neighbors (yielding a payoff of 4CC + 4CD) and the corresponding D-paragon will have three C neighbors and five D neighbors (yielding a payoff of 3DC + 5DD).

X=4; Y=3



Thus, for a Cooperator to survive in a star cluster, the following would need to be true:

$$4CC + 4CD \geq 5DD + 3DC$$

Using the parameters above, we may re-express this condition as:

$4(F+V) \geq 5F + 3(F + V + G)$ which simplifies to: $V \geq 4F + 3G$

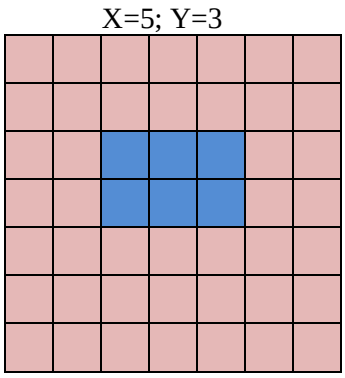
For Cooperation to be stable in star clusters, the value of exchange (V) must again be great relative to the premium for unilateral defection (G) or the penalty for unilateral cooperation (F). *A star cluster of 5 Cs is slightly less sensitive to Fear but* more *vulnerable to Greed, compared to weak square clusters.* See that a star is an *optimal* 5-C cluster. That is, no other configuration of 5 Cs is more stable than a star. Unless otherwise specified, we will always mean a star when we refer to a 5-C cluster.

PROPOSITION 5a: Cooperation can survive in *star* clusters if and only if: $V \geq 4F + 3G$

If Fear and Greed are low enough that star clusters can survive at all, then the D neighbors in contact within the star will convert to Cs and the cluster will immediately become a 9-C strong rectangle cluster. Then cooperation will continue to diffuse more broadly. Thus, we never see star clusters of cooperation at equilibrium.

PROPOSITION 5b: Cooperation can spread from *star* clusters to yield strong rectangle clusters, if and only if: $V \geq 4F + 3G$

Given a 3x2 'weak rectangle' of Cooperators, the weakest link's C-paragon will have five C neighbors and three D neighbors (yielding a payoff of 5CC + 3CD) and the corresponding D-paragon will have three C neighbors and five D neighbors (yielding a payoff of 3DC + 5DD).



Thus, for a Cooperator to survive in a weak rectangle, the following would need to be true:

$$5CC + 3CD \geq 5DD + 3DC$$

Using the parameters above, we may re-express this condition as:

$$5(F + V) \geq 5F + 3(F + V + G) \quad \text{which simplifies to:} \quad V \geq (3/2)F + (3/2)G$$

For Cooperation to be stable in 3x2 clusters, the premium for mutual cooperation (V) must be great relative to Greed or Fear. However, it need not be as great as it needed to be for weak squares or stars. *Notably, Fear is much less important for weak rectangle clusters than it is for minimal squares or stars.*
See that a 3x2 rectangle is an *optimal* 6-C cluster. That is, no other configuration of 6 Cs is more stable than this one. Unless otherwise specified, we will always mean a 3x2 rectangle when we refer to a 6-C cluster.

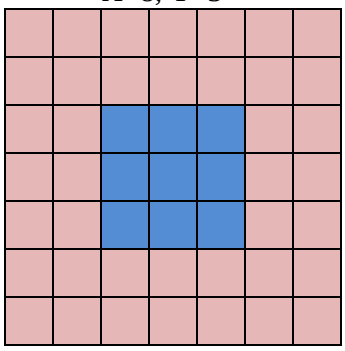
PROPOSITION 6a: Cooperation can survive in *weak rectangle* clusters if and only if: $V \geq 3/2F + 3/2G$

Now let us consider the spreading of cooperation from weak rectangle clusters. *Seen from the perspective of the Ds, this boundary represents the point where a D neighbor of a large cluster of Cs (i.e. a flat edge of Cs at least 3 wide and at least 2 deep) will perform as well those C neighbors. When Fear and Greed are below this (low) joint boundary, the neighboring Ds will convert to cooperation and thus cooperation will spread laterally from flat edges of clusters. This critical value allows cooperation to easily grow across the population. This boundary distinguishes the regions of the parameter space where cooperation diffuses broadly from where it cannot do so. Thus, we call it the* diffusion boundary. *This result is equally sensitive to Fear and Greed.*

PROPOSITION 6b: Above the boundary at $V > (3/2)F + (3/2)G$, cooperation can spread from flat edges into fields of defectors. We call this the *Diffusion Boundary*.

A Cooperator on a smooth edge of a rectangular cluster of at least 3x3 Cooperators has at least one Cooperator neighbor (in the center of the cluster) who is entirely surrounded by Cooperators (yielding a payoff of 8CC). Such protected C neighbors will serve as C-paragons for the entire cluster. By contrast, any D-paragon will have three C neighbors and five D neighbors (yielding a payoff of 3DC + 5DD).

X=8; Y=3



Thus, for a cooperator to survive in strong rectangle clusters, the following would need to be true:

$$8CC \geq 3DC + 5DD$$

Using the parameters above, we may re-express this condition as:

$8(F + V) \geq 5F + 3(F + V + G)$ which simplifies to: $V \geq 3/5G$

This is a very weak condition because Fear may be arbitrarily large and Greed may be greater than the Value of exchange. In our experiments, G is in (0, V] and F is in (0,V], so $G \leq 5/3V$ over this entire space, and we can conclude that an isolated cluster of at least 3x3 Cooperators will *always* be robust in the conditions underlying the behavioral surfaces in the article. Very high Fear will have no effect on the stability for any smooth cluster of at least 3x3, because all Cs on the perimeter will have access to an interior C-paragon. If $G > 5/3\ V$, then cooperation cannot survive in a 3x3 cluster (or in any other structure). See that a 3x3 square is an *optimal 9-C cluster*. Unless otherwise specified, we will always mean a 3x3 square when we refer to a 9-C cluster. Similarly, an *optimal 8-C cluster* is a 9-C cluster with one corner removed and an *optimal 7-C cluster* is a 9-C cluster with two corners removed.

PROPOSITION 7a: Cooperation can survive in *strong rectangle* clusters of at least 3x3 if $V \geq 3/5G$ Thus, the stability of such large clusters does not depend on Fear.

Although a 3x3 cluster is much more robust to invasion than a 2x3 cluster, they have the same tendency to spread cooperation along flat edges. This is because the cooperators on the perimeter of the cluster still have no more than 5 cooperator neighbors, which is the same as for the 2x3 cluster. This yields proposition 7b:

PROPOSITION 7b: Cooperation can spread from *strong rectangle* clusters (or larger smooth clusters) under the same *diffusion boundary* as specified in proposition 6b.

Until now we have illustrated the stability of particular structures, and the spreading of cooperation from those structures. Readers may have noticed that the critical difference between the clusters that were unstable and those that were stable was the way that the arrangement of the cluster may protect insiders from outsiders, such that Ds have less access to Cs than Cs have to each other. This observation leads us to a more general characterization:

PROPOSITION 8: A C-cluster will survive only if its weakest link's C-paragon has more C neighbors than the weakest link's D-paragon does.

Again, we focus on the weakest link of the cluster, because this allows us to identify the conditions where an entire cluster can be stable. Let agent *i* represent the weakest link in a cluster. As before, let X represent the number of C neighbors for *i's* C-paragon and let Y represent the number of C neighbors for *i*'s D-paragon. For *i* to *remain* a C on the subsequent round (and thus for the cluster to survive), the following inequality must hold:

$$X \times CC + (8 - X)CD \geq Y \times DC + (8 - Y)DD$$

Substitute the values in terms of parameters F, V, and G:

$$X(F + V) \geq Y(F + V + G) + (8 - Y)F \qquad \text{which simplifies to} \qquad X \geq \frac{V + G}{V + F}Y + \frac{8F}{V + F}$$

We know that G < V + F by the definition of the PD (i.e. 2CC > CD + DC), so we can re-express this constraint as G = V + F + *e* (where *e* represents some positive number) and then substitute V + F + *e* for G in the inequality above, to yield:

$$X \geq \frac{2V + F + e}{V + F}Y + \frac{8F}{V + F}$$

Because V, F, and e are all positive, we know that 2V + F + *e* > F + V. (See that we could simplify that inequality to V + e > 0, which must be true.) Because the numerator is greater than the denominator, we know that the multiplier on Y in the inequality above is strictly greater than 1. Because the quotient is greater than 1, and the term on the right is also positive, we know that if the inequality above is true, then X must also be greater than Y. Therefore, we can say that regardless of the values of F, V, and G (within the ranges defined by the PD), if this inequality is satisfied, then X must be greater than Y. In other words, if *i* is to remain a C, then *i*'s C-paragon must have more C neighbors than does its D-paragon.

This describes a hard constraint on the structures that can be stable in this embedded social dilemma game:

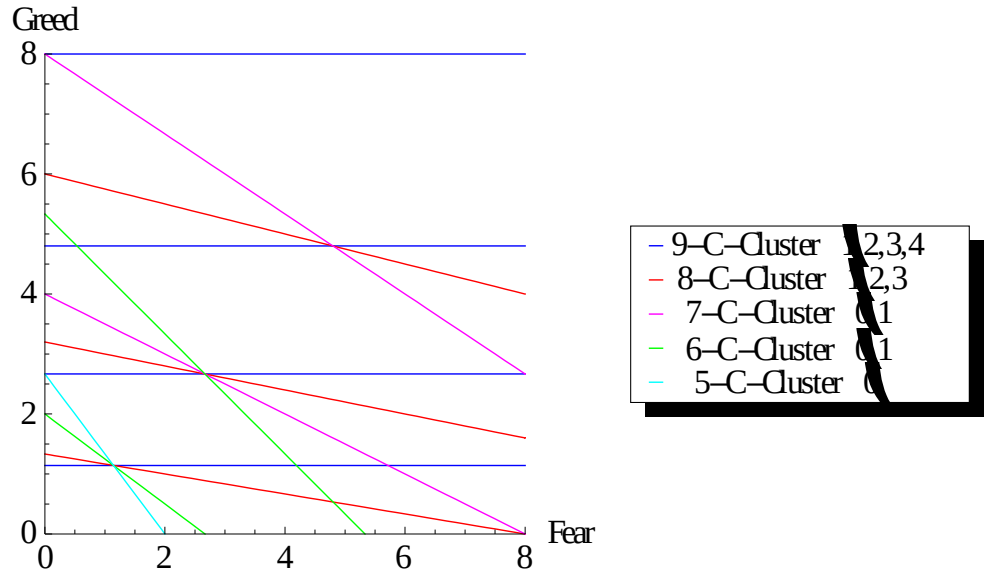
Unless there is at least one C that is relatively protected from Ds – and is accessible to all C agents in a cluster – then any peripheral agents lacking access to such a protected agent will be lost to the cluster, and the cluster will unravel. If there is no such protected core member, the cluster cannot survive at all.

MAP OF THE PARAMETER SPACE – QUALITATIVE BOUNDARIES

We use the term *s*-cluster to mean an *optimal* configuration of $s \leq 9$ cooperators; that is, a configuration of *s* cooperators that is maximally robust when surrounded by defectors. We have shown that an isolated 9-C cluster (which we have called a strong rectangle) will be stable for the entire space that we consider. Paradoxically, this stability can be undermined by stray Cs outside the cluster, which may improve the performance of neighbor Ds. If outside Cs jointly improve the performance of the D-paragon relative to the C-paragon in the cluster, we call them 'satellites' of the cluster. Thus, a set of payoff boundaries in payoff space where optimal s-clusters are sensitive to *k* satellites will distinguish regions where cooperation is generally more or less robust. In fact, mapping this set of boundaries is sufficient to understand the equilibrium level of cooperation over the entire parameter space investigated.

We may define this set of boundaries in the space of Fear and Greed (for $5 \leq s \leq 9$ and $0 \leq k \leq s - 5$) with a single inequality: $V > \frac{(3+k)G + (9-s)F}{s-k-4}$ In other words, if this payoff inequality is satisfied, an *s*-cluster is stable despite *k* satellites boosting the performance of Ds at its weakest link. Figure B-1 below gives a map of these boundaries. The number(s) in parentheses represents the number of satellites at the weakest link. There are 12 such boundaries in the payoff space that we examine, ranging from a boundary for a 9-C cluster (with one satellite) at the top of the figure below to a boundary for a 5-C cluster (with zero satellites) in the bottom left corner.

Figure B-1. Map of Boundaries in Payoff Space.



The four horizontal blue lines at G = 8, G = 4.8, G = 2.67, and G = 1.14 represent the payoff boundaries where strong rectangles become stable with 1, 2, 3, and 4 satellites, respectively. (Strong rectangles are stable with 0 satellites whenever G < 13.33, so over this entire space.) More C satellites make C clusters

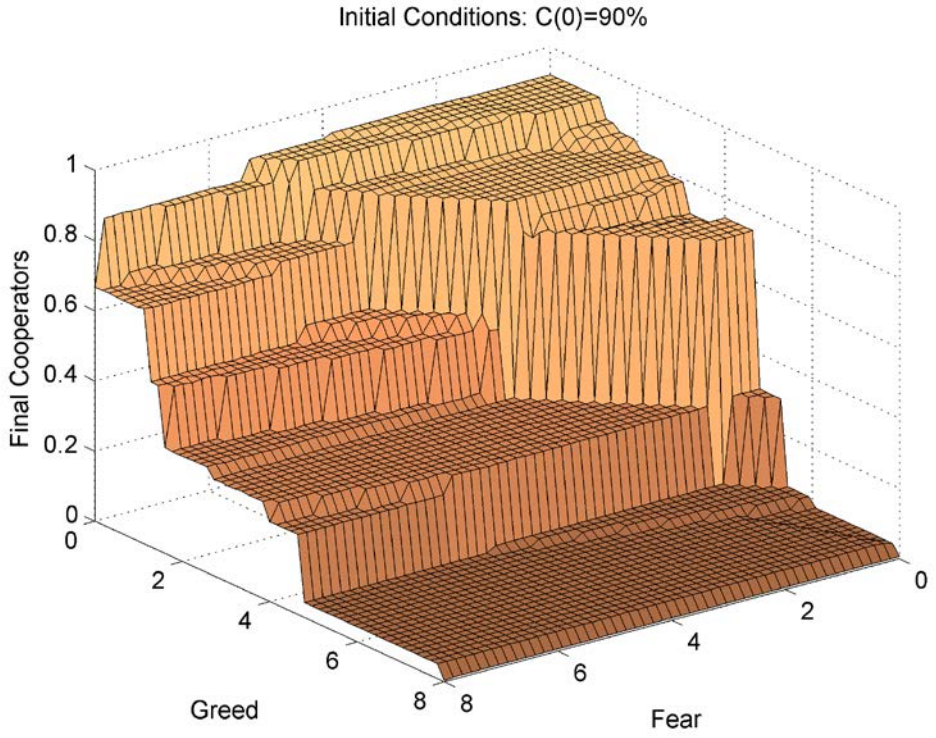
more sensitive to Greed (i.e. the cluster will dissolve at lower levels of Greed). Again, strong rectangles always remain immune to Fear, because they have one or more interior protected C-paragons.

An 8-C cluster is a strong rectangle, but missing one corner. The three red lines shown represent stability boundaries for 8-C clusters with 1, 2, and 3 satellites. (Such a cluster is stable over the entire region with 0 satellites and never stable with 4 satellites.) Lacking a completely-protected interior C role-model, 8-C clusters are slightly sensitive to Fear – that is, less stable at higher levels of Fear.

The two magenta lines represent the boundaries where a 7-C cluster (with 0 or 1 satellites) becomes stable. The two green lines represent the stability boundaries for *weak rectangles* (6-clusters) with 0 and 1 satellites. A detailed analysis of the boundary for weak rectangles with 0 satellites (the 'cliff') will be discussed below. A single cyan boundary in the lower left corner applies to a 5-C cluster, which is fragile and thus never stable with any satellites.

The boundaries above represent fundamental constraints on the dynamics of the model, and thus should help us diagnose the shape of the behavioral surface. Now let us compare the boundaries above to the surface reported in paper, and reproduced as Figure B-2 below:

Figure B-2. Long-Run Cooperation by Greed by Fear, with 90% Initial Cooperators.



(Note that the axis scaling is inverted from the previous line plot to this surface plot, to make the patterns visible.) See that the main diagonal "cliff" on the surface represents the boundary where weak rectangles become stable. As explained in the paper, when G is below this boundary, cooperation is able to diffuse

across the population, because flat edges of will outperform neighboring flat edges of defectors. The opposite is true when G is above this boundary.

As Greed decreases, the small horizontal steps at G = 8 and G = 2.67 and the large steps at G = 4.8 and G = 1.14 represent the increasing stability of strong rectangles to satellite Cs as Greed decreases. If Greed is low enough, a strong rectangle will continue cooperating even if some C satellites boost the performance of the D-paragon at the weakest link. When G > 8, a strong rectangle is fragile and can be destroyed by even a single C satellite anywhere on its perimeter. When 4.8 < G < 8, a strong C rectangle can be destroyed by two C satellites. Some of these boundaries are also meaningful seen from the perspective of D clusters. *When Greed increases above 2.67, a flat edge of Ds can expand in both directions, although it will immediately collapse to a single row again.* Recall that clusters of Ds (which are harmed by their proximity to each other) cannot survive exposure to C clusters as Greed decreases.

Two of the large steps are at a slight angle: One intersects the left Greed axis at G = 1.6 and slopes gradually to intersect the right Greed axis at G = 3.2. The other intersects the left Greed axis at G = 0 and slopes gradually to intersect the right Greed axis at G = 1.33. We have already interpreted these boundaries as representing the stability of 8-C clusters to 2 and 3 satellites, respectively. Seen from the perspective of Ds, the same boundaries represent the conditions where elbow-shaped 3-clusters of Ds and 4-clusters of Ds, respectively, cannot survive.

The 7-C cluster and 5-C cluster configurations are rare and relatively unstable, and the corresponding magenta and cyan boundaries represent only tiny ripples on the behavioral surface.

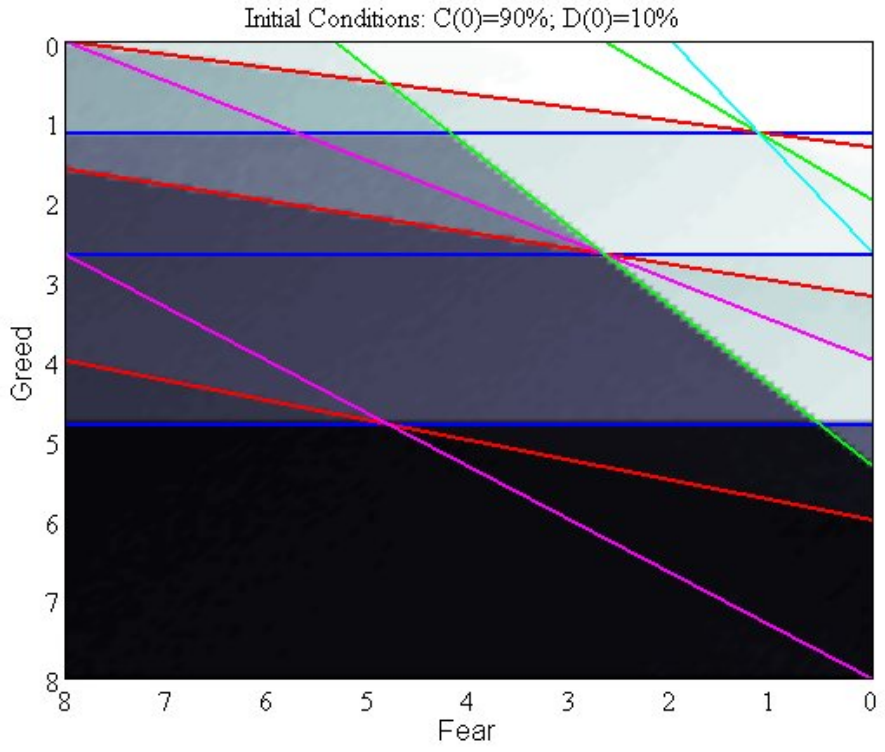
The shape of the surface is determined by the intersection of these various boundaries. See that cooperation generally thrives in the area behind the “cliff.” At the rightmost corner of the cliff, however, there is a tiny region where cooperation plummets because strong C rectangles become fragile and clusters of four Ds are able to grow, when 3V/5 < G < 2V/3 - F (i.e. 4.8 < G < 5.33 - F). In this region, cooperation can still spread along flat boundaries but the resulting large cooperator clusters erode as they near each other. That is, Ds sandwiched between these growing C clusters perform very well and colonize the C clusters, resulting in chaotic dynamics in this tiny region as discussed in the paper.

There is an intriguing dip in the surface (yielding a non-monotonic effect of Greed) just above Greed=3, when Fear is near zero. Above a certain value of Greed – (2V-F)/5 (or around 3.2 at F=0) – a square of four Ds will become stable in a field of Cs, although large C clusters will push large D clusters back along square edges (because Ds perform poorly in large clusters). As a result, Ds will be reduced to strips (which benefit from interacting with Cs on both sides), and their immediate neighbors cycle between C and D. However, increasing Greed a little more – to (V - F)/2 (or around 4.0 at F=0) – changes the structural dynamics of diffusion, such that large C clusters advance on rounded edges (i.e. the cluster corners break down as they spread). The resulting structures are highly irregular and preserve more pockets of stable cooperation. As a

result, cooperation rises as Greed increases into that region, until it hits the boundary at G = 4.8, where C clusters become more fragile and cooperation drops precipitously.

To aid in interpretation, the previous 2 figures are consolidated in a single image below. Here the analytical boundaries are superimposed on the behavioral surface for the mean cooperation at equilibrium (represented visually as shading rather than surface height). This allows readers to easily compare the location of the boundary lines with the mean cooperation at equilibrium for a given region of the parameter space.

Figure B-3 Long-Run Cooperation by Greed by Fear, with 90% Initial Cooperators.



PAYOFF CHANGES: IRREVERSIBLE INTERVENTIONS.

Some interventions (i.e. changes) to Greed or Fear have permanent effects on cooperation even after the changes are fully reversed, that is, some changes in payoff space lead to one-way transitions. In this context, we distinguish two types of irreversible interventions: A one-way downward intervention takes place when an increase in Greed or Fear diminishes cooperation and the loss cannot be reversed by restoring the prior payoff value. A one-way upward intervention occurs when a decrease in Greed or Fear enhances cooperation and the gain cannot be reversed by restoring the prior payoff value. Importantly, both types of irreversible intervention demonstrate how some specific configurations of Cooperators mapped in figure B-1 condition the evolution of cooperation *after* a change at this boundaries takes place.

First consider changes in the dynamics of cooperation after an intervention takes place at the edge of the primary cliff, which we have called the diffusion boundary. We already established in proposition 6b that cooperation can spread from flat edges into fields of defectors if $V > (3/2)F + (3/2)G$. If there is at least a C-cluster of 3 x 2 (i.e. a weak rectangle) and there are no satellites. A slight reduction in Greed or Fear that crosses this boundary will activate spreading dynamics that diffuse cooperation over the grid. As a result, small clusters become larger clusters of cooperators, and there will be at least one Cooperator neighbor entirely surrounded by other Cooperators. Such a protected C neighbor will serve as C-paragon for the entire cluster. Importantly, as established in proposition 8, such an emergent C-cluster will always be stable in the conditions underlying the behavioral surface in this article since the weakest link's C-paragon will have more C neighbors that the weakest link's D-paragon. Therefore, a slight change in Fear and/or Greed that crosses the diffusion boundary will not erode the level of cooperation achieved in the round before the intervention. At the same time, As such, the diffusion boundary is a one-way upward boundary.

Contrary to the one-way upward nature of the diffusion boundary, the four horizontal blue lines in the consolidation region of figure B-1are all one-way downward boundaries. These are the payoff boundaries where strong rectangles become stable with 1, 2, 3, and 4 satellites. These are one-way downward boundaries because a descending step, that is, an increase in Greed at the edge of any of these boundaries, will cause a loss in cooperation that cannot be recovered even if Greed were to be set to its minimum right after the intervention. This process is, therefore, the mirror image of the dynamics described in the paragraph above. For example, the key local configuration that explains the nature of the small horizontal step at $G = 8$ is the strong rectangle with 1 satellite. In this context, setting $G > 8$ will cause all the Cooperators in the neighborhood of the D-paragon to become defectors, thus reducing the original configuration (strong rectangle with 1 satellite) to a weak rectangle in the round following the intervention. After such a change, reversing the intervention, that is, setting $G \leq 8$, simply cannot cause the weak rectangle to evolve into a strong rectangle again. As such, this is a one-way downward boundary.

GLOSSARY OF TERMS USED IN THIS ANALYSIS

Cluster – A cluster refers minimally to a set of contiguous agents – where each agent is connected to at least one other agent in the cluster – of the same type (e.g. C), and surrounded by at least two layers of agents of the opposite type (e.g. D). We focus on clusters of 9 or fewer agents and their surrounding neighborhoods. Because diffusion occurs through local influence in the model, we can develop a sufficient description of model behavior for populations of any size by analyzing the dynamics in patches as small as 11x11 (constituted by a cluster, its direct neighbors, and its neighbors' neighbors).

Paragon – Agent *i*'s C-Paragon refers to the highest-scoring C in agent *i*'s neighborhood (including *i* itself) and *D*-Paragon refers to the highest-scoring D in agent *i*'s neighborhood. Agent *i* compares these two paragons to assess the performance of the two strategies, in selecting its own strategy for the next round.

Weakest Link – The *weakest link* in a cluster is the agent most apt to convert to the opposite type (e.g. a C changes to a D). The weakest link *i* in a C cluster will be the C whose D-paragon scores highest relative to its C-paragon. If this weakest link will continue cooperating, the entire cluster will survive into the next round.

Cluster Paragon – We may integrate the two previous concepts: The C-paragon and D-paragon of the weakest link in the cluster may be regarded as the C-paragon and D-paragon of the entire cluster, as it is the performance comparison of these two agents that determines immediate cluster survival.

Optimal s-Cluster – A cluster of *s* agents may be configured in a variety of ways. Because in this analysis we are interested in identifying conditions where *no* cluster of *s* agents can survive, we focus our attention on the *optimal* configuration of agents (i.e. most robust when surrounded by agents of the opposite type), for a given cluster size *s*. (If we show that cooperation cannot survive in an optimal *s*-cluster, it obviously cannot survive in any other *s*-cluster.) For example, an optimal 4-C-cluster is a square of 4 Cs (surrounded by a mass of Ds), where the performance comparison of the C-paragon and D-paragon of the cluster is most favorable to survival of the cluster. In this case, a 2x2 square is a the unique optimal 4-C-cluster, although optimal clusters are not always unique.

Satellites – A 'satellite' cooperator is a C within two steps of a C-cluster that does not improve the performance of the cluster's C-paragon, but improves the performance of its D-paragon. For example, if *j* is a D-paragon adjacent to a C-cluster, and one or more Cs appear *behind j*, those satellites will improve *j*'s performance and may endanger the cluster without even coming into contact with it. Unless otherwise specified, we discuss *isolated* clusters (with zero satellites).